DEVELOPMENT AND VALIDATION OF OBSERVATIONAL SCHEDULE FOR PRIMARY SCHOOL MATHEMATICS CLASS IN SOUTH EDUCATION ZONE OF CROSS RIVER STATE

Mrs Alice E. Asim,
Akpan, Joseph E.
and
Eni I. Eni

Abstract

This study involved the development and validation of an Observational Schedule for Primary School Mathematics Classroom interactions. Twelve primary 6 classes selected through stratified proportionate sampling technique were studied in Cross River South Education Zone. Sixty four (64) observable items constituted the instrument which focused on classroom interactions in mathematics. The instrument had two parts with seven subscales - Lesson Presentation, Use of Instructional Materials, Classroom Management and Evaluation (Teachers' part). Pupils' Activities, Pupil–Teacher Interactions, and Pupil - Pupil Interactions formed the Pupils' part. The validity of the schedule was established by subjecting the instrument to expert judgment by Mathematics Educators, Measurement and Evaluation Experts and a trial testing in twelve primary 6 Mathematics classrooms. Inter-rater reliability coefficients ranged from 0.52 to 0.97 for the sub scales. There was high internal consistency (0.67 to 0.99) among items in the subscales and full scale which revealed the relative contribution of the items within the subscales to the validity of the developed instrument. However, the usefulness of the instrument in evaluating teaching and learning of Mathematics in primary schools was ascertained by test-retest method with a coefficient of stability of 0.99. It was recommended among others, that the instrument be used in data collection in Mathematics teaching in primary schools.

Keywords: Observational schedule, primary schools, mathematics, teaching.

Introduction

Mathematics is a key subject in the school system because it is fundamental to the learning of the sciences and meaningful living in society. It has become an entry requirement into institutions of higher learning in Nigeria. Experiential knowledge has shown that all is not well with mathematics teaching and learning. To correct any anomaly in this area, it is necessary to look at mathematics teaching at the foundational level, that is, the primary school. It is at this level that pupils come in first formal contact with mathematics teaching. Research has shown knowledge gaps in teachers in different aspects of teaching and assessment (Asim, Ijente & Bassey, 2010). Mathematics experiences for this category of learners may have a long lasting effect on their achievement in and attitude to mathematics (Jones, 2013; Morajkar, 2016). For any meaningful intervention, data needs to be collected about interactions in mathematics classrooms. Nevertheless, different modalities of measurement like interviews or questionnaire administration may not truly reveal what happens during mathematics teaching. This informed the decision to develop an observational schedule meant for primary school mathematics classroom interactions. However, the quality of an instrument item is judged by its validity, reliability and practicability. Validity here reflects how well an instrument measures what it is intended to measure (Akpan, 2012). On his own part, Joshua (2012) states that validity is the extent to which measurements are useful in making decisions relevant to a given purpose. Asim (2012) holds that validity refers to the extent to which an item measures what it is designed to measure. Validation of primary school mathematics observational schedule, therefore, refers to a determination of how well the items in the schedule actually measure what they purport to measure. It refers to the appropriateness, meaningfulness and usefulness of the references made from the results of the items in observational schedule (Armon, 2006).

A well validated observational schedule must show appropriateness for the intended purposes for the subjects for which the observational schedule is designed. Dyer (1995) opined that the content of the observational schedule must be suitable for the subjects and must ensure the achievement of the purpose of designing it. The items in the schedule must be subjected to thorough scrutiny by experts in the subject area before running a pilot study and making necessary amendments to the observation system or procedure.

Observation may be unstructured or structured. The former is most useful in exploratory and descriptive research while the latter is used when behaviours of interest are known and this type of observation often involves the use of observational schedule (Asim, 2013). Observational schedule is used in research employing observational methods to study behaviour and interaction in a structured and systematic fashion. It consists of predefined systems for classifying and recording behaviours and interactions as they occur, together with sets of rules and criteria for conducting observation and for allocating events to categories (Duffy, Wairent, & Walsh (2001). According to these authors, observational schedule is used to collect systematic and reliable data that can be used in research study.

There are five principal ways of entering data onto a structured observational schedule. They include: event sampling, instantaneous sampling, interval recording, rating scale and duration recording (Joshua, 2012; Akpan, 2012; Amaefuna, 2014). Event sampling, also known as a sign system, requires a tally mark to be entered against each statement each time it is observed. This method is useful for finding out the frequencies or incidence of observed situation or behaviour, so that comparison can be made. If it is important to know the chronology of events, then, it is necessary to use instantaneous sampling, sometimes called time sampling.

Duration recording is very suitable for continuous behaviours rather than a single event. What is recorded in this case is the type of behaviour with also an indication of the duration of the behaviour. For example, a child may remove her shoes for a twenty-minute period, a child may delay starting to do any writing for ten minutes, again a single behaviour but which continues for a longer time than each of the interval or instantaneous recording. The observer is driven by the event, not the frequency of the observation. It means that the observer needs to structure the recording schedule to indicate the total duration of the single continuous behaviour. Rudner (1994) adds that the observational schedule is sufficiently reliable to permit a stable type(s) of behaviour of the individuals in the target group. From the ongoing, reliability of the observational schedule is the ability of the observational schedule to consistently provide the needed types of behaviour or activity from the observed group.

Observational schedules are used in research employing observational method to study behaviour and interaction in a structural and systematic fashion (Obemeata, Ayodele &Araromi, 1999). According to Akpan (2012), it is used to collect systematic and reliable data that can be used in research study for reliable decision making or finding solution to problematic situations.

This method of data collection is of significance especially when there are difficulties in obtaining relevant information through self-report, because the subjects are unable to communicate (with infants or confused adults) or provide sufficiently detailed information about complex interaction patterns (Joshua, 2012).

Purpose of the study.

This study had the following purposes:

1 Development of an observational schedule for primary school mathematics class.

2 Application of the developed observational schedule to evaluate teaching-learning of Mathematics in primary schools to confirm its appropriateness for use to evaluate teaching and learning of Mathematics in primary schools.

Research questions

The following research questions were posed to give direction to this work on the development and validation of observational schedule for primary school mathematics class:

i. What is the relative contribution of each item to the developed observational schedule?

ii. What are the reliability coefficients of the subscales and the whole scale of the developed observational schedule?

iii. To what extent is the developed observational schedule useful in evaluating teaching and learning of Mathematics in primary schools?

Methodology

The type of design used in this work is the action research. This method is suitable for a work of this type because the research does not need the manipulation of independent variables; rather, it seeks to develop a skill approach relying on actual observations and behavioral data. Action research is practical and directly relevant to an actual situation or classroom situation.

The sampling procedure used in this work was stratified proportionate random sampling to ensure fair representation. Selection by proportion was computed both for the schools and primary six classrooms located in the different Local Government Areas within the zone of the study. The study sample comprised pupils and teachers of twelve primary six Mathematics classrooms selected from twelve public primary schools in the zone.

The researchers with the assistance of some experts in the discipline of Mathematics developed the instrument. Relevant works, (such as Flander's Interaction Analysis Categories – Flander, 1970; Eggleston's Science Teaching Observation Schedule - Egglestone, 1975; Procidia Social and Behavioural Science, 2010) were reviewed and ideas were incorporated wherever they were applicable while developing the instrument. There was strict guidance in the construction of the instrument. That was so because all the items in the instrument were thoroughly scrutinized by both Mathematics educators as well as measurement and evaluation experts.

The instrument was made in two parts: teacher's activities part and pupils' activities part. All the items in the instrument were behaviours expected in a typical Mathematics class. They were obtained from Mathematics teachers, text on method of teaching Mathematics and University of Calabar Teaching Practice Evaluation Form. Teacher's activities part had four behaviour categories namely: lesson presentation, use of teaching aids, classroom management and evaluation. The pupils' activities part had three behaviour categories namely: pupils' behaviours, pupils'/teacher's behaviours and pupil/pupil's behaviours. Each category contained the types of behaviour to be observed in a Mathematics classroom. Items in each category were clearly stated, void of overlapping and also covered the target behaviours. Items in each category were ordered from zero to five indicating the intensity of the behaviours observed.

One of the researchers and one other Mathematics teacher (two raters) practiced to become proficient in entering observed behaviours correctly into suitable categories as they occurred. The practice was done in two primary six Mathematics classes which were not part of the sample. This training lasted for three weeks. Thereafter, the researcher and the trained Mathematics teacher trial - tested the instrument in the selected public schools. At the beginning of each lesson, they entered the class and sat at a position where they could see clearly both the teacher and the pupils. They then recorded each behavior observed on the schedule as it occurred during the lesson. Each of the twelve classes was observed once and the entire exercise lasted for the period of two months.

The data collected were used to calculate consensus estimate of interrater reliability of the entire scale class by class of each school using Cohen Kappa's method and the range of the reliability coefficients of the instrument from the twelve schools was obtained to know how reliable the instrument was. Also, the reliability coefficients of the sub-scales of the instrument were calculated using Cronbach Coefficient Alpha to see how each subscale contributed to the reliability of the developed observational schedule. To further examine the relative contribution of the items within subscales in the developed observational schedule, class by items matrix was formed. Scores obtained from this matrix were grouped into odd and even scores. The scores obtained were used to calculate the reliability of the instrument using split-half method and the result was corrected using Spearman Brown Prophesy formula.

In determining the extent of usefulness of the developed instrument in evaluating teaching and learning of Mathematics, scores obtained from the two raters were used to calculate reliability coefficient using Pearson's Correlation Coefficient to see how reliable the instrument was for use in evaluating teaching and learning of Mathematics.

Results And Discussion

This work was designed to develop an observational schedule for primary school mathematics class. The result of the final validation of the instrument is presented. The presentation of the results is done in two parts which had to do with the development of the instrument.

The items in the instrument were relevant to the measuring task and captured what they were designed to measure as shown in Tables 2 and 3. Apart from the scrutinized vetting by three measurement experts, the instrument was trial tested in twelve mathematics classes for data collection. Every item that was found irrelevant was discarded and wrongly framed items were restructured correctly to further strengthen the validity of the instrument. The total items finally maintained were 64 out of 74 initial items.

The purpose of calculating the reliability estimate was to determine the consensus estimate of interrater reliability of the instrument using Cohen Kappa's method. That showed the extent of agreement between the two raters in using the same instrument to rate the same mathematics class. The range was seen to be 0.45. That is presented in Table 1

Beside, the estimates of internal consistency of the items for the subscales and the full scale were also calculated to determine how relevant or how contributory each item was to the developed instrument using Cronbach Coefficient Alpha. Subscale by subscale reliability value was also determined to see the standing of subscales in the developed instrument. That is shown in Table 2.

In addition to that, class by item matrix was formed to further ascertain the relative contribution of the items to the developed instrument. Scores obtained from this matrix were grouped into odd and even scores. The scores were used to calculate the reliability coefficient of the instrument using split-half method and the result obtained was corrected using Spearman Brown Prophesy formula.

Also, the usefulness of the developed instrument in evaluating teaching and learning of

TABLE 1: Consensus estimates of interrater reliability of the two raters class by class

Classes	No of items	Kappa
1st	64	0.52
2nd	64	0.92
3rd	64	0.70
4th	64	0.92
5th	64	0.80
6th	64	0.97
7th	64	0.90
8th	64	0.71
9th	64	0.60
10th	64	0.71
11th	64	0.79
12th	64	0.60

Kappa = $\frac{0-E}{1-E}$, where O is observed percentage of agreement, E is expected percentage of agreement or chance agreement and 1 is perfect agreement. So, 1- E represents the maximum amount that the two raters could have improved above chance agreement. The numerator of kappa equals the percentage of agreement that was actually observed between raters (O) minus the percentage of agreement expected by chance (E). That is O – E. This represents the amount the raters actually did improve beyond chance agreement. So, kappa is equal to the amount of improvement that raters actually showed above chance, divided by the maximum amount of improvement that they could have shown.

The result of computation showed that the range of kappa for the twelve classes was from 0.52 to 0.97. The breakdown class by class is as follows:

i. First class

In the first class, kappa was 0.52. That showed a moderate agreement between the two raters which indicated acceptable reliability of the instrument.

ii. In the second class, kappa was 0.92. That was seen to be closed to perfect agreement between the raters which showed that the instrument was reliable.
iii. In the third class, kappa was 0.70. That gave acceptable reliability of the instrument.
iv. The fourth class showed that kappa was 0.92. A very high agreement, showing that the instrument was reliable.
v. For the fifth class, kappa was 0.80. It showed that the instrument was reliable since the agreement between the raters was very high.
vi. In the sixth class, kappa was 0.97. Very closed to perfect agreement indicating a very high reliability of the instrument.
vii. Kappa in the seventh class was 0.90 which still showed high reliability of the instrument.
viii. The eighth class had kappa of 0.71. That indicated an acceptable agreement and a high reliability of the instrument.
ix. In the ninth class, 0.60 was seen to be kappa. That was moderate agreement and by that the instrument was considered reliable.
x. The tenth class had kappa of 0.71. It was acceptable agreement and indicated the instrument reliable.
xi. In the eleventh class, the instrument was seen reliable since kappa was 0.79.

The twelfth class had kappa to be 0.79. The agreement was high between the raters. The instrument was considered reliable.

TABLE 2: Estimates of internal consistency of items within subscales and full scale

	Subscales part I	No of items	Alpha
A	Lesson presentation	18	0.91
B	Use of instructional material	4	0.89
C	Class management	5	0.99
D	Evaluation	8	0.93
	Subscale part II		
A	Pupils' activities	13	0.72
B	Pupils'-teacher activities	6	0.64
C	Pupil/pupils' activities	10	0.67
	Full scale	64	0.98

The computation showed that items within the subscales had range of reliability from 0.64 to 0.99. The breakdown of the results subscale by subscale is as follows:

Part I

A Lesson presentation:

This subscale had 18 items and estimate of 0.91 showing that the items contributed well to the development of the full instrument.

B Use of instructional material:

There were 4 items in this subscale with a reliability estimate of 0.89.

C Class management:

This subscale had 5 items with a reliability estimate of 0.99 showing that the items contributed highly well to the development of the instrument.

D Evaluation:

Items in this subscale were 8 and had reliability estimate of 0.93.

Part II

A Pupils' activities:

In this subscale, items were 13 and the reliability estimate was 0.72.

B Pupils'/teacher's activities

This subscale had 6 items and estimate of 0.64

C Pupil/pupil's activities:

10 items were developed in this subscale and it had reliability estimate of 0.67.

Moreso, the full scale with 64 items, had reliability estimate of 0.98

Reliability involves the extent to which measuring instrument is measuring some behaviours in a systematic and therefore repeatable way (Armon, 2006). This is a principal demand for acceptable observational schedule as well as all instruments. The reliability of the observational schedule for primary school mathematics class is revealed in the reliability coefficients.

In as much as the problem of incessant poor performance in mathematics is an issue, the need for developing an observational schedule to observe the practice of teaching and learning of mathematics in the primary schools is imperative. Based on some key principles guiding development of observational schedule as opined by Dyer (1995), the items in the structured observational schedule were constructed and subjected to mathematics educators, measurement and evaluation experts and the supervisor for vetting. Also, it was subjected to trial test in two primary schools with three arms in each of the two schools to ensure content validity (Treanor, 2010). Nine items that were not measuring what was expected in a mathematics classroom setting were removed. Also, some items that were ambiguous and unclear as to what they were supposed to measure were restructured as the case may be. The entire items were reduced to a manageable number of 64 variables: a sufficient number for the purpose of validity and to cause data entry reliable. Four unfit items were reconstructed before the instrument was trial tested in the selected twelve mathematics classrooms. Total item in the instrument was 64 after which nine items were dropped.

Besides, the extent to which the different behaviours were displayed in a typical Mathematics class was observed with regard to the chosen rating values (0 to 5) in the instrument. The rating values ranged from 0 for 'not observed behaviour' to 5 for 'observed and excellently executed behaviours'. The extent of the different behaviours displayed in a typical Mathematics class was determined by calculating the percentage score of each of the rating values of the observed behaviours class by class by the two raters. Particularly, the percentage scores of the two extreme rating values, 'not observed behaviour' and 'observed and excellently executed behaviour' ranged from 23.4% to 51.6% and 0% to 33.8% respectively. The rating value, 'not observed' took the highest percentage score among other rating values in the twelve Mathematics classes observed. It showed that most of the expected behaviours in the instrument to be observed in a typical Mathematics class, that would have enhanced good performance in Mathematics, were not seen displayed in the Mathematics classroom. Also, the rating value, 'observed and excellently executed' had the least or the lowest percentage score among other rating values in the instrument. That, again, showed that most of the expectedly observed behaviours in the instrument were not excellently executed during the Mathematics class. Those were hints to incessant poor performance of Mathematics in primary schools. This, as a matter of fact, calls for urgent attention of government, education stalk holders and primary schools' Mathematics teachers for remedy.

Conclusion

For a good and quality observational schedule to be developed, some key principles guiding the development of such instrument must be followed. It was based on these principles that this instrument was developed. Not less importance was the need to subject the items of the instrument to vetting and analysis to ensure the validity and reliability of the developed instrument. In this particular study, the items were subjected to experts to ascertain that they measure what they were designed to measure and trial tested for data collection. The data collected were analyzed to establish how reliable the instrument was.

Recommendations

The following recommendations are made based on the scope and findings of this study:

(1) Mathematics teachers should be trained and retrained on the use of this observational schedule.

(2) Educational researchers should be made aware of this observational schedule and encouraged to use in conjunction with other methods of data collection about mathematics teaching in primary schools

(3) Government and education stakeholders should use this instrument as a measuring tool to examine the effectiveness of teaching and learning of mathematics in the classroom and therewith plan and organize training workshop and seminars for mathematics teachers on how the subject should be taught in a hope of ensuring good performance.