Working Paper Series 2018:01

Title: **Sampling items from lengthy scales using Item Response Theory (IRT); The case of the High Performance Organizations (HPO) Scale**

Date: **15 February 2018**

Author(s): **Robert Goedegebuure, PHD**

Keywords: **Item Response Theory; HPO**

CONTENTS

SAMPLING ITEMS FROM LENGTHY SCALES USING ITEM RESPONSE THEORY (IRT)

The case the High Performance Organizations (HPO) Scale

INTRODUCTION

Researchers in the field of organizational behavior, especially when employing quantitative research through surveys, often prefer using existing scales and standards. The logic of this preference is twofold. First of all, existing data collection instruments have been tested for validity and reliability by the initiator, and by followers who – ideally – have confirmed the validity of the measurement instrument and thus have lent support to the theoretical model underlying the questionnaire. Secondly, using existing instruments enables researchers to compare their own empirical results to those obtained in similar studies in other settings. The practice also has its disadvantages. A practical disadvantage is that the objective of the research goes well beyond the aim of the scales being used, and as a consequence the overall data collection of which the existing scale is a component part, becomes lengthy. Even though the evidence of the impact of the length of questionnaires on the response rate and the quality of response is mixed (see: Bogen, 1996; Shalqvist et al, 2011), a more compelling argument for shortening the questionnaire is the relevance of the retained items and the enhanced parsimony and efficiency of the final questionnaire, rather than the reduced length of the questionnaire per se.

Unfortunately, most articles introducing or using the scales limit themselves to addressing the validity and reliability of the scales while neglecting issues that have a prominent place Item Response Theory (IRT): item difficulty; item discrimination; item information; and test information. It would be good practice for researchers to have a critical look at the scales used, preferably using IRT. Failure to do so, inevitably leads to the prolonged use of inefficient scales, reduced quality of research and (at least according to the majority of empirical studies on the issue) reduced response rates.

In this paper we will use the data from a study done in Nepal, among employees and managers of three organizations in government owned or controlled industries. In assessing the performance of the three organizations, the researchers have made used of an extended version of the High Performance Organizations (HPO) model. The HPO model has been developed by De Waal (2007), and has been used and tested in various countries and industries (e.g. De Waal, 2010; De Waal et al. 2009; De Waal et al, 2014). In the context of our research in Nepal, the HPO model has been extended, by adding several dimensions of organizational performance that are not, or not explicitly, part of De Waal's model. The focus of this paper, however, is on an assessment of the original HPO scale, and its items.

Prior to this study, to the best of our knowledge, there is no research that has critically examined the HPO questionnaire using IRT. The objective of this paper, is to fill this gap. A general objective of this paper is to show how IRT can be used to sample a relatively small sample of items from a multi-item questionnaire in such a way that the amount of information is kept. A specific objective is to do so for the HPO questionnaire, in order to guide future researchers who are considering to make use of it in their studies.

LITERATURE REVIEW

A Brief Introduction to Item Response Theory

Item response theory (IRT) can be defined as an applied statistical technique for describing information in data obtained via a data collection instrument with respect to the items that are part of it and the overall performance of the instrument (Reckase, 2009; Raykov & Marcoulides, 2018). The fundamental concept in IRT is the relationship between a construct being evaluated, and the probability of a certain response to an item given a person's with a certain score on the construct. This concept is reflected by the so-called item characteristic curve (ICC). ICCs differ from one item to the next. The key properties of items are known as difficulty and discrimination which are estimated in IRT models. Difficulty, or item location, represents the location of an item on the scale. For a binary (or binary coded) item, the difficulty is the location on the scale where a person is expected to succeed. Discrimination is related to the slope of the ICC. It represents how fast the probability of success changes with scores near the item's difficulty. Item with large discrimination value can better distinguish between low and high levels of the latent trait. IRT has been extensively applied to mainly educational science where the objective is to create tests that distinguish, for example, students with sufficient knowledge or ability to pass a test. Increasingly, IRT is also applied to studies in behavioral, social and organizational sciences, where researchers make use of *polytomous* items (e.g. Likert scales). Our interest is in the latter.

The High Performance Organizations Model

De Waal's model (De Waal, 2007) consist of five factors (or dimensions), which are measured with 35 items. The five dimensions are labeled continuous innovation (CI; 8 items); openness & action orientation (OAO; 6 items); management quality (MQ; 12 items); employee quality (QEMP; 4 items); and long-term orientation (LTO; 5 items). Although the analysis of our HPO data suggests another structure than De Waal's five-factor structure (Goedegebuure, forthcoming), for the sake of ease of illustrating the application of IRT, we will use the five-factor structure as a basis, and treat CI; OAO; MQ; QEMP; and LTO as five unidimensional constructs. Our objective is to assess these five unidimensional scales, and to make a selection of a subset of items out of the 35 items in such a way that the information contained in the full set of items is retained. Of course, de Waal's 35 items are a sample from an even bigger population of items. But, again for the sake of illustration the use of IRT, we assume that de Waal's 35 items cover the concept of organizational performance.

DATA ANALYSIS

Introduction to the Data

Our research makes use of data on three Nepalese organizations is government owned or controlled sectors of the Nepalese economy. Within each of the three organizations, 100 employees and managers have filled out questionnaires; the total sample size is therefore 300. The HPO set of 35 questions is one part of the total questionnaire of around 80 questions. The answer scale used in the HPO questionnaire, and in our questionnaire, is a 10-point scale (from 1=does not apply at all to 10=applies completely).

Continuous Improvement (CI)

CI is measured by 8 items (see the questionnaire in annex 1).

For (ordinal) polytomous items, IRT offers a variety of models. The partial credit model (PCM) is considered appropriate for settings where items require successive completion of a number of tasks. However, application of PCM is not restricted to this analysis of component tasks, as one can conceptualize the category boundaries in our 10-point scale as steps that respondents have to clear to score themselves on that item. While PCM assumes a constant discrimination parameter across all the items that make up the scale, this assumption is relaxed in the Generalized PCM (GCPM). A parsimonious version of the PCM is the Rating Scale Model (RSM) that requires all items to have the same number of categories. Although the rating scale in our questionnaire is a 10-point scale for all items, the RSM (as implemented in STATA 15 which is the software used for data analysis in our research) won't work whenever some categories for some items are not ticked by any respondent. Lastly, the Graded Response Model (GRM) which is explicitly used for Likert scales, and does not have the category restriction of the RSM.

The selection of the best model to use can be made on statistical grounds. While the PCM is a special case of (or nested in) the GCPM, it can be tested whether the general version is significantly outperforming the restrictive version. RSM and GRM, however, are different kinds of models, and therefore formal likelihood-ratio tests can not be applied. Generally used criteria for comparison, are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), that take into account the number of free parameters in the fitted models.

For the 8 items of CI, all categories of the 10-point have been ticked by respondents, and therefore the RSM for these items can be fitted, and the AIC and BIC can be computed. The GRM, with the lowest AIC and BIC, is the best to use from a statistical point of view, for CI. We will use the GRM model for all five factors of the HPO model.

Table 1. Goodness-of-fit statistics for the CI-scale

Model	Obs	Log_likelihood	df	AIC	BIC
PCM	300	-3,727	73	7,600	7,870
GPCM	300	-3,653	80	7,466	7,762
RSM	300	-3,802	17	7,637	7,700
GRM	300	-3,607	80	7,374	7,670

The results of applying the GRM to the 8 items of CI, are depicted in table 2. In the table we have only listed the difficulty coefficients for **v1**, in order to reduce the output and, more importantly, since the same information is easier to interpret using graphs.

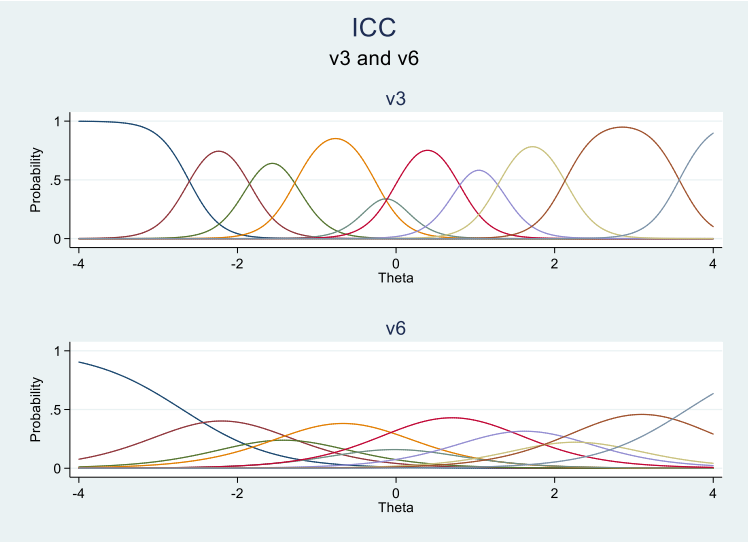
From table 2 we can see that the discrimination coefficients range from 1.72 for **v6** to 5.08 for **v3** (higher scores are better). We will show the ICCs for these two variables.

Table 2. GRM model for eight-item CI-scale

		Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
v1	Discrim	2.11	0.19	11.30	0.00	1.74	2.48
	Diff						
	>= 2	-2.68	0.27	-9.89	0.00	-3.21	-2.15
	>= 3	-1.46	0.14	-10.48	0.00	-1.74	-1.19
	>= 4	-0.95	0.11	-8.64	0.00	-1.17	-0.74
	>= 5	-0.25	0.09	-2.76	0.01	-0.43	-0.07
	>= 6	-0.03	0.09	-0.36	0.72	-0.21	0.14
	>= 7	0.69	0.10	6.72	0.00	0.49	0.89
	>= 8	1.16	0.12	9.39	0.00	0.92	1.40
	>= 9	1.96	0.18	11.11	0.00	1.61	2.30
	10	2.95	0.31	9.55	0.00	2.35	3.56
v2	Discrim	4.76	0.46	10.43	0.00	3.86	5.65
v3	Discrim	5.08	0.51	9.88	0.00	4.07	6.09
v4	Discrim	4.25	0.38	11.06	0.00	3.50	5.00
v5	Discrim	2.42	0.21	11.45	0.00	2.01	2.84
v6	Discrim	1.72	0.16	10.55	0.00	1.40	2.04
v7	Discrim	2.43	0.21	11.33	0.00	2.01	2.85
v8	Discrim	2.45	0.22	11.13	0.00	2.02	2.88

In the graph below, the ICCs of the two variables are depicted. The high discrimination coefficient of item **v3** is reflected by the relatively steep slopes at the difficulty levels for each category. Higher scores on the latent construct (CI) correspond to distinctly higher probabilities from 1 to 10 on the item **v3**. In contrast, for item **v6** the distributions of probabilities are relatively flat. Over relative wide ranges of scores on the latent construct, there is some probability of a certain score on the item, reflecting the low level of discrimination.

Figure 1. ICC of items v3 and v6



As a consequence of the differences in discrimination scores, not all items carry the same amount of information. Ideally, in social-economic research, scales give information over a wide range of the latent constructs they represent, thereby allowing to correlate scores on these constructs to other constructs. From the graph below we see that most of the information provided by the 8-item scale, stems from only three items: **v2**, **v3** and **v4**. Information is highest in the range from, approximately, -2.5 to +2 scores on

the underlying dimension (labeled *theta* in the graph). This is confirmed by figure 3, that combines all information contained in the eight individual items in the overall test information.

Figure 2. Item Information Functions of v1-v8

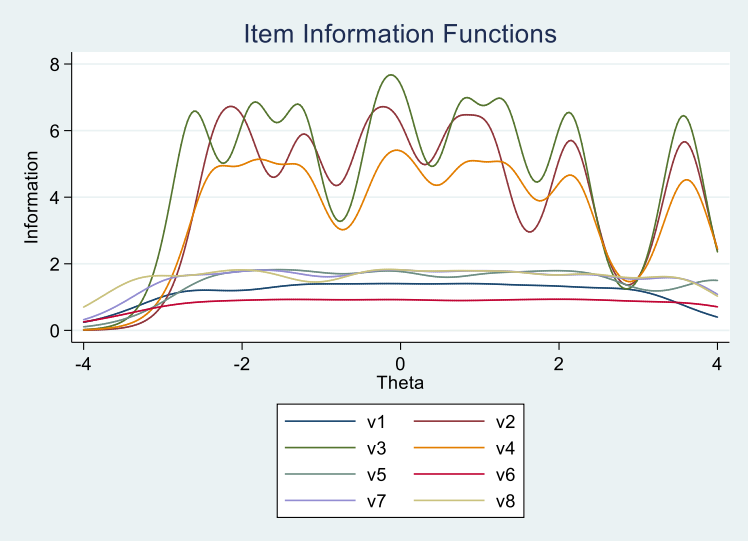
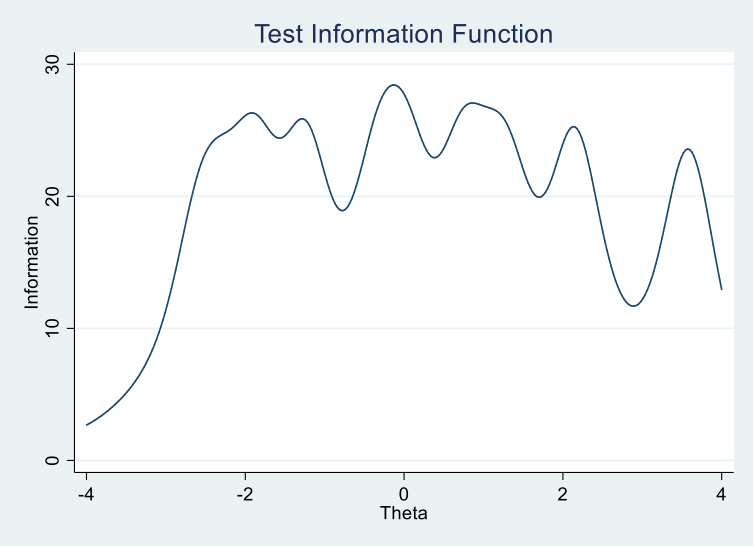
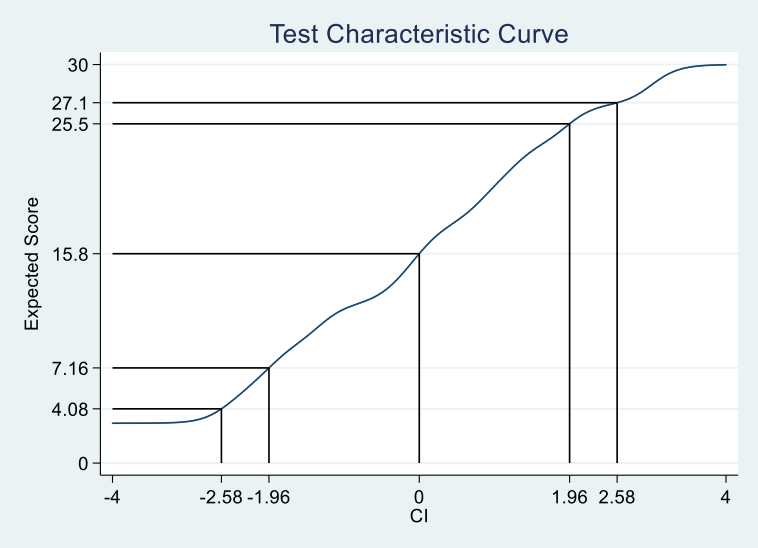


Figure 3. Test Information Function of CI



Given the item discrimination coefficients, and item information, we can without loss of information reduce the CI-scale from 8 items to just 3 items: **v2**, **v3** and **v4**. The test characteristic curve (TCC) shows how the scores on the latent construct (CI) are related to scores on the original items. The latent construct is normally distributed, with a mean of zero. Of the respondents, some 95% have scores from -1.96 to +1.96, which corresponds to scores of 4 to 27 on the original items. A total score of 27 is equivalent to, for example, scores of 9 on each of the three items. In between these scores, the relationship between the observed score and the latent score is approximately linear.

Figure 4. Test Characteristic Curve for CI



In short, sampling items **v2**, **v3** and **v4** from De Waal's 8-item score on CI, would produce results that are similar to scorers obtained using all eight items.

Long-term Orientation (LTO)

While CI contains some well-discriminating items that enable us to measure an acceptably broad range of the underlying construct, we use the LTO items (v31-v35) to illustrate that not all scales have that property. In this respect, a warning to researchers adopting existing questionnaire and scales like the HPO questionnaire is in place, since traditional measures like Cronbach's alpha can be deceptive. The alpha statistic for the 5 items of the LTO scale is equal to 0.76, which is according to commonly accepted norms, acceptable to good (Kline, 2000; DeVellis, 2012). However, all 35 items both within and between the five dimensions are quite highly correlated. We suspect that this is due to response bias and acquiescence effects (see Goedegebuure, forthcoming). As a consequence, the average of alpha scores of random samples of five items out of 35, is 0.816, and only 7.4% of these samples have alpha scores lower than 0.76. Against that benchmark the internal coherence of the items **v31**-**v35** is disappointingly low!

Skipping the tabular output of the GRM-model for the five LTO-items, the ICCs of the items with the lowest and highest discrimination (**v32** and **v35**, respectively) are shown in figure 5. The interpretation of the top part of figure 5 is that the probability for scores on the underlying LTO dimension of up to just above zero, is at an item score of 6. Actually, no respondents have uses item scores below 6. As a consequence, the item hardly provides information on the lower end of the LTO dimension. Figure 6 zooms in on **v32**. It shows that only scores of 8 and above on the item provide information on the higher end of the scale – but less information than already provided by item **v35** with its higher ability to discriminate at this part of the scale.

In contrast, **v35** provides information over a broader range of LTO. Still, very high scores on LTO (LTO>+2) are covered by item scores of 10 only, and as a consequence it is hard to make a distinction between respondents at the high end of the scale.

Figure 5. Item Characteristic Curves for items v32 and v35

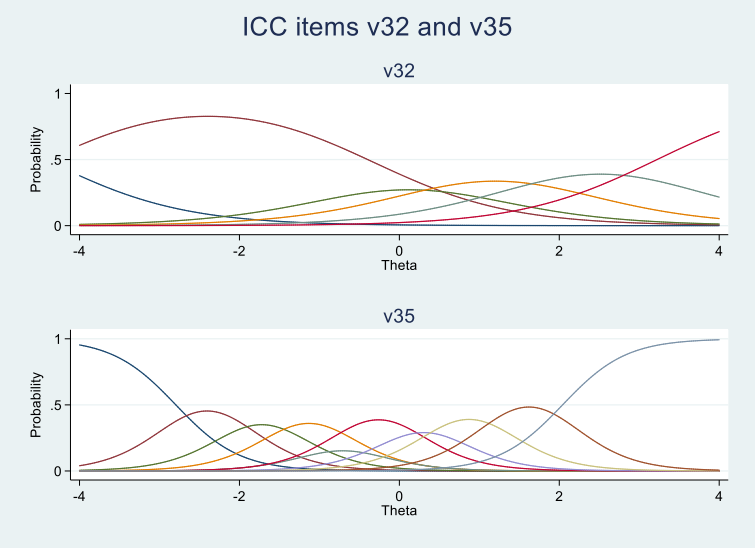
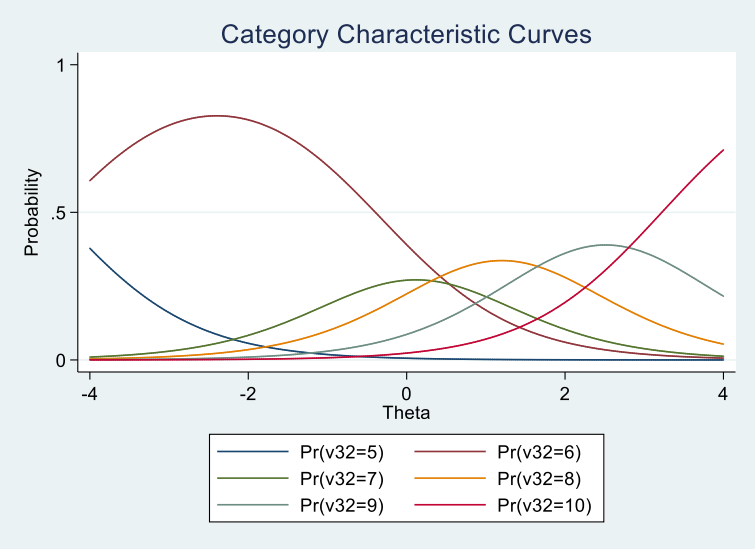
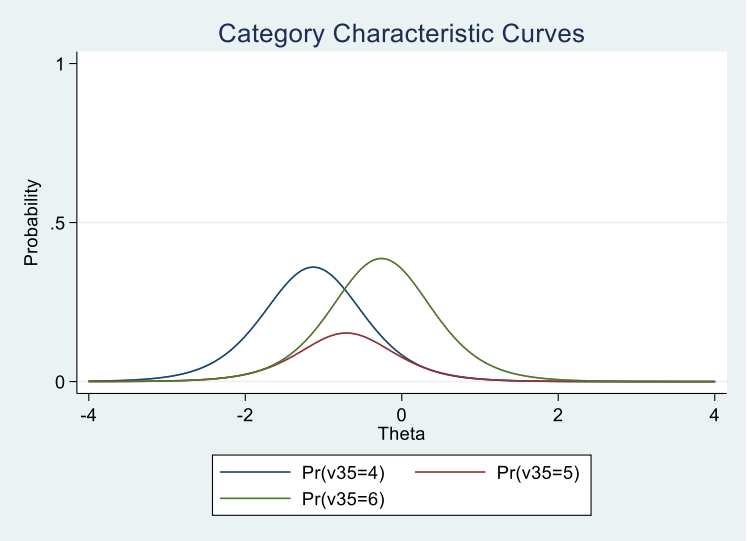


Figure 6. Item Characteristic Curve for items v32



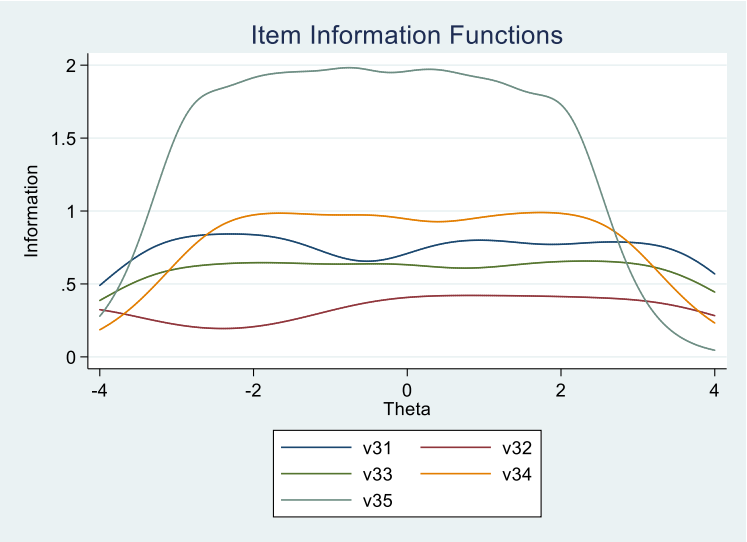
Typical for all 35 items in the questionnaire, is that the score of 5 is not informative: not at any point of the underlying dimensions (*theta*) is an item score of 5 the most likely score. This is an indicator that the 10-point scale suggest a level of accuracy that is overly ambitious. On the one hand, the items provide little information at the extremes (outside the range of -2 < *theta* < +2), and within that range a rating scales with less categories would suffice. A score of 5 might be interpreted as the midst of the scale (which is not quite the case, of course, as 5.5 is in the middle of the scale) and be sought by people who have no clear view on the matter; as a consequence the distribution for scores of 5 is much flatter than scores in the neighboring categories 4 and 6 (see figure 7).

Figure 7. Item Characteristic Curve for items v35 (item scores of 4-6)



In order to decide on the beat items to retain, we use the Item Information Functions, as in figure 8. While the bulk of the information in the middle range (from, say, -3 to +2) is provided by **v35**, it performs poorly above *theta*>+2. Item **v31** could be retained to cover this part of LTO presuming we have confidence that scores on 9 versus 10 on this item are telling.

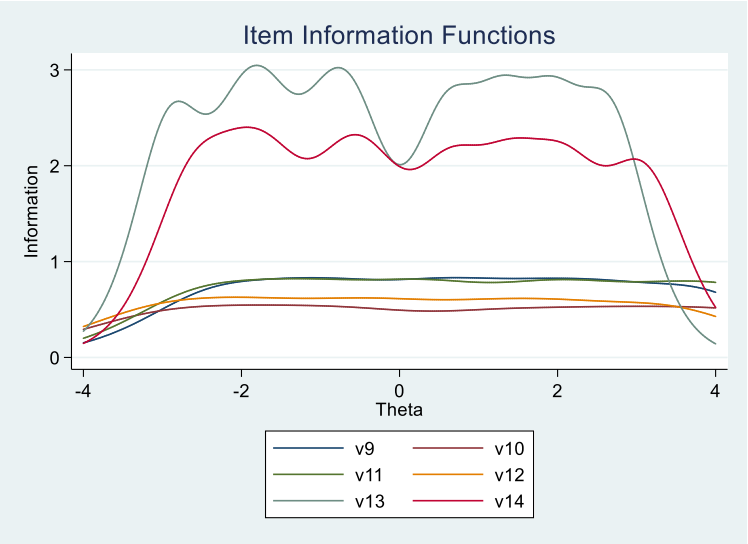
Figure 8. Item Information Functions for v31-v35



Openness & Action Orientation (OAO)

Following the same procedures for the OAO dimension, we suggest keeping **v13** and **v14**. All other items have low abilities to discriminate between respondents on this dimension, and hence add little information. Interestingly, **v13** performs somewhat on the lower end of the OAO scale, while **v14** is better able to distinguish between respondents at the higher end of the scale.

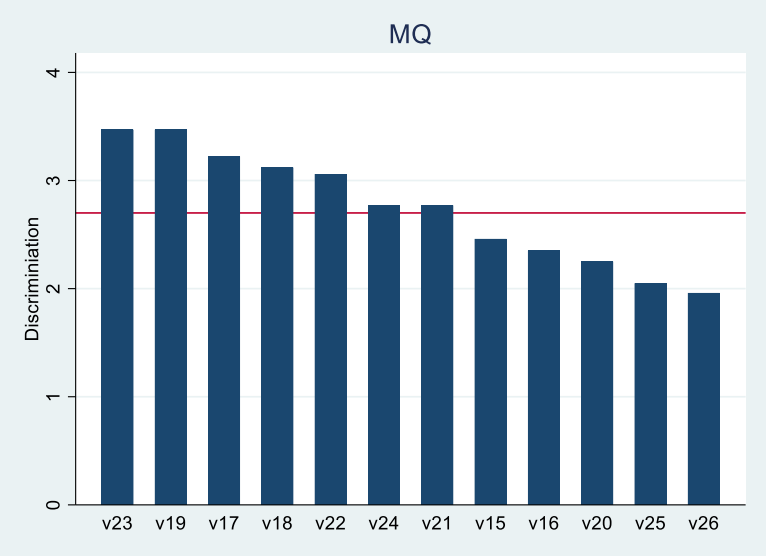
Figure 9. Item Information Functions for v9-v14



Management Quality (MQ)

Although no less than 12 items of the HPO model are meant to measure MQ, according to the IRT procedure used in our research, only four items (**v19**, and **v21**-**v23**) are informative. The selection was done in a sequence of steps, based on inspection of discrimination coefficients and TIF curves. Initially, the six items above with discrimination coefficients above the red line were kept, after which a GRM with fewer items was estimated (figure 10).

Figure 10. Discrimination Coefficients for items v15-v26



In some steps, items with relatively low discrimination coefficients were kept, since they added information at the extremes of the MQ scale. Of the four retained items, the information curves of **v19** and of **v22** seem to be encapsulated by **v21** and **v23**; the information curves run parallel to, but at a somewhat lower informational level, and therefore they can be left out without losing a lot of information (see figures 11 and 12).

Figure 11. Item Information Functions for MQ

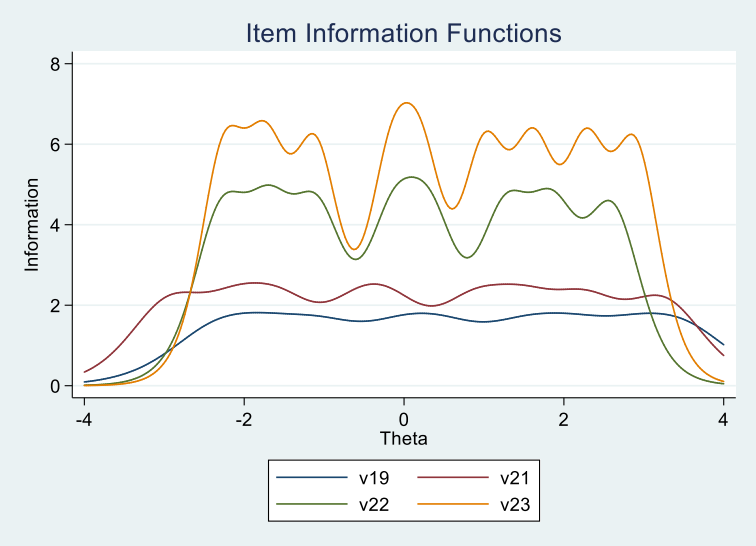
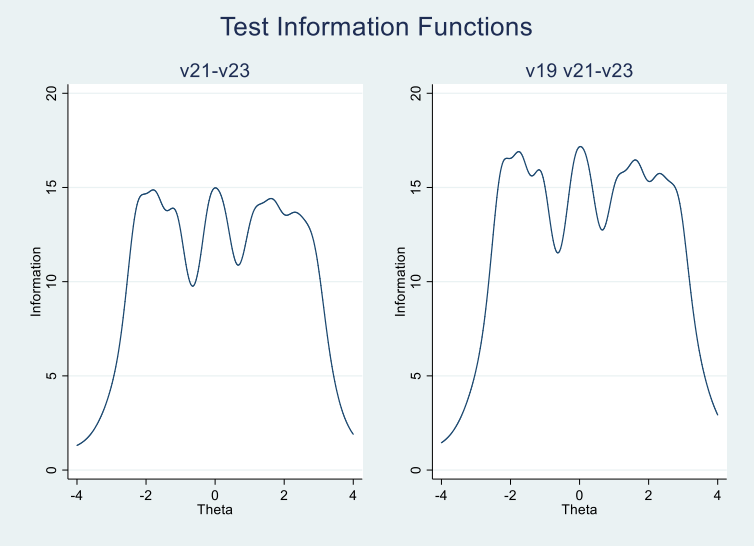


Figure 11. Test Information Functions MQ



Employee Quality (QEMP)

Lastly, QEMP has two items that provide most of the information on the construct: **v27** and **v28**. The item information functions and the test information function, are depicted in figure 12 and 13 below. The test information function is almost identical to (or dominated by) the information contained in **v28**.

Figure 12. Item Information Functions v27-v30

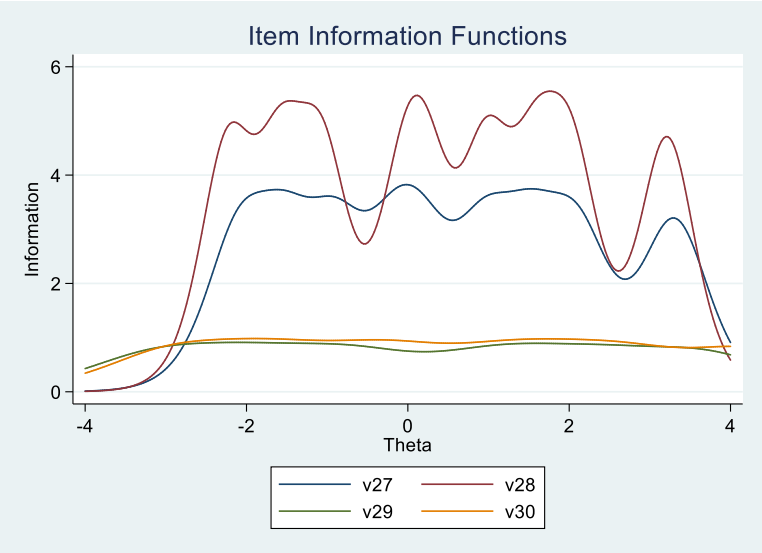
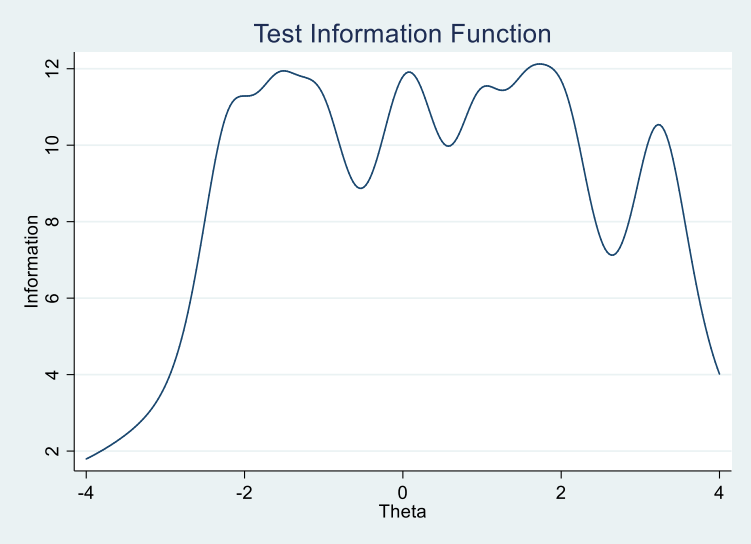


Figure 13. Test Information Function QEMP



CONCLUSIONS

In this paper we have illustrated the use of IRT, from the perspective of researchers adopting existing data collection instruments (scales and questionnaires). The illustration was done using the HPO questionnaire, consisting of 35 items that are assumed te measure the five factors of the HPO model. The HPO questionnaire was part of a questionnaire used to measure performance in three government owned or controlled Nepalese organizations (N=300).

Although the 35-item HPO questionnaire has been used for many years, in several countries, industries and organizations, our research indicates that the majority of the items can be easily left out without a loss of information.

In general, our findings serve as an invitation to researchers to critically review any standard instruments used in their studies, as it will shed light on the general usability of the scale, the informational value of the scale items, and efficient sampling of items.

Specifically, our findings suggest that our research could have done with a much smaller number of items for the HPO part of the study.

Scale	**Items (see annex 1)**		**Kept items**	
Continuous Improvement	8	v1-v8	3	v2-v4
Openness and Action Orientation	6	v9-v14	2	v13-v14
Management Quality	12	v15-v26	4	v19; v21-v23
Employee Quality	4	v27-v30	2	v27-v28
Long-term Orientation	5	v31-v35	2	v31; v35
Total	**31**		**13**	

In addition, our analyses suggest that the 10-point scale suggest an accuracy of measurement that has not been achieved in our sample. Moreover, the “midpoint” of the scale provides no information, probably due to the fact that people who have weak feelings about items are inclined to use a score of 5. A 10-point scale

could be useful if the items in the questionnaire were able to provide information on the (very) low and high ends of the scale, but this is – at least in our sample – not the case. The 10-point scale covers quite well the range of, broadly speaking, scores of -2 to +2 on the unobserved variables, but outside that range there's little information to accurately score the respondents. This is due to the fact that items are quite similar in difficulty. A better scale on (dimensions of) HPO, could benefit from adding items with differing levels of difficulty while leaving out items with similar levels of difficulty and low discriminations.

REFERENCES

Bogen, K. (1996). The effect of questionnaire length on response rates: A review of the literature. *Proceedings of the Section on Survey Research Methods*, Alexandria, VA. 1020-1025.

Kline, P. (2000). *The handbook of psychological testing* (2nd edition). London: Routledge

Raykov, T. and Marcoulides, G.A. (2018). *A Course in Item Response Theory and Modeling with STATA.* STATA Press

Reckase, M.D. (2009). *Multidimensional Item Response Theory.* New York: Springer

Sahlqvist, S., Song, Y., Bull, F., Adams, E., Preston J., Ogilvie, D. (2011). Effect of questionnaire length, personalisation and reminder type on response rate to a complex postal survey: randomised controlled trial. *BMC Medical Research Methodology,* 2011, 11-62

Waal, A.A. de (2007), The Characteristics of a High Performance Organization. Business Strategy Series, August

Waal, A. de, Duong, H. and V. Ton (2009), High Performance in Vietnam: The Case of the Vietnamese Banking Industry, Journal of Transnational Management, 14:179–201

André de Waal (2010), Achieving high performance in the public sector. What needs to be done? Public Performance & Management Review, 34, 1: 81–103

Waal, A. de and Tan Akaraborworn, C. (2014), Is the high performance organization framework suitable for Thai organizations?, Measuring Business Excellence, 17, 4: 76-87

ANNEX 1: HPO QUESTIONNAIRE

v1	CI: Our organization has adopted a strategy that sets it clearly apart from othe
v2	CI: In our organization processes are continuously improved
v3	CI: In our organization processes are continuously simplified
v4	CI: In our organization processes are continuously aligned
v5	CI: In our organization everything that matters to the organization's performanc
v6	CI: In our organization both financial and non-financial information is reported
v7	CI: Our organization continuously innovates its core competencies
v8	CI: Our organization continuously innovates its products, processes and services
v9	OAO: My manager frequently engages in a dialogue with employees
v10	OAO: Organizational members spend much time on knowledge exchange and learning f
v11	OAO: Organizational members are always involved in important processes
v12	OAO: My manager allows making mistakes
v13	OAO: My manager welcomes change
v14	OAO: Our organization is performance driven
v15	MQ: My manager is trusted by organizational members
v16	MQ: My manager has integrity
v17	MQ: My manager is a role model for organizational members
v18	MQ: My manager applies fast decision making
v19	MQ: My manager applies fast action taking
v20	MQ: My manager coaches organizational members to achieve better results
v21	MQ: My manager focuses on achieving results
v22	MQ: My manager is very effective
v23	MQ: My manager applies strong leadership
v24	MQ: My manager is confident
v25	MQ: My manager is decisive with regard to non-performers
v26	MQ: My manager always holds organizational members responsible for their results
v27	QEMP: My manager inspires organizational members to accomplish extraordinary res
v28	QEMP: Organizational members are trained to be resilient and flexible
v29	QEMP: Our organization has a diverse and complementary workforce
v30	QEMP: Our organization grows through partnerships with suppliers and/or customer
v31	LTO: Our organization maintains good and long-term relationships with all stakeh
v32	LTO: Our organization is aimed at servicing the customers as best as possible
v33	LTO: My manager has been with the company for a long time
v34	LTO: New management is promoted from within the organization
v35	LTO: Our organization is a secure workplace for organizational members