Chapter 6

Lessons Learned while Developing a Romanian Version of the *Mill Hill Vocabulary* Test*

John Raven**

Abstract

Whereas Raven's *Progressive Matrices* tests have repeatedly been shown to have impeccable test properties in many different cultures, it has proved remarkably difficult to develop a range of *Mill Hill Vocabulary* tests for cross-cultural use. This has more implications for the use of tests on a cross-cultural basis than might at first sight appear since, on the face of it, nothing could be simpler than generating equivalent sets of words for use in different cultures. The present paper shows that even generating parallel versions of the same IRT-based test, in open-ended and multiple-choice formats, is fraught with difficulties. For example, the introduction of apparently acceptable distractors into multiple-choice versions of items which function effectively in open-ended format can destroy them. The relative merits of alternative computer programs for carrying out the requisite analyses are assessed, and most found wanting.

* A version of this chapter is in electronic form on the Web Psych Empiricist http://www.wpe.info/papers_table.html

** The data on which this paper is based were collected by Anca Dobrean and Camelia Rusu with the assistance of Mircea Comşa, Robert Balazsi and numerous students. The questions the study sought to answer were raised by Camelia Rusu. The analyses were carried out by Joerg Prieler and Jean Raven.

Introduction

Raven's *Mill Hill Vocabulary* (MHV) tests (of which there are several versions derived from one basic version) were developed in 1938/39 to measure Spearman's "*reproductive*" ability alongside Raven's *Progressive Matrices* (RPM) tests - which measure Spearman's *eductive* ability.

The development of the scale is described in some detail in the relevant section of the *Manual* (Raven, J., Raven, J. C., & Court, 1998). As with the RPM, a graphical version of *Item Response Theory* was used to check whether the words scaled properly, and in the same way, in each of the versions of the test (specifically when in open-ended and multiple-choice format) and, in particular, whether information contained within the distractors in the multiple-choice versions interfered with the rank order of difficulty of particular words. The causes of any variation in the shapes of the *Item Characteristic Curves* (ICCs) across versions were investigated and the items either replaced or corrected. As with the RPM, these ICCs were plotted separately for children from different socio-economic backgrounds.

When revising the sequence of words, and later some of the words themselves, in response to changing word usage and cross-cultural (particularly UK - US) differences in word usage in the 1970s, use was made of information on both the average difficulty of the words in different cultures (specifically, Australia, the US, and the UK) and in the shapes of the item characteristic curves (Raven, 1981).

These questions re-surfaced in the context of the development of the Romanian version of the test.

More specifically, could one, using the, in some ways, more sophisticated computer programs that had been developed in the interim show that the words really had the same order of difficulty when presented in multiple-choice compared with open-ended format; could one demonstrate that the words which had been developed for the "parallel" versions of the test really had the same difficulty as, and functioned in the same way as, those they were thought to parallel; and could one demonstrate that, where words were of very different difficulty in the open-ended and multiple-choice forms of the test, this was due to the presence of certain distractors.

Due to the difficulties of replicating Raven's original procedures using the methods developed by Fischer (see Raven, Prieler, & Benesch, 2005) and even generating sets of 84 3-pl ICCs of the form developed by Benesch and Prieler (see above paper) using a DOS version of BILOG

(which had been shown to approximate the Raven/Fischer curves) an attempt was made to use an alternative program - RUMM - which had been promoted as a solution to most of our prayers.

Unfortunately, (i) collection of the relevant data from a nationally representative sample of Romanians was delayed due to technical reasons and (ii) we were unwilling to invest in a full version of RUMM without an assurance that it would answer all our questions. Accordingly we used the demonstration version of RUMM, which is limited to 99 respondents and 16 items.

Data drawn from a subsample of Romanian respondents for 16 items were therefore assembled for the analyses to be reported here.

At this point it is necessary to say a little more about the MHV itself.

There are two Forms of the *Senior* version of the test, which will be our concern here.

Each of these consists of two Sets of 34, hopefully parallel, words, known as Set A and Set B.

In Form I, Set A words are presented in open-ended (OE) format and Set B in multiple-choice (MC) format.

In Form II, Set A words are in multiple-choice format and Set B in open-ended format.

To check whether either of the sets of words are parallel when presented in the same or different formats two samples of individuals are required, one of whom has taken Form I and the other Form II.

Question 1: *Can RUMM generate 33 sets of ICCs yielding as much information as the sets of 3-pl ICCs shown in Figure 12 of Raven, Prieler, and Benesch?*

Answer: "No". Figures 6.1 and 6.2 show that the only ICCs we found ways of plotting using RUMM were the equivalent of 1-pl ICCs and give no indication of variation in slope (i.e. the discriminatory power of the items) or the effects of distractors (often misleadingly subsumed under the words "guessing" or "chance") on the shapes of those curves. Thus it is impossible from them to derive any insights into item functioning or how to correct malfunctions.

Question 2: *Do the same items have similar difficulty when presented in open-ended and multiple-choice format?*

Answer: "No". Using classical (i.e. not IRT-based) indices of item difficulty, it is clear from Table 6.1 that items A20, A22, A26, B22,

Figure 6.1. *Mill Hill Vocabulary Scale*: *Romanian Version*
Items 13-28 (Preliminary Data)
RUMM Item Characteristic Curves
Set A, Multiple-Choice

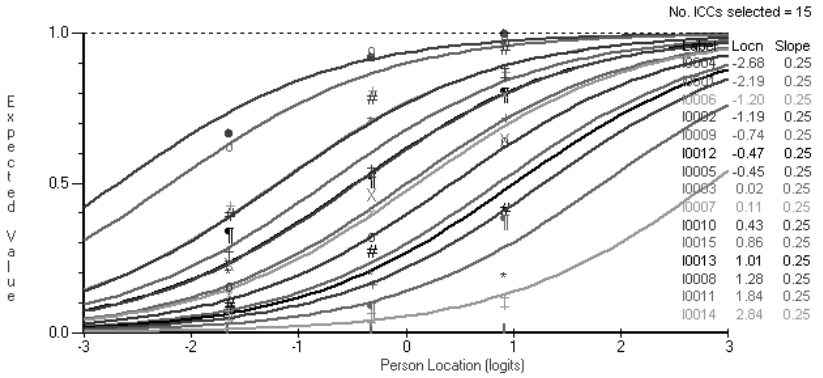
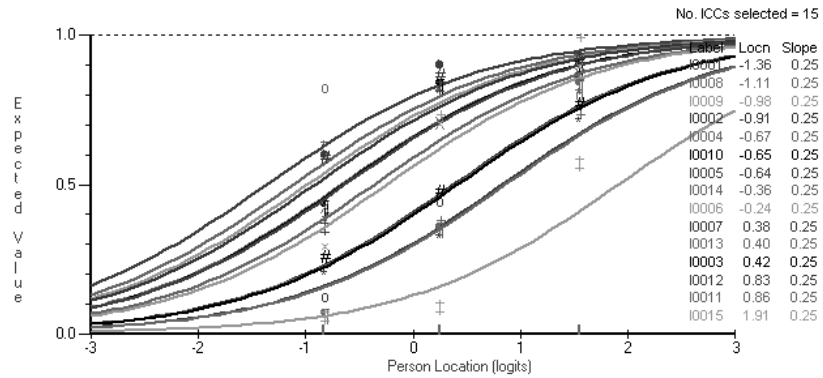


Figure 6.2. *Mill Hill Vocabulary Scale*: *Romanian Version*
Items 13-28 (Preliminary Data)
RUMM Item Characteristic Curves
Set A Open-Ended



B23, and B24 are very much more difficult in multiple-choice format. Something is distracting those who really know the answer!

Out of the sub-set of items included in this study, only B18 is easier in multiple multiple-choice format, presumably because of information contributed by the distractors.

Table 6.1 also shows the words that are not of equivalent difficulty in the parallel Set in the same format.

Table 6.2 shows that some words (e.g. A16, A26 and B23) have very different discrimination indices (i.e. item-total test correlations) in open-ended and multiple-choice formats. Some, e.g. A20, have poor discrimination in both formats.

Table 6.1. *Mill Hill Vocabulary Scale: Romanian Version*
Items 13-28 (Preliminary Data)
Comparative Item Difficulties of the Same Word in Multiple-Choice and Open-Ended Format and of "Parallel" Word in Other Set

	Item Difficulties (% Correct)			
	Set A		Set B	
	MC	OE	MC	OE
Item	Sample 2	Sample 1	Sample 1	Sample 2
13	83	82	59	76
14	68	75	85	60
15	41	48	44	23
16	86	72	44	42
17	52	68	35	43
18	68	61	57	29
19	40	48	42	31
20	16	79	84	76
21	56	75	65	57
22	33	70	34	73
23	13	39	16	65
24	51	38	34	60
25	23	48	83	87
26	4	64	75	64
27	27	22	32	11
28	30	17	23	20
n	92	93	93	92

Key: Difficulty of OE word very different from MC.
Difficulty of Set A word very different from Set B equivalent in same format.

Question 3: *Do the Distractor Characteristic Curves help us to find out whether some distractors confuse people who know the answer?*

Answer: "Yes". But we should first look at Figure 6.3, which presents the item distractor curves results for a reasonably well-functioning item - A15. It will be seen that choice of the correct answer increases with total score while choice of distractor 4 falls away. Thus the item works as it should, although some distractors attract no one and thus have no function.

Figure 6.4 shows that Item A17 behaves even better.

Figure 6.5, relating to item A20, shows something different. Choice of option 3 (the correct answer) falls with increasing total score, as does

Table 6.2. *Mill Hill Vocabulary Scale*: *Romanian Version*
Items 13-28 (Preliminary Data)
Comparative Item Discriminative Power of the Same Word in Multiple-Choice and Open-Ended Format and of "Parallel" Word in Other Set

	Item-Total Correlations			
	Set A		Set B	
	MC	OE	MC	OE
	Sample 2	Sample 1	Sample 1	Sample 2
Item	r_{it}	r_{it}	r_{it}	r_{it}
13	.40	.38	.50	.62
14	.50	.35	.51	.56
15	.31	.47	.51	.38
16	.46	.26	.60	.43
17	.51	.59	.52	.53
18	.43	.56	.67	.48
19	.56	.62	.56	.61
20	.13	.17	.62	.59
21	.64	.46	.38	.48
22	.45	.47	.42	.47
23	.42	.62	.20	.41
24	.40	.46	.26	.34
25	.35	.34	.31	.47
26	.10	.47	.55	.49
27	.58	.53	.39	.42
28	.42	.41	.32	.28
n	92	93	93	92
Cronbach Alpha	0.699	0.736	0.749	0767

choice of distractor 1, while choice of distractor 4 (a wrong answer) increases with total score.

Figure 6.6, relating to item A26, shows that distractor 1 attracts almost everyone and deflects them from the correct answer.

Figure 6.7, relating to item B22, again shows a well functioning item.

Figure 6.8, relating to item B23, again reveals how choice of a wrong answer can increase dramatically with total score.

Figure 6.3. *Mill Hill Vocabulary Scale*: *Romanian Version*
Items 13-28 (Preliminary Data)
RUMM Item Distractor Characteristic Curves
Item A15

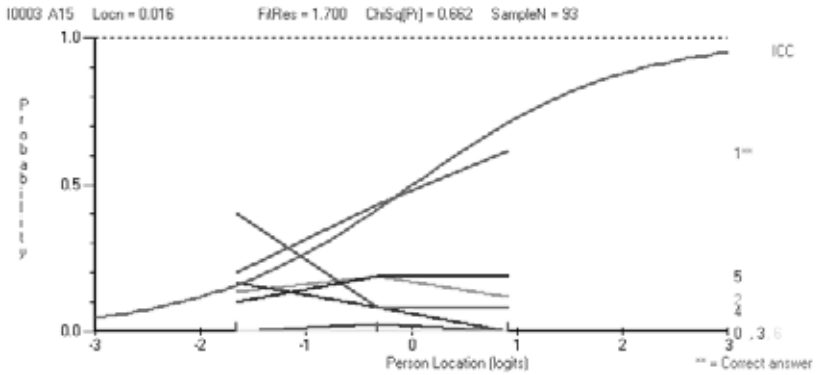
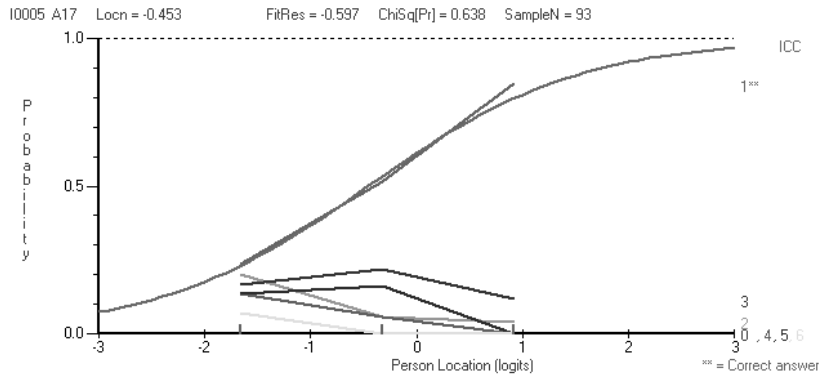


Figure 6.4. *Mill Hill Vocabulary Scale*: *Romanian Version*
Items 13-28 (Preliminary Data)
RUMM Item Distractor Characteristic Curves
Item A17



What do sets of 3-pl ICCs tell us?

The information assembled in Raven, Prieler, and Benesch (2005) shows that the original (1935) Raven ICCs reveal a great deal more about item functioning than do modern 1-pl ICCs … which reveal almost nothing. Nevertheless that paper shows that 3-pl ICCs do a good job of approximating the Raven curves. Although most of the differences between the 1- and 3-pl curves in that paper are not striking, it must be

Figure 6.5. *Mill Hill Vocabulary Scale*: *Romanian Version*
Items 13-28 (Preliminary Data)
RUMM Item Distractor Characteristic Curves
Item A20

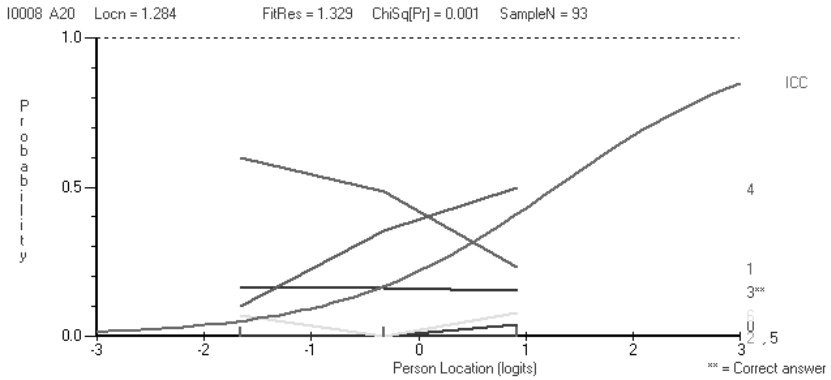
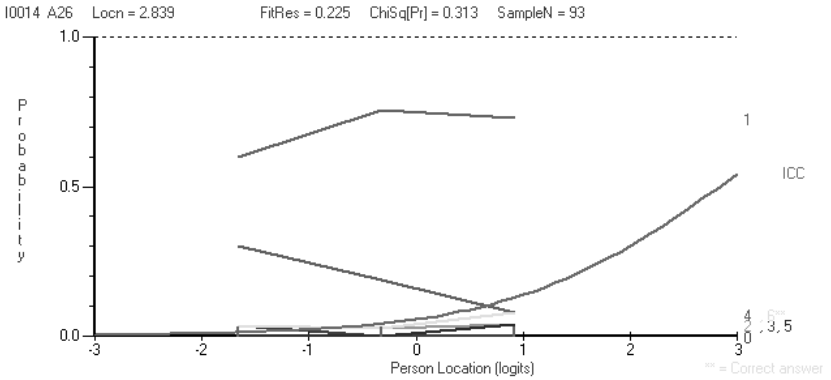


Figure 6.6. *Mill Hill Vocabulary Scale*: *Romanian Version*
Items 13-28 (Preliminary Data)
RUMM Item Distractor Characteristic Curves
Item A26



noted that *all* the data relate to good items which had been extensively worked over. This is why few serious defects can be discerned in the plot of 60 3-pl SPM*Plus* ICCs. The data presented in Figure 6.9 for the preliminary Romanian data for Set A in multiple-choice format (the RUMM 1-pl ICCs for which were discussed earlier) tell a different story, although it must immediately be reiterated that this is purely a methodological study based on a small sub-sample of people and items. It

Figure 6.7. *Mill Hill Vocabulary Scale*: *Romanian Version*
Items 13-28 (Preliminary Data)
RUMM Item Distractor Characteristic Curve
Item B22

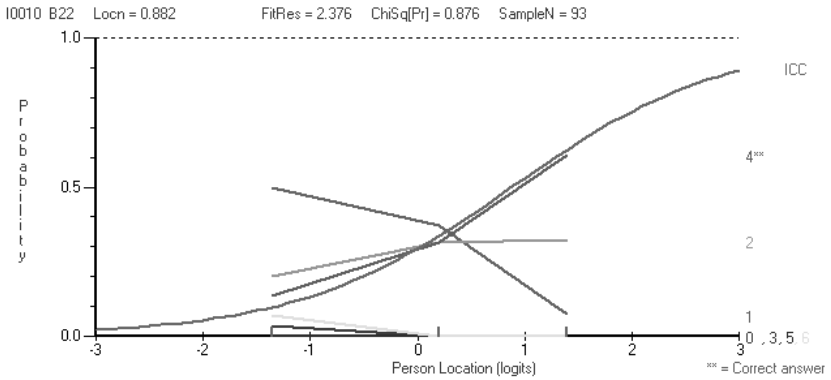


Figure 6.8. *Mill Hill Vocabulary Scale*: *Romanian Version*
Items 13-28 (Preliminary Data)
RUMM Item Distractor Characteristic Curve
Item B23

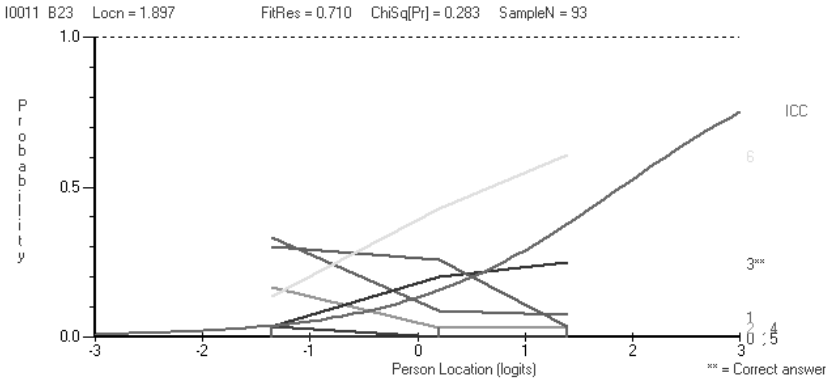
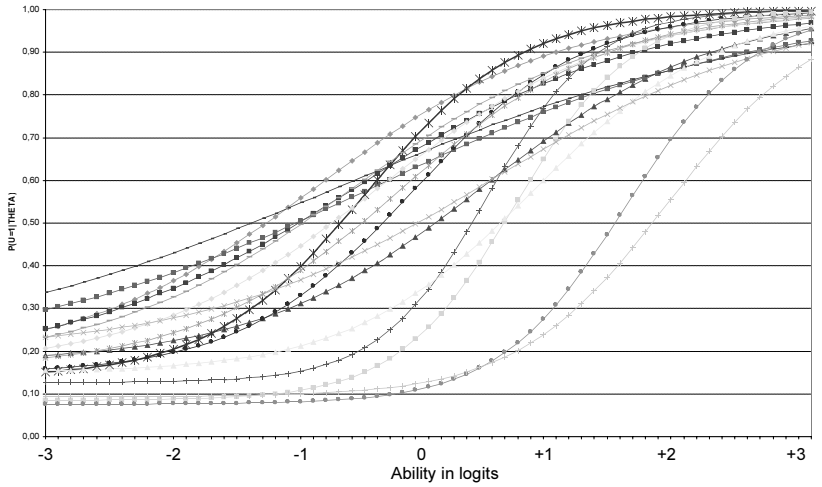


Figure 6.9. *Mill Hill Vocabulary Scale*: *Romanian Version*: *Preliminary Data*
Set A, Multiple-Choice
Items 13-28; Sub Set of Respondents
3-pl Item Characteristic Curves



is therefore extremely unlikely that the results displayed in Figure 6.9 will replicate when the full data set becomes available. Nevertheless, from a methodological point of view, the results are striking.

Although it is not possible in that Figure to identify which curve belongs to which item, it is obvious from their ICCs that some of the items are functioning very poorly. Their ICCs cross those for *all* the other items. Far too many low ability people get these items right and far too many high ability people never get them right. In other words, there is something about these items which leads low ability people to select the correct answer and something which prevents the more able from choosing it. As we have seen, plotting the *Item Distracter Characteristic Curves* enables one to become even clearer about what, exactly, was the problem with the items.

So, can the sub-tests be considered "parallel"?

It the course of discussions of the implications of the data presented above, it was suggested that the tests might nevertheless be considered "parallel" if the graphs of their *Test Information Functions* (TIF) were similar.

Test Information Function curves plot the quality of the diagnostic information provided by differences between test scores at different levels

of ability. Thus, if, as is commonly the case, the TIF curve is roughly Gaussian (often described as "normal"), it means that the test discriminates well among those of moderate ability but does a poor job among those with high or low ability. Thus, if one of the uses the data are to be put to is to differentiate among those who have been referred as potential candidates for Special or Gifted education, this is not exactly desirable. Thus, contrary to what might be expected, the ideal shape for a test information function curve might be rectangular or even bimodal (see Hambleton et al., 1991 for a fuller discussion.)

The *Test Characteristic Curves* (TCCs) for the 4 variants of the subset of MHV items and respondents discussed here are shown in Figure 6.10 and the *Test Information Function* curves in Figure 6.11.

The *Test Characteristic Curves* are not dramatically different, although it can hardly be said that they are "the same".

The same cannot be said for the *Test Information Function* curves. So it would seem that the deficits in the tests and the differences between them do show up here. It is therefore just possible that a "simple" comparison of the TIFs for different tests would enable one to decide whether they are to be considered interchangeable or not … but it would be of little help in deciding what to do about any differences that might be revealed.

****

Concluding Cautionary Note: *The above conclusions are entirely tentative*: *they are based on an analysis of a sub-set of items using a minute sub-set of the data that will become available. Absolutely no substantive conclusions should be drawn about the quality of the Romanian MHV which will eventually be published.*

Nevertheless, they heavily underline the importance of the methodological questions that were raised and indicate the analyses that would be required to answer them using the data from the full sample of respondents and items.

Figure 6.10. *Mill Hill Vocabulary Scale*: *Romanian Version*: *Preliminary Data*
Test Characteristic Curves
Items 13-28; Subset of Respondents
Set A, Multiple-Choice
Set A, Open-Ended
Set B, Multiple-Choice
Set B: Open-Ended

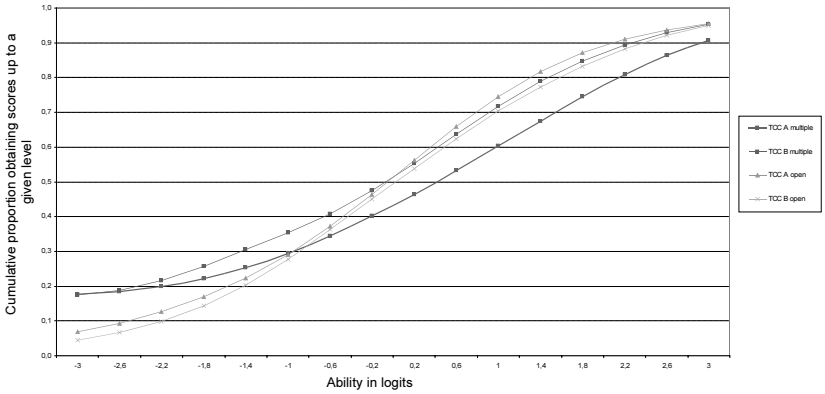
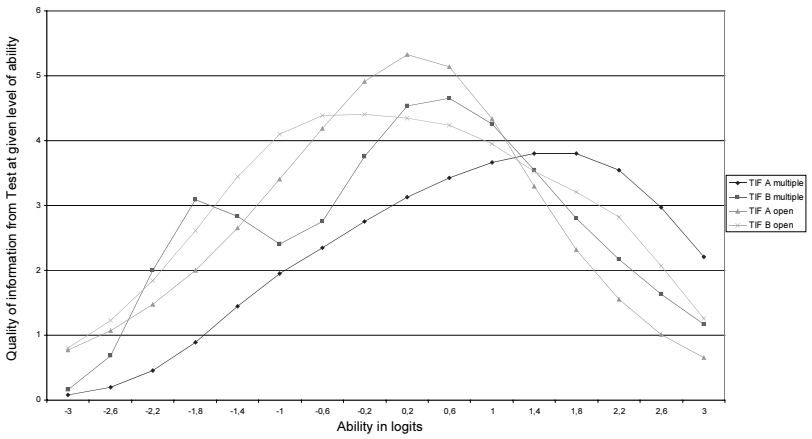


Figure 6.11. *Mill Hill Vocabulary Scale*: *Romanian Version*: *Preliminary Data*
Test Information Function Curves
Items 13-28; subset of respondents
Set A, Multiple-Choice
Set A, Open-Ended
Set B, Multiple-Choice
Set B: Open-Ended



References

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Press.

Raven, J. (1981). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research Supplement No.1*: *The 1979 British Standardisation of the Standard Progressive Matrices and Mill Hill Vocabulary Scales, Together With Comparative Data From Earlier Studies in the UK, US, Canada, Germany and Ireland*. San Antonio, TX: Harcourt Assessment.

Raven, J., Prieler, J, & Benesch. M. (2005). A replication and extension of the item-analysis of the Standard Progressive Matrices *Plus*, together with a comparison of the results of applying three variants of Item Response Theory. http://wpe.info/papers_table.html Updated in the previous chapter of this book.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 5*: *The Mill Hill Vocabulary Scale*. San Antonio, TX: Harcourt Assessment.