Using Factor Analysis Procedures to Validate Score Reporting Practice of Large Scale Examinations: Establishing the baseline

Abstract

Performance of candidates in large scale examinations is often reported using a composite score that represents an aggregation of several components of a subject. The components are meant to reflect the fact that subjects are made up of different topics or modalities and each modality is assessed by means of a subset of items. The subsets of items measure a candidates' knowledge with respect to the specific domain. However, more often than not, the construct validity or psychometric independence of each specific domain has not been empirically defined although the domain has intuitive meaning. Factor analysis can be used to make sure that the score reporting practice as indicated by the number of domains is supported by the underlying factor structure. In this paper, Social Studies and Science final examinations test scores were used as dependent variables to extract underlying dimensions. The co-variance matrix for each of the two subjects was submitted to a principal component analysis with Varimax rotation to produce factor loading. The results indicated a unidimensional factor structure for Social Studies and a three component model for Science. The findings were used to evaluate the adopted score reporting structure for each of the two subjects.

Key words: construct validity, score reporting practice, exploratory factor analysis, confirmatory factor analysis, scree plot.

The inferential nature of educational measurement means that the construct of interest to educational researchers are not directly observable. Researchers have to identify metric variables that are sensitive and reliable enough to measure a variety of unobservable constructs. For example, constructs like problem solving, language proficiency, depression, and anxiety have to be measured indirectly using self-report rating scales and a variety psychometric instruments. Mathematics for example might be subdivided into problem solving and computation: In this case, each modality is operating as a subscale and candidates receive a separate score for each dimension. However, empirical evidence has to be provided to support the hypothesized one-to-one relationship between the score reporting practice and the structural nature of the construct being measured. In Botswana, primary school learners

Validity as a unified concept

Under the traditional conceptualization of validity, validity is thought to be made up of three elements; that is face validity, content validity and criterion related validity. According to Nitko (1996), content validity evidence refers to "How well the assessment tasks represent some defined domain of content" (p. 46). Content validity is match more rigorous as items in the test are specifically linked to objectives in the syllabus to make sure that each item targets a specific area of the subject matter. This would ensure that all major areas or domains of a subject are targeted. In an event that some area of the content are not target then the examination runs the risk of exhibiting low content validity. However, the traditional representation of validity is now slowly be replaced by an approach that views validity as a unified concept. Messick (1996) defines validity as "…an overall evaluative judgment of the degree to which empirical evidence

and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment' (p. 6). The author subscribes to a unified concept of validity that is characterized by six elements that must be considered when assessing a measurement scale; these are content, substantive, structural, generalizability, external, and consequential aspects of construct validity. Taken together, the six aspects of validity should function as validity criteria for evaluating psychometric scales (Messick, 1989). Specifically, structural validity demands that the internal structure of the measurement scale be consistent with the theoretical structure of the construct that the scale purports to measure (Messick, 1989). According to Gu, Turkan and Gomez (2015);

> A test's internal structure (or dimensionality) refers to the latent factor structure that underlies observed test performance. The internal structure of the test summarizes the patterns of responses by specifying the nature and number of underlying factors as well the relationships among them (p. 1)

Therefore, construct validation of measurement scales is much more imperative as it is concerned with the extent to which a scale actually conforms to the structure of a construct of interest.

Rationale for the study

When criterion referencing testing (CRT) was introduce in Botswana in 1997 (National Development Plan 8, 1997), each examinable subject was assessed by means of dimensions that were generated by a panel of subject experts. Although the dimensions had intuitive meaning, there were no studies done to test their psychometric validity. The rationale for the current research is to empirically establish the structural validity of the dimensions for social studies and

science using 2004 data. The year 2004 was chosen as a way of establishing a baseline that will serve as a reference point for future research work.

Conceptual framework

Assessment instrument such as tests and examinations are often made of several metric variables that are designed to measure a construct. The variance within each metric variable can be partitioned into systematic and error variance. While the error variance is a random factor, the systematic variance is attributable to the underlying causative agent referred to as a dimension or a factor. Structural equation modelling procedures (i.e., principal component analysis or maximum least squire) group the metric variables according to their co-variance to generate meaningful patterns or dimensions (Rummel, 1967). The extracted dimensions, if each is identified by several high loading metric variables, form the measurement model of the instrument. Fig 1 depicts a situation where 60 metric variables in an English Language test are grouped into four dimensions; the four dimensions are correlated and each dimension is identified by 15 questions. This model provides a conceptual framework for the current study.

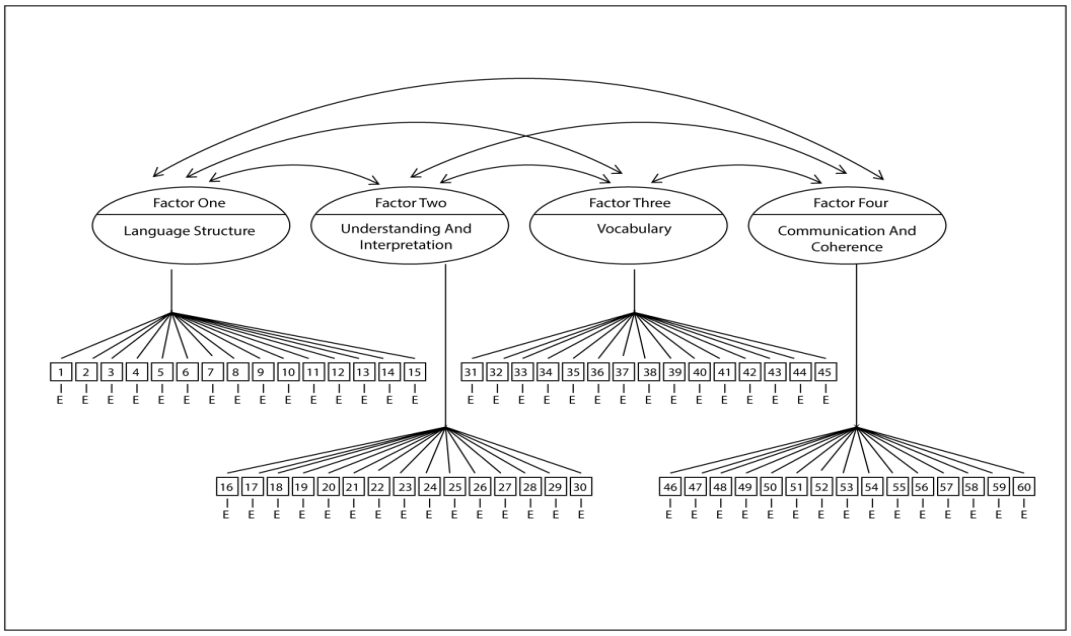

Fig 1: Factor analysis conceptual model

Statement of the problem

Score reporting practice adapted by assessment institutions should be backed by empirically determined measurement model. Lack of a well-established structural model is a cause of concern since the reported scores might be an over-estimation or underestimation of the actual measurement model of the subject being assessed. An over-estimation actually means that candidates are graded on a dimension that is not substantively meaningful while under-estimation results in a situation where candidates are not credited for having demonstrated mastery of an essential skill. In both of the two scenarios (over-estimation and under-estimation) the assessment process not only lacks validity but it is also unfair to the learners.

Purpose of the study

The purpose of the study is to apply principal component analysis procedures to determine the underlying factor structure of two subjects (social studies and science). The rotated solution will also assist towards generation of appropriate labels for each factor.

Research Questions

The analysis will be directed towards answering the following research questions;

- How many factors explain performance of candidates in PSLE social studies?
- How many factors explain performance of candidates in PSLE science?
- How many items have significant loading on each of the extracted factors?

The determination of the most reliable underlying factor for each subject will ensure that the grades assigned to candidates reflect what they know, understand and can do. Secondly, the diagnostic potential of the examination papers will be enhanced as feedback information provided to learners and teachers will be packaged according to the recovered dimensions. Teachers will be able to identify concepts that need more attention in the classroom and design appropriate remedial instructions.

Literature

Researchers in the field of Counselling and Human Services have carried out interesting factor analytic studies to understand the dimensionality of different self-report scales. An example is a study done by Breidenbach and French (2012) to confirm the dimensional structure of the Brigance Comprehensive Inventory of Basic Skills II (CIBS-II; Brigance, 2010). The scale was developed to "identifying student achievement, identify and monitor academic strengths and weaknesses, obtaining data to support referrals, and reporting progress for individual educational plans" (Breidenbach & French, 2012, 478). The scale has nine subtests which reduce to five dimensions namely Basic Reading, Reading Comprehension, Mathematics, Written Expression

and Listening Comprehension. Though the test enjoys wide applications in the learning and teaching environment due to the fact that it can be administered by classroom teachers; the construct validity of the test was in question (Breidenbach & French, 2012). Presentation of a clear factor structure of the test was meant to enhance the construct validity of the instrument and ensure appropriateness of different interpretations of scores generated from the test. Breidenbach and French (2012) employed a confirmatory factor analysis (CFA) to assess the internal structure of the instrument. The research question for the study was, "To what extent does scores from the standardization sample of the CIBS-II support the composite scores structure provided by the publishers?" The participants in the study were drawn from four main regions of the USA (i.e., South, Midwest, West and Northeast). A total of 1411 participants were included in the study; the sample was divided into two subgroups for cross-validation purposes.

At the beginning of the study, the researchers specified four competing models; Model 1 hypothesized a single factor structure which essentially argued for a single score reporting practice. Model 2 tested for the existence of five underlying factors that corresponded to the five subtests (i.e., Basic Reading, Reading Comprehension, Mathematics, Written Expression and Listening Comprehension). Model 3 allowed the first order factors in Model 2 to load on a higher order factor while Model 4 was a three factor model created by assuming that Basic Reading, Reading Comprehension and Written Expression will load on a single factor (Breidenbach and French, 2012). Multiple indices were used to evaluate model fit; for a model to be judged as being the best fitting one, all the fit indices had to be satisfied and the residual for good-fitting model should exhibit normality characteristics. The chi-square test was particular useful in this case because Model 3 was nested within Model 2.

The results showed that Model 1 and Model 3 had very poor fit while Model 4 produced an inadmissible final solution where the correlation between Reading and Writing factors was greater than 1 (Heywood case). Model 2 met all the fit criteria but the only major problem was the structure coefficients between subtest and large residual variance indicating non-normality. The high intercorrelation between subtests was an indication of possible presence of a high order factor; so, the researcher decided to re-specify the model (Model 2A) allowing covariance of errors between highly related subtest (e.g., Spelling and Word Recognition). The re-specified model produced excellent fit indices as residual were approaching normality (Breidenbach & French, 2012). The cross validation study using the other half of the sample also indicated that Model 2A was a good fit to the data. Therefore, the researchers were able to successfully demonstrate the fidelity of the five score reporting policy thus showing that the construct validity evidence of the test was high.

Another informative study was conducted by Alavi, Kaivanpanah and Nayernia (2011). These researchers investigated the factor structure of the Tehran English Proficiency Test (UTEPT) the purpose of which was to "test if the underlying factor structure obtained in a data driven manner corresponds with the proposed structure of the test." (Alavi, Kaivanpanah and Nayernia, 2011, p. 27). The test was designed to assess the examinee's knowledge of grammar, vocabulary, and reading comprehension. The implicit assumption in this case was that performance in the test can be differentiated into three distinct components thus justifying the award of three separate scores for each modality. Thus it would be possible to extract three separate factors corresponding to the three sections of the test.

The sample of the study comprised 850 participants randomly sampled from a population of 3000 students. The data was first tested for conformity to multivariate normality, the

normality assumption was met as the "skewness and kurtosis were within the recommended limits" (Alavi, Kaivanpanah and Nayernia, 2011, p. 39). The analysis of the estimated higher order model was assessed for data fit using multiple fit criteria involving both statistical and practical indicators. The chis-square goodness of fit was acceptable and the three practical fit indexes showed excellent fit. The conclusion reached was that the UTEPT data showed very good fit to a model with three first-order factors and a higher order factors. The three first-order factors are Reading, Vocabulary, and Structure. The factor loadings for the measured variables loaded significantly on their respective target factors. Therefore, the scoring structure of UTEPT can be said to possess construct validity as the test sections have fidelity to the structure of the construct. The empirical evidence and theoretical rationale justify the intended score interpretations and use (Messick, 1989).

Kuriakose (2011) carried out a study to examine the underlying factor structure of English Language Development Assessment (ELDA). ELDA is a high stake language proficiency test that is used by several states in the USA for purposes selection and placement. The test not only provides information that can be used for instructional intervention but it is a direct implementation of the No Child Left Behind Act of 2001. According to Kuriakose (2011) "Standardized assessment results became an integral part of accountability after NCLB (2001) mandated that all states develop an assessment system aligned to the state standards and required that all students be tested…" (p. 2). A number of states in America (i.e., Arkansas, Iowa, Louisiana, Nebraska, South Carolina, Tennessee, and West Virginia) adopted ELDA as a standard assessment tool for their learners. However, the states applied different grading procedures thus creating a situation where there was lack of consistency in the interpretation and use of scores across states. The purpose Kuriakose's research was to analyze data from these

states in order to establish a common underlying factor structure that would then inform grading and reporting practices in all the participating states.

ELDA assesses language proficiency in the four main areas of reading, writing, speaking, and listening. The test has two components; the first component is administered to leaners in grade 3 to 5 (n = 4577) and the second component is administered to the 9 to 12 grade cluster (N = 2330). This stratified sample allowed the researcher to test the hypothesis about the generalizability of the recovered factor structure across proficiency levels. CFA modelling procedures based on the maximum likelihood estimation was used to recover salient factors in the ELDA data. Four different models were hypothesized and subsequently tested and model fit was evaluated using several goodness-of-fit indices. The first model tested hypothesized a single language proficiency factor where all measured variables were constrained to load on only one language factor. The second model explained the covariance within the measured variables in terms of two correlated factors; the first factor was created by combining listening and speaking modalities while the second factor was created by combining reading and writing domains. The model fit indices indicated a poor fit for both the single factor and the two-correlated factor models.

The third hypothesized model argued for the existence of a higher order language ability factor while the fourth model was a bi-factor solution that allowed the higher order factor to have a direct influence on the metric variables. The fit statistics for the bi-factor model were superior when compared to the fit statistics for other models. Generally, the results of the study showed that English language as assessed by ELDA had a hierarchical and multidimensional underlying structure as well as the fact that the four modalities are equally weighted. The findings from the

study had far reaching implications as some of the states in the USA were reported to be using a grading system that give more weighting to reading and writing (Kuriakose, 2011).

In Botswana, primary school leaving candidates write two separate final English Language examination papers (e.g. Multiple choice paper with 60 items and a continuous writing component made up of composition and letter writing modalities). Each continuous writing modality is assessed by means of 10 criteria. The maximum score that a candidate can obtain for continuous writing is 20 and all the 20 criteria are assumed to be tapping on a single language use construct. The categorization of all the 20 criteria under one dimension implies a unidimensional structure for continuous writing. The unidimensional model has not been verified empirically and there is no theoretical justification of its use. Mogapi (2016) carried out an EFA study to establish the dimensionality of the continuous writing component using 2003 PSLE data. The study derived the sample from four educational regions within South Central District in Botswana. Simple random sampling procedures were applied at the level of schools. A total of 22 schools were randomly sampled from the South Central District making a total sample of 1800 candidates.

Exploratory Factor Analysis estimation procedures with Varimax rotation were used to identify covariance within the 20 metric variables used to assess composition and letter writing proficiency. Multiple model fit criteria were used to extract reliable factors that best represents the language proficiency construct. The suitability of the data for factor analysis was assessed by means of Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) and Bartlett's Test of Sphericity. The Kaiser criterion (K1 rule) and the Scree Plot were used to extract reliable factors from the matrix. Four factors satisfied the KI rule and their cumulative variance explained was 48.68% (Mogapi, 2016). The Scree plot tended to corroborate the four factor solution.

Since the analysis converged on four dimensions, the next logical stage was to find an appropriate name for each of the dimensions. Dimension 1 was named 'Logical Development of Ideas' because most of the items that load on this dimension dealt with the ability of the candidate to put ideas in a logical and coherent manner. The Dimension 2 was named 'Communication of Feelings' as the majority of items associated with the dimension required the candidate to show their feelings and/ or emotions. Dimension 3 and Dimension 4 were labelled as 'Correct Use of Language Devices' and Appropriate Register'. Dimension 3 deals with correct use of adjectives and conjunctions while Dimension 4 examines the ability to write an address and salutation correctly when writing a letter. Mogapi (2016) made a conclusion to the effect that: "Research on the dimensionality of language proficiency strongly points to the multidimensional nature of the language proficiency construct." (p. 18). It would appear that reporting four separate dimensions is more in line with the theoretical structure of the construct than the current two dimensions. However, the validity of the four factor solution for English Language remains to be confirmed as the study only used 20 continuous writing items out of a total of 80 items. Further research in this area is encouraged.

Methodology

Studies that use factor analysis can either be categorized as exploratory or theory driven; in some cases both exploratory and data driven approaches are integrated within the same study (e.g., John, Cho, Ling, Steinberg, & Stone, 2008). The current study follows the exploratory route to determine the number of underlying dimensions using examinations scores for social studies and science subjects. However, the unit of analysis in the case is the school and not the individual candidate. Using the school as the unit of analysis minimizes random error associated with each candidate but at the same time reduces variability with the group thus increasing the likelihood of under-factoring. The 2002 Ministry of Education report shows that there were a total of 770

primary schools in Botswana. Due to the fact that factor analysis is a large sample technique, purposive sampling was used to select schools with more than 30 leaners. This sampling procedure leads to the selection of 150 schools that were distributed across all 10 educational districts in the country.

Instrumentation

Every year primary school leaving candidates sit for final examination covering five subjects; English, Mathematics, Setswana, Science, and Social Studies. All examinations papers have 60 multiple choice questions with the exception of English and Setswana papers (the two subjects have an additional written paper). The current study only used examinations results for social studies and science for year 2004. School level data was used to uphold the principle of anonymity and confidentiality of information. The data does not reflect names of candidates and school are identified by serial numbers to make it impossible to link a set of data with a particular school. The covariance matrices for social studies and science scores were subjected to principal component analysis with Varimax rotation. The two data sets were tested for their suitability for factor analysis using Bartlett's Test of Sphericity and Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy. According to Williams, Brown, and Onsman (2010), "The KMO index ranges from 0 to 1, with 0.50 considered suitable for factor analysis. The Bartlett's Test of Sphericity should be significant ($p<.05$) for factor analysis to be suitable." (p. 5). The number of reliable underlying dimensions were determine using the eigenvalue value greater than one criteria (Kaiser, 1960) as well as an assessment of the scree plot (Cattell, 1960; Horn, 1965). Factors that occurred before the point of infection were judged to have substantive meaning.

Ethical Considerations

The current study presents an analysis of final examinations scores for social studies and science. To maintain the ethical principles of confidentiality and anonymity, the school was used as the unit of analysis. This was meant to protect the identity of learners who responded to the questions. Secondary, the data were aggregated at a national level to generate intercorrelation patterns and factor loadings. Therefore, the results cannot be directly linked to individual leaners or sampled schools but shows the dimensional structure of each subject.

Data Presentation and Discussion

The 2004 social studies data were subsequently analyzed using principal component analysis techniques with Varimax rotation. According to Fig. 1, the KMO value is .922 indicating a marvellous factorability index. In other words, factor analyzing the data will result in extraction of a considerable amount of variance in the matrix. The Bartlett's test has a significant value of .000, this is a strong indication that the sample intercorrelation matrix did not come from a population in which the intercorrelation matrix is an identity matrix. The analysis finally converged on one dominant factor that accounted for 43.39% of the variance in the covariance matrix. However, there are several smaller factors that satisfy the eigenvalue greater that one rule. The substantive reliability of these smaller factors is very low as they are identified by few items; retaining such factors for further analysis may lead to overestimation of the underlying factor structure. The scree plot (Figure 2), on the other hand, clearly shows the existence of one dominant factor thus supporting a unidimensional hypothesis. Thus, it can be tentatively concluded that social studies subject as tested using the 2004 examination paper is a unidimensional construct. The evidence as indicated in Fig. 1 and Fig. 2 is not in line with a score reporting practice for social studies where candidates were given three scores in the areas

of knowledge, understanding, and application. This calls for revision of the dimensions with a view to formulating domains that will capture the cognitive categories inherent in the subject.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.922
Bartlett's Test of Sphericity	Approx. Chi-Square	6935.003
	df	1770
	Sig.	.000

Fig 2: KMO an Bartlett's Test

Table 1: Total variance explained for social studies paper

	Component	Initial Eigenvalues		
		Total	% of Variance	Cumulative %
Raw	1	3315.290	43.395	43.395
	2	440.469	5.765	49.160
	3	325.756	4.264	53.424
	4	265.561	3.476	56.900
	5	215.822	2.825	59.725
	6	205.884	2.695	62.420
	7	188.281	2.464	64.884
	8	168.138	2.201	67.085
	9	153.603	2.011	69.095
	10	139.295	1.823	70.919
	11	129.464	1.695	72.613

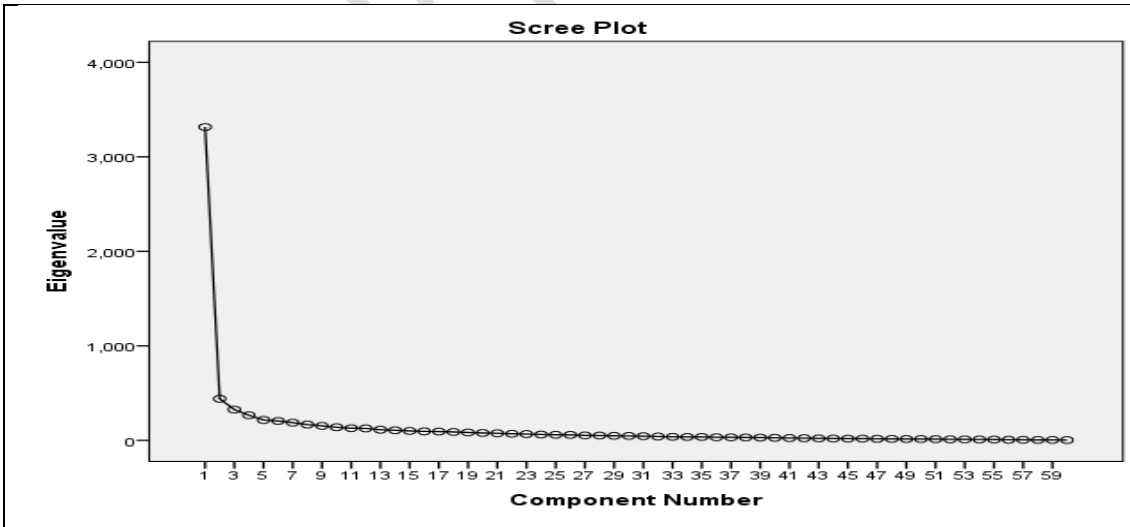

Fig. 3: Social Studies Scree Plot

Science Results

The 2004 science scores were also subjected to exploratory factor analysis; these data were also assessed using KMO and Bartlett's test of sphericity to determine the suitability of the matrix. The values shown in Figure 2 for both indictors were within acceptable margins. The final solution for science indicated three very distinct factor accounting for 34.30%, 22.40% and 5,50% of the variance in the matrix respectively. Although there are 13 additional dimensions that could be recovered using the engine value greater that one rule, they all appear to lack theoretical relevance and as such can be dropped. The three factor solution suggested by the data is aligned to the score reporting structure that was used to general the science examination in 2004 divided the content into three dimensions. The dimensions are mentioned as Knowledge, Understating and Application. A conclusion can be tentatively made that the 2004 science examination had high construct validity.

However, there is a noticeable disparity in the way the items are distributed across the three extracted dimensions. Dimension 1 and Dimension 2 are indicated by 32 items and 22 items respectively while Dimension 3 has only 1 item that have a significant loading on it. Under Dimension 3, candidates are required to demonstrate ability to transfer scientific concepts to a new situation, use information to identify trends, process information from a variety of sources and make appropriate conclusions. These skills cannot be adequately assessed by a single item. Construct underrepresentation normally occurs in a situation where few items are used to assess a broad domain. Using a single item or few items to assess a broad domain such as application may also result in scores that fail to differentiate between students who have mastered the content and those with partial mastery of the subject matter. Generally, a conclusion can be

reached to the effect that the 2004 PSLE science scores reflect the underlying structure of the construct. The only major concern is the apparent underrepresentation of Dimension 3.

Current

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.913
Bartlett's Test of Sphericity	Approx. Chi-Square	6985.129
	Df	1770
	Sig.	.000

Fig 4: Apparent underrepresentation of Dimension 3

Table 2: Total variance explained for the science paper

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	4124.055	34.295	34.295
2	2692.246	22.388	56.683
3	659.451	5.484	62.167
4	353.067	2.936	65.103
5	318.881	2.652	67.755
6	247.267	2.056	69.811
7	220.403	1.833	71.644
8	215.089	1.789	73.433
9	196.785	1.636	75.069
10	175.600	1.460	76.529
11	172.216	1.432	77.961
12	155.592	1.294	79.255
13	145.217	1.208	80.463
14	131.376	1.093	81.555
15	129.135	1.074	82.629
16	116.237	.967	83.596
17	112.191	.933	84.529

Fig 5: Science Scree Plot

Table 3: Science Dimensions

Science Dimensions								
Dimension 1						**Dimension 2**		**Dimension 3**
11(.855)	16(.830)	17(.820)				21(.850)	40(.84	09(.991)
10(.790)	Xx(765),	29(.750)				**22(.732)**	**53(.714)**	
46(.745)	44(.744)	37(.733)				**57(.669)**	**20(.668)**	
34(.718)	58(.710)	13(.692)				**55(.622**)	52(.615)	
47(.691)	**30(.678)**	**42(673)**				38(.607)	**59(.605)**	
28(.666)	6(.641**)**	**xx(.629)**				Xx(.583)	51(.568)	
32(.614)	**50(.600**)	31(.594)				41(.555)	33(.549)	
18(.582)	43(.570)	**14(.561)**				**48(.519)**	49(.510)	
54(.557)	24(.551)	**12(.536)**				26(.496)	39(.491)	
7(.523)	5(.501)	5(.479)				35(.465)	19(.458)	
60(. 471)	23(.459)					45(.458)	08(.449)	

Conclusion

Dimensions for each subject are usually generated by a panel of subject experts and are useful assessment instruments as they provide some guidance as to how the subject should be assessed. To enhance the inferential meaning of test scores, the psychometric validly as well independence

of each dimension has to be established through scientific investigation. In this paper two subjects were used to illustrate the validation process. In the case of social studies, the three subject dimensions were not confirmed by the data. Therefore, there is need to review the dimensions so as to come up with those that have substantive meaning. The analysis of science data yielded a much more prosing situation; three dimensions were extracted from the covariance matrix. Although Dimension 1 and Dimension 2 were strongly represented, Dimension 3 was only indicated by just one item. The reliability and construct validity of the item can only be enhance by making sure that more items tapping on the construct are developed during the item construction stage. Measurement experts have been advised to always evaluate their measurement instruments or scales and use the evidence collected to answer a range of validity questions. One of the questions posed by Messick (1993) is: "Does our way of scoring reflect the manner in which domain processes combine to produce effects and is our score structure consistent with the structure of the domain about which inferences are to be drawn or predictions made?" (p. 2). Both quantitative and qualitative data are required to address this concern.

Recommendations

Only two subjects out of a total of five examinable subjects were included in the study. There is need for a large scale study to establish baseline factor structure of all examinable subjects and ensure that the adopted score reporting model for each subject is correctly aligned to the empirical structure of the subject. The obtained covariance matrices can also be tested for group invariant to further enhance the consequential validity of the interpretations derived from the final examination scores.

. **References**

Alavi, Kaivanpanah, & Nayernia. (2011). Factor structure of a written English proficiency test: A structural equation modeling approach. *Iranian Journal of Applied Language Studies*, *3*(2), 28-50.

Brigance, A. H. (2010). *Comprehensive Inventory of Basic Skills—II.* North Billerica, MA: Curriculum Associates.

Cattell, R. B. (1966). The scree test for the number of factors. Multivariate Behavioral Research. 1, 245- 276.

Gu, L., Turkan, S., & Gomez, P. G. (2015). Examining the internal structure of the test of English for Teaching (TEFT). New Jersey: Educational Testing Service.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30(2), 179-85.

John W. Young, Yeonsuk Cho, Guangming Ling, Fred Cline Jonathan Steinberg, & Elizabeth

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20,* 141-151.

Kuriakose, A. (2011). The factor structure of English language development assessment: A confirmatory factor analysis. Unpublished doctoral dissertation, Arizona State University, Arizona.

Messick, S. (1989).Validity. In R. L. Linn (Ed.), Educational measurement(3rd ed., pp. 13-103). New York: Macmillan.

Messick, Samuel. (1993). Foundation of validity: Meaning and consequences in psychological assessment. Educational Testing Service, Princeton: New Jersey.

Mogapi, M. (2016). Establishing the assessment model for English language continuous writing component. International Journal for Scientific Research in Education, 9 (1), 7-19. Retrieved from: http://www.ijsre.com/assets/vol.%2C-9(1)-mogapi%2C-m.-o..pdf

Republic of Botswana, National Development Plan 8, 1997/8-2002/3. Ministry of Finance and Development Planning. Gaborone: Government Printer.

Nitko, A. J. (1996). *Educational Assessmwent of students*. (2nd ed.). Prentice-Hall: New Jersey.

Reublic of Botswna, Ministry of Education. (2002). Primary school leaving examination results. Gaborone: Governmnt Printer.

Rummel, R. J. (1967). Understanding factor ananlysis. *The Journal of Conflict Resolution, 11 (4),* 444- 480. Retrived from: http://jcr.sagepub.com/content/11/4/444.extract

Williams, B., Brown, T., & Onsman, A. (2010). Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine, 8*(3). Retrieved from http://ro.ecu.edu.au/jephc/vol8/iss3/1