Direct RESEARCH JOURNALS
Direct Research Journal of Social Science and Educational Studies
Vol.6 (2), pp. 37-43, March 2019
ISSN 2449-0806
DOI: https://doi.org/10.26765/DRJSSES.2019.1974
Article Number: DRJA2843351974

http://directresearchpublisher.org/journal/drjsses/
Research Paper

Behavioural Model for Measuring Teachers' Qualities in Uganda: Rasch Analysis

Pius Ochwo

University of Kisubi, P.O Box 182, Entebbe – Uganda.
Author E-mail: bropius@gmail.com

Received 25 February 2019; Accepted 17 March, 2019

The purpose of this study was to develop a behavioral frequency measure of teacher quality for upper primary schools (i.e., grades 5 to 7) in Uganda. The main research questions addressed the following: What are the psychometric properties of a newly developed measure of teacher quality in Uganda? The study utilized a descriptive survey design, and on the basis of purposive technique, 36 teachers revealed information through the study questionnaire and interview guides. Using Rasch modeling, the results rendered a 38-question measure focusing on four domains (1) teacher planning and preparation, (2) classroom environment, (3) teacher instruction, and (4) teacher professional ethics. It was found that the 38-item Teacher Quality Measure (TQM) produced acceptable psychometric properties (e.g., Coefficient α = 0.88). The study concluded that a psychometrically sound behavioural model of teacher quality can be developed for teachers in Uganda.

Keywords: Behavioural Model, Teachers' Qualities, Rasch Analysis, Uganda

INTRODUCTION

One of the most popular ways of measuring Teacher quality in Uganda today, is through academic achievement, and is usually determined by the scores from the State exam. The higher the grades, the more it is assumed that the teacher is effective and the greater excellence attributed to the school (Owen 2005:312). The Uganda National Examination Board (UNEB) is mandated to measure academic achievement in Uganda. The Primary Leaving Examination (PLE) can rightly be said to be a "measure" of teacher quality to a greater extend in Uganda, as it monitors and evaluates academic achievement. Acana (2005), however, observes that Primary Leaving Examination is basically intended for the selection of kids for the Secondary School level, and other post Primary Vocational Institutions. A number of research done on students' achievement, indicate that PLE cannot be the best measure of students' achievement or teacher effectiveness, as other factors such as illness may prevent a brilliant pupil from taking the test or performing as expected. Globally, there exists some research that has attempted to develop measures of teacher quality, or also labeled as teacher performance or effectiveness. It has been challenging to quantify and measure the construct and researchers have used several different methods including observational evaluation rubrics, portfolios, and value-added calculations for student achievement (Raudenbush, 2004; Schacter and Thum, 2004; Tucker *et al*., 2003). Several large-scale teaching surveys exist that focus on measuring reform-oriented practices or enactment of curriculum from the national center for education statistics (NCES), the trends in international mathematics and science study (timss), and the surveys of enacted curriculum (sec). However, as expected, there are concerns about using self-report surveys of teacher performance (Rasavi, 2000), although some research has shown a high correlation between surveys and observation (Mayer, 1999). More and diverse research is needed to examine the reliability and validity of these measures. The aim of this study was to develop a portable behavioural measure for teacher quality.

METHODOLOGY

The study used a Descriptive survey design which enabled the study to use both quantitative and qualitative

paradigms. Across- sectional design is a type of design used to obtain information about opinions, attitudes, preferences, practices and concerns of a cross-section of a group of people, and uses the results to generalize to the larger population (Creswell 2009). The study population was Uganda with a population of 34.6 million, and the target population was Wakiso district as well as Kampala Metropolitan City which have approximately 2.4 million people (UBOS. 2017). The study was conducted in schools within Kampala and Wakiso districts of Uganda. These districts were selected because they consist of a variety of schools such as rural and urban, varying class sizes, and children of diverse backgrounds. Thirty-six (36) respondents were sampled using purposive sampling technique.

Informal interviews were conducted with several primary school teachers and administrators in Uganda. Interview questions were developed from an extensive literature review on teacher quality and effectiveness. Using a data reduction technique (i.e., qualitative data analysis (QDA); Caudle, 2004), the interviews were transcribed and common themes emerged that were analogous to Danielson, (2006) four domains. Questions were compiled under each of the above domains to create a pilot measure – the teacher quality measure (TQM) containing 80 items. This preliminary measure was distributed to five Ugandan teacher volunteers who were asked to review the measure for several validity criteria (fowler, 2002). The volunteers included two females and three male educators with an average of 20 years of experience in their profession/field. More specifically, these reviewers included: (1) an Associate Dean in the School of Education (i.e., 30 years of experience), (2) an Associate Registrar (i.e., 20 years of experience), (3) a teacher of English (i.e., 8 years of experience), (4) a retired principal (i.e., 40 years of experience), and (5) a teacher of Math (i.e., 2 years of experience). The feedback provided by the reviewers was used to improve the survey. The revised survey was administered to teachers in the Kampala and Wakiso districts in Uganda. One teacher in each of the schools in each district was asked to contact the teachers and ask them to fill out the TQM. Participants had approximately one month to respond (i.e., in October of 2018), and the survey took approximately 20 minutes to complete.

RESULTS

Qualitative data analysis

After transcribing the interviews to develop items for the Teacher Quality Measure (TQM), four major themes emerged, which eventually became different sections of the TQM. These themes were analogous to Danielson, (2006) measures relating to quality teaching which included: (1) lesson preparation, (2) classroom environment, (3) instruction, and (4) professionalism. Based on these major themes, questions were developed to represent the full spectrum of the construct.

Quantitative data analysis

Descriptive statistics

The Teacher Quality Measure (TQM) contained 38 items and was completed by 36 upper primary (i.e., 5th - 7th grade) teachers from Wakiso District in Uganda. The Likert scale for the measure ranges from 0 (i.e., Few) to 3 (i.e., Almost All), which corresponds to 0% to 25% of the time and 76% to 100% of the time, respectively. The total possible point on this measure is 114. The mean of the 38 items from the final analysis sample was 83.3 (SD = 13.4). The range was 70 with a minimum score of 32 and a maximum score of 102. Histograms and skewness and kurtosis statistics revealed normally distributed data.

Rasch summary statistics

Prior to interpretation of the individual item and person data, appraisal of whether the data fit the model is required. Table 1 presents overall information about whether the data showed acceptable fit to the model. The first overall statistic to consider is separation, the index of spread of the person positions or item positions. For persons, separation is 2.61 for the data at hand (real), and is 2.78 when the data have no misfit to the model (model). If separation is 1.0 or below, the items may not have sufficient breadth in position. In that case, it is critical to reconsider what having less and more of the construct means, and upon revision, add items that cover a broader range. Item separation for these data was 2.28, a slightly smaller continuum than for persons. It is typical to find larger separation values for items than for persons. This is usually a function of the data having a smaller number of items and a larger number of people. However, in this analysis, there were 38 items and 36 people. Overall, the model had a separation value greater than 2, which indicates that true variability among items was larger than the amount of error variability. A larger item separation is preferable and perhaps a larger sample may remedy this. Note that the mean for items was 0.0. The mean of the item logit position is always arbitrarily set at 0.0, similar to a standardized (z) score. The person mean here was 1.30, which suggests these items were easy, on average, for persons to “agree with.” The persons had a higher level of the trait than the items did. If the person mean was -1, -2, or +1 or +2, this would indicate that the items were potentially too hard or too easy for the sample. Thus, a person mean of 1.30 suggests that the items were too easy. Following the examination of item means, mean infit and outfit for person and item mean squares were investigated.

Table 1a. Summary statistics of 36 measured persons – Winstep output

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	88.0	38.0	1.30	0.25	1.04	0.1	1.01	0.0
S.D.	12.9	0.2	0.81	0.10	0.29	1.3	0.47	1.3
MAX.	112.0	38.0	3.93	0.71	1.91	3.6	2.58	4.2
MIN.	50.0	37.0	-1.26	0.18	0.71	-2.2	0.54	-1.9

REAL RMSE 0.29, True S.D. 0.76, separation 2.61, Person reliability 0.87
MODEL RMSE 0.27, True S.D . 0.76, separation 3.78, Person reliability 0.98
S.E. of Person mean = 0.14
Person Raw Score-to-measure Correlation = 0.93
CRONBACH ALPHA (KR-20) Person raw score reliability = 0.88

Table 1b. Summary Statistics of 38 measured items – Winstep ouput

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	83.3	36.0	0.00	0.25	1.04	-0.1	1.01	0.0
S.D.	13.4	0.2	0.68	0.05	0.38	1.3	0.47	1.3
MAX.	102.0	36.0	2.12	0.42	1.04	3.6	2.58	4.2
MIN.	32.0	35.0	-1.44	0.19	0.58	-2.2	0.56	-1.9

REAL RMSE 0.27, true SD 0.62, separation 2.28, item reliability 0.84
MODEL RMSE 0.26, true S.D. 0.63, separation 2.50, item reliability 0.86
S.E. of item mean = .11
UMEAN=.000 USCALE=1.000
Item raw score-to-measure correlation = -0.98
1367 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 24796.85 with 1292 d.f. p=.0000.

Mean infit and outfit is expected to be 1.0. For both, they were 1.04 and 1.01, respectively. Related to this, the mean standardized infit and outfit are expected to be 0.0. In the current model, they were 0.1 and 0.1 for persons, and 0.1 and 0 for items. Overall, the data fit the model somewhat better than expected, which may signal some redundancy (i.e., possibly redundant items). According to Bode and Wright (1999), the standard deviation of the standardized infit is an index of overall misfit for persons and items. Using 2.0 as a cut-off criterion, both persons (i.e., standardized infit SD = 1.0) and items (i.e., standardized infit SD = 1.3) showed little overall misfit. Here the data evidenced acceptable fit overall. This is in contrast to the overall Chi-Square test for this model, which was significant. This indicates that the Rasch model does not fit these items well (X^2 = 2479.85, *df* = 1,292, *p* = .000).

Rasch response scale

Table 2 contains information about how the response scale was used. For these data, the response scale was 0 (Few), 1 (Many), 2 (Most), and 3 (Almost All). The step logit position is where a step indicates the transition from one rating scale category to the next (e.g., from a 2 to a 3). The “Observed Count" is the total of times the category was chosen from all items and all persons, whereas the "Observed Average" is the average of logit positions modeled in the category. It should increase by category value, and the current model demonstrated this. For example, persons responding with a 0 had an average measure (-0.01) much lower than those responding with a 1 (0.47). There was no significant misfit for the categories as the misfit indices (i.e., mean square misfit) for the categories were below 1.5. "Sample Expected" is the best possible value of the average logit position for these data. These values should not be very different from the observed averages, and for these data they were not. The cut-off value for infit and outfit mean squares is 1.0, and they were close to this value. Step (i.e., Structure) calibration is the logit calibrated difficulty of the step. This is shown in (Figure 1). These values should increase with category value, and for the current model they did. It can also be observed that the standard error, which is a measure of uncertainty around the step calibration, were all near 0 for each step (i.e., 0.13, 0.08, and 0.06).

Rasch probability curves

A final way of examining step use is via probability curves. These curves display the likelihood of category selection (Y-axis) by the person-minus-item measure (X-axis). If all categories are utilized as expected, each

Table 2a. Summary of category structure (38 measured items) from the teacher quality measure

CATEGORY		OBSERVED		OBSERVED SAMPLE		INFIT	OUTFIT	STRUCTURE CALIBRATION	CATEGORY MEASURED
LABEL	SCORE	COUNT	%	AVERAGE	EXPECTED	MNSQ	MNSQ		
0	0	74	5	-0.01	-0.13	1.14	1.35	NONE	-2.02
1	1	160	12	-0.47	-0.48	0.99	0.88	-0.59	-0.58
2	2	392	29	0.98	1.03	0.90	0.86	0.14	0.54
3	3	741	54	1.79	1.77	1.02	1.02	0.73	2.08

Table 2b. Continuation of summary of category structure (38 measured items) from the teacher quality measure

CATEGORY LABEL	STRUCTURE		SCORE TO MEASURE			50% PROB.	COHERANCE M->C C->M		ESTIMATED DISCRITE
	MEASURED	S.E							
0	NONE		-2.02	-INF	-1.33		63%	67%	
1	0.59	0.13	-0.58	-1.33	0.03	-1.00	32%	20%	0.83
2	0.14	0.08	0.54	0.03	1.34	0.04	40%	64%	0.95
3	0.73	0.06	2.08	1.34	+INF	1.03	79%	66%	1.07

M->C = Does Measure imply Category?
C->M = Does Category imply Measure?

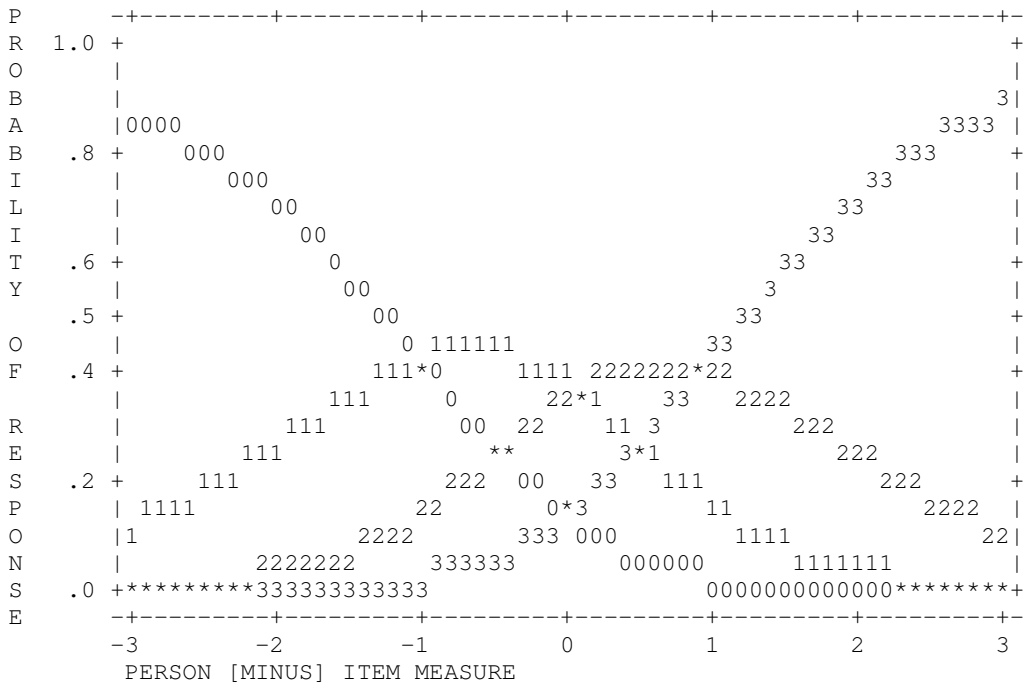

Figure 1. Category probabilities: MODES - Structure measures at intersections

category value will be the most likely at some point on the continuum (Figure 1), and there will be no category categories were utilized; however, it can be seen that category 1 is almost eclipsed by 0 and 2, which might indicate that respondents are not using the response scale as intended, or that the “Many” category might not be necessary.

inversions where a higher category is more likely at a lower point than a lower category. For these data, all the

Rasch item misfit diagnostics

Table 3 presents item misfit diagnostics. Measure is the logit position of the item, with error being the standard

Table 3. Item Statistics: Misfit Order – Winsteps Output.

ENTRY		TOTAL MODEL			INFIT		OUTFIT		PTMEA	EXACT ATCH	
NUMBER	SCORE	COUNT	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	OBS%	EXP%
26	76	36	0.45	0.21	2.02	3.6	2.58	4.2	A 0.00	27.8	44.5
15	94	36	-0.54	0.28	1.62	1.8	2.49	2.8	B 0.03	52.8	66.5
38	95	36	-0.62	0.29	2.04	2.5	1.76	1.6	C 0.21	61.1	67.6
18	89	36	-0.19	0.25	1.92	2.7	1.74	1.8	D 0.32	38.9	53.7
25	32	36	2.12	0.21	1.69	2.6	1.66	2.2	E 0.32	27.8	43.4
17	71	36	0.66	0.20	1.21	1.0	1.52	1.9	F 0.38	38.9	42.9
4	100	36	-1.14	0.36	1.43	1.1	.97	0.1	G 0.32	83.3	80.3
1o	86	36	-0.02	0.23	1.31	1.2	1.23	0.8	H 0.32	55.6	51.7
32	98	36	-0.90	0.33	1.25	0.8	0.89	-0.1	I 0.39	77.8	75.8
5	84	36	0.09	0.23	1.22	0.9	1.14	0.5	J 0.36	38.9	50.6
27	79	36	-0.39	0.21	1.04	0.3	1.15	0.6	K 0.38	38.9	46.4
6	92	36	0.32	0.27	1.00	0.1	1.09	0.4	L 0.32	61.1	63.0
35	97	36	-0.80	0.31	1.09	0.4	0.93	0.0	M 0.30	72.2	72.0
8	63	36	0.96	0.19	1.06	0.4	1.08	0.4	N 0.42	44.4	41.1
13	80	36	0.28	0.21	1.08	0.4	1.01	0.1	O 0.52	50.0	46.4

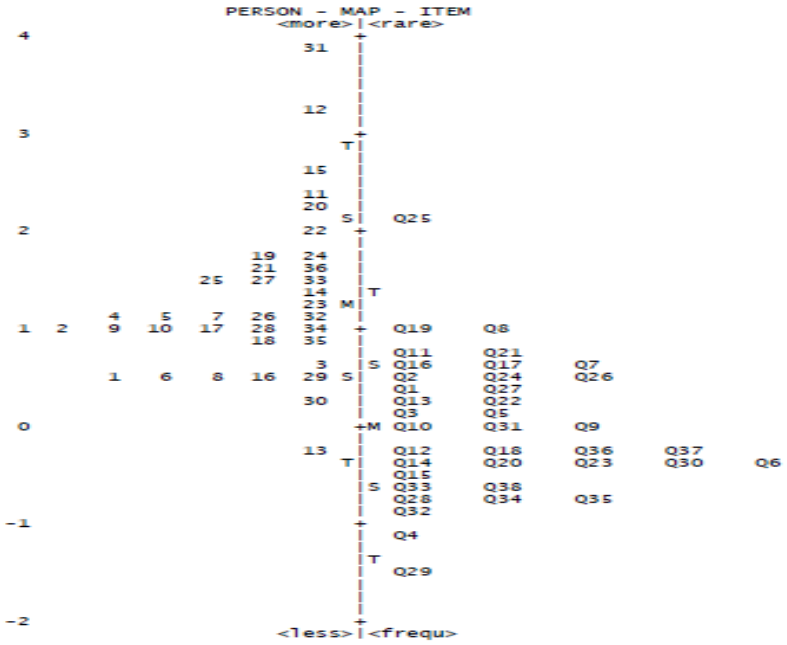

Figure 2. The map of persons and items for the 38-item Teacher Quality Measure (TQM). The distribution of person positions is on the left side of the vertical line and items on the right.

error of measurement for the item. Item fit for the model was determined by the infit and outfit diagnostics and the point measure correlations. Infit "…is a t standardized information-weighted mean square statistic, which is more sensitive to unexpected behavior affecting responses to items near the person's measure level" (Linacre, 2009, p. 252). Outfit "… is a t standardized outlier-sensitive mean square fit statistic, more sensitive to unexpected behavior by persons on items far from the person's measure level" (Linacre, 2009, p.252). A point measure correlation is the correlation between the item score and the measure, which should be positive (Linacre, 2009). Infit and outfit values of less than 2 were considered acceptable (Smith, 2004). Infit for all items on this measure were either less than 2 or extremely close to 2 (e.g., Item 26 [2.02] and Item 38 [2.04]), and were therefore acceptable. Most outfit measures were less than 2 as well. Items 15 (i.e., "I treat each of my pupils'

equally.") had an outfit of 2.49, and Item 26 (i.e., "I do not teach to a variety of learning styles.") had an outfit of 2.58. A vast majority of the sample endorsed the highest category on the scale for these items. Thus, these items were very easy to endorse, which could account for the high outfit value. A correlation below 0.15 indicates a potentially misfitting item, and the values are preferably between 0.3 and 0.5. Point measure correlations for this model ranged from 0.00 and 0.60. Unsurprisingly, the two lowest correlations were with the items listed above. As mentioned above, a point measure correlation below .15 indicates a potentially misfitting item, and these two items were already suspicious due to their large outfit values. None of the other items had correlations below 0.15.

Rasch map of Persons and items

The map of persons and items are shown in (Figure 2). The distribution of person positions is on the left side of the vertical line and items on the right. Each "X" represents one person in this figure. "M" marks the person and item mean; "S" is one standard deviation away from the mean; and "T" is two standard deviations away from the mean. To determine variability, item measure values were investigated using the item/person map for this model. The degree to which these items are targeted at the teachers was investigated. As seen above, the scale appeared to be applicable for its purposes. The items were normally distributed; however, some items (e.g., Items 25 and 29) were separated from the others on the far ends of the scale. A few items appear at the same logit value, indicating some redundancy. Fortunately, there were no large gaps in between the items, which could indicate that more items need to be added to address the full range (i.e., content domain) of the construct. The map shows many persons appearing above where the items are targeted. The items covered a range from -1.5 to 2 logits in difficulty, which is narrower than the range of about -.5 to 4 for persons. That is, most teachers found most items easy to endorse (i.e., endorsing the highest category on the Likert scale), with the exception of Item 25 (i.e., "I have pupils prepare a written summary of the day's lesson"). Overall, the above evidence indicates that more difficult items to endorse may need to be added in future studies to fill in the range of the construct measured, or some items should be considered for elimination based on misfit diagnostics or placement on the variable map.

DISCUSSION

The 38 items TQM produced acceptable psychometric properties (e.g., Coefficient α = 0.88). The current results from the Rasch model agreed with ones previously obtained under Classical Test Theory by Kyllonen and Kim, (2005). The item and person-separation reliabilities were high, indicating excellent psychometric quality for the TQM. All items were highly correlated with the total score, suggesting substantial item homogeneity. The item-fit measures indicated that the 38 items represented well the specified contents of the TQM. However, the main concern of this final version of the TQM is that most of the persons appeared above the mean on the vertical ruler, and the items covered a narrower range compared to the persons. This indicates that more "difficult" items may need to be added in the future. Also, at some points on the scale there were items at the same position on the vertical ruler, which indicated that these items may be redundant.

Generally, past studies do not show that to date flexible, efficient measure exists with the potential to evaluate teacher quality. Teacher qualification with a main focus on content knowledge seems to have been one of the ways teacher quality/effectiveness was measured. Many of the studies investigating measures of teacher quality have sought to use teacher certification scores as a proxy for content knowledge, with the results generally showing a positive relationship (Greenwald et al.,1996; Rowen et al., 2002; Ferguson and Ladd, 1991). These measures, however, are general measures of content that do not inform the types of knowledge or ability that a teacher requires to be an effective teacher. Other measures have looked specifically at performance on instruments designed to test a teacher's knowledge for teaching and found a significant and positive relationship to student achievement (Hill et al., 2005). Although there appears to be evidence of a link between content knowledge and achievement, the type of content knowledge that is assessed is dependent on the instrument or measure.

Conclusion

The objective of the study aimed to add to the pupil achievement research base by creating a measure of teacher quality. The hypothesis was that a psychometrically sound measure of teacher quality can be developed. The results rendered a 38-question TQM focusing four domains: (1) teacher planning and preparation (2) teacher classroom environment, (3) instruction, and (4) teacher professionalism reliable (α = 0.88). Therefore, school teachers who use TQM more frequently are able to diagnose teacher quality problems with greater specificity, and use that feedback to improve their pupils' scores. Future research, however, is encouraged to continue to define the theoretical network of relationships that may exist, and continue to further validate the scores on a measure of teacher quality.

Author`s declaration

I declare that this study is an original research that was carried out by me and I agree to publish it in the journal.

REFERENCES

Acana S (200). Uganda National Examinations Board. Paper presented at the 32nd Conference for Educational Assessment, Singapore.

Alabama schools. In H. F. Ladd (Ed.), Holding schools accountable: Performance-based reform in education, pp. 265–298. Washington, DC: The Brookings Institution. Alexandria, VA: Association for Supervision and Curriculum Development Approaches (3rd ed.). Los Angeles, CA: Sage Publications.

Bode RK, Wright BD (1999). Rasch measurement in higher education. In Higher Education: Handbook of theory and research (pp. 287-316). Springer Netherlands.

Caudle SL (2004). Qualitative data analysis. In Wholey JS, Hatry HP, Newcomer KE (Eds.), Handbook of practical program evaluation (2nd ed., pp. 417 – 438). San Francisco, CA: Jossey-Bass.

Creswell JW (2009). Research design: Qualitative, quantitative, and mixed methods

Danielson C (2006). Enhancing professional practice: A framework for teaching.

Ferguson RF, Ladd HF (1996). How and why money matters: An analysis of

Fowler FJ, Jr (2002). Survey research methods (3rd ed.). Thousand Oaks, CA: Sage Publications.

Greenwald R, Hedges L, Laine R (1996). The effect of school resources on student achievement. Review of Education Research. Pp.361-396.

Hill H, Rowan B, Ball D (2005). Effect of Teachers' Mathematical Knowledge for Teaching and Student Achievement. American Education Research Journal, 371.

imply for statistical practice? Journal of Educational and Behavioral Statistics, 29(1):121-129.

Kyllonen PC, Kim S (2005). Personal qualities in higher education: dimensionalityof faculty ratings of students applying to graduate school. Paper presented at the annual meeting of the American Educational Research Association, Montréal, Canada.

Linacre J (2009). A User's Guide to Winsteps. Program Manual Guide 3.68.0.

Mayer RE (1999). The promise of educational psychology: Learning in the content areas. Upper Saddle River, NJ: Prentice Hall.

Owen K (2005). Substantive communication of space mathematics in upper primary uestionnaire use in occupational stress research (Discussion Papers in Accounting and Management Science, 01-175) Southampton, UK. University of Southampton.

Raudenbush SW (2004). What are value-added models estimating and what does this

Razavi T (2001). Self-report measures: an overview of concerns and limitations of

Rowen B, Correnti R, Miller R (2002). What large-scale survey research tells us about teacher effects on student achievement. Teachers College Record, Pp.1525-1567.

Schacter J, Thum YM (2004). Paying for High and Low Quality Teaching. Economics of Education Review, 23: 411-430.

school. In H. L. Chick & J. I. Vincent (Eds.), Proceedings of the 29th annual conference of the International Group for the Psychology of Mathematics Education (Vol. 4, pp. 33–40). Melbourne: PME

Smith RM (2004). Introduction to Rasch measurement: Theory, models and applications. Jam Press.

Tucker PD, Stronge JH, Gareis CR, Beers C S (2003). The efficacy of portfolios for teacher evaluation and professional development: Do they make a difference? Educational Administration Quarterly, 39(5):572-602.

Uganda Bureau of Statistics (2017). The National Population and Housing Census 2014 – Area Specific Profile Series, Kampala – Uganda.