COMPARISM OF CUT-OFF SCORES USING TWO SETTING METHODS

Iweka Fidelis, (Ph.D) and Tei-Firstman R.I

Department of Educational Psychology Guidance and Counselling, University of Port Harcourt, Nigeria

ABSTRACT: *The objective of the study is to identify if there is significant difference between using the Angoff method and the norm-referenced methodin the setting of cut off scores in school setting. The study made use of 80 (JSS 3) Basic 9 students from Nembe Local Government Area of Bayelsa state. The sample was drawn through simple random sampling method. The design of the study was comparative analysis. A forty-item multiple choice objective test on mathematics which were tested for goodness of fit using the Big step software was used. The internal consistency which was determined by Cronbach alpha was 0.64 while two research questions directed the conduct of the work. Percentages, intra-class correlation coefficient (ICC) were used to analyses the data collected. The comparative analysis made between the setting of cut off score using the Angoff and norm- referenced method has significant difference. Recommendations are made to supplementing the Angoff method with additional data from alternative methods to improve the appropriateness when setting performance standards in school settings. The Angoff method should be use as it is considered to be defensible, easy to apply, easy to explain to the policy makers who may ultimately set the passing score and it has been found to be extremely replicable across panels.*

KEYWORDS: Comparism, Settings, Method, Scores

INTRODUCTION

Background to the Study

Secondary education is the period when students are prepared for higher education and useful living within the society (National Policy on Education, 2004).With the increase in technology advancement and the nation's goal of becoming one of the twenty more developed economies by the year 2020, students need to be well equipped to cope with the changes in the society. There is the need therefore to increase the number and competence of student's entry and succeeding in science technology in higher institution. This can only be achieved if students have achieved a desired level of proficiency.

One of the recent trends in educational assessment is the rebirth or rekindled interest in criterion referenced test. The evolution of pass/fail marking system, with its emphasis on absolute standards, has also contributed to this approach to student assessment (Payne in Iweka 2014).

Schools are experiencing increased pressure to use results from assessment programmes to identify students who do not have the needed skills to graduate from school or who may have problems in "the next level" and may benefit from instructional activities beyond those provided in their regular classroom. These policies are often based, in part on student's test performance. Students with scores lower than minimum passing score are classified as needing instructional interventions beyond what the regular classroom teacher can provide. These minimum passing scores are often determined by using the Angoff standard setting method and other methods.

In licensure and certification testing, it is a common practice to use a criterion referenced approach to set the pass point (cut off score)for an examination. This approach provides a defensible rationale for identifying a cutoff score. The main rationale behind criterion-referenced cutoff scores is that one must be able to distinguish between candidates who can demonstrate sufficient knowledge to be licensed or certified and those who cannot.

A criterion referenced standard is a predetermined standard of performance that shows how the individual has achieved a desired level of performance (Orluwene, 2012). In order to decide whether an individual has achieved a desired level, a standard or cutoff score must be determined. Cusimano (2006) defined standard- setting as the process of deciding what is good enough. Kane (2004) stated that the passing score is a point on the observed- score scale where as thestandard is a conceptual boundary on the true- score scale between acceptable and non-acceptable performance, or in other words, a standard is the boundary between those who perform well enough and those who do not. Norcini (2003) standard are generally classed as absolute (Criterion based) or relative (norm based). According to Boursicot and Robert (2006) an absolute standard determines the pass and fail outcome by how well a candidate performedand he/she is usually judged against an arbitrarily set external standard. Hence, it is independent of the performance of the group. A relative standard on the other hand, compares how well the examinees have performed compared to others who took the test and hence the outcome is dependent on the performance of the group.

The outcome of assessment is determined by the standard setting method used. There is a wide range of standard- setting methods but themost popular ones are the norm referenced and the criterion referenced method. The most used and researched, criterion referenced method of standard setting, is the Angoff method (Boursicot & Roberts 2006).

In establishing a standard, there is a need to establish the possible and appropriate cut off scores to give a fair stance between the student’s ability and predication of their performance.

Angoff (1971) inadvertently introduced a method for standard setting that is, using the amount of attention devoted to it in the research context as an indicator, one of the most commonly used method of setting standard today.

The original method has been modified in different ways by researchers (Hambleton & Plake, 2005; Impara & Plake, 2007) in an attempt to improve it. Berk (2006) published a Consumer’s guide to standard Setting techniques, which include a set of criteria to be used to assess standard setting methods. He also assessed various cutoff score setting procedures including five Angoff type methods.

The variants in Angoff methods can be classified as item judgment methods. Each item on a test is assessed in terms of how likely, minimally acceptable or competent candidates (those who would barely meet mastery standards) are to answer that item correctly (Ricker, 2002). The Angoff method, in its basic form, is seemingly a very simple process. Perhaps its simplicity should not be surprising, given that it arose from foot note in a book chapter (Angoff 1971, p. 515). A group of judges are each asked to (independently) think of a group of minimally competent candidates who would border on the mastery/non mastery cut off. The most typical instruction is for judges to think of 100 candidates who would ‘just barely” meet the performances criteria.

Angoff first proposed the method, his instruction was to think of only one candidate. However with the exception of Impara & Plake (2007), the hypothetical pool of candidates is used.

The judges, working independently, then estimate what proportion of that sample of minimally acceptable candidate would answer eachitem in the test correctly. These p-values are summed and usuallydenoted as the Minimum Passing Level for judge (MPL). The minimumpassing level represents an individual judge cut score for the test. Themean of these cut scores is the final cut scores for the test. Thestandard error can also be calculated for the cut scores, a lowerstandard error is desirable since it denotes better agreement among the judges (and less uncertainty about where the true-cut score should lie).This method does not just apply minimally competent candidates, but could also be used to create a cut score for any grouping within the population. Tiratira, (2009), Angoff methods could be used to set a cut score for standard of excellence on a test. In this case, judges would be required to conceptualize a group of minimally excellent examinees.

However, another way of establishing cut off score is through the norm referenced. The norm-referenced method are easy to use andunderstand, can easily be explained to trainees and variations in testdifficulty are automatically corrected for as the pass mark is influenced by the performance of the examinee cohort (Verhoeve, Verwijen, Muijtjen, Scherpher & Vander, 2002). The draw backs of these methods are that examinees will always fail irrespective of their performance, students deliberately influence the pass score and that the pass score isnot known in advance (Verhoeve et al 2002). On the other hand, themain advantage of the Angoff method of standard setting are that, it is widely used in a range of certifying examination and that it is rather well supported by research evidence, (Norani 2003). However, it is not without pit falls. It can be very labour intensive and time consuming (Borsicot, 2006). Research has also shown that judges often find it difficult to accurately conceptualize a border line candidate.

It is important to have an understanding of how arbitrary the judgment involved in decision-making of standards setting can be. George (2006) views that all standard setting methods that involved judges making arbitrary decision are fundamentally flawed. Others however argued that although all standard setting method require human judgment, they can be made with careful deliberation and hence be fair and reasonable. Norcini (2003) stated that although all standards are judgmental, the credibility of each standard varies, depending on who sets the standard and the methods they used (Norcini & Guille, 2002). The validity of a test is determined as much by the method used to set the standard as by the test content itself.

Dowing, Tekian & Tudkowsky, (2006) argued that all standards are ultimately policy decision and that there is no “gold” standard for a passing score. A cut score takes into consideration different levels of performance. Thus, by definition, it is criterion referenced, because, this standard corresponds to a measure of what would be considered as a minimally acceptable performance, it can vary widely, depending on the job and/or on the specific criterion of performance level (D’ Almerida, 2006). What is key is the process of setting the standard. The four Key principles that under pin the process of standard setting are that it is systematic, reproducible, absolute and unbiased.

METHODS OF STANDARD SETTING

Overview of the Angoff Method.

In its basic form, the Angoff (1971) method of setting cut scores entails asking a group of judges to examine each item on a test and estimate what proportion of a target group of examinees will answer each item correctly. The target group of examinees are those who are on the borderline between the competent and the incompetents. Often judges are instructed to envision a hypothetical group of 100 target examinees and directed to estimate how many of the hypothetical group will answer each item correct. These item performance estimates for the target group are summed across item to obtain each judges cut score. The judges cut scores are averaged to obtain the estimate minimum passing score (mps) that a minimally competent candidate (MCC)' would obtain. In a school setting, the language "minimally competent candidate" has little meaning. Instead some schoolssubstitute phrases such as just competent student' or "barely proficient student'. Extensive research on the Angoff method has resulted in numerous modifications as follows:

1. Providing extensive training for judges in the process of both identifying the target examinee and in estimating their performance

2. Providing actual performance to judges (often along with the impact-percent passing or failing associated with the judges initialcut score)

3. Including more than one opportunity to estimate examinee performance (Plake & Impara, 2007).

Other modifications which are less pervasive include permitting judges to discuss their ratings or perspectives after an initial rounded of item performance. Estimation and requiring judges to estimate performance for a category of examinees in addition to the target examinees (e.g. estimating performance for the average examinee in addition to the target examinee. The studies reported in this paper used the variation described in Impara & Plake (2007) in which judges made dictomous estimate of examinee performance. Recent research suggests that some standard setting judges, make item performance estimates that systematically differ from actual performance.

Borderline Group Method

This method of setting a cut score can be accomplished in several ways. The method described below is a modification of Living Stone & Zieky (2002).

In school setting, teachers may be provided with a list of students in their class who can take the test on which a cut score is to be set. After providing teachers with the description of the test content, teachers are directed to make global estimate of their students into categories such as "below proficient, "proficient" and above proficient." Each of these categories would be definedoperationally within the content of the test context (the definition) which may be drawn up by a committee of teachers or by central office staff. After making initial classifications, the teachers are asked to go back through the list and (for example) indicate which is proficient and below.Proficient students are on the borderline between these two categories. For the borderline group method, students who are identified in their final classification comprise the borderline group. Test performance of students in this group serves as the basic data for the setting of cut off score. This is not the only way to identify the students who are in the borderline group, but it is a strategy that has been shown to work in school setting, (Impara&Plake2007).

All classifications must be competent prior to teachers knowing the scores of their students on the test. It may be done prior to testing or it may be done in conjunction with the Angolf method at the time item performance estimates are made. An advantage of this method is that it is a task-that is consistent with the literature

Advanced Impact Method

Dillon in Impara & Plake (2007) cited this method as another method that might be used to set the cut-off score and is to simply ask teachers what percentage of their current students are ready to graduate or are eligible for promotion to the next grade or who qualify for extra instructional assistance prior to asking teachers to estimate the percentage of their students who “qualify”. It is important to define what it means to qualify so that all teachers are using the same basis for their estimate. The question could be asked at the same time as theteachers are classifying their students into global categories to be used for the borderline group methods. If this is done at the same time with the Angoff, but prior to round one (1) of Angoff item performance estimate or it could be just prior to providing actual performance data prior to round two of the Angoff ratings.

Averaging across teachers will result in an estimate of the percentage of students in the school who are eligible for instructional intervention.After administering the test, accumulates percentage distribution can be used to find the score that identifies the appropriate percentage of students. One might expect the outscore obtained by this method to be near the cut score set by either the Angoff method or borderline group method to the extent that this estimate is more extreme, it may reflect a boundary point. There is a risk that this method may result in some deflation of the appropriate value if there is some belief by the teacher that the percentage of qualified students will reflect badly on them.

Similarly, if the teacher defined the student’s in need in a different way then, the percentage of students in need may be highly inflated (almost any students not grasping all the concept in the content area may be classified as being in need of extra instruction). For this reasons, this method should not be the only method employed. It should be used as a supplement to other method and extreme values may need to be discounted.

Standard from the Item Response Theory (IRT)

IRToriginated and developed in psychology and sociology in the 1940s and 1950s and the first half of 1960s. it was first formalized in the work of Lord and Novick (1968) to allow the evaluation of both student abilityand item properties, such as item discrimination capacity. The popularity came much more later in the 1970s with the inception of computers (Vander linden & Hambleton, 2010). It was initially used for dichotomous responses alone until very recently when polychromous models are formulated.

IRT was originally developed to overcome the limitation of CIT. A major part concerning the official work was provided in the 1960s (Wiberg2004). It was a probabilistic model for expressing the relationship between an individual’s response to an item called traits or ability measured by the instrument (Reeve 2002). It relates item performance to the ability measured. The ability is denoted by IRT uses the ability scale to determine the amount of latent trait an individual possess. The ability scale is an interval scale with a midpoint of one. It ranges from plus infinity to minus infinity, but practical consideration is usually from -3 to + 3 (Reeve, 2002). It is assumed that every examinee answering an item has an ability to get a score that places him/her on the ability scale and so at every ability level there is a probability of answering an item correctly P (ø). Hence, IRT is of the idea that the probability of answering an item correctly is a function of the item characteristics and the ability of the construct the person possesses. The relationship between the probability of correct response and the ability is called item response function, an s-shaped curve called the item characteristics curve (ICC) is used to describe the item response function.

There are three parameters used to describe the ICC. The difficulty parameter of an item which describes where the item function in the ability scale is on the ability scale where the probability of correct response to an item response to an item is 0.5. The more difficult an item is, the higher an examinee's ability must be in order to answer the item correctly. Items with high p-values are hard items, which most examinee's including those with low ability, will have at least a moderate chance of answering correctly. Discrimination parameter a, describes how an item can differentiate between examinee having ability above the item location. This reflects the steepness of the ICC in its middle section; the steeper the curve, the better the item candiscriminate. High discrimination indicates that the higher-scoring examinees tends to answer the item incorrectly.The guessing parameter c is introduced in the model to account for the performance of low ability examinees on multiple choice items, where the low ability students can choose the correct answer by guessing. C values are usually used to the reciprocal of the number of answer choices; a c-value for a four point item will be 25. The c-value can be influenced by random and norm random factors, it is also called the item lower asymptote, because the ICC does not get lower than c no matter how low the students ability.

Standard from IRT

Another way of establishing cut off scores is through the item response theory (IRT). When judges are asked to assess the probability of candidate correctly answering an item, they are in essence determining the difficulty of the item. In effect, the Angoff rating estimates the ability level denoted as θ in item response theory (IRT) of a minimally acceptable examinee (Kane, 1987). Taube (1997) extended this idea by using judges ratings to work backward to calculate b(difficulty) parameters for each item using a Rach IRT model, given by:

P (ø)I = ${}^{1}/_{(1 + \exp\,(-D(\emptyset - b)}$ where P (ø)i is the probability of an examinee with a given 0 correctly answering item I, and D is a scaling constant equal to 1.7.

Instead of calculating the sum of the item probabilities as the cut off score the mean item difficult was calculated.

Standard from the Normative Performance

In establishing the standard of minimally acceptable performance using this method, the performance of students who are already certified as masters of objective are used. These groups are tested on the given objective. The raw scores corresponding to any possible percentage of the students score distribution should be selected to serve as the standard of minimally acceptable performance (Orluwene, 2012).

Statement of Problem

Standardized tests are made to increase "accountability" among educators and students. As such students are expected to meet some standard of proficiency that the tests are designed to assess. Ideally, this standard will be an embodiment of learning objectives. The standard should represent "mastery" of learning objectives or some level of basic proficiency necessary to move unto the next level or function in real world. In effect, establishing a standard can be conceptualized as policy making that has an impact on everyone involved in the testing procedure. In the Nigerian scenario, admission to top Universities, Polytechnics, colleges of education and even secondary schools has become highly competitive and difficult because of the entry level requirements such as that of admission test based on norm cut off scores. One of such test procedure is that, which involves cut off scores to determine who will be qualified to enter a certain college or university after taking an admission test.

In establishing a standard, there is a need to establish the possible and appropriate cutoff scores to give a fair stance between student's ability and prediction of their performance. There are several methods of establishing cut off scores. Methods for setting cut off have suffered several attacks mainly, because regardless of continuous efforts to improve standard-setting methodology, deciding what is appropriate remain very much of subjective judgment. Again finding the appropriate balance between passing those who should fail and failing those who should pass continues to haunt people involved in setting cut off scores. On the other hand, there have been several controversies in the use of cut off scores in the setting of passing scores especially in the admission test of colleges of universities. The problem of the study therefore, is "what could be the most appropriate method to use in the setting of cut off scores". This problem has been a matter of concern to many research workers and people who have vested interest in education. Hence the researcher realizes the importance of making empirical study to answer this kind of inquiry.

Aim and Objectives

The aim of the study is to make a comparative analysis of establishing cut off scores using two standard setting methods. In specific terms, the objectives of the study are:

i. Comparing the norm reference and the Angoff methods of setting cut off scores

ii. To estimate inter rater reliability of the Angoff method.

Significance of the Study

The study is designed to compare the norm-referenced and the Angoff methods of setting cut off scores in deciding what is good enough, hence it could be of great importance to students, teachers, school administrators and other stake holders in education.

The findings will be usedin the future to establish an appropriate method of setting cut off scores in schools and colleges thereby, giving a fair stance between students ability and prediction of their performance. It will also help in making decisions about students who needed to be "relooped." That is students who might be eligible for instructional interventions beyond what would normally be available at regular classroom. These interventions might include recommendations for extra moral class, specially designed after school programmes (other than a special education classification to try to bring the student's performance up to standard).

The findings from the study could guide, refocus, redirect efforts of institutions of learning towards finding the appropriate balance between passing those who should pass in their admission" processes and those who should fail.

The findings-would also provide empirical data as well as information in this area that could stimulate researchers for further studies.

Research Questions

To effectively carry out this study the following questions are posed:

To what extent do the different standard-setting methods result in different standard?

2. What is the inter-rater reliability of the modified Angoff method?

Scope of the Study

The study concentrated on only one academic subject area (mathematics), and a school in Nembe Local Government Area of Bayelsa State was used. There has been a long series cut score usage in this school system, in the admission of their students. This study does not involve itself with students in Primary, Senior secondary and Tertiary schools. The study focused on junior secondary school and other not specified above are not part of the study. Basically, test items are constructed from table of specification from the broad types of tests but only multiple choice test type was use. The test was developed by the researcher. The test consists of approximately 40 multiple choice items. It is designed to assess a set of learning outcome defined by the school. The test was developed and pilot tested within the district. The psychometrics characteristics of the test were of sufficient qualify to justify its administration and use as one element in the identification of students. The choice of basic 9 was that it is an important class not only because of the large number of students who sit for examination but also because of the fact that selection into senior secondary schools in the state and in 'other states even beyond is highly premised on the performance of students in the class.

METHODOLOGY

The objective of the study is to make a comparative analysis of setting cut off scores using the Angoff method and norm referenced method.

A sample of 80 JSS 3 students from government secondary school Nembe in Bayalsa state of Nigeria were selected through simple random technique and used for the study.

The responses were used to establish the norm for the test using the norm referenced method, to establish cut off score for the Angoff method. There were four teachers from the same college who were used as judges to examine the test items. The norm referenced method of standard setting was applied to the raw scores of the students on the multiple choice items of the test

A panel of raters (Teachers) also set the standard using the Angoff method for the same multiple choice questions. The researcher compared the pass/fail rates derived from the norm referenced and Angoff method and also assessed the interrater reliability of the Angoff method.

In the norm referenced method, the standard was determined by plotting the raw scores on a graph then avoiding the tail to exclude outliers and thereafter calculating an adjusted mean. In the modified Angoff method, a panel of four judges participated in the standard setting exercise. All four were experienced in teaching the students and also, were familiar with the curriculum and included a good mix of race and gender. A consensus on the definition of a minimally acceptable student, that is, borderline candidate was reached. Bearing that definition is mind, each rater judged each item and the probability that a borderline candidate would answer the item correctly. All ratings were collected and the mean of each rater total judgmentscores on all forty items were calculated. The mean score indicates in the rater's judgment, is the score that a minimally competent candidate would obtain. Comparision of the two methods, Angoff and norm reference method was done by looking at their percentage agreement, which was determined by calculating the percentage of (Students) that gets the same result (pass or fail) by the 2 different methods. The inter-rater reliability of the Angoff method was checked by using intra class correlation coefficient (ICC) employing the average method of reliability and two-way random effects model. The researcher used this model because the judges were a random sample (of all possible judges) and the questions were also a random selection.

Instrument

A test developed by the researchers was utilized for the comparison of the cut off scores using the Angoff method and the norm-referenced method. The test measures a person's ability to acquire learning. The questions covered topics prescribed in the curriculum. The test had 40 items in mathematics. The goodness of fit of the items and the ability of the persons taking test was done using big step software.

The goodness of fit was used in the study to identify the items that are not fitting for the test because the researcher want to make sure that before using the test for comparison of cut off scores it has established the goodness of fit. The goodness fit is a statistical output that will help us to see the match between the ability of students to the items as of difficulty or easiness. The person taking the test might have a high ability but the test item is easy, so there was no matching between ability of the person who took the test and the test item.

Prior to the setting of cut off scores, the test was first administered to eighty students, item analysis, reliability and norm were established. The test was reliable with crobach alpha of 0.64.

GOODNESS OF FIT USING THE IN MSQ AND IN ZSTD

ENTRY IN.MSQ INZSTD

ENTRY	IN. MSQ	INZSTED
Item 1	1.00	.00
Item 2	1.02	.25
Item 3	1.05	.75
Item 4	1.00	.00
Item 5	.93	-1.33
Item 6	1.00	.00
Item 7	1.03	.41

Item 8	1.00	.00
Item 9	1.00	.00
Item 10	1.00	.00
Item 11	111	.133
Item 12	1.00	.00
Item 13	1.16	.98
Item 14	1.00	1.00
Item 15	1.11	2.19
Item 16	1.00	.00
Item 17	1.4	.71
Item 18	1.00	.00
Item 19	.96	-22
Item 20	1.00	.00
Item 21	1.05	.35
Item 22	1.00	.00
Item 23	1.01	.22
Item 24	1.00	.00
Item 25	.87	-2.39
item 26	1.00	.00
Item 27	.88	-1.77
Item 28	1.00	00
Item 29	.84	-2.90
Item 30	1.00	.00
Item 31	.96	-.76
Item 32	1.00	.00
Item 33	1.00	.00
Item 34	.87	-2.77
Item 35	1.00	.00
Item 36	.99	.18
Item 37	1.00	1.0
Item 38	1.06	1.17
Item 39	1.00	.00
Item 40	1.08	1.11

NOTE: IN MSQ Means Square Fit

INZSTD = Standardized Fit

The table shows the goodness of fit of the test. There are three columns in the table.

The entry pertains to the item number of the test. The IN MSQ pertains to the Means Square fit and ZSTD pertains to standardized fit. An item which has a MSQ of less than 1.3 and ZSTD fit of less than 2.0 is a good fit. In terms of MSQ, we can see that all the items are in good fit, while in terms of ZSTD number 15,25,29 and 34 are not in good fit. Majority of the test items (36) items have a good fit: This is important in the setting of cut off

Scores because it assures that the test is good for the purpose ofsetting cut off scores.

RESULTS

Inter-rater reliability (Angoff method).

Four Judges (raters) were involved in creating the scores for the Angoff method. Average scores of the four raters were found to be 50.00, therefore, the pass mark was set at 50.00. The intra class correction co-efficient (ICC) measured the inter-rater reliability of the four judges. The ICC calculated as the average measure of reliability and by using two way random effects model was 0.64. This indicates a good inter-rater reliability.

The result of the goodness of fit test shows that only four items are misfitting. This indicates that majority of the test items fit or matches the ability of the person who took the test. This is important in this empirical study since it will take away the doubt that the comparism will not be appropriate for a reason that the ability of the person who took the test does not fit the item. Therefore, the goodness of fit test will strengthen the test to be a good instrument to compare the difference in setting of cut off scores using Angoff method and the norm-reference method.

Calculation of Pass Score

To calculate the pass score using the norm-referenced method, we plot the raw scores on a graph and then excluded the extreme 5O% to avoid the influence of outliers. The pass score was set at mean minus ISD.

The pass rates with the norm-referenced method was 85% (68/80) with cut off at (54) and that of the Angoff method was l00% (80 out of 80)with cut off at (50).

The percentage agreement between the Angoff and norm-reference method was 85% (95% confidence interval). Norm-referenced (mean, minus 1.0.SD) as the pass/fail cut off scores was entirely arbitrary (though it is common practice among educationalist).

DISCUSSION/CONCLUSION

The pass rate with the norm-referenced method was 85% and that of the Angoff method was l00% and the percentage agreement between the two was 85%, 95% confidence interval or stated simply, these two standard setting methods yielded different standards. There were significant differences in the out-come of these two standard methods, as shown by the difference in the proportion of candidates that passed/failed the test. The modified Angoff method was found to have moderate inter-rater reliability also. This result concurs with the findings of (Downing, et at, 2006). Verhoren et al (2002) compared pass/fail rate in undergraduate medical assessment and found them to be different. This finding is similar to that reported in previous studies by (Humphrey &Macfayden, 2002).

Although it is now fairly well established that different standard setting methods result in different pass scores, they can be made credible, defensible and acceptable by ensuring the credibility of Judges and using of a systematic approach to collect their judgment.

The result also showed inter-rater reliability of the standard determined by the Angoff method as moderately good (0.64). Wayne etal (2005) noted very good inter rater reliability (ICC 0.88) for theAngoff method. A reliability coefficient of .8 or more is considered satisfactory in high state examination. In trying to compare findings on the reliability of the Angoff method; there is no consensus on the definition of the modified Angoff method. Meaning that no correction was appropriate to guessing, no group discussion and no reality check was done.

Conventional standard setting methods such as the norm-referenced method are arbitrary, whereas the modified Angoff method of standard setting is more objectiveand has good inter-rater reliability. Although the proportion of the candidates that passed/failed variesconsiderably with the method of standard setting used, there was an agreement between the Norm-referenced and the Angoff method ofstandard setting. The Angoff method has self-evident face validity as it replaces grossly arbitrary methods with a reasoned standardized method open to inquiry.

Nonetheless, there is need to further investigate the statistical characteristics of the Angoff method in order to establish its limits and strength. The number of judges/raters who participated in the Angoff method of standard setting was small and this was borne in mind while interpreting the findings. There is no clear consensus among researchers on the most appropriate number of raters/judges, however recognized that a larger panel size would have probably yielded more valid findings. Again the choice of mean minus ISD as the pass/fail cutoff score in the norm-referenced method of standard setting was entirely arbitrary. Downing et al (2006) highlights the key aspect to consider in selecting judges as their content expertise, familiarity with the examinees and good balanced in gender, ethnicity and the panel fulfilled all these requirements. Also the careful attention paid to selecting judges for the Angoff method, as the passing score established is only as credible as the judges.

REFERENCES

Anastasis, A and Urbina, S (2000). *Psychological testing (7th ed.).*New York. Macmillian Publishing Company

Angoif, W.H (1971).*Scales, norms and equivalent scores in R.L Thordike (ed.) Educational measurement (2nd ed).* Washington D.C. American Council on education.

Berk, E.O. (2006). Standards for educational and psychological tests. USA: Freeman and company.

Boursicot K, and Roberts T. (2006). Setting standard in a professional higher education course: Defining the concept of the minimally competent students in performance based assessment at the level of graduation from medical school. *Higher education quarterly, doi 10.1046/,13 65-2923- 2000.00690 X*

Cusimamo M (1996). Standard-setting in medical education. *Acad med. doi 10.1097/0000 1888-199610000-00062 (Publmed) (cross Ref)*

D'Almeida, M (2006).*Standard-setting procedures to establish cut scores for multiple choice criterion reference tests in the field of education*. A comparison between Angoff and Id matching methods: University of British, Colombia.

Downing, S.,Tekian, A. &Tudkowsky R.C. (2006). Procedures for establishing defensible absolute passing defensible absolute passing scores on performance examination in health profession Education. *Teaching and Leaving in medicine 18:50-57 doi: 10.1207 I S 15328015. tim 180-11 (Pubmed) (Crossed)*

George, T. (2006). Statistical methods for psychology. London: International Thomsom Publishing Inc.

Hambleton, R. K &Plake, BIS. (2005). Using an extended Angoifprocedures to set standards on complex performance assessments.*Applied measurement in Education, 8, 41-55*

Humphrey-Murto S., and Macfaden J.C (2002). Standard setting: A comparison of case author and modified borderline group methods in small scale O.S.C.E. *Academic medicine*.

Impara J.C, and Plake BS (2007). Teacher's ability to estimate item difficult: A test of the assumption in the Angoif standard setting method. *Journal of educational measurement.*

Iweka F (2004).*Comprehensive Guide to test construction and administration*. Omoku: Chifas

Kane, M (2004). Validating the performance standard associated with passing scores. *Review of Educational Research.*

Kaufman, D.M., Maun K.V, MuijtjensAmm, and Vader VleutenCpm (2010). A comparison of standard setting procedures for an OSCE in undergraduate medical education.

Kpolovie, P.J (2002). *Test, measurement and evaluation in education*. Nigeria: Emhail.

Lawley, D. N. (1943). On problems connected with item selection and testconstruction. *Proceedings of the royal society of Edinburgh, 61, 273-287.*

Lawley, D. N. (1944). The factorial analysis of multiple item tests. *Proceedings of society of Edinburgh, 63, 74-82.*

Lord, F.M. (1953). The relation of test score to the trait underlying the test. *Educational and psychological measurement, 13, 517-548*

Maurer, T.J, Raju, N.S & Collins, WC (1998): pear and subordinate performance appraisal measurement equivalence.*Journal of Applied Psychology.*

National Policy on Education (2004). FederalRepublicof Nigeria- Lagos. EGN. Press

National Research Council (1999). Setting reasonable and useful performance standard in Pelligrino W, Jones LR, Mitchel K.J. editor. *Grading, the national report card: evaluating NAEP and transforming the assessment of educational process Washington DC*: National Academy Press.

Norcine: J. J, and Shea J.A. (1997): The credibility and comparability of standards applied measurement in Education.

Norcini, T. &Guille, C. (2002). Educational Measurement and testing (2nded) USA: Simon &Schister Inc.

Norcini, T. (2003). Application of item response theory to practical testing problems. Berkley: University of Carlifonia Press.

Orluwene, G.W (2012). *Introduction to test theory and development processes*. Port Harcourt: Chris-Rom integrated services.

Pabico, A p (2008): Improved English proficiency among Filipino adults. surprising, *Available online: http :www.pcj.org//blog/p2330.*

Reeve, C.T. (2002). The influence of test context on item difficulty. Educational and Psychological measurements, vol. 36, 329-337.

Ricker K.L (2002). Setting cut scores: critical review of Angoff and modified Angoff methods unpublished manuscript. University of Canada.

Stone, P and Zieky, C.A (2002). Accessing students critical thinking ability. Journal of educational evaluation. 4(16) 36-40.

Stone, P. and Zieky., C.A (2002). Assessing students' critical thinking ability. Journal of educational evaluation. 4(16).

Talente, G., Haist S.A, and Wilson J..P (2003): A model for setting performance standard for standardized patient examination, evaluation and the Health profession. Doi: 10.1177101632703258105

Taube K. T. (1997).*The incorporation of empirical item difficulty data in the Angoff standard setting procedure*. Evaluation and the Health Professions.

Tiratira, S.O (2009). Test of mathematical abilities, New York: Austin Tx Pro.

Tucker, L.R (1946). Maximum validity of a test with equivalent items, psychometrika, 11, 1-13

Vander Ven, AH.G.S(1980): Introduction to scale. New Yoke Wiley in VANDER LINDER W. J & HAMBLETON R.K (2010): *Hand book of modern item response theory (Eds) NY: Springer*

Verhoeven B.H, Verwijnen G.W, MuijtjenAmm, Scherpbier AJJA, and Vand.er Vleuten CPM (2002): and expertise for an Angoffstandard setting procedure in progress testing. Item writers compared to recently graduated students. *Medical education. Doi 10.1046/D.1365-2923.2002.01301.x*

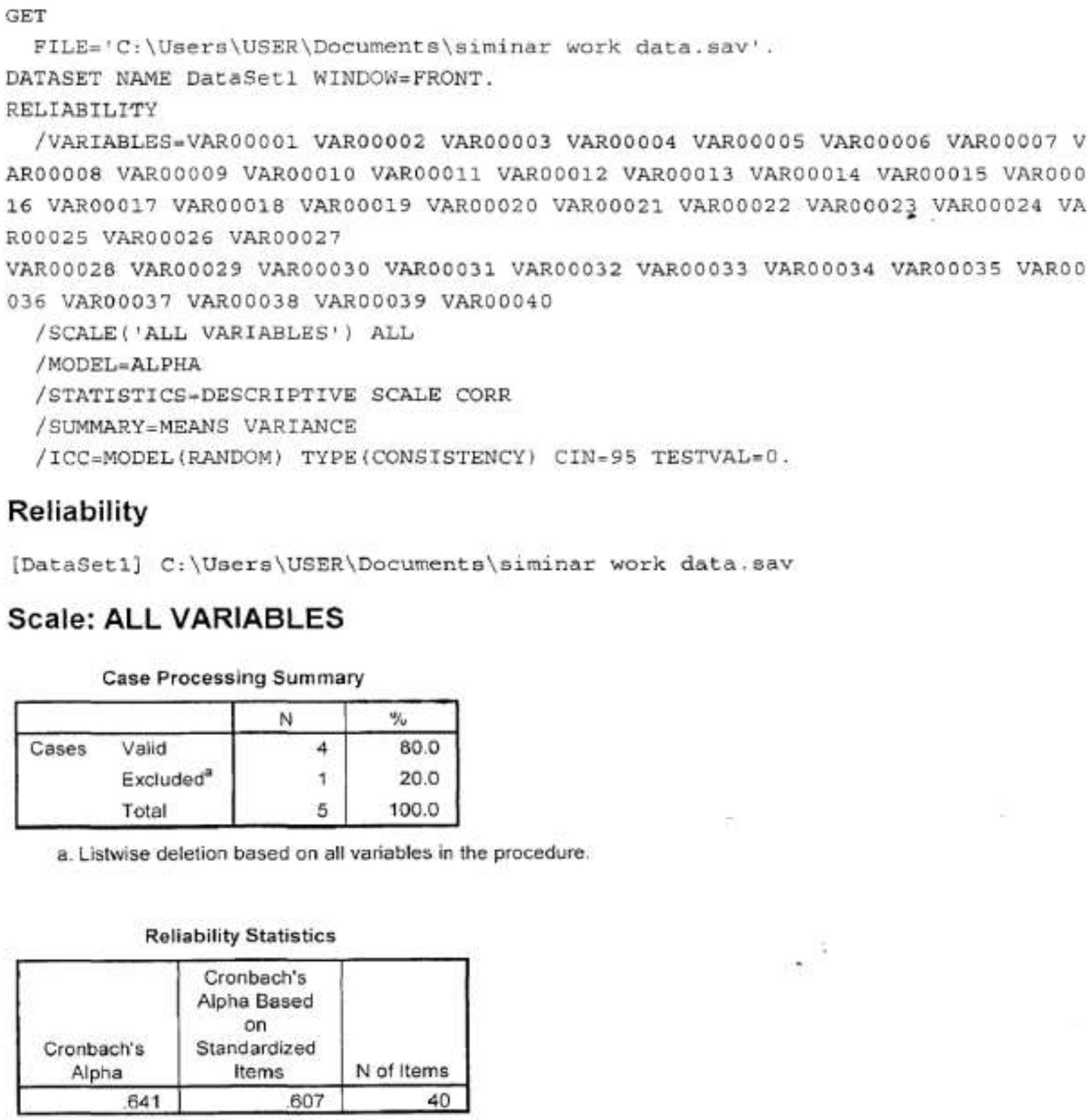
Wayne DB, Fudala M., Buherj, Feingass J, Wada LD, MC GrghieW.C (2005): K Comparism of two standard setting methods for advanced cardiac life support training Academic medicine. doi: 101097 1000 1885-20051000 1-000 18.

Wiersma, W &Jurs, S.G (1990): Educational measurement and testing. boston: allyn and bacon.

Zieky M.J Livingstone S.A (1982) manual for setting standards on the basic skills assessment test, Princeton, N.J; Education testing.

Zieky, M. (2001): So much has charged: How the setting of courses has evolved since the 1980s. in G.Izek (ed), setting performance standards. Cosncepts methods and practices. NJ. Lawrence Er / baum.

APPENDICES



```
GET
  FILE='C:\Users\USER\Documents\siminar work data.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
RELIABILITY
  /VARIABLES=VAR00001 VAR00002 VAR00003 VAR00004 VAR00005 VAR00006 VAR00007 V
AR00008 VAR00009 VAR00010 VAR00011 VAR00012 VAR00013 VAR00014 VAR00015 VAR000
16 VAR00017 VAR00018 VAR00019 VAR00020 VAR00021 VAR00022 VAR00023 VAR00024 VA
R00025 VAR00026 VAR00027
VAR00028 VAR00029 VAR00030 VAR00031 VAR00032 VAR00033 VAR00034 VAR00035 VAR00
036 VAR00037 VAR00038 VAR00039 VAR00040
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /STATISTICS=DESCRIPTIVE SCALE CORR
  /SUMMARY=MEANS VARIANCE
  /ICC=MODEL(RANDOM) TYPE(CONSISTENCY) CIN=95 TESTVAL=0.
```

Reliability

```
[DataSet1] C:\Users\USER\Documents\siminar work data.sav
```

Scale: ALL VARIABLES

Case Processing Summary

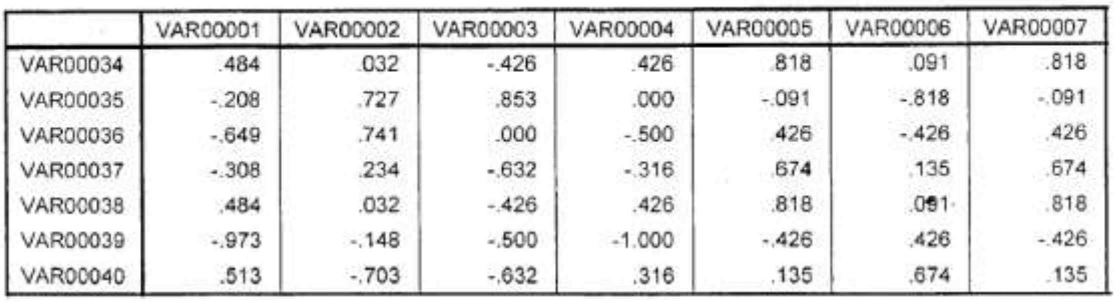
		N	%
Cases	Valid	4	80.0
	Excluded[a]	1	20.0
	Total	5	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

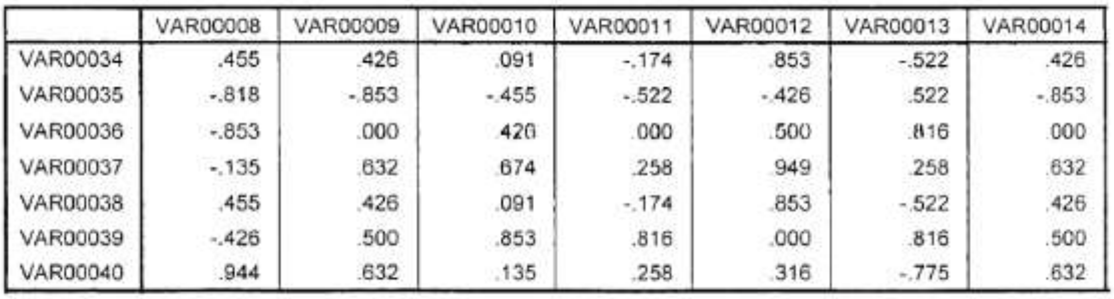
Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.641	.607	40

Inter-Item Correlation Matrix



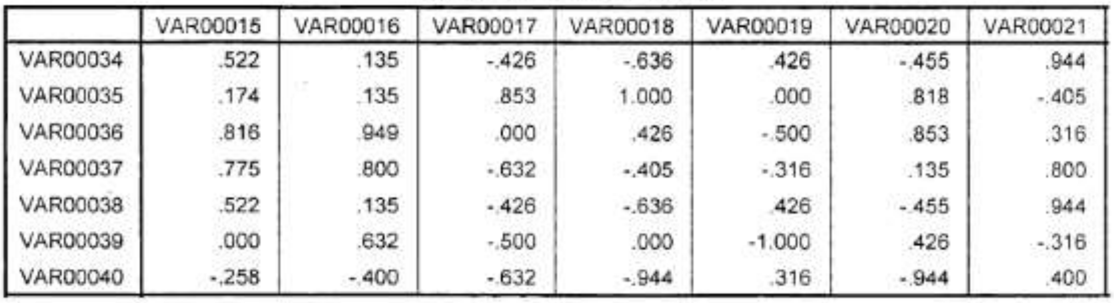
	VAR00001	VAR00002	VAR00003	VAR00004	VAR00005	VAR00006	VAR00007
VAR00034	.484	.032	-.426	.426	.818	.091	.818
VAR00035	-.208	.727	.853	.000	-.091	-.818	-.091
VAR00036	-.649	.741	.000	-.500	.426	-.426	.426
VAR00037	-.308	.234	-.632	-.316	.674	.135	.674
VAR00038	.484	.032	-.426	.426	.818	.091	.818
VAR00039	-.973	-.148	-.500	-1.000	-.426	.426	-.426
VAR00040	.513	-.703	-.632	.316	.135	.674	.135

Inter-Item Correlation Matrix



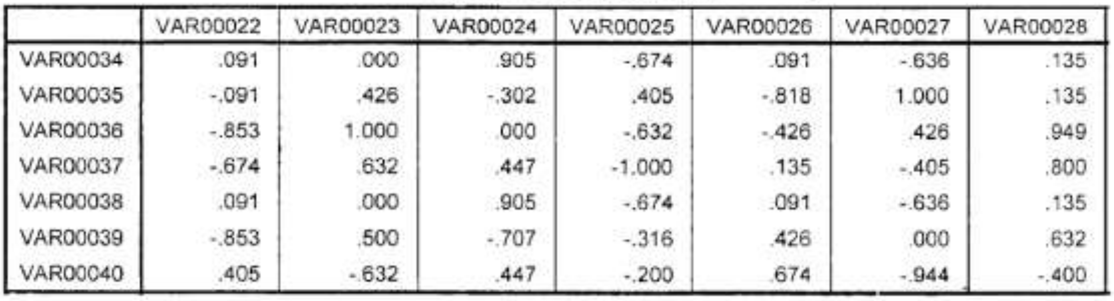
	VAR00008	VAR00009	VAR00010	VAR00011	VAR00012	VAR00013	VAR00014
VAR00034	.455	.426	.091	-.174	.853	-.522	.426
VAR00035	-.818	-.853	-.455	-.522	-.426	.522	-.853
VAR00036	-.853	.000	.426	.000	.500	.816	.000
VAR00037	-.135	.632	.674	.258	.949	.258	.632
VAR00038	.455	.426	.091	-.174	.853	-.522	.426
VAR00039	-.426	.500	.853	.816	.000	.816	.500
VAR00040	.944	.632	.135	.258	.316	-.775	.632

Inter-Item Correlation Matrix

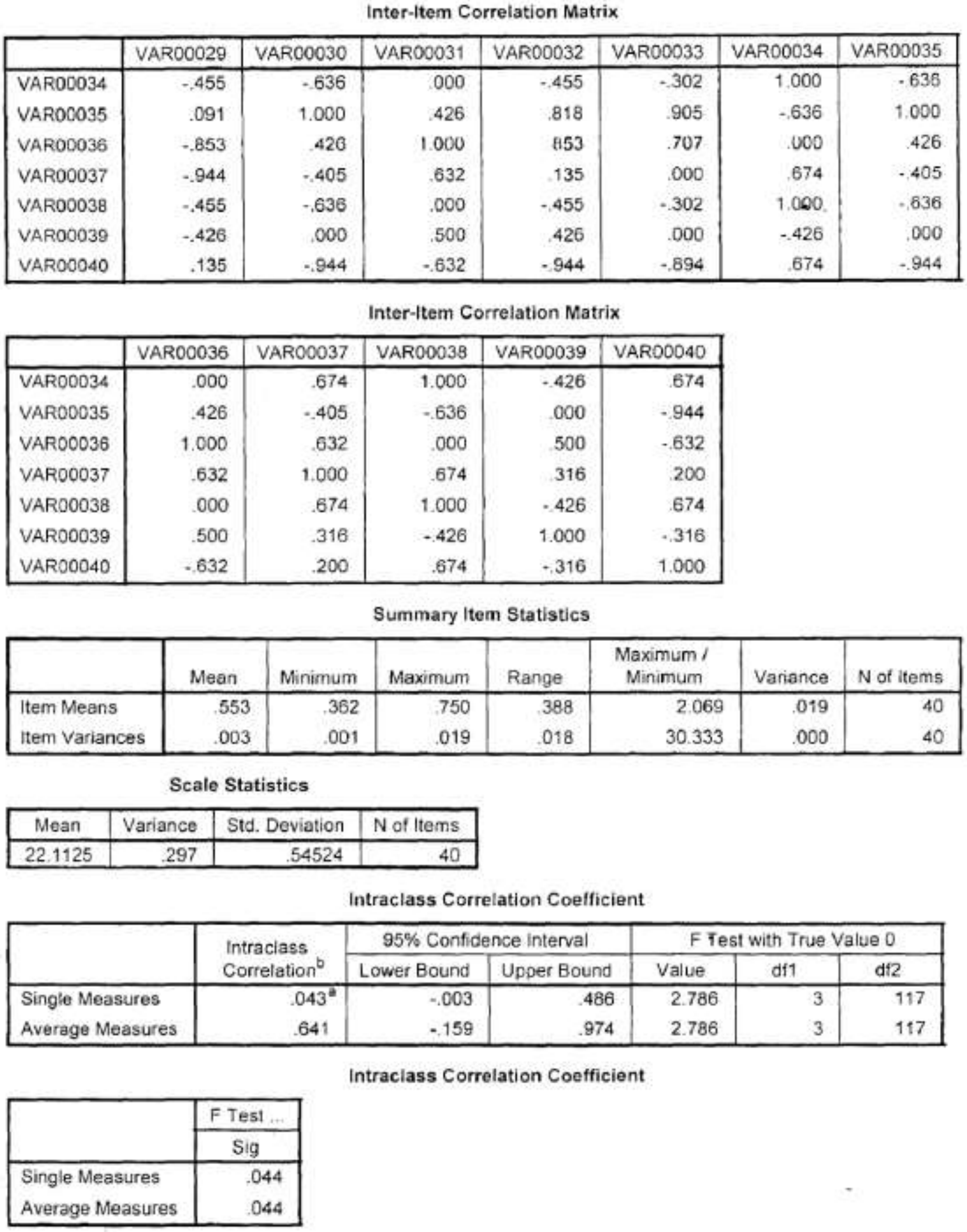


	VAR00015	VAR00016	VAR00017	VAR00018	VAR00019	VAR00020	VAR00021
VAR00034	.522	.135	-.426	-.636	.426	-.455	.944
VAR00035	.174	.135	.853	1.000	.000	.818	-.405
VAR00036	.816	.949	.000	.426	-.500	.853	.316
VAR00037	.775	.800	-.632	-.405	-.316	.135	.800
VAR00038	.522	.135	-.426	-.636	.426	-.455	.944
VAR00039	.000	.632	-.500	.000	-1.000	.426	-.316
VAR00040	-.258	-.400	-.632	-.944	.316	-.944	.400

Inter-Item Correlation Matrix



	VAR00022	VAR00023	VAR00024	VAR00025	VAR00026	VAR00027	VAR00028
VAR00034	.091	.000	.905	-.674	.091	-.636	.135
VAR00035	-.091	.426	-.302	.405	-.818	1.000	.135
VAR00036	-.853	1.000	.000	-.632	-.426	.426	.949
VAR00037	-.674	.632	.447	-1.000	.135	-.405	.800
VAR00038	.091	.000	.905	-.674	.091	-.636	.135
VAR00039	-.853	.500	-.707	-.316	.426	.000	.632
VAR00040	.405	-.632	.447	-.200	.674	-.944	-.400



Inter-Item Correlation Matrix

	VAR00029	VAR00030	VAR00031	VAR00032	VAR00033	VAR00034	VAR00035
VAR00034	-.455	-.636	.000	-.455	-.302	1.000	-.636
VAR00035	.091	1.000	.426	.818	.905	-.636	1.000
VAR00036	-.853	.426	1.000	.853	.707	.000	.426
VAR00037	-.944	-.405	.632	.135	.000	.674	-.405
VAR00038	-.455	-.636	.000	-.455	-.302	1.000	-.636
VAR00039	-.426	.000	.500	.426	.000	-.426	.000
VAR00040	.135	-.944	-.632	-.944	-.894	.674	-.944

Inter-Item Correlation Matrix

	VAR00036	VAR00037	VAR00038	VAR00039	VAR00040
VAR00034	.000	.674	1.000	-.426	.674
VAR00035	.426	-.405	-.636	.000	-.944
VAR00036	1.000	.632	.000	.500	-.632
VAR00037	.632	1.000	.674	.316	.200
VAR00038	.000	.674	1.000	-.426	.674
VAR00039	.500	.316	-.426	1.000	-.316
VAR00040	-.632	.200	.674	-.316	1.000

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	.553	.362	.750	.388	2.069	.019	40
Item Variances	.003	.001	.019	.018	30.333	.000	40

Scale Statistics

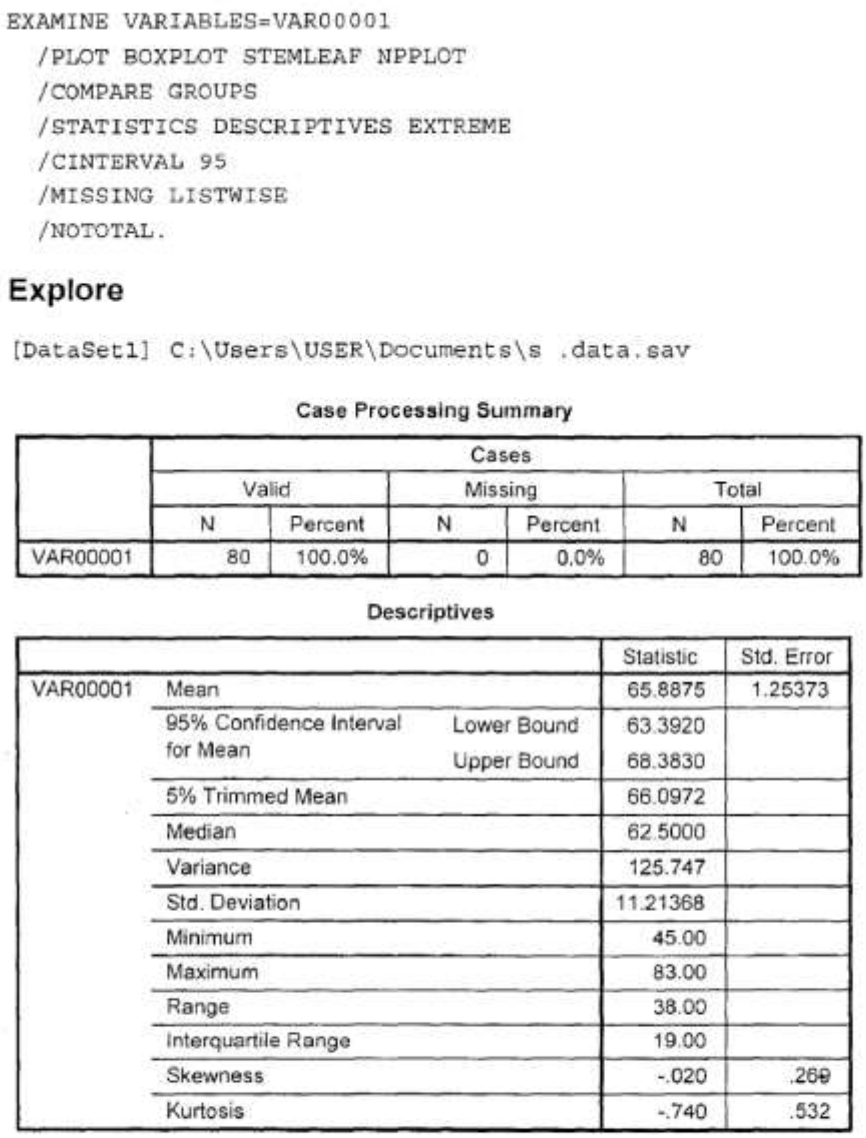
Mean	Variance	Std. Deviation	N of Items
22.1125	.297	.54524	40

Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0		
		Lower Bound	Upper Bound	Value	df1	df2
Single Measures	.043[a]	-.003	.486	2.786	3	117
Average Measures	.641	-.159	.974	2.786	3	117

Intraclass Correlation Coefficient

	F Test ...
	Sig
Single Measures	.044
Average Measures	.044



```
EXAMINE VARIABLES=VAR00001
  /PLOT BOXPLOT STEMLEAF NPPLOT
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES EXTREME
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

Explore

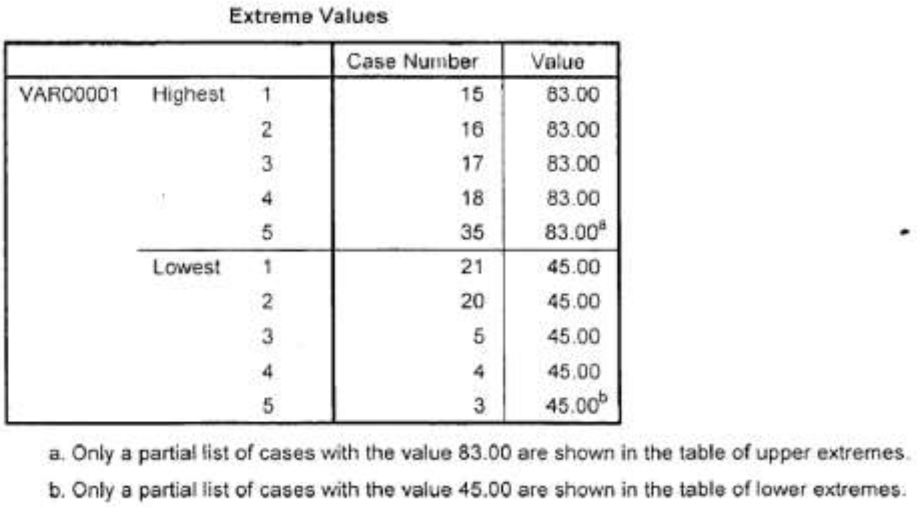
```
[DataSet1] C:\Users\USER\Documents\s .data.sav
```

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
VAR00001	80	100.0%	0	0.0%	80	100.0%

Descriptives

			Statistic	Std. Error
VAR00001	Mean		65.8875	1.25373
	95% Confidence Interval for Mean	Lower Bound	63.3920	
		Upper Bound	68.3830	
	5% Trimmed Mean		66.0972	
	Median		62.5000	
	Variance		125.747	
	Std. Deviation		11.21368	
	Minimum		45.00	
	Maximum		83.00	
	Range		38.00	
	Interquartile Range		19.00	
	Skewness		-.020	.269
	Kurtosis		-.740	.532

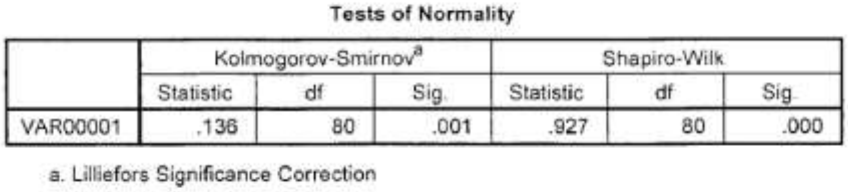


Extreme Values

			Case Number	Value
VAR00001	Highest	1	15	83.00
		2	16	83.00
		3	17	83.00
		4	18	83.00
		5	35	83.00[a]
	Lowest	1	21	45.00
		2	20	45.00
		3	5	45.00
		4	4	45.00
		5	3	45.00[b]

a. Only a partial list of cases with the value 83.00 are shown in the table of upper extremes.

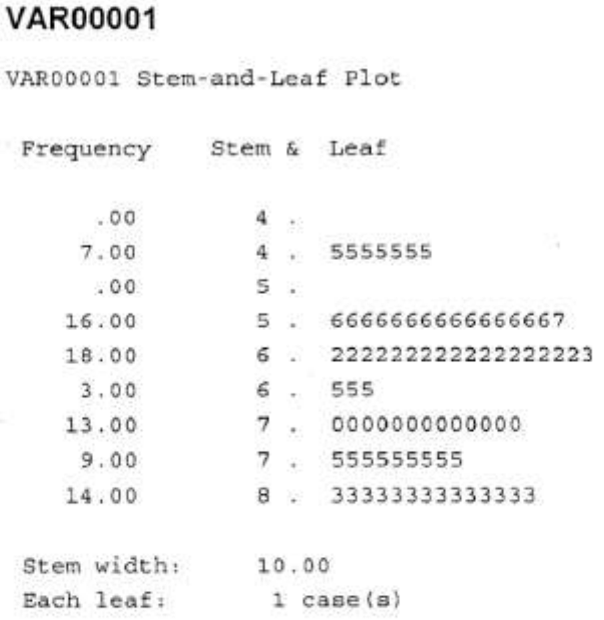
b. Only a partial list of cases with the value 45.00 are shown in the table of lower extremes.



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
VAR00001	.136	80	.001	.927	80	.000

a. Lilliefors Significance Correction



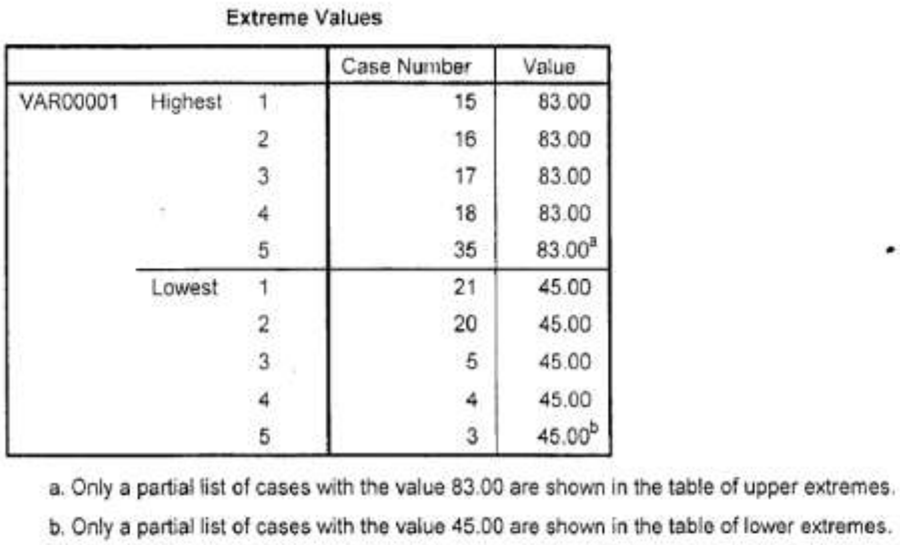
VAR00001

```
VAR00001 Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

      .00        4 .
     7.00        4 .  5555555
      .00        5 .
    16.00        5 .  6666666666666667
    18.00        6 .  222222222222222223
     3.00        6 .  555
    13.00        7 .  0000000000000
     9.00        7 .  555555555
    14.00        8 .  33333333333333

 Stem width:      10.00
 Each leaf:        1 case(s)
```

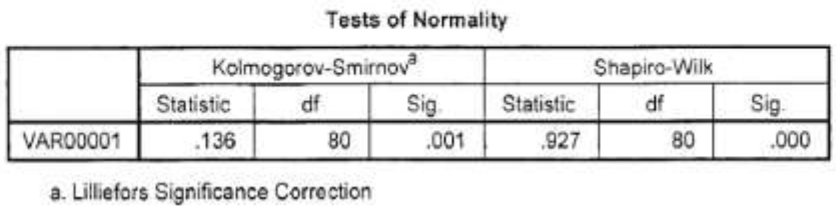


Extreme Values

			Case Number	Value
VAR00001	Highest	1	15	83.00
		2	16	83.00
		3	17	83.00
		4	18	83.00
		5	35	83.00[a]
	Lowest	1	21	45.00
		2	20	45.00
		3	5	45.00
		4	4	45.00
		5	3	45.00[b]

a. Only a partial list of cases with the value 83.00 are shown in the table of upper extremes.

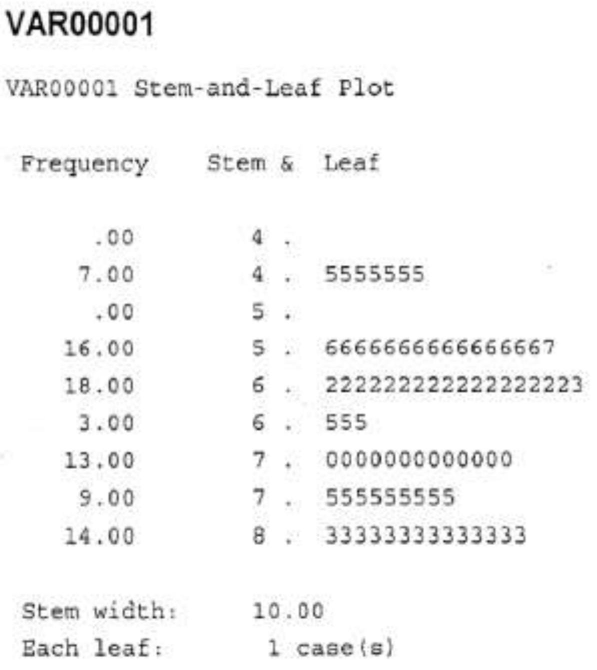
b. Only a partial list of cases with the value 45.00 are shown in the table of lower extremes.



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
VAR00001	.136	80	.001	.927	80	.000

a. Lilliefors Significance Correction



VAR00001

```
VAR00001 Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

      .00        4 .
     7.00        4 .  5555555
      .00        5 .
    16.00        5 .  6666666666666667
    18.00        6 .  222222222222222223
     3.00        6 .  555
    13.00        7 .  0000000000000
     9.00        7 .  555555555
    14.00        8 .  33333333333333

 Stem width:     10.00
 Each leaf:       1 case(s)
```