Using rasch model to identify differential item functioning of teachers' job satisfaction scale with respect to gender

Metu, Ifeoma Clementina (Ph.d)
Department of Educational Foundations
Nnamdi Azikiwe University, Awka

Abstract:
Making sure that a test or scale is not biased is one of the essential factors to consider in selection and use of psychological test. This means that it is important that a scale is fair to all respondents in a population. 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 gender, 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 Win steps 3: 80 Rasch software (Linacre, 2014), was used to analyze the data. From the result, some Q/ the items in the scale functioned differently. 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.

KEYWORDS: Rasch model, Threshold parameter, Differential Item Functioning

Introduction

Social science researchers mostly use scales to measure latent limits like anxiety levels, attitudes, or 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. 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 (l) interpreting raw scores as measures; raw scores have little inferential value and are not interval measures and are usually affected by missing values: meaning that concerning attitudes 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 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. 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, inl 960. The model specifies that for an item to be answered correctly, it depends on the ability of a person or how strong his attitude (O) 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 data 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 Bemstein (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 inl 978. 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 a statement or item than persons having less of the latent trait. Rasch model is based on the 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 methodology, a 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 it 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 this affect group performance either positively or negatively. Hambleton (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 with other items, if any item differ 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 and Panchalakesan in Pallant and Tennant (2007), 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 ndicators; 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<O. 05 as noticeable. The DIF items will be excluded from the scale.

One group that their pattern of response to personality scales should be of interest is gender. A set of characteristics that differentiates males from females is called gender. The difference may vary depending on the context used. Researchers and theoretical literatures (Lippa, 2010; Wood & Eagly, 2002; and Weisberg, DeYoung, & Hirsch, 2011, show that there are differences of males and females especially when responding to personality scales. There is a lot of literature concerning differences in personality and social behaviour. Therefore, it is important to understand gender differences in personality so as to assess both sexes fairly. This study is using Rasch model to discover if differences in group performance are because of invariance. This is to make sure that a measure is assessing the same latent trait between gender; male and female. This will also establish that 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 gender?

Method

The purpose of this study is to use Rasch (IRT) model to identify differential functioning items of Teachers' Job Satisfaction Scale with respect to gender. The design of the study is a 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 7,303 teachers in all the public secondary schools in Enugu State. This number is made up of 2,568 males and 4,735 females. 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 375 males and 597 for females. A draft instrument, Teachers' Job Satisfaction Scale (TJSS) of 90 items was developed by the researcher. The instrument is grouped into 6 subscales or clusters. Each of the items calls for a graded response to each statement and is expressed in 4 categories. The instrument was found to be adequate by experts. 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 researcher 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 (l) 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 researcher liaised with the principals of the schools whose teachers were used for the study for the distribution and collation of the questionnaires.

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 research question.

Results Research Question

To what extent do the items function with respect to gender (male and female)?

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

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

Item	Male	Female	DIF Contrast	Probability
Item		Female	DIF	Probability
Number	DIF Measure	DIF measure		

1	-1.26	-1.78	**.52**	**.000**
2	-.30	.40	-.10	.255
3	-.12	-.06	-.06	.223
4	.17	.28	-.11	.080
5	-.96	-1.05	.09	.298
6	-.10	-.77	**.67**	**.014**
7	-1.01	-1.07	.07	.115
8	-1.28	-1.28	.00	.419
9	-.65	-1.07	.42	.417
10	1.00	.54	.46	.140
11	-1.54	-1.54	.00	.691
12	.24	.44	-.22	.109
13	.94	1.09	-.14	.106
14	.89	.89	.00	.957
	.89			.000
				.255
				223
		.28		.080
		-1.05		.298
		— .77		.014
		-1.07		.115
		-1.28	.00	.419
		-1.07	.42	.417
10 .54 .46 140 11 -1.54 .00 .691				
12		.44 -.22	109	
13		1.09	106	
14		.89 .00	.957	
15		.36	.53	.017
16	.92	.41	.51	.036
17	.10	.10	.00	.666
18	.82	.51	.31	.144
19	1.32	1.54	-.21	.439
20	.21	.18	.03	.993
21	-.35		-.08	.433
22	1.10	1.07	— .07	.677
23	.58	.08	.63	.026

24	.04	.04	.00	.711
25	-.49	-.41	-.09	.118
26	1.39	1.41	-.02	.457
27	-.24	-.24	.oo	.876
28	-.53	-.63	.09	.137
29	.70	26	.44	.202
30	-.79	-.72	-.06	.489
31	-.64	-.75	.12	.329
32	.21	21	.oo	.893
33	-.44	-.44	.oo	.276
34	-.96	-.96	.oo	.873
35		-.34	-.11	.524
36	-.67	-.56	**-.11**	.603
37	.71	.71	.oo	.485
38	-.35	-.49	.14	.583
39	-.31	-.33	.02	.970
40	-.60	-1.26	.66	.019
41	-.48	-1.05	.57	.029
42	-1.19	-1.25	.06	.499
43	-.55	-1.07	.52	.ooo
	-.68	-1.23	.55	.ooo
45	.88	.76	.12	.170
46	1.31	.99	.32	.400
47	1.45		.58	.034
48	1.22	1.65	-.44	.210
49	.05	-.21	.26	.066
50	1.20	1.25	-.05	.354
51	1.82	1.68	.14	.436
52	.41	.21	.20	.075
53	.79	.95	-.16	.068
54	1.56	1.45	**.11**	.121
55	1.56	1.06	.51	.039

Table I: shows how the items function with respect to gender. Bond and Fox (2015) suggested that DIF contrast that is greater than 0.5 and probability that is less than 0.05 will signify DIF on the groups' performance. DIF indicators based on the studied groups which are: (l) DIF Contrast > 0.5, and (2) p < 0.05. Hence, the researcher detects DIF using DIF contrast greater than .5 logits and 0.05 as showing noticeable and significant difference respectively.

From the above table, noticeable gender DIF could be observed in Il items whose gender DIF contrast was above .5 logits. e.g., items 6, 23, 40, 41, 43, 44, 47, and 55 with DIF contrasts .52, .67, .53, .51, .63, .66, .57, .52, .55, .58, and .51 respectively. For these items, their logit values were above 0.5 and probability values equally less than 0.05 (.000, .014, .017, .036, .026, .019, .029, .000, .000, .034, and .039) respectively. This suggests significant DIF effects. These I I items represent 20% of the items.

Discussion

The purpose of the research question is to find out how the different items of the TJSS function with respect to gender (males and females). As seen from the table, gender DIF could be observed in I I items whose gender DIF contrast was above .5 logits. This can be seen in items l, 6, 23, 41, 43, 44, 47, 55 whoselogit values were above .5. The items may be tapping a secondary factor over-and-above the one of interest (job satisfaction). This number represents 20% of the items. All of the I I items also have their p- values less than .05 also suggesting significant DIF effects. In other 10 items i.e., 8, 11, 14, 17, 24, 27 32, 33, 34, and 37, the DIF contrast was .00 meaning that the items have equal strength for male and female teachers. Put together, 80% of the TJSS items (44 items) function identically among the two groups since their item measures are equally positive or negative. The Il items with noticeable and significant DIF effect will be kept aside for further investigation concerning item bias. This means that the Il items will be excluded from the scale. This is in consonance with the study by Royal (2010) that detected two items that were potentially problematic, DIF wise and were consequently discarded. The result is also in agreement with the study carried out by Ariffm, Idris, and Ishak (2010), whose findings detected 13 items with DIF effects. In other 10 items i.e., 8, 11, 14, 17, 24, 27, 32, 33, 34, and 37, the DIF contrast was .00 meaning that the items have equal strength for male and female teachers.

Conclusion

Gender DIF effects were observed in a small percentage of the items (20%), for male and female teachers. This figure represents eleven items which are excluded from the scale. Apart from that, the other 44 items (80%) have equal strength for male and female teachers. The 44 items are retained for the scale. Therefore, out of the 90 initial items, 35 items were discarded during factor analysis while Il items were discovered to have noticeable DIF effects and were removed. Forty-four (44) items remain for the Teachers' Job Satisfaction Scale.

Recommendation

It was recommended that IRT-based DIF statistical techniques be adopted by psychometricians so that a scale will be fair to all respondents in a population.

References

Andrich, D. (1978). Application of a psychometric rating scale model to ordered categories which are scored with successive integers. Journal of Applied Psycho logy.

Arrafin, S.R., Idris, R., & Ishak, M.N. (2010). Differential Item Functioning in Malaysian Generic Skills Instrument (MyGSI). Journal Pedidican Malaysia. 35(1), 1-10.

Bond, T.G., & Fox, C.M. (2015) (3rd ed.). Applying the Rasch model: Fundamental Measurement in the Human Sciences. Lawrence Erlbaum Associates Inc.

Hambleton, R.K., Swaminathan, H. & Rogers, J.H. (2006). Fundamentals of Item Response Theory: Sage.

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 (E ds.). Introduction to Rasch Measurement.. JAMPress.

Lippa, R.A. (2010). Gender differences in personality and interests: when, where and why? Social and Psychology Compass. 4, 1098-1110.

Nunnaly, J.C. & Bernstein, I. (1994). Psychometrics Theory, Ed.3: McGrawHill.

Pallant , J.F. & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and depression Scale (HADS).

BrJC1in 46, 1-18.

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.

Royal, K. (2010). Making Meaningful Measurement in Survey Research: A demonstration of the utility of the Rasch model. IRT Applications, 28, 1-16. http://www.airweb.org/images/irappa 28. pdf.

Smith, R.M. (2004). Fit Analysis in Latent Trait Measurment 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.

Weisberg, Y.J.., DeYoung, C.G.., Hirsch, J.B. (201 1). Gender Differences in Personality across the Ten aspects of the Big Five.. frontiers Media S A (CH)/https://www.frontiersin.org/articl es/10.3389/fpsyg.2011.00178/fu11 wood, W., & Eagly, A.H. (2002). A cross cultural analysis of gender behaviour: Implications for the origin of sex differences. Psychol. Bull. 128, 699727.