Building Instrument with Construct Validity in Mind: Our Malaysian Experiences

Mohd Zali Mohd Nor
mohd.zali@my-newstar.com
MyRasch
(Malaysia)

ABSTRACT: *In many studies in Malaysia, several issues in providing evidence toward construct validity have been observed. In most problem cases, studies only provided face validity, factor analysis and reliability index yet claimed their instruments have sufficient construct validity. Another issue was a mix-up on assessment-type and perception-type items. Finally, insufficient sampling and targeting during pilot fail to provide empirical evidence on content validity. This paper presents the construct validity requirements according to Messick's construct validity framework and proposes several methods to deal with the above issues.*

KEYWORDS: **Construct Validity, Rasch Measurement Model.**

➔ Received 28/05/2023 ➔ Revised manuscript received 28/06/2023 ➔ Published 30/12/2023.

1. Introduction

The definition of instrument construct validity has several connotations which are very similar. APA Standard (1985) defines Construct validity as "*appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores*". Meanwhile, Messick (1989) describes it as "(*Construct) Validity is an overall evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment*". Basically, in layman term as we normally use, construct validity answers to the question "*does your instrument measures what it's supposed to measure?*" Without construct validity, the interpretation of scores from an instrument has little meaning.

However, interpretations of the sufficient requirements for construct validity vary amongst student researchers in Malaysia. Based on the author's experiences, many students were observed not having sufficient evidence to support construct validity on their tested instruments. Many just relied on factor analysis and reliability indices to claim construct validity, where, in fact, validity should be argued, demonstrated, and proved with more than just these.

This paper presents Messick's framework for construct validity, and evidence for each framework aspect could be provided to support the claim for construct validity for an instrument. In most cases, analyses using the Rasch Measurement Model (RMM) are utilized to present the evidence.

2. Messick's construct validity framework

Messick proposed six aspects of construct validity to provide "*evidence and rationales supporting the trustworthiness of score interpretation in terms of explanatory concepts that account for both test performance and score relationships with other variables*" (Messick, 1995). These are Content, Substantive, Structural, Generalizability, External and Consequential, described as follows:

- Content - Evidence of content relevance, representativeness, and technical quality. (Messick, 1995). In other words, answers to the question, "*Do test items appear to be measuring the construct?*"
- Often, students usually provide face validity as evidence. However, no further empirical evidence was presented to substantiate the face validity. In this paper, a method shall be discussed on how to provide such evidence.
- Substantive - Theoretical rationales for the observed consistencies in test responses. Normally, consistencies are reported via reliability indices. Reliability alone does not imply validity. Validity requires arguments (Bond, 2015).

- Structural - The extent to which interrelationships of dimensions measured by the test correlate with the construct and test scores (Messick, 1995). Sound instrument structure could be demonstrated using RMM unidimensionality test, Item and person fit analyses and category probability.
- Generalizability - Score properties and interpretations generalize to and across demographics, time, and places (Messick, 1995).
- Consequential - Implication of score interpretation as a basis of action and potential consequences of test use (Messick, 1995). Consequential validity provides the scope of potential risks if the scores are invalid or inappropriately interpreted. This aspect is the least reported by the student but could be important to minimize the risks of using the instrument.

In this paper, the first four aspects of Messick's construct validity shall be briefly discussed, and several guidelines are proposed for student researchers to consider when planning for pilot studies.

3. Providing evidence toward Messick's construct validity aspects

3.1. Content Validity

Other than face validity to provide initial verification of items, data from the pilot study could also be used to provide evidence on whether the content of the instrument is supposed to measure the intended latent trait, especially for an assessment instrument. For example, consider the following rubric of an instrument measuring Information Security Maturity Model (ISMM):

In order to prove that the content measures organizations are measured according to the appropriate levels of maturity, the following have to be planned before the pilot study takes place:

A rubric of clear levels of maturity, sub-domain and requirements for each level should be established. This will provide the basis for item building for the instrument.

Samples should be conveniently selected to include all the levels. This means the researcher should know the samples. Demography is important to ensure sufficient samples are collected.

After pilot data is collected and analyzed, evidence of whether the content measures the intended level could be ascertained using the Wright Map. For example:

The above Wright Map shall tell us whether the selected organizations (or samples) with known levels behave correctly on the items from the rubric. For example, samples with known Level 5 should be at the top, and samples with known Level 1 should be at the bottom. Similarly, the map also shows us to observe whether the items behave correctly on the known samples. This gives

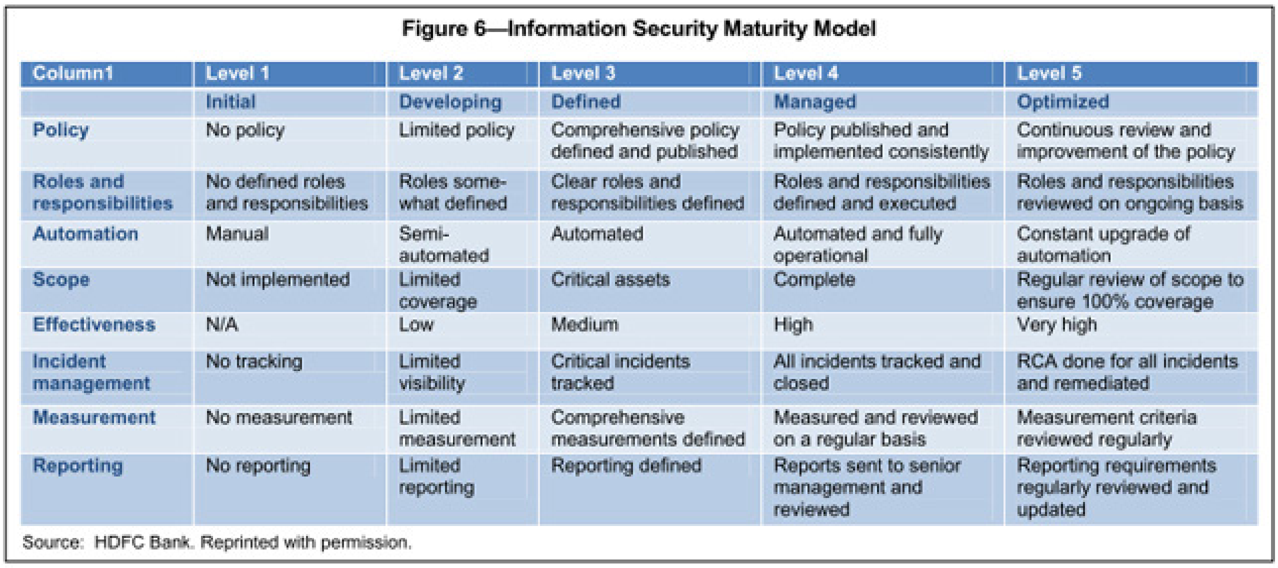


Figure 6—Information Security Maturity Model

Column1	Level 1	Level 2	Level 3	Level 4	Level 5
	Initial	Developing	Defined	Managed	Optimized
Policy	No policy	Limited policy	Comprehensive policy defined and published	Policy published and implemented consistently	Continuous review and improvement of the policy
Roles and responsibilities	No defined roles and responsibilities	Roles somewhat defined	Clear roles and responsibilities defined	Roles and responsibilities defined and executed	Roles and responsibilities reviewed on ongoing basis
Automation	Manual	Semi-automated	Automated	Automated and fully operational	Constant upgrade of automation
Scope	Not implemented	Limited coverage	Critical assets	Complete	Regular review of scope to ensure 100% coverage
Effectiveness	N/A	Low	Medium	High	Very high
Incident management	No tracking	Limited visibility	Critical incidents tracked	All incidents tracked and closed	RCA done for all incidents and remediated
Measurement	No measurement	Limited measurement	Comprehensive measurements defined	Measured and reviewed on a regular basis	Measurement criteria reviewed regularly
Reporting	No reporting	Limited reporting	Reporting defined	Reports sent to senior management and reviewed	Reporting requirements regularly reviewed and updated

Source: HDFC Bank. Reprinted with permission.

Figure 1. Information Security Maturity Model

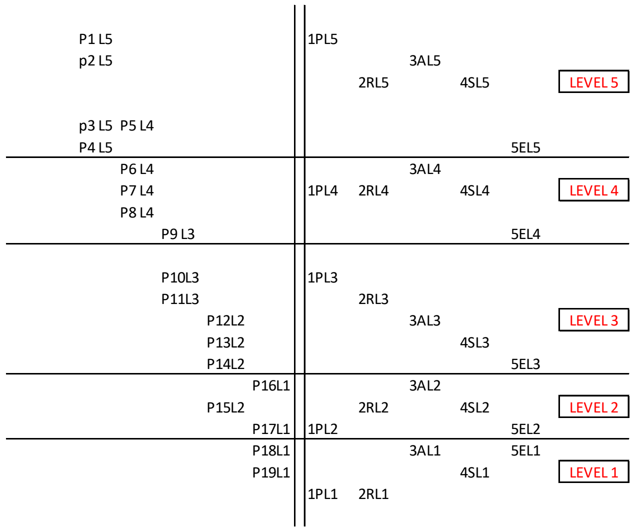


Figure 2. Wright Map

empirical evidence on whether the instrument's content measures what it should measure.

For perception-type instruments, sometimes transforming into a rubric may not be possible. However, it is still possible for experts to endorse the rank of items according to difficulty or agreeableness. This endorsement could then be compared to the results of pilot data.

3.2. Substantive

Students often encountered cases where the reliability of pilot data was low. These are often caused by insufficient items to separate samples and vice-versa. In the below examples (Figure 3), the left Wright Map indicates the items are separating only 1 sample. This results in very low person reliability. Meanwhile, the middle and the right maps have items that cover the full spectrum of samples and are able to separate the samples properly.

In order to increase the reliability of instruments, it is recommended that samples be conveniently chosen to test the full spectrum of items. Again, demographics are an important consideration when planning for the pilot. If an instrument consists of test items with varying difficulties, the samples should also be of varying abilities to check whether the items are functioning well. If sample abilities are known, then this is a bonus, as the items could be checked against person abilities later.

3.3. Structural

The following three areas need to be addressed to ensure the instrument has good structural validity and could be used in the final instrument. First, all items measure the same latent trait. This is tested using Unidimensionality test (Table 1), which indicates if some items indicate measuring secondary dimension. However, further investigation on these items is necessary to confirm unidimensionality.

Second, investigation of the quality of item and person is necessary to ensure all items are fit to measure the latent traits of persons. An example of item fit analysis is shown in Table 2 below.

Third, a good category structure ensures that the ordering of measures is consistent with the ordering of categories (Adam et al., 2012). A good category structure would have the category probability curve as per Figure 4 below. If disordered thresholds are exhibited, most of

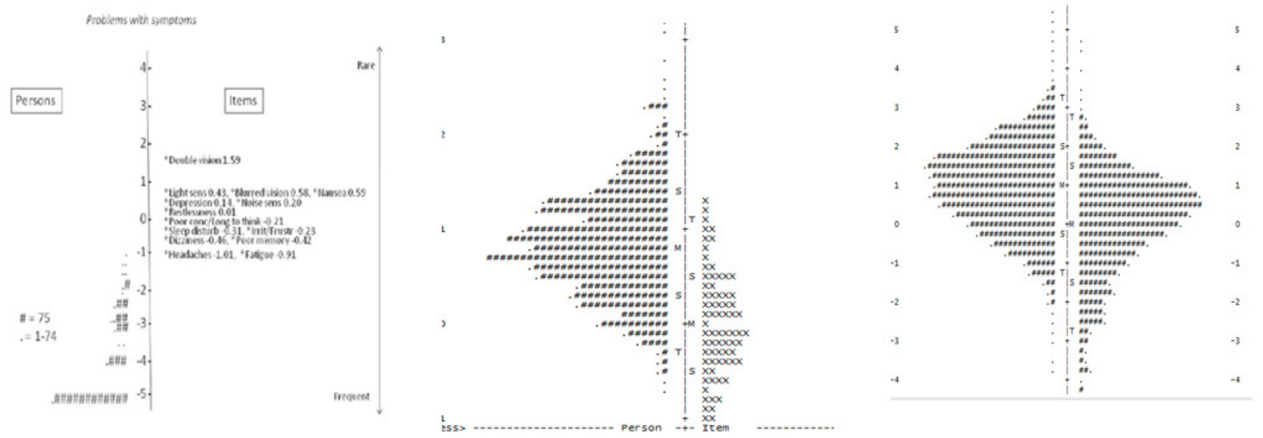
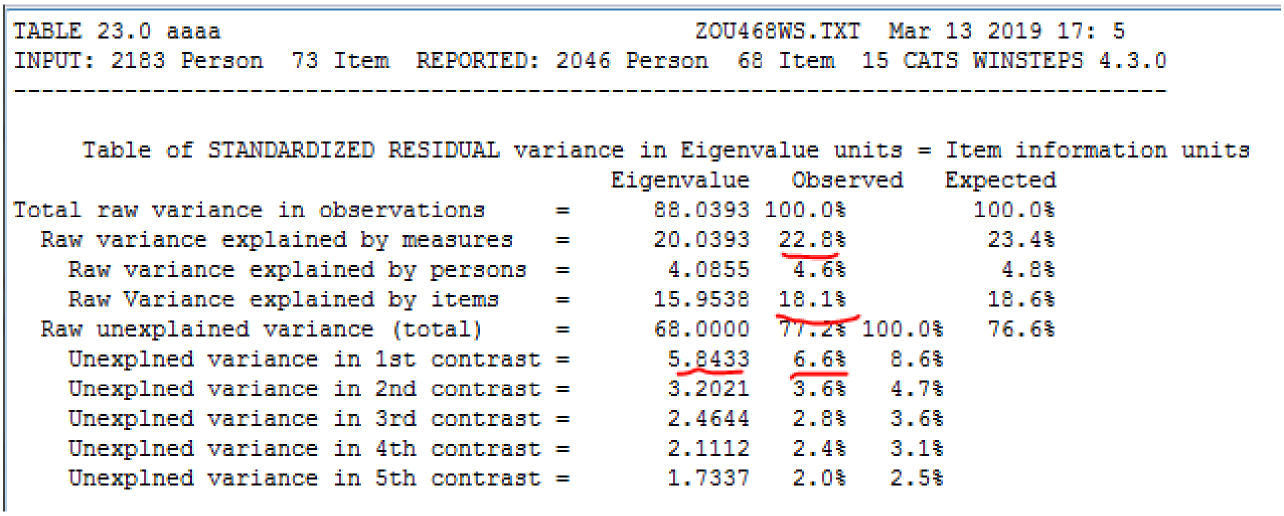


Figure 3. Sample distributions

Table 1. Unidimensionality test



```
TABLE 23.0 aaaa                                        ZOU468WS.TXT  Mar 13 2019 17: 5
INPUT: 2183 Person  73 Item  REPORTED: 2046 Person  68 Item  15 CATS WINSTEPS 4.3.0
--------------------------------------------------------------------------------

     Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = Item information units
                                                  Eigenvalue  Observed   Expected
Total raw variance in observations     =      88.0393 100.0%         100.0%
  Raw variance explained by measures   =      20.0393  22.8%          23.4%
    Raw variance explained by persons  =       4.0855   4.6%           4.8%
    Raw Variance explained by items    =      15.9538  18.1%          18.6%
  Raw unexplained variance (total)     =      68.0000  77.2% 100.0%   76.6%
    Unexplned variance in 1st contrast =       5.8433   6.6%   8.6%
    Unexplned variance in 2nd contrast =       3.2021   3.6%   4.7%
    Unexplned variance in 3rd contrast =       2.4644   2.8%   3.6%
    Unexplned variance in 4th contrast =       2.1112   2.4%   3.1%
    Unexplned variance in 5th contrast =       1.7337   2.0%   2.5%
```

Table 2. Item fit

Item	Item measure	MNSQ infit	MNSQ outfit
Double vision	1.59	0.98	0.94
Nausea	0.59	1.08	1.03
Blurred vision	0.58	1.17	1.32
Light sensitivity	0.43	1.06	0.93
Noise sensitivity	0.20	1.06	1.13
Depression	0.14	0.86	0.75
Restlessness	0.01	0.88	0.70
Poor concentration/Longer to think	–0.21	0.92	0.79
Irritability/Frustration	–0.23	0.98	0.90
Sleep disturbance	–0.31	1.08	1.10
Poor memory	–0.42	0.96	1.00
Dizziness	–0.46	1.09	1.07
Fatigue	–0.91	0.85	0.85
Headaches	–1.01	1.18	1.17

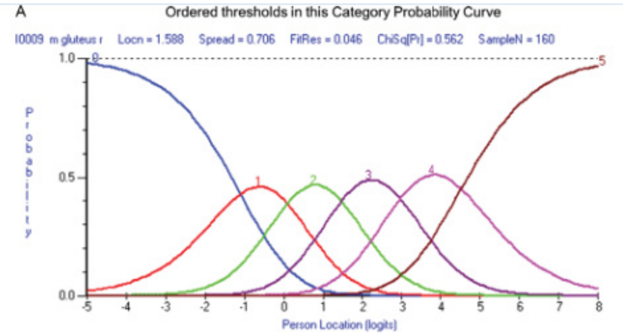
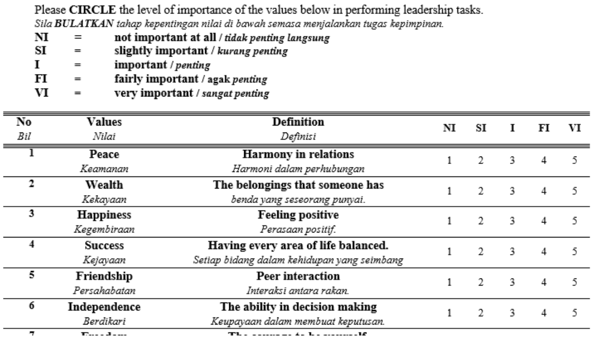


Figure 4. Category probability curve

the categories may need to be investigated, and categories should be reduced or collapsed.

Another issue of category structure faced by student researchers is mixing perception scales and assessment scales. Perception scales often measure what we feel or perceive, which may not be what we are. These are not necessarily the same thing. For example (see Figure 5), if one says very important in wealth, it may not mean he/she is wealthy. Hence, this may be incorrect if the researcher tries to establish a person profile based on perception scales.



Please **CIRCLE** the level of importance of the values below in performing leadership tasks.
Sila ***BULATKAN*** *tahap kepentingan nilai di bawah semasa menjalankan tugas kepimpinan.*

NI	=	**not important at all /** ***tidak penting langsung***
SI	=	**slightly important /** ***kurang penting***
I	=	**important /** ***penting***
FI	=	**fairly important / agak** ***penting***
VI	=	**very important /** ***sangat penting***

No *Bil*	**Values** *Nilai*	**Definition** *Definisi*	**NI**	**SI**	**I**	**FI**	**VI**
1	**Peace** *Keamanan*	**Harmony in relations** *Harmoni dalam perhubungan*	1	2	3	4	5
2	**Wealth** *Kekayaan*	**The belongings that someone has** *benda yang seseorang punyai.*	1	2	3	4	5
3	**Happiness** *Kegembiraan*	**Feeling positive** *Perasaan positif.*	1	2	3	4	5
4	**Success** *Kejayaan*	**Having every area of life balanced.** *Setiap bidang dalam kehidupan yang seimbang*	1	2	3	4	5
5	**Friendship** *Persahabatan*	**Peer interaction** *Interaksi antara rakan.*	1	2	3	4	5
6	**Independence** *Berdikari*	**The ability in decision making** *Keupayaan dalam membuat keputusan.*	1	2	3	4	5
7	**Freedom**	**The courage to be yourself**					

Figure 5. Sample perception items

Recommendations for good instrument structure would be to have a sound rubric of the instrument framework established to ensure that all items are indeed measuring the latent trait. Also, proper scales or categories should be established.

3.4 Generalizability

In most cases, students only check for demographic bias after a pilot study has been conducted. Often, the results could not be substantiated due to insufficient samples and incomplete demographics. Planning demographics early is very important to avoid the above issues.

4. Summary

Table 3 below summarizes the basic areas to consider when planning the pilot study with construct validity in mind.

Table 3. Validity aspects and requirements

	1 - Content	2 - Consistency	3 - Structure	4 - Generalizeability	5 - External	6 - Consequential
Demography		√		√		
Targeting	√	√	√			
Proper Scale	√		√			
Rubrics	√					

Planning for sufficient demography shall help in consistency and generalizability aspects. A good spectrum of known samples (targeting) shall be beneficial to provide evidence for content, consistency, and structure aspects. Proper scales or categories are helpful for content and structure aspects. Meanwhile, good rubrics are beneficial to provide empirical evidence for content aspects.

References

Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement*, *72*(4), 547-573.

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, *50*(9), 741.