IDENTIFYING DIFFERENTIAL ITEM FUNCTIONING OF TEACHERS' JOB SATISFACTION SCALE WITH RESPECT TO LOCATION USING THE RASCH (IRT) MODEL

[1]**Metu, Ifeoma Clementina (Ph.D) (**ifeomametu2015@gmail.com)
[2]**Eleje, Lydia Ijeoma (Ph.D) (**li.eleje@unizik.edu.ng)
[3]**Njideka G. Mbelede (Ph.D) (** ng.mbelede@unizik.edu.ng)
[4]**Nneka C. Ezeugo (**nc.ezeugo@unizik.edu.ng**)**

[123&4]Department of Educational Foundations, Faculty of Education, Nnamdi Azikiwe University, Awka Anambra State, Nigeria.

Abstract

It is important that a scale is fair to all respondents in a population. This is one of the essential factors to consider in selection and use of psychological test. Item Response Theory (IRT) standards show that a scale should be independent of the properties of the sample. Differential Item Functioning (DIF) means the difference between psychometric properties of an item between groups that have the same ability. Specifically, this research determined differential Item Functioning of Teachers' Job Satisfaction Scale (TJSS) with respect to location, using the Rasch model. The sample comprised of 972 teachers from 36 public secondary schools in eight Local Government Areas in Enugu state of Nigeria. The researcher developed a 90- item instrument. This was trial tested and factor analysis was run but only 55 items survived. In order to answer the research question, Conditional Maximum Likelihood Estimation Technique of the Winsteps 3:80 Rasch software (Linacre, 2014), was used to analyze the data. From the result, some of the items in the scale functioned differently with respect to location. This is an indication of DIF effects. It means that some items are not measuring what they are expected to measure. It was recommended that psychometricians should adopt IRT techniques so that a scale will be fair to all respondents in a population.

Key words: Rasch model, Threshold parameter, Differential Item Functioning

Introduction

Scales are usually employed by social science researchers to measure latent traits like anxiety levels, attitudes, or science ability. To get the final score in such scales, item responses are scored and summed. For researchers to construct such measures in recent time, they employ two primary measurement theories which are classical Test Theory (CTT), and Item Response Theory (IRT). Researchers measure latent traits indirectly using test or survey because traits are naturally unobservable. It is worthy of note that unobservable traits should be assessed in this way because they have great influence on how persons react to survey items. Since it is difficult to get a perfect measure to assess how a person reacts to a set of test items that relates to an underlying measure, researchers try to create scores that are approximately at the level of the hidden trait possessed by the person (Bond & Fox, 2015). Both CTT and IRT can be used as tools to achieve this, but according to Sharkness and DeAngelo (2011), apart from having a common purpose, the two measurement systems have significant differences in their modeling processes, and also in their assumptions about the nature of the construct to be measured. CTT predicts the result of psychological testing such as test takers' ability and difficulty of an item. Classical test analysis shows that there is a link between the observed test score, the sum of the true score, and the error score. This means that the theory portends that observed test score is true score added to some error. CTT requires simple mathematical analysis which is easy to interpret. However, the theory has some limitations which include (1) interpreting raw scores as measures; raw scores have little inferential value and are not interval measures and are usually affected by missing values. Therefore, they cannot be compared for conclusions. (2) psychometric properties of instruments under CTT are sample-based in nature i.e. the properties depend on the set of items and sample of the respondents from which the data was collected.

Furthermore, CTT assumes that errors of measurement remain the same for all respondents and as such is constant across trait range, but items should differentially affect standard error of measurement (SEM) depending on their relationship to the trait level.

On the other hand, IRT according to Iweka (2018) is known as a probabilistic theory since it deals with the probability of possible response to items in a test. IRT is based on the idea that the chance of getting correct answers to an item depends on the person and item parameters. This means that people that possess greater level of the trait being measured are more likely to respond positively or correctly to an item. Although trait level and item difficulty are separate issues in IRT, they are essentially related. In fact, item difficulty or threshold is perceived in terms of trait level (Metu, 2020). Specifically, when an item is difficult to endorse, it means that it requires a respondent that is at a higher level of the trait being measured for it to be answered correctly or to be responded to positively but an easy item or easy to endorse item needs only a respondent with a low trait level to be responded to at a higher category. An important feature of the IRT modeling approach is that the parameters of the persons do not depend on the parameters of the items, and vice versa. Also in IRT, precision at each level of the construct being measured is assessed using standard error of measurement (SEM). This implies that each person and item parameter estimate is accompanied by its SEM, meaning that measurement is more precise.

There are many IRT models; amongst them is the Rasch model. This model was proposed by Georg Rasch, in1960. The model specifies that for an item to be answered correctly, it depends on the ability of a person or how strong his attitude (ϴ) is, and the location/ threshold or difficulty of the item, only. Rasch proposed this simple logistic model as a basis for constructing objective measures since he saw the need to define the difficulty of an item to be independent of the population and ability of a person to be independent of the items he has solved. When the Rasch model is used on an attitude scale where higher scores mean agreement with the attitude statement, ability of a person shows how respondents support the item while item difficulty means how easy or hard it is to agree with the item. Bond and Fox (2015), explained that with Rasch model, raw scores can be converted into equal interval units of measurement called log odd units (logit). Bond and Fox also stated that the ability of the scale to detect the level of the attribute is a way of measuring the reliability. Supporting this, Nunnaly and Bernstein (1994), stated that if different populations are used to measure the same construct in a different environment, ability produced should remain the same.

Rasch Rating Scale Model is the particular Rasch model used for rating scale data. This was developed by Andrich in 1978. This model is most suitable for rating scale data (e.g. Likert-scale data), because it places on a scale, the relationship between agreeability with a statement and chance of an item response. This means that persons with higher amount of a latent trait (job satisfaction), are more likely to positively endorse a statement or item than persons having less of the latent trait. Rasch model is based on principle of fundamental measurement and as such will address the weaknesses in CTT. That is why the model was chosen for this study; to identify differential functioning items.

Differential item functioning is an item analysis technique in psychometric bias analysis. DIF occurs when persons from different groups show varying degree of success on an item or where they endorse an item differently after they have been matched on the construct the item is meant to measure. This means that if different group of testees (e.g. male and female), have been observed to be almost at the same ability level, it is expected that their performance will be similar on test items administered to them, irrespective of which group they belong. The most important thing about DIF techniques is that test takers from different groups are matched according to their scores and then the tehnique finds out how the different groups performed on each test item to know whether one particular group is having a peculiar problem with any of the items. Most often DIF occurs because test items contain extraneous variables that are irrelevant to the construct under investigation and these affect group performance either positively or negatively. Hambleton *et al.* (2006) suggests that any item that is detected to function differently is dissimilar because it does not function in unison in different subgroups. Therefore, DIF analysis is designed to identify items that do not reflect similar functions when given to groups with roughly the same capability. In the past 40 years IRT-based DIF statistical techniques has been developed and used to identify items that function differently among similar groups.

One great advantage of the Rasch model procedure is that it develops item difficulty ratings separately for each group while removing the effect of person ability. This means that when comparing item difficulty estimates, the differences in person effects are removed. When data is fitted to the Rasch model, the scale is expected to work in the same way, no matter the group that is assessed. Therefore, the chance of being able to affirm an item or perform a task for persons on the same level of ability should be the same irrespective of the group involved. Supporting this view, Smith (2004) posited that assessment of DIF can give important information about fairness of measurement instruments across gender, age groups and locations. That is, assessment of DIF helps to find out whether items in a scale function in unison with respect to groups. However, using Rasch modeling to investigate differential functioning items is strictly on the threshold or location parameter. This is to maintain sum score sufficiency.

The threshold parameter which is the difficulty parameter of an item shows how difficult it is to agree with a statement or to indicate any category in the ordinal rating scale. This means that, for example, a teacher will be at a high level of job satisfaction to tick or endorse "strongly agree" for a statement that is difficult to endorse or difficult to agree with. That is, a teacher needs to possess a higher trait level of the construct job satisfaction in order to agree strongly with an item whose threshold value is high. According to Smith, in comparison to other items, if any item differs in its ability to differentiate respondents, it is said to be a misfitting item. For Rasch model such item is considered biased and is flagged off or discarded from the rest of the items.

For Wright & Panchalakesan in Ike et al. (2021), a DIF contrast that is less than 0.5 logits is DIF negligible and unimportant but values greater than 0.5 logits show that the difference is noticeable. Linacre (2012) also suggested that DIF contrast with the value of 0.64 logits and probability less than 0.05 will show clearly that the item function differently between the groups. Again, Bond and Fox (2015) gave out as DIF indicators; DIF contrast that is greater than 0.5 and p <0.05. Based on the above suggested criteria, DIF items for this study were identified using DIF contrasts >0.5 logits and p<0.05 as noticeable. The DIF items will be excluded from the scale.

One factor that has great influence on workers' perception of work environment and also affects their job satisfaction is location. A location can either be mainly rural or urban with each having its specific characteristics. An item in a scale may favour people settled in either of the two locations whereby they respond positively to it while it may not favour the other group. A lot of studies have been carried out on school location as a factor that determines teachers' job satisfaction; however, these studies are based on CTT. This study is using Rasch (IRT) model to discover whether differences in groups are based on true score differences or because of invariance. This is to make sure that a measure is assessing the same latent trait across locations; urban and rural. This will also establish that the items in a scale are functioning in unison across groups of interest.

Research Question

To what extent do the items of the Teachers' Job Satisfaction Scale function with respect to location?

Method

The purpose of this study is to develop Teachers' job satisfaction scale by running factor analysis on the item responses and using Rasch (IRT) model to identify differential functioning items with respect to location.

The design of the study is combination of survey and instrumentation. The study was conducted in Enugu state. Enugu state is one of the five (5) states in the South East geopolitical zone of Nigeria. The state is made up of 17 Local Government Areas (L.G.A) which are classified into six (6) education zones by the State Post Primary Schools Management Board- PPSMB (2019). Enugu state was chosen among the five states in the South East for the study through simple random sampling (balloting). The population of the study comprised of 7,303 teachers in all the 145 public secondary schools in Enugu state. This number is made up of 2,568 males and 4,735 females. There were 4,098 teachers from urban and 3,205 from the rural area. The data was supplied by the Planning, Research and Statistics (PRS) Department of the Post Primary School Management Board, Enugu.

Multi-stage sampling procedure was employed to draw a sample of 972 secondary school teachers from 36 sampled schools. This number is made up of 555 from urban and 417 from rural location. A draft instrument, Teachers' Job Satisfaction Scale (TJSS) of 90 items was developed by the researchers. The instrument was grouped into 6 subscales or clusters. Each of the items called for a graded response to each statement and is expressed in 4 categories of "strongly agree" (4), "Agree" (3), "Disagree" (2), "Strongly disagree" (1). The instrument was validated by experts and found to be adequate and reliable. It was trial-tested on 50 teachers that are not from the population under study.

Furthermore, in order to ensure that the items in the instrument are valid and adequate as well as exact representatives of the various constructs, the responses of the trial testing of individual items were subjected to factor analysis. From the Rotated Component Matrix, the items loaded on four factors. The researchers adopted a criterion of .350 minimum factor loading standard as recommended by Schuster and Milland (1978) for accepting an item in terms of item loadings to a factor. Twenty-three (23) items were found to be factorially impure as they could not load highly on any of the four (4) factors while 12 items were found to be factorially complex as they loaded on more than one (1) factor. Thus, 35 items were dropped after factor validation while 55 items emerged for the TJSS at that stage.

The 55 item instrument was distributed to a sample of 972 secondary school teachers. The researchers liaised with the principals of the schools whose teachers were used for the study for the distribution and collation of the questionnaire.

A DIF analysis output from Rasch Rating Scale Model software WINSTEPS 3: 80 (Linacre, 2014) was used to analyze the data in order to answer the research question**.**

Results

Research Question

To what extent do the items of TJSS function with respect to location (urban and rural)?

To answer this research question, DIF measures according to location, contrasts and probability levels were presented in the table below:

Table 1: Differential Item Functioning (DIF) scores with respect to Location

Item Number	**Urban DIF Measure**	**Rural DIF measure**	**DIF Contrast**	**Probability**
1	-1.67	1.48	-.19	.059
2	.41	.28	.13	.056
3	-.16	.03	-.19	.107
4	.21	.27	-.06	.631
5	-1.17	-.81	-.35	.400
6	-.82	-.63	-.19	.220
7	-1.14	-.92	-.22	.110
8	-1.28	-1.28	.00	.061
9	-1.16	-1.08	-.08	.322
10	.63	.05	**.58**	**.014**
11	-1.58	-1.51	-.07	.270
12	.48	.21	.27	.070
13	1.10	.95	.15	.054
14	1.02	.73	.29	.100
15	.42	.33	.09	.139
16	.35	.56	-.22	.200
17	.14	.04	.10	.418
18	1.42	.69	**.73**	**.003**
19	1.60	1.29	.31	.301
20	.18	.18	.00	.965
21	-.33	-.27	-.06	,571
22	1.30	.84	.46	.410
23	.02	.09	-.08	.169
24	.02	.07	-.06	.173
25	-.49	-.37	-.13	.602
26	1.48	1.34	.14	.069
27	-.21	-.28	.07	.830
28	-.62	-.55	-.06	.635
29	.52	.01	**.51**	**.032**
30	-.74	-.77	.02	.893

31	-.79	-.60	-.18	.207
32	.24	.18	.05	.708
33	-.49	-.39	-.10	.083
34	-1.02	-.89	-.13	.058
35	-.35	-.41	.05	.762
36	.69	.18	**.51**	**.047**
37	.66	.78	-.12	.868
38	-.52	-.32	-.20	.337
39	-.42	-.23	-.19	.104
40	-1.15	-1.15	.00	.587
41	-,99	-1.21	.22	.131
42	-1.22	-1.24	.02	.994
43	-.97	-.89	-.08	.131
44	-1.03	-.94	-.09	.830
45	.78	.84	-.06	.557
46	.89	.34	**.55**	**.028**
47	.84	.84	.00	.639
48	1.61	1.32	.28	.061
49	-.13	-.07	-.07	.675
50	1.31	1.13	.18	.219
51	1.71	1.76	-.06	.842
52	.22	.38	-.15	.077
53	.89	.89	.00	.739
54	1.34	1.71	-.37	.202
55	1.34	1.25	.09	.117

Table 1 showed the results of how the items function with respect to location. Bond and Fox (2015) suggested these DIF indicators based on the studied groups which are: (1) DIF Contrast > 0.5, and (2) p < 0.05. Hence, the researcher detected DIF using DIF contrast greater than .5 logits and p< 0.05 as showing noticeable and significant difference respectively. From the above table, noticeable location DIF could be observed in 5 items whose location DIF contrast was above .5 logits. e.g. items 10, 18, 29, 36, and 46 with DIF contrasts .58, .73, .51, .51, and .55 respectively. For these items, their logit values were above .5 and probability values equally less than 0.05 (.014, .003, .032, .047, and .028) respectively. The 5 items will be excluded from the scale. In other 5 items like items i.e. 8, 20, 40, 47, and 53, the DIF contrast was .00 meaning that the items have equal strength for urban and rural location teachers.

Discussion

The purpose of the research question was to find out how the different items of the TJSS function with respect to location (urban and rural). As seen from the table, Location DIF could be observed in 5 items e.g. items 10, 18, 29, 36, and 46. For these items, their logit values were above .5 and p values equally less than 0.05. The 5 items represent 9% of the items. This means that the 5 items do not function equally for rural and urban school teachers. For example, Item 10 is *"Amount of work I do exceed available time"*. This item may not function in unison for teachers in urban and rural areas because of difference in school population. Mostly, schools in urban areas are over populated. Most teachers from urban location when responding to this item may indicate "strongly agree" because of the number of the classes they teach, the class exercises, tests and assignments they grade on daily bases while teachers from rural areas that do not have population problem may indicate "strongly disagree". It should be noted that probability of being able to do a task or affirm an item, for persons at the same "ability" level should remain the same across groups. The implication is that "item 10" favours one group; the item maybe tapping a secondary factor (population) over-and-above the one of interest (job satisfaction). Therefore the 5 items with DIF effects were excluded from the scale. This is in consonance with the study by Madu (2012) that detected 11 items that were potentially problematic, DIF- wise and were consequently discarded. The result is also in agreement with the study carried out by Ike *et al*. (2021), whose findings detected and flagged off 13 items with DIF effects. In other five (5) items i.e. items 8, 20, 40, 47, and 53, the DIF contrast was .00 meaning that the items have equal strength for both groups (urban and rural school teachers). Put together 91% of the TJSS items (50 items) function identically among the two groups since their item measures are equally positive or negative. These 50 items are retained for the scale.

Conclusion

Location DIF effects were observed in a small percentage of the items (9%) for urban and rural area teachers. This figure represents five items. Apart from that, the other 50 items (91%) have equal strength for teachers in the urban and rural areas. The 50 items are retained for the scale. Therefore out of 90 initial items, 35 items were discarded after factor analysis while five (5) items were discovered to have noticeable and significant DIF effects and were also removed. Fifty (50) items remain for the Teachers' Job Satisfaction Scale.

Recommendations

The following recommendations are made:

1. That the IRT- based DIF statistical techniques be adopted by psychometricians so that a scale will be fair to all respondents in a population.
2. The scales should be used to measure latest traits such as anxiety levels, attitudes and abilities among others.

References

Andrich, D. (1978). Application of a psychometric rating scale model to ordered categories which are scored with successive integers. *Journal of Applied Psychology, 2(4), 40-45*

Bond, T.G., & Fox, C.M. (2015). *Applying the rasch model: Fundamental measurement in the human sciences.* Lawrence Erlbaum Associates Inc.

Hambleton, R.K., and Jones S.C (2013). Item Response Theory models and testing practices: current international status and future directions. *European Journal of Psychological Assessment.* 13(1, 21-28).

Ike ,J.E., Ene, C. C., Ojobo, B., Ani, M.I., Metu, I.C., Ugwu, E.C., Ezegwu, Agugoesi, J.O (2021).

Assessment of differential item functioning to detect gender-biased items in economics multiple choice questions in senior secondary school certificate. *Journal of Critical Reviews,* 3(1) 516-523.

Iweka, F. (2018) Use of differential item functioning (DIF) analysis for bias analysis in test construction. *International Journal of education, Learning and development.* 6(3).80-91.

Linacre, J.M. (2014). A user’s guide to winsteps/ministeps rasch model program: MESA Press Loomis.

Linacre, J.M. (2012). *Estimation methods for rasch measures.* Chapter 2 in E.V. Smith & R.M. Smith (Eds.). Introduction to rasch measurement.: JAM Press.

Metu, I.C. (2020). Using rasch model to identify differential item functioning of teachers’ job satisfaction scale with respect to gender. *AJB-SDR* 2(2). 59-65.

Madu, B.C. (2012). Using transformed item difficulty procedure to assess ender-related differential item functioning of multiple choice mathematics items administered in Nigeria. *Research on Humanities and Social Sciences.* 2(6) 41-55.

Nunnaly, J.C. & Bernstein, I. (1994). Psychometrics theory, Ed.3: McGraw-Hill.

Rasch, G. (1960/1980). Probabilistic models for intelligence and attainment tests (expanded edition). Chicago, IL: *British Journal of Mathematics and Statistical Psychology, 19, 49-57.*

Smith, R.M. (2004). Fit analysis in latent trait measurement models. *Journal of Applied Measurement.* 1(2): 199-218.

Sharkness, J., & DeAngelo, L. (2011). Measuring students’ involvement; A theoretical comparison of CTT and IRT. Eric.ed.gov/?id= EJ930336.

FAMILY SOCIAL CAPITAL ASSOCIATION WITH LEARNING OUTCOMES OF STUDENTS IN NNAMDI AZIKIWE UNIVERSITY HIGH SCHOOL, AWKA, ANAMBRA STATE, NIGERIA.