CREATING QUALITY PRECISION INSTRUMENT: VFMA_35 FOR THE CONDUCT OF VALUE FOR MONEY AUDIT USING THE RASCH MODEL

Razimah Abdullah
Edith Cowan University, Australia

Mohd Saidfudin Masodi
Accounting Research Institute
Universiti Teknologi MARA, Malaysia

ABSTRACT

Just like an examination paper, the reliability of a survey questionnaire as a measuring instrument is crucial if we want a good and valid instrument. The length of test and rating scale or distractor operational definition are among the determinants that will affect the validity of the instrument. However, how do we assure that we are measuring what is to be measured? This paper investigates the construct validity of an instrument named, VFMA_35 by applying the Rasch analysis. The precision of the instrument is established from the item measurement standard error (SE). Analysis using Rasch creates a functional precise instrument which has the unique predictive ability of measurement and is operable to meet all the requirements of prudent measurement.

Keywords: *Audit, Quality, Survey, Value for Money, Performance measurement*

Introduction

Effectiveness of survey with rating scales is gaining increasing importance in determining respondent's true state of mind. Mindsets are intangible and care is needed to make meaningful inference from responses. In most cases where such 'scales' are used, responses to a number of items are typically summed to yield a total score intended to locate the respondent on a continuum from less to more on the variable of interest. An ordinal rating scale is achieved by assigning integral numerals (e.g., 0 - 1 - 2 - 3) to descriptive response categories (e.g., none - mild - moderate - severe) as a means of partitioning the underlying latent quantitative continuum into successively increasing (or decreasing) amounts of the variable. A higher number, for example, choosing a 2 instead of 1, does not equate to twice the degree of the attribute. Usually, the raw score approach from such numeric ratings may appear simple and straight forward, but the appropriateness and legitimacy rest on certain fundamental assumptions that are often overlooked. First, for respondents to be able to communicate their positions accurately, the descriptive response category labels need to have distinct and unambiguous meanings that reflect the discriminations of the rating scale (Lopez, 1996). Second, for arithmetic operations, such as summation of integral numerals assigned to response categories, the magnitudes that successive categories represent need to be of equal interval (Stevens, 1946; Merbitz et al., 1989).

Another issue is that of the number of questions in a questionnaire that fits the purpose of testing the underpinning theory. Even for a general questionnaire, most people are totally lost as to where they should begin. A person may begin by adopting and adapting a similar survey related to the area of study but the issue on the number of items remains vague. Not only are the number of questions required but the survey instrument must also have a certain level of difficulty; hence, the test length and the discriminating power of items used are equally important. It is a known fact that the raw score results from tests or ordinal responses of a questionnaire varies with the test length whilst the percentage of correct responses in a test varies with item difficulty. Even though, sample size does not matter, Mok and Xu (2011) insist that the model used must be right. The issue of what is deemed to be the prudent test length for a specific purpose continues to be debated.

This paper offers an explanation on the parameters that ought to be considered in constructing a reliable instrument. There are a set of priorities that must be observed to assure the precision of measurement is not sacrificed. The resulting instrument has the ability to generate meaningful and accurate information or interpretation. Potential loss of important information cannot be compromised in building quality instrument to gather appropriate data. The sufficiency of measuring the required attributes must be within tolerable range of specification so that reasonably good decisions can be made. Bear in mind that in measurement, we are actually making some conclusion which may affect someone's future and we should be responsible enough in pursuing due care. For this purpose, the Rasch model is deployed vigorously to achieve an effective instrument construct with the necessary precision as it has the ability to clean the instrument from any item misfit and prevent potential data defects.

The Rasch model has been used extensively in the field of education (Green et al., 1984; Griffin, 2007; A Rashid et al., 2008), health sciences and psychology (Pallant & Tennant, 2007; Steinmeyer & Möller, 1992; Tennant et al., 2004). In finance, the Rasch model has been used to measure the financial capability of mutual fund investors and their awareness of the terms and risks of mutual fund investments (Pellinen et al., 2011) as well as the severity of gambling problems through the measurement of gambling symptoms (Strong & Kahler, 2007) whilst in management, it has been used to evaluate the types of employee participation in workplace decision-making and the degree of involvement by employees in such activities (Drehmer et al., 2000). The Rasch model has also been used in education for School Opinion Surveys to establish parent and student satisfaction benchmarks as a means to track changes in client satisfaction with public education (Bond & Fox, 2007, p. 186). The Rasch model is now extended in this paper to evaluate the performance of internal audit activities.

The Rasch model measures latent traits; latent trait models for personal measurement have been demonstrated to identify what happens when a person attempts an item. We will now be able to determine whether the internal audit process, specifically the individual audit task undertaken, is easy or difficult. Using the Rasch model changes the paradigm of measurement which is commonly accomplished by counting the frequency in the numbers of successfully achieved tasks, which can lead to invalid assessment thus invalid measurement (Wright, 1977).

This paper presents a brief description of the instrument and the response rate. It also explores the statistical analysis of the responses comprising item fit, unidimensionality, differential item functioning, rating scale validity, and person-item map. This is followed by a discussion on the creation of a better fitting instrument. Finally, a concluding observation is made on the quality of the instrument based on the results of the research and the further uses for the instrument.

Instrumentation

In any measurement, the appropriateness of instrument used is crucial. A foot long laboratory thermometer is sufficient to establish whether a kettle of water is truly boiled at the altitude of 4,000m during a mountain climb. In this case, precision does not matter but is sufficient for the intended purpose. Things would change greatly when you want a thermometer fit for an open heart surgery. Certainly you need a more precise instrument. Life is at stake and it is important to have the correct instrument in place. Cost is no more the issue; precision and reliability overrides all. Reliability of an instrument or the standard error of measurement matters most. Fisher et al., (2010, p.1) argue that "reliability coefficients indicate the proportion of total variance attributable to differences among measures separated along a quantitative continuum by a testing, survey, or assessment instrument." But reliability or statistical validity is not the only consideration. Wright & Stone (1999, p. 167) maintain that "validity deals with the meaning of inferences drawn from test scores." Because of inferences, other validity issues such as content, construct, predictive and fit should not be overlooked. A majority of these concerns will be explored in this paper when measuring the ability to perform internal audit.

Internal audit is one of the components of corporate governance. Corporate governance plays an important role in organisations, more so following the global financial crisis and emphasis on corporate performance throughout the world. Internal audit practices have been identified to influence internal audit performance (Fadzil et al., 2005; Dittenhofer, 2001). For example, Fadzil et al. (2005) acknowledged that risk assessment is affected by audit work, audit program and audit reporting. As pointed out by Dittenhofer (2001), internal audit is a complicated process: its achievement is through the accomplishment of internal audit objective.

The internal audit process can be divided into four dimensions: planning (P), execution or fieldwork (E), reporting (R), monitoring and follow-up on findings (M). The most recent study on internal audit in Malaysia was on the measurement of auditing practices using the IIA standards as questionnaire items (Fadzil et al., 2005). The questionnaires were sent to internal auditors and audit committees of companies listed in Bursa Malaysia. Fadzil et al. (2005) concluded that three more components apart from objectivity be added to SPPIA namely, a) audit reviews, b) audit programs and c) audit reporting. The attribute professional proficiency by SPPIA represents a trait resulting from education, training, experience and competency. As such, in this study, professional proficiency is included under competency when internal audits are being planned. The factors for determining internal audit performance will be based on the activities undertaken in the various stages of the internal audit, starting from planning to reporting audit findings.

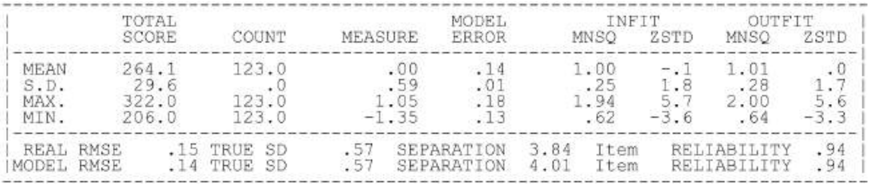
A questionnaire on Value for Money Audit or VFMA-35 of 35 items with four scales (0 *Not Practised* to a shade of 1-3 as *Least* to *Readily Practised*) was developed to measure the effectiveness of the internal audit process. The data obtained would help in interpreting the behaviours which are deemed as being the competency shown by the respondents in undertaking the task in the internal audit process. The criterion or content of the survey instrument is based on the internal audit best practices (Beckmerhagen et al., 2004; Dittenhofer, 2001; Fadzil et al., 2005; ISO19011:2002). For face validity, the items are commented upon by experts from the accounting and auditing professions. A pilot test was carried out before administering the survey to the internal auditors in Malaysia. 123 responses (31.06%) were received from a total of 396 distributed questionnaires. The outcome of the survey will give the measure of the quality of internal audit practices as reflected by the person mean, $\mu_{\text{PERSON,}}$ being the Maximum Likelihood Estimate (MLE) in the Rasch analysis. Baghaei (2008) used the Rasch model as a construct validation tool. The author argues that trustworthiness of score meaning and the underlying interpretation is vital in construct validity. The minimum sample size for using Rasch is 30 (Bond & Fox, 2007, p. 43; Linacre, 1994).

Statistical Analysis

A total of 4,270 data points arising from 123 respondents on 35 items are analysed using Winsteps, a Rasch analysis software. Items difficulty and person measure locations or person ability are expressed in logits through the transformation of the raw score percentage into its success-to-failure ratio or odds which is then converted to its natural log. The Rasch analysis provides indicators or statistics of how well the items fit within the underlying construct. The results yield a Chi-Square value of 7840.09 with 4112 degree of freedom. The test raw score Cronbach-α registers a reliability of 0.92 which allows for further analysis of the instrument.

The goodness of the survey instrument is described by the precision or errors in the item difficulty estimates and person ability estimates, item fit, person fit, and, reliabilities of person and item estimates. Table 1 shows the summary of items estimates with the mean defaulted at 0 logits. Item reliability is a high 0.94 on a 0 to 1 scale, similar to interpreting Cronbach-α, which is transformed to item separation index of 3.84, indicating very good sufficiency of item range (Fisher, 2007). The order of items is replicable across the other sample. The internal auditors' performance or ability estimates (Table 2) mean of +1.30 logits indicated that the tasks are comparatively easy. The maximum item measure is +1.05 logits (SE=0.18) while person ability is at a high +5.50 logits (SE=1.01). Despite the very good reliability, more difficult items, however, need to be introduced for that large gap of 4.45 logits. Nevertheless, there is sufficient item for the easy task where the minimum item measure is at -1.35 logits against the person measure of -0.89 logits.

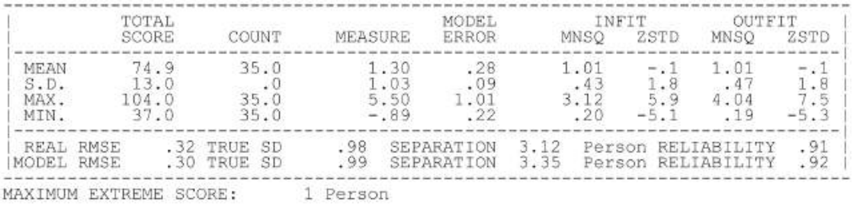
Table 1: Summary of Measured 35 Items



```
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ  ZSTD   MNSQ  ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     264.1     123.0         .00     .14      1.00   -.1   1.01    .0 |
| S.D.      29.6        .0         .59     .01       .25   1.8    .28   1.7 |
| MAX.     322.0     123.0        1.05     .18      1.94   5.7   2.00   5.6 |
| MIN.     206.0     123.0       -1.35     .13       .62  -3.6    .64  -3.3 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .15 TRUE SD     .57  SEPARATION  3.84 Item   RELIABILITY  .94 |
|MODEL RMSE    .14 TRUE SD     .57  SEPARATION  4.01 Item   RELIABILITY  .94 |
-------------------------------------------------------------------------------
```

The instrument has a small measurement model error of 0.14 (Table 1) and capable of yielding a good person separation of 3.12 (Table 2) but the Infit MNSQ SD=0.43 is slightly high. Both items and person Infit MNSQ and z-std values is close to the ideal 1 and 0, giving an indication on the goodness of fit of the instrument measuring what is to be measured. The summary statistics also indicated that only 1 person has responded to a scale of 3 for all items. The maximum extreme score by itself does not reveal much useful information about the person.

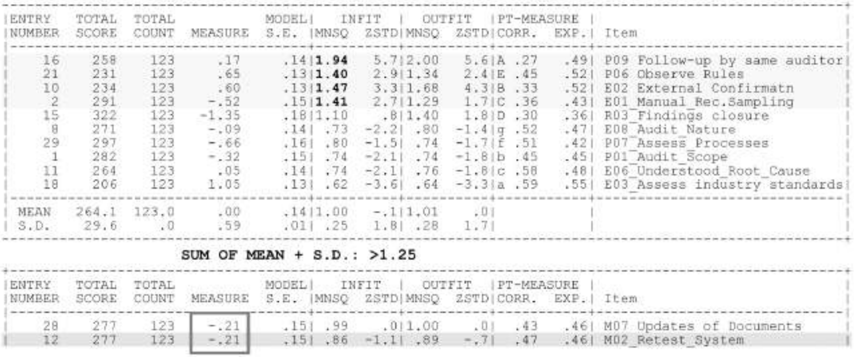
Table 2: Summary of Measured 122 Persons



```
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      74.9      35.0       1.30      .28      1.01    -.1   1.01    -.1 |
| S.D.      13.0        .0       1.03      .09       .43    1.8    .47    1.8 |
| MAX.     104.0      35.0       5.50     1.01      3.12    5.9   4.04    7.5 |
| MIN.      37.0      35.0       -.89      .22       .20   -5.1    .19   -5.3 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .32 TRUE SD     .98  SEPARATION  3.12  Person RELIABILITY  .91 |
|MODEL RMSE    .30 TRUE SD     .99  SEPARATION  3.35  Person RELIABILITY  .92 |
-------------------------------------------------------------------------------
MAXIMUM EXTREME SCORE:       1 Person
```

Item Fit. Generally, the items in the survey have a positive Point Measure Correlation and a small measurement error, with a mean of SE=0.14. An item is a misfit when it has a larger MNSQ than the sum of the mean of MNSQ and SD; in this case 1.25. Table 3 shows that Items 2, 10, 16 and 21 are misfits with both MNSQ > 1.25 and z-std > +/-2. Further scrutiny of the items showed that Item 12: M02-*Retest System* and Item 28: M07 –*Update of documents* have the same measure of -0.21 logits. The respondents identified the items as measuring the same task. An item whose MNSQ is nearer to 1 and z-std nearer to 0 is deemed a better fit; Item 28 is maintained and Item 12 is deleted while content validity is preserved.

Table 3: Consolidated Item Misfit



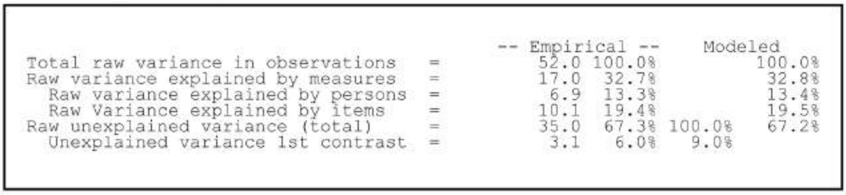
```
-------------------------------------------------------------------------------------------------+
|ENTRY   TOTAL  TOTAL           MODEL|   INFIT    |  OUTFIT   |PT-MEASURE |                         |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| Item                    |
|------------------------------------+----------+----------+-----------+-------------------------|
|    16    258    123      .17   .14|1.94   5.7|2.00   5.6|A .27   .49| P09 Follow-up by same auditor|
|    21    231    123      .65   .13|1.40   2.9|1.34   2.4|E .45   .52| P06 Observe Rules       |
|    10    234    123      .60   .13|1.47   3.3|1.68   4.3|B .33   .52| E02 External Confirmatn |
|     2    291    123     -.52   .15|1.41   2.7|1.29   1.7|C .36   .43| E01_Manual_Rec.Sampling |
|    15    322    123    -1.35   .18|1.10    .8|1.40   1.8|D .30   .36| R03_Findings closure    |
|     8    271    123     -.09   .14| .73  -2.2| .80  -1.4|g .52   .47| E08_Audit_Nature        |
|    29    297    123     -.66   .16| .80  -1.5| .74  -1.7|f .51   .42| P07_Assess_Processes    |
|     1    282    123     -.32   .15| .74  -2.1| .74  -1.8|b .45   .45| P01_Audit_Scope         |
|    11    264    123      .05   .14| .74  -2.1| .76  -1.8|c .58   .48| E06_Understood_Root_Cause |
|    18    206    123     1.05   .13| .62  -3.6| .64  -3.3|a .59   .55| E03_Assess industry standards|
|------------------------------------+----------+----------+-----------+-------------------------|
| MEAN  264.1  123.0      .00   .14|1.00   -.1|1.01    .0|           |                         |
| S.D.   29.6     .0      .59   .01| .25   1.8| .28   1.7|           |                         |
-------------------------------------------------------------------------------------------------+

                   SUM OF MEAN + S.D.: >1.25

+------------------------------------------------------------------------------------------------+
|ENTRY   TOTAL  TOTAL           MODEL|   INFIT    |  OUTFIT   |PT-MEASURE |                         |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| Item                    |
|------------------------------------+----------+----------+-----------+-------------------------|
|    28    277    123     -.21   .15| .99    .0|1.00    .0|  .43   .46| M07 Updates of Documents |
|    12    277    123     -.21   .15| .86  -1.1| .89   -.7|  .47   .46| M02_Retest_System       |
```

Unidimensionality. To satisfy unidimensionality, the items in the instrument must measure the same composite of abilities; the performance of value for money audit. The principal component analysis of the residuals in Rasch shows that the raw variance explained by measures of 32.7% closely matches the expected target of 32.8% (Table 4). Although this meets the minimum requirement of 20% in acceptable calibration by Reckase (1979), the more stringent threshold of 40%, as sought by Conrad et. al. (2010) is not achieved. Nevertheless, the unexplained variance in the first factor of 6.0%, rated the instrument as good (Fisher, 2007).

Table 4: Standardized Residual Variance (in Eigenvalue Units)

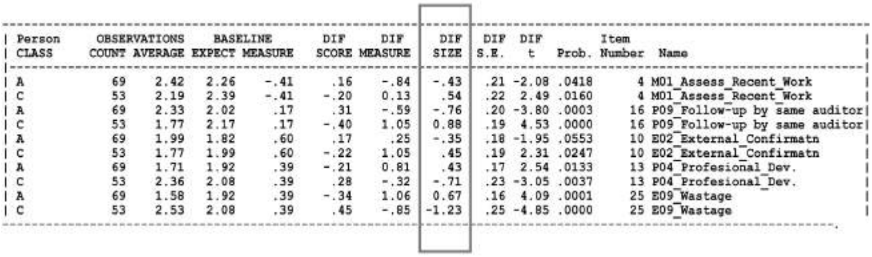


```
                                           -- Empirical --     Modeled
Total raw variance in observations     =      52.0 100.0%          100.0%
Raw variance explained by measures     =      17.0  32.7%           32.8%
  Raw variance explained by persons    =       6.9  13.3%           13.4%
  Raw Variance explained by items      =      10.1  19.4%           19.5%
Raw unexplained variance (total)       =      35.0  67.3% 100.0%    67.2%
  Unexplained variance 1st contrast    =       3.1   6.0%   9.0%
```

Local dependence test for the largest standardised residual correlation yields a very good outcome where none of the items breach the 0.7 limit, indicating item independence in the instrument.

Differential Item Functioning. A good item shall have invariant estimates across sub-groups of persons (Bond & Fox, 2007). A DIF contrast shows that the items need to be investigated; whether different inferences are made about the underlying construct. Table 5 shows five items: P04-*Professional Development,* P09-*Follow-up audit by same auditor*, E02-*External Confirmation,* E09-*Wastage* and M01-*Assess recent works in follow up audit* register both DIF > 0.5 and t > +/- 2.

Table 5: Consolidated DIF: Private Sector vs Public Sector

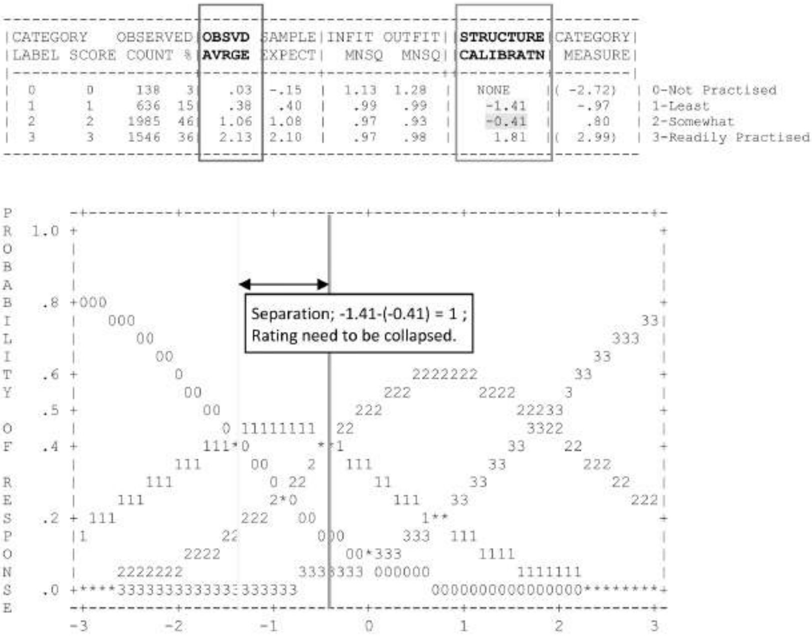


```
-------------------------------------------------------------------------------------------------------------
| Person   OBSERVATIONS    BASELINE        DIF     DIF     DIF    DIF  DIF          Item                     |
| CLASS   COUNT AVERAGE EXPECT MEASURE   SCORE MEASURE    SIZE   S.E.   t    Prob. Number  Name              |
|------------------------------------------------------------------------------------------------------------|
| A          69    2.42    2.26    -.41     .16    -.84    -.43    .21 -2.08 .0418      4 M01_Assess_Recent_Work |
| C          53    2.19    2.39    -.41    -.20    0.13     .54    .22  2.49 .0160      4 M01¯Assess¯Recent¯Work |
| A          69    2.33    2.02     .17     .31    -.59    -.76    .20 -3.80 .0003     16 P09¯Follow-up by same auditor|
| C          53    1.77    2.17     .17    -.40    1.05    0.88    .19  4.53 .0000     16 P09¯Follow-up by same auditor|
| A          69    1.99    1.82     .60     .17     .25    -.35    .18 -1.95 .0553     10 E02¯External_Confirmatn |
| C          53    1.77    1.99     .60    -.22    1.05     .45    .19  2.31 .0247     10 E02¯External¯Confirmatn |
| A          69    1.71    1.92     .39    -.21    0.81     .43    .17  2.54 .0133     13 P04¯Profesional_Dev.    |
| C          53    2.36    2.08     .39     .28    -.32    -.71    .23 -3.05 .0037     13 P04¯Profesional¯Dev.    |
| A          69    1.58    1.92     .39    -.34    1.06    0.67    .16  4.09 .0001     25 E09¯Wastage            |
| C          53    2.53    2.08     .39     .45    -.85   -1.23    .25 -4.85 .0000     25 E09¯Wastage            |
-------------------------------------------------------------------------------------------------------------.
```

DIF size ranges from -1.23 to 0.88. The t-value for E09-*Wastage* is conspicuously high at +4.09 logits for internal auditors in public sector, in contrast with those in private sector at -4.85 logits. Items 10 and 16 have already been identified earlier as misfits. Items 4, 13 and 25 can now be removed to make the instrument clear of DIF items.

Rating scale validity. Scale calibration is crucial in any measurement. The validity of the scale ultimately affects measurement precision due directly to thinner spread of responses across response categories. Enough data in each category is required for stable estimates. Usually, collapsing problematic categories with adjacent more functioning categories is done to improve variable definition or clarification to the data. The verification process in Rasch is linked to the threshold values between each rating scale. The observed average increases steadily and consistently from 0.03 to 2.13 logits, indicating consistency in response pattern; persons with higher ability endorse the higher categories, and those with lower ability endorse the lower categories (Figure 1). As a guide, collapsing the data should create a more uniform frequency distribution. The Rasch-Andrich Threshold is where the transition of decision making occurs from one scale to another. This is captured in the structure calibration column where Linacre recommends that the difference in threshold should be 1.4 logits apart but not exceeding 5 logits (Bond & Fox, 2007, p. 224). If the separation is less than 1.4, then it is recommended to collapse the affected ratings into one and split the rating scale if it is more than 5.

It was found that the separation between rating 1 and 2 needed to be collapsed since the separation was less than 1.4. If the Infit MNSQ SD is found to be smaller and yields a larger person separation, then the new scale of ‘012’ will be taken instead.



```
-------------------------------------------------------------------
|CATEGORY   OBSERVED|OBSVD SAMPLE|INFIT OUTFIT||STRUCTURE|CATEGORY|
|LABEL SCORE COUNT %|AVRGE EXPECT|  MNSQ  MNSQ||CALIBRATN| MEASURE|
|-------------------+------------+------------++---------+--------|
|  0     0   138   3|   .03  -.15|  1.13  1.28||  NONE   |( -2.72)| 0-Not Practised
|  1     1   636  15|   .38   .40|   .99   .99||  -1.41  |   -.97 | 1-Least
|  2     2  1985  46|  1.06  1.08|   .97   .93||  -0.41  |    .80 | 2-Somewhat
|  3     3  1546  36|  2.13  2.10|   .97   .98||   1.81  |(  2.99)| 3-Readily Practised
-------------------------------------------------------------------
```

```
P      -+---------+---------+---------+---------+---------+---------+-
R  1.0 +                                                             +
O      |                                                             |
B      |                                                             |
A      |                                                             |
B   .8 +000                                                          +
I      |   000                                                     33|
L      |      00                                                333  |
I      |        00                                            33     |
T   .6 +          0                        2222222          33       +
Y      |           00                    222       2222     3         |
    .5 +             00                222              22233          +
O      |               0 11111111  22                  3322          |
F   .4 +             111*0       **1                 33    22        +
       |          111     00    2   111            33        222     |
R      |       111          0 22       11        33             22   |
E      |    111             2*0          111   33                 222|
S   .2 + 111             222   00           1**                      +
P      |1              22        000      333  111                   |
O      |           2222             00*333        1111               |
N      |    2222222            3333333 000000         1111111        |
S   .0 +****3333333333333333333              0000000000000000********+
E      -+---------+---------+---------+---------+---------+---------+-
       -3        -2        -1         0         1         2         3
```

Separation; -1.41-(-0.41) = 1 ;
Rating need to be collapsed.

Figure 1: Structure Calibration of Rating Scale

Person-Item Map. The statistics in Rasch analysis culminates in the person-item or variable map (Figure 2). The map provides the simplest meaning to the data and shows internal auditors' abilities on the left and items difficulty on the right. This will be the premise of the instrument construct validity acceptance. It is important that the measures used in assessing the effectiveness of the internal audit function are appropriately targeted to the internal auditors being assessed. Poorly targeted measures can result in floor or ceiling effects. The person-item difficulty map shows that there are some easy internal audit tasks or activities. Apart from items M02 and M07, all the other items are in a good hierarchical order or well targeted to the internal auditors with an item measuring range of 2.40 logits. However, as earlier detected, a more difficult task is required for internal auditors with high abilities.

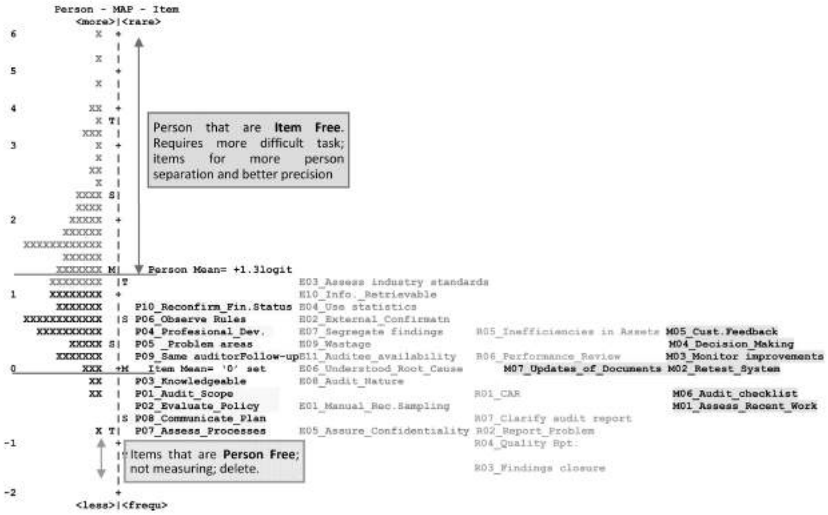


Figure 2: Person-Item Map: VFMA-35

Discussion

Several iterations are done by collapsing the data and deleting the items identified as misfits and having DIFs. A better fit instrument is finally constructed, showing marked improvements across various fit statistics including the MNSQ, z-std, reliability, SE and variance measured. The results in Table 6 are the cleaned values followed by the original VFMA-35 values which are shown in *italics*.

Reliability	Item	: 0.95 (*0.94*)	Person: 0.91(*0.91*)	Person Separation: 3.14(*3.12*)
Infit MNSQ SD	Item	: 0.16 (*0.25*)	Person: 0.32(*0.43*)	
Mean Error	Item	: 0.19 (*0.14*)	Person: 0.36(*0.28*)	
PCA Variance Measured		: 42.5%(*32.7%*)	Unexplained 1st Contrast: 5.9%(*6.0%*)	

Table 6. Cleaned VFMA-26 Instrument Construct Properties

The cleaned instrument now has better reliability and wider item scale range of 3.29 logits against person range of 7.14 logits as compared to the original VFMA-35 of 2.40 logits and 6.89 logits, consecutively. The increase in item scale range by 0.89 logits, a marked increase by 37.08%, is remarkable. The PCA of explained variance also improved tremendously to 42.50%

thus passing the more stringent limit of the 40% used in Rasch analysis by Conrad et. al. (2010) in determining unidimensionality.

The use of Rasch analysis is a useful tool in construct validation of the instrument. Construct validity encompasses other validity issues such as content validity and face validity. An instrument or a test is invalid if it does not measure what it is intended to be measured. Validity is tied to expert opinions in the subject matter. The new VFMA-26 instrument shows that item length is secondary in measurement. It is the sufficiency of good items to cover the expected range of difficulty that influences item reliability. The fit statistics indicate which data are useful and which data are not.

However, reliance on the final statistics should not be the overarching decision to reach the goal for a good instrument. Face validity has determined that items identified as misfits and to be removed, such as P04 Professional development, E02 External confirmation, E09 Check on wastage and M02 Retest of system, are essential in internal audit activity and these items should remain. The internal audit process follows a certain sequence, from planning, to execution, reporting and monitoring. By removing these items to create a more precise instrument, it is assumed that these tasks are unimportant. But without them, the whole internal audit process will not be effective. Instead of deleting the item entirely, an appropriate action would be in identifying a more difficult item that relates to the type of item measured.

Another controlling factor is the measurement SE; how precise should the measurement be. A key point to emerge from the Rasch analyses of item response data is that the measurement error is revealed across the measured range. With a known measurement error, judgments can then be made on the interpretations of the response data and the corresponding proposed actions for which the instrument is created.

The mean measurement error is influenced by how well or poorly targeted the instrument is to the test sample. If the purpose of the instrument is to act as a screening tool in finding who is below some threshold risk level, these errors might not matter, but errors near the threshold matter greatly. If the intent is to estimate population parameters rather than individual achievement, larger errors at the individual level, as shown in this survey, can be tolerated.

Conclusion

A precise instrument is as precise as the purpose of using it. Statistics is used to give support to the precision of the instrument: goodness of fit to the Rasch model, item misfits, unidimensionality and invariances. In measurement, we aim to estimate the ability of individuals on a scale. Item difficulty and person ability pictured in the person-item map, estimated on a logit scale, has some associated error. Estimation error decreases when persons and items are better targeted. The items and the scale used should also be able to distinguish abilities that indicate 'less to more' progression. The results presented here indicated that the unidimensional assumption is present and that the survey instrument is of quality, and precise in measuring the performance of value for money audit.

The results are important as they show that this instrument is appropriate and is the right instrument to measure internal audit activities. Further, this instrument can be used in conjunction with the evaluation of the internal audit function, which may include the evaluation of interaction of the audit committees with internal auditors, and the prevalence of collaborative activities with others charged with compliance or quality assurance. Future work in this area will focus on evaluating the internal audit function and its impact on corporate governance.

Acknowledgements

The author expresses her gratitude to Nik Mohd. Hasyudeen Yusoff, Executive Chairman, Audit Oversight Board at the Malaysian Securities Commission and Nur Hayati Baharuddin, Technical Director of the IIA Malaysia, for their assistance in reviewing the research instrument. She also acknowledges the Edith Cowan University Postgraduate Research Scholarship for her PhD study.

References

A Rashid, R., Abdullah, R., Ghulman, H. A., & Masodi, M. S. (2008). Application of Rasch-based ESPEGS model in measuring generic skills of engineering students: A new paradigm. WSEAS *TRANSACTIONS on Advance in Engineering Education, 5*(8), 591-602.

Baghaei, P. (2008). The Rasch Model as a construct validation tool. *Rasch Measurement Transactions, 22*(1), 1145-1146.

Beckmerhagen, I. A., Berg, H. P., Karapetrovic, S. V., & Willborn, W. O. (2004). On the effectiveness of quality management system audits. *The TQM Magazine,* 16, 1, 14-25.

Bond, T.G., & Fox, C.M. (2007). *Applying the Rasch Model: Fundamental measurement in human sciences* (2nd ed.). New Jersey: Lawrence Erlbaum Associates.

Conrad, K. J., Riley, B. B., Conrad, K. M., Chan, Y.-F., & Dennis, M. L. (2010). Validation of the crime and violence scale (CVS) against the Rasch Measurement Model including differences by gender, race, and age. *Evaluation Review, 34*(2), 83-115.

Dittenhofer, M. (2001). Internal auditing effectiveness: an expansion of present methods. *Managerial Auditing Journal, 16*(8), 443-450.

Drehmer, D. E., Belohlav, J. A., & Coye, R. W. (2000). An exploration of employee participation using a scaling approach. *Group & Organization Management 25*(4), 397-418.

Fadzil, F. H., Haron, H., & Jantan, M. (2005). Internal auditing practices and internal control system. *Managerial Auditing Journal, 20*(8), 844 – 866.

Fisher, W. P., Jr. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions,* 21, 1, 1095.

Fisher, W. P., Jr., Elbaum, B., & Coulter, A. (2010). Reliability, precision, and measurement in the context of data from ability tests, surveys, and assessments. *Journal of Physics: Conference Series, 238*(1), 1-6.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement 21*(4), 347-360.

Griffin, P. (2007). The comfort of competence and the uncertainty of assessment. *Studies in Educational Evaluation 33*(1), 87-99.

ISO 19011:2002 Guidelines for Quality and/or Environmental Management Systems Audits. International Organization for Standardization, Geneva.

Linacre, J. M. (1994). Sample Size and Item Calibration Stability. *Rasch Measurement Transactions, 7*(4), 328.

.

Lopez , W. (1996). Communication validity and rating scales. *Rasch Measurement Transactions*, 10:1 p.482-3.

Merbitz, C., Morris, J., & Grip, J.C. (1989). Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation*, *70*(4), 308-312.

Mok, M. M. C., & Xu, J. K. (2011). *Using multidimensional Rasch to enhance measurement precision: Initial results from simulation & empirical studies*. Paper presented at the Pacific Rim Objective Measurement Symposium (PROMS), NIE Singapore.

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology 46*(1), 1-18.

Pellinen, A., Törmäkangas, K., Uusitalo, O., & Raijas, A. (2011). Measuring the financial capability of investors: A case of the customers of mutual funds in Finland. *International Journal of Bank Marketing, 29*(2), 107-133.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational and Behavioral Statistics, 4*(3), 207-230.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677-680.

Steinmeyer, E.M., & Möller, H.-J. (1992). Facet theoretic analysis of the Hamilton-D scale. *Journal of Affective Disorders 25*(1), 53-61.

Strong, D. R., & Kahler, C. W. (2007). Evaluation of the continuum of gambling problems using the DSM-IV. *Addiction, 102*(5), 713-721.

Tennant, A., McKenna, S. P., & Hagell, P. (2004). Application of Rasch analysis in the development and application of quality of life instruments. *Value in Health, 7*(1), S22-S26.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement,* 14(2), 97-116.

Wright, B., & Stone, M. (1999). *Measurement Essentials* (2nd ed). Wilmington, Delaware: Wide Range, Inc.