Psihološka obzorja / Horizons of Psychology, 17, 3, 5-24 (2008)
© Društvo psihologov Slovenije 2008, ISSN 1318-187
Znanstveni empiričnoraziskovalni prispevek

An empirical comparison of Item Response Theory and Classical Test Theory

Špela Progar[1] and Gregor Sočan[2]*
[1]Mirna Peč, Slovenija
[2]University of Ljubljana, Department of Psychology, Ljubljana, Slovenia

Abstract: Based on nonlinear models between the measured latent variable and the item response, item response theory (IRT) enables independent estimation of item and person parameters and local estimation of measurement error. These properties of IRT are also the main theoretical advantages of IRT over classical test theory (CTT). Empirical evidence, however, often failed to discover consistent differences between IRT and CTT parameters and between invariance measures of CTT and IRT parameter estimates. In this empirical study a real data set from the Third International Mathematics and Science Study (TIMSS 1995) was used to address the following questions: (1) How comparable are CTT and IRT based item and person parameters? (2) How invariant are CTT and IRT based item parameters across different participant groups? (3) How invariant are CTT and IRT based item and person parameters across different item sets? The findings indicate that the CTT and the IRT item/person parameters are very comparable, that the CTT and the IRT item parameters show similar invariance property when estimated across different groups of participants, that the IRT person parameters are more invariant across different item sets, and that the CTT item parameters are at least as much invariant in different item sets as the IRT item parameters. The results furthermore demonstrate that, with regards to the invariance property, IRT item/person parameters are in general empirically superior to CTT parameters, but only if the appropriate IRT model is used for modelling the data.

Key words: item response theory, classical test theory, measurement invariance, psychometrics

Empirična primerjava teorije odgovora na postavko in klasične testne teorije

Špela Progar[1] in Gregor Sočan[2]
[1]Mirna Peč
[2]Univerza v Ljubljani, Oddelek za psihologijo, Ljubljana

Povzetek: Teorija odgovora na postavko (TOP) temelji na modelu odnosa med latentno lastnostjo in odgovorom na postavko ter posledično omogoča neodvisno ocenjevanje parametrov postavk in oseb ter lokalno oceno napake merjenja. Kljub tem teoretičnim prednostim TOP v primerjavi s klasično testno teorijo (KTT) pa rezultati empiričnih raziskav pogosto ne pokažejo sistematičnih razlik med parametri po KTT in TOP ter med stopnjo invariantnosti parametrov po KTT in TOP. V pričujoči empirični študiji smo skušali, na podlagi podatkov iz Mednarodne raziskave trendov znanja iz matematike in

[*] Naslov/Address: doc. dr. Gregor Sočan, Univerza v Ljubljani, Oddelek za psihologijo, Aškerčeva 2, 1000 Ljubljana, tel.: +386 1 241 11 84, fax.: +386 1 42 59 301, e-mail: gregor.socan@ff.uni-lj.si

naravoslovja (TIMSS 1995), odgovoriti na naslednja raziskovalna vprašanja: (1) Kako primerljivi so parametri postavk in oseb po KTT in TOP? (2) Kako invariantni so parametri postavk KTT in TOP preko različnih skupin oseb? (3) Kako invariantni so parametri postavk in oseb po KTT in TOP preko različnih skupin postavk? Ugotovili smo, da so parametri oseb in postavk po KTT in TOP zelo primerljivi, da so parametri postavk po KTT in TOP podobno invariantni v različnih skupinah oseb, da so parametri oseb po TOP bolj invariantni preko različnih skupin postavk ter da so parametri postavk po KTT, ocenjeni iz različnih skupin postavk, v enaki meri ali celo bolj invariantni od parametrov postavk po TOP. Rezultati so pokazali tudi, da so na splošno parametri postavk po TOP, v smislu invariantnosti parametrov, empirično superiorni parametrom po KTT, vendar le v primeru, ko se uporabljeni model TOP v zadostni meri sklada s podatki.

Ključne besede: teorija odgovora na postavko, klasična testna teorija, merska invariantnost, psihometrija

CC = 2200

Item-response theory (IRT) appears to be the currently prevailing paradigm within the psychometric theory. However, this is only partially reflected in the psychometric practice: with an important exception of educational measurement, most psychological measuring instruments still appear to be based on the classical test theory (CTT). The aim of this study was to examine some possible causes of this gap, besides conservatism of the psychometric practitioners. We tried to provide a partial answer to two commonly asked questions: “To what extent does the choice of the test construction method influence the properties of the resulting test?” and “To what extent are the theoretical advantages of the IRT reflected in empirical superiority of the resulting measurement instruments?” In the sequel, we shall first review some important properties of both paradigms and summarize the available empirical evidence.

The main advantage of CTT is its simplicity: the estimation of the measured trait is performed simply by adding the scored responses to items. CTT does not involve truly latent variables: despite the fact that the true score is not empirically observable, it can be defined operationally as the average score on the infinite number of equivalent repetitions of the measurement process (see, for instance, Lord and Novick, 1968). Accordingly, within the CTT framework, the question of model validity is almost never addressed. The basic equation of CTT, the additive decomposition of the observed score into the true score and the random error, is a tautology rather than a model and can not possibly be subjected to a test. Groups of assumptions, called “measurement models” (most important among them being the congeneric and the tau-equivalent model, respectively) do play a role in CTT, but they often affect optimality rather than validity: for instance, if the test items depart largely from the essential tau-equivalence, coefficient alpha will remain the lower bound to reliability, but will be a relatively inefficient lower bound. On the other hand, the assumptions which do affect the validity of the results—for instance, the assumption of uncorrelated errors––are mostly impossible to be tested empirically for practical reasons. Finally, some parts of CTT are based on a model which is incorrect *a priori*: for instance, although the reliability can be defined as the squared correlation coefficient between the true and observed scores, this relationship is not linear almost surely whenever the range of possible scores is bounded (which is almost always the case).

On the other hand, IRT is a model-based paradigm: it starts with modelling the relationship between the latent variable being measured and the item response. The aim is, at least in principle, to find an accurate model rather than a robust approximation. The question of model fit therefore always plays an important role in an IRT analysis. An important feature of the modelling approach is that the parameters of the persons do not depend on parameters of the items, and vice versa. Consequently, parameters of the persons are invariant across items, and parameters of the items are invariant in different populations of persons. However, these properties depend on the validity of the item-response model. Since real data never perfectly fit to some reasonably parsimonious abstract model, questions can arise about how large the lack of fit can be so that one can still rely on these properties, and whether IRT is always superior to CTT with regard to accuracy and invariance of parameter estimates. Little evidence is available so far regarding these questions, partly perhaps because the model fit has traditionally been treated dichotomously rather than in sense of the degree of fit.

Apart from the invariance problem, another relevant question is whether one should expect IRT and CTT to produce substantially different tests when selecting items from a larger pool. A common feature of both approaches is to discard items to which the answers are not related, or are even negatively related to the test score. In classical test construction (see, for instance, Nunnally and Bernstein, 1994), the primary item selection criterion is the corrected item-total correlation. Additionally, the spread of item difficulties should be large enough so that the discrimination power is not concentrated around the average scores only. Because of the stress on the item discrimination indices, classical tests can be expected to have both relatively high internal consistency and high unidimensionality[1]; the former may be quantified by means of coefficient alpha or some other similar reliability coefficient, and the latter by means of the proportion of common variance, explained by the first common factor. On the other hand, IRT test analysis does not involve the notion of reliability, at least not as defined in CTT. So an IRT-based test may not have a very high internal consistency; instead, it will not tend to concentrate the discrimination power around average score – in fact, the concentration of the discrimination power can be set according to the aim of the test constructor. In any case, somewhat different products should be expected depending on the paradigm used in the process of test construction.

[1] It should be noted that the relationship between reliability and unidimensionality is intricate; for a discussion, see Ten Berge and Sočan (2004).

Recently, several studies compared CTT and IRT, mostly with regard to the comparability of the item and person parameters. For instance, Fan (1998) in his study, based on real data, found very high correlations both among the person parameters (correlations were in all instances higher than .96) and among the item difficulties (correlations higher than .90). He further found no evidence of a higher invariance of the IRT item parameters in comparison to the CTT item parameters. Similar conclusions were reached by Courville (2005) in another empirical study.

An obvious limitation of empirical studies is that the parameter values can not be manipulated and that the true values are not known. MacDonald and Paunonen (2002) conducted a Monte Carlo study in which they controlled the spread of item difficulty and item discrimination. Similarly to Fan (1998) and Courville (2005), they found high correlations between the CTT and the IRT difficulties. The discrimination indices, however, correlated highly only when the spread of discriminations was large and the spread of difficulty values was small. Moreover, the CTT discrimination estimates were in some conditions (i.e., at a large spread of difficulties) less accurate than the IRT estimates.

The abovementioned studies did not attempt to construct a test by selecting items from a larger item pool and to subsequently investigate the invariance of the parameters across such item sets, although this may be a fairly typical situation in practice. Our study was therefore set into the item selection framework. Within this framework, it attempted to address the following standard questions:

- How comparable are CTT and IRT based item and person parameters?
- How invariant are CTT and IRT based item parameters across different participant groups?
- How invariant are CTT and IRT based item and person parameters across different item sets?

Method

Data source

The data used in our study are from the Third International Mathematics and Science Study (TIMSS 1995) administered in 1994 and 1995 to the third and fourth grade students in 45 different countries. Mathematical achievement was measured with 102 and science achievement with 97 different items, respectively. Items were divided in 26 exclusive clusters, which were distributed into 8 different test booklets. Most of the items (approx. 80%) were multiple-choice items, while others required short or more elaborated answers. Complete international database is available on *http://isc.bc.edu/timss1995i/Database.html*.

Since test booklets consisted of different clusters of items, each student responded just to a subset of items, which represents a problem for data analysis using CTT. Therefore only items from two different test booklets were selected and analyzed for the purpose of this study, i.e. items from booklet number 5 for the mathematical achievement and booklet number 6 for the science achievement. Booklet number 5 consisted of 39 mathematical items (two items were eliminated, because they were not administered in Latvia) and booklet 6 consisted of 37 science items as shown in Table 1. These items represent math and science item pools. Items were dichotomous, except for three mathematical items and one science item, which were subsequently dichotomized for the purpose of this study.

Table 1. *Structure of mathematical and science item pools.*

	Math Item Pool	Science Item Pool
Multiple-choice items	25	26
Short answer items	7	3
Elaborated answer items	7	8
Total	39	37

Participants

From the complete international database only students from six European countries (Hungary, Latvia, Netherlands, Norway, Scotland, and Slovenia) that responded to items from booklet 5 and 6 were selected. The structure of students that responded to mathematical and science items is shown in Tables 2 and 3.

Table 2. *Structure of students that responded to mathematical items (from booklet number 5).*

	Lower grade				Upper grade			
	Female		Male		Female		Male	
	N	%	*N*	%	*N*	%	*N*	%
Hungary	175	18.7	199	19.5	189	19.6	182	18.3
Latvia	125	13.3	126	12.3	135	14.0	143	14.4
Netherlands	152	16.2	201	19.7	162	16.8	152	15.3
Norway	134	14.3	132	12.9	132	13.7	149	15.0
Scotland	199	21.2	192	18.8	197	20.5	213	21.5
Slovenia	152	16.2	171	16.7	148	15.4	153	15.4
Total	937	100.0	1021	100.0	963	100.0	992	100.0

Note. 35 students for which gender was not indicated are not included in this table.

Table 3. *Structure of students that responded to science items (from booklet number 6).*

	Lower grade				Upper grade			
	Female		Male		Female		Male	
	N	%	*N*	%	*N*	%	*N*	%
Hungary	195	20.2	175	18.0	187	19.4	182	18.2
Latvia	130	13.5	127	13.1	139	14.4	137	13.7
Netherlands	172	17.9	176	18.1	162	16.8	156	15.6
Norway	134	13.9	142	14.6	124	12.9	150	15.0
Scotland	178	18.5	195	20.0	199	20.6	214	21.4
Slovenia	154	16.0	158	16.2	153	15.9	163	16.3
Total	963	100.0	973	100.0	964	100.0	1002	100.0

Note. 49 students for which gender was not indicated are not included in this table.

Data analysis

Data from both the math and the science item pool were analyzed using both CTT and IRT procedures. Missing value analysis was conducted using EM (expectation-maximization) algorithm in *SPSS 11.0 for Windows* before item analysis (the percentage of students that had at least one missing response was 0.3 % in the math item pool and 0.5 % in the science item pool).

Within the CTT framework the following parameters were computed:

- person parameter as the proportion of correct answers,
- item discrimination parameter as the corrected point-biserial correlation,
- item difficulty parameters as the proportion of correct responses to particular items and
- alpha reliability coefficient and the greatest lower bound to reliability (GLBR). The latter was computed according to the algorithm proposed by Ten Berge, Snijders and Zegers (1981).

The greatest lower bound to reliability (GLBR) is the highest value which is certainly not higher than the actual reliability in the sample. It is therefore the most accurate conservative reliability estimate possible. GLBR does not have a closed-form solution and can only be estimated by numerical algorithms.

The IRT analysis was conducted by means of *BILOG-MG 3.0* (Zimowski, Muraki, Mislevy, & Bock, 1996). All IRT estimations were obtained using the marginal maximum likelihood (MML) method with normal prior distribution, which is the default for *BILOG-MG*.

Within the IRT framework the following parameters were computed:

- person parameter (commonly known as the *theta* value),
- item discrimination (slope) parameter (the *a* value),

- item difficulty (location) parameter (the *b* value).

Information about the item and model fit was obtained through the differences in log likelihoods between different unidimensional logistic models (1PL, 2PL, and 3PL models) and through comparisons between expected and empirical item characteristic curves (ICC's; χ^2 test and the graphical method). Unidimensionality of the tests was analyzed with minimum rank factor analysis (MRFA; Ten Berge and Kiers, 1991), which is the single available factor analytic method making possible the determination of the proportion of the common variance, explained by some number of common factors. In contrast to this, other similar methods only make possible to evaluate the proportion of the total variance, explained by the common factors, which is however not directly associated with unidimensionality. In our case, the proportion of the common variance of the items, explained by the first common factor, was taken as a measure of unidimensionality.

Comparability of CTT and IRT person and item parameters

Comparability of the CTT and the IRT person parameters was assessed by correlating the proportion correct and theta values for participants in both math and science item pool. Comparability of the CTT and the IRT item parameters was assessed by correlating the CTT and the IRT based item difficulty parameters and item discrimination parameters.

Invariance of CTT and IRT item parameters across different participant groups

The degree of the CTT and IRT item parameters invariance was assessed by comparing the item parameter estimates within each measurement framework across two or more different participant groups: (1) lower and upper grade, (2) male and female, (3) Hungary, Latvia, Netherlands, Norway, Scotland, and Slovenia. Comparison included correlations between item parameter estimates in different groups (grade, gender, country) and comparing means of item parameters (paired *t*-test and ANOVA).

Invariance of CTT and IRT person and item parameters across different item sets

After both the CTT and the IRT item parameters had been obtained for all items in both item pools, four new tests (two mathematical and two science tests) of 12 items were constructed. For the construction of new CTT tests highly discriminative items with approximately normally distributed item difficulties were selected. For the new IRT tests highly discriminative and informative items with item difficulty parameters across wide range of theta values were selected (the target test information function was high through all levels of latent variable). Since the item pools were rather small, only psychometric information was considered when constructing new tests.

Item and person parameters in these four new tests were compared with the original parameters, which were obtained with all items in both item pools. Comparison included correlations between the *new* and the *original* parameter estimates within each measurement framework as well as comparing mean differences of person/item parameters (paired *t*-test).

Results and Discussion

IRT Model Fit Assessment

The main goal of our study was to investigate both the comparability of CTT and IRT item parameters and the invariance of CTT and IRT item parameters in different conditions. As we were especially interested in item discrimination parameters, which usually represent the most important criteria for item selection, but also appear to be the most unstable item parameters, the 1PL model that would only allow the inspection of item difficulty parameters was not of interest. We were also not interested in the 3PL model for two reasons: (1) the 3PL model with a unique lower asymptote for each item would lead to different meaning of item difficulty for each item in the test and to different meaning of item difficulty for the same item in different item sets, and this would bias the results of comparison of item parameters across different item sets; (2) the 3PL model with a common lower asymptote for all items was not reasonable because not all items in our item pools were multiple-choice items.

The assessment of the IRT model fit indicates that the 2PL model comparing to the 1PL model fits to the data significantly better for all tests (the differences in *–2 log likelihoods*) and also that the 3PL model fits to the data better than the 2PL for two tests only. According to the χ^2 test empirical ICC's differ significantly from expected ICC's in 2PL model for only some items in the math item pool, but for more than half of the items in the science item pool (Table 4). Since the χ^2 test in *BILOG-MG* can only be computed for tests with more than 20 items, only graphical comparison of expected and empirical ICC's could be made for new tests; the comparison showed that practically all items in new tests fit the 2PL model.

Table 4 also presents the main results of the (uni)dimensionality analysis. We can conclude that the assumption of unidimensionality holds to a reasonable extent in all new tests and in the math item bank. In the case of science item bank, however, the assumption of unidimensionality is clearly violated, which is probably the reason for a poor fit of many science items to the unidimensional 2PL model according to the results of χ^2 test. The unidimensional 2PL model is obviously not fully adequate for the science item bank (a multidimensional model would be appropriate), but was nevertheless used for modelling these data as well. It should be emphasized at

Table 4. *Analysis of unidimensionality, reliability and IRT model fit for 2PL model for both item pools and for the new tests.*

	Math			Science		
	Item Pool	New CTT test	New IRT test	Item Pool	New CTT test	New IRT test
% common variance	37.9	42.4	42.6	33.4	35.8	35.5
% explained common variance	51.7	73.0	72.4	38.4	57.1	57.7
Alpha reliability	0.893	0.833	0.819	0.819	0.724	0.710
GLB to reliability	0.920	0.864	0.857	0.865	0.784	0.775
% of misfitting items (χ^2 test)[a]	29.7	N/A	N/A	51.3	N/A	N/A

Note. N/A = not available

[a] χ^2 test cannot be computed for new tests which consist of less than 20 items.

this point that all results involving the science item bank consequently demonstrate invariance of item/person parameters in the case of fairly heterogeneous test and a rather poor IRT model fit.

Comparability of CTT and IRT person and item parameters

The first goal of our study was to assess the comparability of CTT and IRT person and item parameters. As can be seen from the results in Table 5, CTT and IRT person parameters correlate very highly in both item pools, indicating that very similar math/science achievement estimates would be obtained regardless of the measurement framework. However, the comparison of the distributions of the CTT and IRT person parameters shows that this indication is only partly true. Since both math and science item banks contain more easy items then hard items, the distribution

Table 5. *Correlations between CTT and IRT person and item parameters.*

	Math Item Pool	Science Item Pool
Person parameters	0.984	0.990
Item parameters		
Item difficulty	0.972	0.922
Item discrimination	0.935	0.831

Note. In order to make the correlations positive, CTT item difficulty parameters were reversed so that higher values relate to more difficult items. Math item pool consisted of 39 items; science item pool consisted of 37 items.

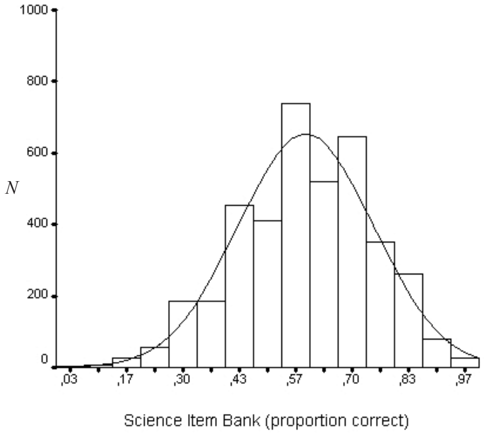

Figure 1. Distribution of proportion correct in science item pool. The curve represents normal distribution.

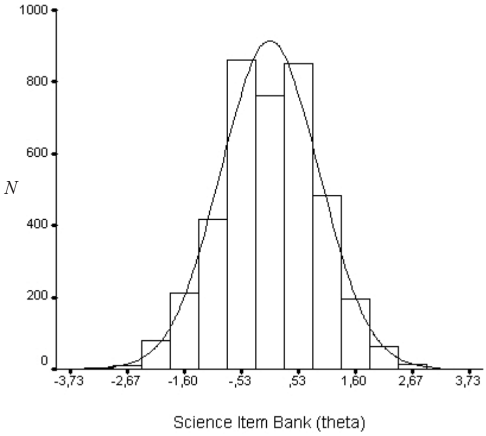

Figure 2. Distribution of theta values in science item pool. The curve represents normal distribution.

of the CTT person parameters in both item banks is negatively skewed. In contrast to this, the distributions of the IRT-based person parameters are more symmetrical, which could be the result of either item parameters being incorporated in the assessment of person parameters or the effect of prior normal distribution. The difference in person parameter distributions (for example, see Figures 1 and 2) could therefore be the result of the fact that the CTT based person parameters depend on the item parameters whereas the IRT person parameters do not, but, since neither true classical nor true IRT scores are known in this study, this conclusion would be unfounded.

The CTT and IRT item parameters correlate very highly, as well as the person parameters (also in Table 5). Item difficulty parameters correlate higher than item discrimination parameters, but the latter correlate very high as well in all conditions. Both the item difficulty and the item discrimination parameters are more comparable in the case of mathematical item pool, where the 2PL model fits the data better.

Invariance of CTT and IRT item parameters across different participant groups

The comparison of CTT and IRT item difficulty and item discrimination invariance across different groups of participants (regarding grade, gender and country) is presented in Table 6. Both CTT and IRT item difficulty parameters were estimated very consistently in all conditions. The highest consistence was found across the gender groups and the lowest one across different countries. The IRT item difficulties are slightly more consistently estimated than the classical item difficulties in most conditions for the math item pool, where the 2PL model fits to the data reasonably well, but less consistently in all conditions for the science item pool, where the model

Table 6. *Comparison of IRT and CTT item parameters invariance across different groups of participants.*

	Math Item Pool			Science Item Pool		
	Grade	Gender	Country	Grade	Gender	Country
Item difficulty						
CTT	0.946	0.979	0.787	0.983	0.983	0.859
IRT	0.953	0.980	0.783	0.973	0.979	0.835
Item discrimination						
CTT	0.894	0.945	0.657	0.770	0.903	0.601
IRT	0.930	0.943	0.694	0.698	0.877	0.514

Note. Correlations between item parameters estimated in two or more groups are presented. In case of country, single measure intraclass correlations (consistency) are shown. Math item pool consisted of 39 items, science item pool consisted of 37 items.

fit is rather poor. These results are coherent with presumption that IRT item parameter invariance can only be expected when the model fits the data (i.e., when the assumptions about the data are valid; Embretson & Reise, 2000; Hambleton, Swaminathan & Rogers, 1991). Considering that the IRT item parameter estimates for science data are not optimal within the unidimensional 2PL model, item difficulties are actually surprisingly invariant across different participant groups.

Both the CTT and the IRT item discrimination parameters are less invariant across different participant groups than the item difficulty parameters. The difference between item difficulty invariance and item discrimination invariance is especially high in the case of science data and for the IRT parameters; it seems that poor model fit has greater impact on the IRT item discrimination estimates than on the IRT item difficulties estimates. Similarly to item difficulty parameters, the consistency of IRT discrimination estimates is higher, compared to classical parameters, for most conditions in the case of math item pool, but lower in all conditions in the case of science item pool.

Another important issue remains to be discussed. The fact that item parameters were consistently estimated in two or more different participant groups does not mean that the item parameters are actually invariant. High correlations between item difficulties indicate that items which are easier, for instance, for students in the lower grade, also tend to be easier for students in the upper grade. The absolute values of item difficulty parameter estimates in the lower and in the upper grade however differ (see Table 8 and Figures 3 and 4 for example). Since students from upper grade probably know more mathematics and science than students from lower grade, this difference is expected for the CTT item difficulties but not for the IRT item difficulties. It should be noted that, since IRT person/item parameter scale is arbitrary, IRT item parameters obtained from different groups of participants should be rescaled before the comparisons between them can be made (Embretson & Re-

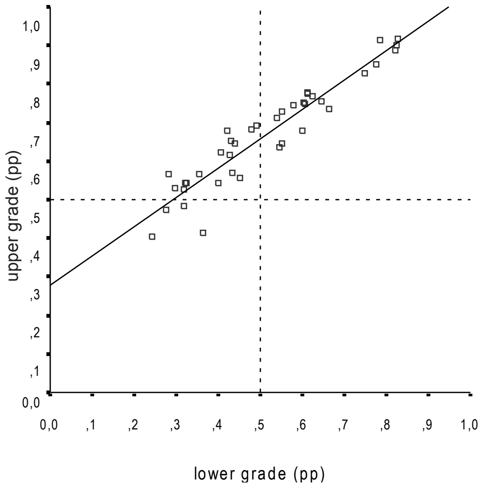

Figure 3. CTT item difficulty parameter estimates (pp) from math item pool in lower and upper grade.

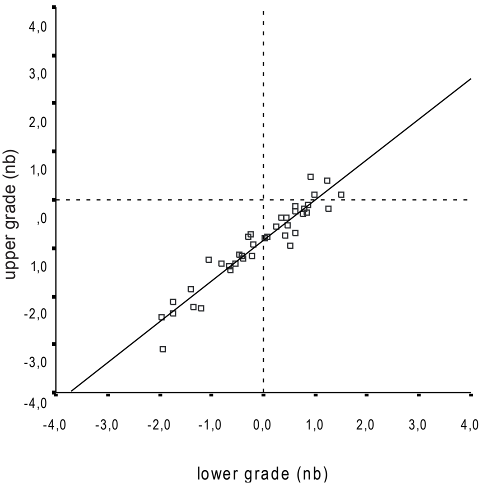

Figure 4. Non-rescaled IRT item difficulty parameter estimates (nb) from math item pool in lower and upper grade.

ise, 2000; Hambleton et al, 1991). But in Figure 4 non-rescaled item difficulties are shown; if these parameter estimates are rescaled, the IRT item difficulties are not only consistent but also (almost) invariant in the absolute sense. For the CTT item parameters, on the other hand, no such rescaling procedure exists.

Invariance of CTT and IRT person and item parameters across different item sets

After the item parameters for all items in math/science item pool had been obtained, two new tests were constructed from each item pool, the new CTT and the new IRT math/science test, respectively. It became evident, when selecting items for new tests, that our item pools were too small and did not have enough highly discriminating items. Consequently new CTT and IRT tests consist of up to three quarters of identical items; although – by design – different item selection criteria were used. This outcome, however, would not be a very likely one if item pools were large enough and of a high quality. The results regarding all four new tests are presented in the sequel but, since new CTT and IRT tests are very similar, the difference between them is not further discussed.

It should also be noted that item content was neglected when selecting items from item pools which resulted in rather different contents and dimensionality of the new tests comparing to the original item pools (see Table 4 for the results of dimensionality analysis). This is particularly true in the case of science data: new science tests contain items from some science achievement fields only and are consequently closer to unidimensional, compared to all items from the pool, which are more diverse in contents.

In theory, IRT item parameters are invariant, not only across different participant groups, but also across different item sets, measuring the same latent variable. Accordingly, the IRT person parameter estimates should be very similar regardless of the item set they were obtained from. The property of item/person parameters invariance has implications for an important IRT application, i.e. computerized adaptive testing, where items are presented in different item sets and different participants usually reply to different items. Classical item discrimination parameters and person parameters, on the other hand, are clearly related to items in item set they are assessed from: the classical true score can only be defined with regard to a particular test; since test scores depend on the properties of the particular items in the test, so do the item-total correlations.

Table 7. *Comparison of IRT- and CTT-based item and person parameters invariance across different item sets.*

	New Math Tests		New Science Tests	
	CTT	IRT	CTT	IRT
CTT person parameters	0.942	0.908	0.870	0.857
IRT person parameters	0.943	0.921	0.883	0.869
Item difficulty parameters				
CTT	same	same	same	same
IRT	1.000	0.997	0.997	0.998
Item discrimination parameters				
CTT	0.975	0.914	0.631	0.700
IRT	0.975	0.885	0.408	0.377

Note. Correlations between the *original* and *new* person/item parameters are presented.

Person parameter invariance in different item sets was assessed by correlating the *original* (from item pools) and the *new* CTT and IRT person parameters and by comparing means of achievement estimates from item pools and new tests. It should be noted that correlations, presented in Table 7, only partially reflect the property of invariance as discussed above as item pools and new tests are not entirely different: all items from new tests are also included in item pools, what had influence on correlations, especially on the correlations between the *original* and *new* classical test scores. Furthermore, item pools contain more items than new tests and the person parameters were assessed from one measurement only (the context in which items appeared was the same). Although, due to this bias, classical test scores appear to be more invariant than they in fact are, IRT person parameters are slightly more invariant when estimated from different item sets than CTT person parameters in all conditions. Correlations between *original* and *new* parameters are, again, higher in the case of mathematical items, because both new math tests measure achievement that is very similar to achievement measured in math item pool (all tests are relatively unidimensional, items are similar in contents), whereas the contents of new science tests is somewhat different comparing to contents of all items in science item bank (also new science tests are unidimensional, but science item pool is rather multidimensional). New science tests obviously do not measure *the same* science achievement (i.e., the same latent variable) as is measured by all items in science item bank; the degree of invariance of person parameters is consequently lesser. Considering this, the correlations between original and new IRT person parameters in the case of science achievement are actually surprisingly high. The results of comparing mean differences in person parameters obtained from item pools and new tests (presented in Table 9) confirm the results on person parameter invariance discussed above. They suggest that very similar person parameter estimates are obtained from item pools and new tests, since mean differences in parameters are very small, and also that IRT person parameters are more invariant when estimated from different item sets in all conditions. It is, again, possible that these high correlations and small differences between original and new IRT person parameters are, above all, the consequence of prior normal distribution used in parameter estimations; within our research design however this remains only speculation.

As previously discussed, IRT item parameters should be invariant in different item sets since they are, in theory, independent of all other items in the test. In CTT only item difficulties are independent of other items in the test, whereas item discrimination always depends on all other items. The comparison of item parameters invariance across different item sets is presented in Table 7. Correlations between *original* and *new* IRT item difficulty parameters are higher than correlations between item discrimination parameters. Very consistent estimates of IRT item difficulty parameters are obtained in new tests and both item pools, despite somewhat different contents of new (especially science) tests and the fact that item difficulty estimates were not optimal in the science item pool. The mean differences in IRT item difficulties across different item sets are also very small in all conditions (see Table 9). Classical item difficulty parameters are the same, as the same data set was used for estimation of *original* and *new* parameters.

The item discrimination parameters are, in comparison to the item difficulties, less invariant in all conditions; furthermore, the classical discrimination parameters are at least as much invariant as the IRT item discriminations (Table 7). Note that classical item discrimination parameters are unjustifiably high since items in item pools and new tests are not entirely different (in comparison of the *original* and the *new* CTT item discriminations the correlations between partly same items were calculated). Both the CTT and the IRT item discrimination parameters correlate lower in the case of science items, which is probably due to the difference in contents of science item pool and new science tests: new science tests measure different and more unidimensional science achievement than all items in science item bank and items are simply not equally indicative for the *old* and the *new* science achievement. Apart from the content differences for science tests, poor invariance of the IRT item discrimination estimates, compared to the classical estimates, is probably also the result of suboptimal *original* item discrimination estimates. Apparently poor model fit for items in science item bank has the same effect on the IRT item discrimination invariance in different item sets than on the IRT item discrimination invariance across different participant groups. Despite the lower correlations of the item discrimination parameters as compared to the item difficulty parameters, the mean differences in item discrimination parameters are in general not significantly different when estimated from different item sets (Table 9) and although the differences between the *original* and *new* IRT item discrimination parameters for science data are generally larger than the differences for math data, these differences are also statistically insignificant, because the standard errors of means are higher in this case as well.

Conclusions

The present study attempted to address the questions of comparability of CTT and IRT person/item parameters, invariance of CTT and IRT item parameters across different groups of participants and invariance of CTT and IRT person/item parameters across different item sets.

Our main conclusions regarding the first question are that the IRT and the CTT person parameters are highly comparable and also that item difficulties and item discriminations are very comparable, with the latter showing somewhat lower comparability. Similar results regarding IRT and CTT parameters comparability were also found by Courville (2005), Fan (1998) and MacDonald and Paunonen (2002).

As for our second research question, the results of our study also support previous findings by Fan (1998) and MacDonald and Paunonen (2002) that CTT and IRT item parameters show very similar invariance property across different participant groups, with item difficulty parameters being more invariant than item discrimination parameters. But, in addition to this, the results of this study also show that IRT item parameters are generally more invariant than CTT parameters in the case of good IRT model fit, whereas CTT item parameters are more invariant in the case of poor IRT model fit. This conclusion is very important for application of IRT in test constructing as it implies (1) that the question of IRT model fit should be addressed very carefully when using IRT for item calibration, and (2) that, if an appropriate IRT model is used for modelling item responses, theoretically superior IRT item parameters would also be empirically superior and could therefore be used for item calibration on non-representative groups of participants. It should be noted here that, although CTT item parameters also proved to be very consistently estimated in different groups of participants, the absolute values of item parameters differ and that, in contrary to the IRT item parameter estimates, there is no rescaling procedure, which would place CTT item parameter estimates from different participant groups on a common scale (Embretson & Reise, 2000; Hambleton et al, 1991).

The invariance property of person and item parameters across different item sets, an object of our third research question, also has an important consequence in test application: if person and item parameters are invariant across different item sets, different items can be used for measuring the same latent variable. Since the same data set was used for parameter estimation in item pools and new tests and, consequently, CTT item difficulties stayed the same, it was not possible to compare the invariance property of item difficulty parameters in both measurement frameworks; however, the IRT difficulty parameters have proven to be very invariant across different item sets. In contrary to theoretical and intuitive expectations, the results of this study also show that the CTT item discrimination parameters are just as or more invariant across different item sets than the IRT discrimination parameters and, as expected, that IRT item discriminations are invariant only if the model fits the data. We also found invariant person parameter estimates across different item sets, with the IRT person parameters being more invariant in all conditions. This is very interesting with regard to the IRT person parameters related to the science data, since the test with all items from the item pool and the new tests are obviously not measuring the same latent trait and since the IRT discrimination parameters are not invariant across item sets, and is therefore possibly only the effect of prior normal distribution used in the MML estimation method for IRT person parameters.

Of course, the present study, as it is an empirical study, has its share of limitations. First of all, item pools were limited in both item difficulties and item discriminations. The characteristics of the particular items that were used in the study could have some effect on the comparability of CTT and IRT person/item parameters and on the person/item parameters invariance properties. Larger item pools, with items varying more in item difficulty and item discrimination, or simulation studies that would manipulate different item characteristics should be used in future studies to determine whether the limited item pool also limits the generalization of the results from this study. Limited item pools were also related to item selection for new tests. An important aim at the beginning of this study was to determine whether the test construction method influences the properties of the resulting test. Unfortunately, because item pools were limited in the number of highly discriminative and informative items, very similar new tests were constructed using both CTT and IRT item selection criteria and possible differences in test construction could not be revealed. The second obvious shortcoming of this study was that the true person and item parameters were unknown. Because of this, only the comparison between CTT and IRT item/person parameters could be made and, for instance, the question whether the difference in distributions of CTT and IRT person parameters for item pools is the result of the effect of the easiness of the test on CTT person parameters or the effect of prior normal distribution on IRT person parameters, could not be answered. The information truly valuable to a psychometric practitioner in this case would, of course, not be that of the comparability of IRT and CTT parameters, but about accuracy of parameter estimates. Again, we propose further investigation, possibly a simulation study that would overcome this limitation.

The fact that majority of the items allowed for guessing was not taken into account explicitly in our analyses. This should not be a source of bias against either paradigm. On one hand, guessing degrades the fit of the IRT models, because the empirical ICC can not be expected to approach zero when the theta value is small. On the other hand, guessing degrades the CTT-based measurement, too, because it narrows the range of item difficulties and because the difficulty parameters are not directly comparable across items with different probability of a random success.

Overall, the findings from this study indicate that CTT and IRT item/person parameters are very comparable and show similar invariance properties when estimated across different participant groups or across different item sets. Despite the high similarity between CTT and IRT parameters, the results also demonstrate that theoretical advantages (invariance property) of the IRT parameters are, at least to some extent, reflected in their empirical superiority, but only if an appropriate IRT model is used for modelling the data, the condition that is difficult to satisfy and even to assess appropriately in practice. Our findings also demonstrate that, with the MML method used for the IRT parameter estimation, model misfit influenced the item parameters invariance but not the person parameters invariance, what suggests that prior normal distribution has a great impact on person parameter estimates. Further investigation, preferably with simulated data with known true values of item/person parameters, is needed to provide missing guidelines about the assessment of IRT model fit, about the extent of model misfit that would still yield sufficiently invariant item and person parameters and to determine the influence of different IRT estimation methods on the accuracy of IRT person/item parameter estimates.

References

Courville, T. G. (2005). *An empirical comparison of item response theory and classical test theory item/person statistics*. Unpublished doctoral dissertation, Texas A&M University. Retrieved November 5, 2007 from http://txspace.tamu.edu/bitstream/handle/1969.1/1064/etd-tamu-2004B-EPSY-Courville-2.pdf?sequence=1.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person parameters. *Educational and Psychological Measurement*, *58*, 357–381.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item re-*

sponse theory. Newbury Park, CA: Sage.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

MacDonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person parameters based on item response theory versus classical test theory. *Educational and Psychological Measurement, 62*, 921–943.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Ten Berge, J. M. F., & Kiers, H. A. L. (1991). A numerical approach to the approximate and the exact minimum rank of a covariance matrix. *Psychometrika, 56,* 309–315.

Ten Berge, J. M. F., Snijders, T. A. B., & Zegers, F. E. (1981). Computational aspects of the greatest lower bound to the reliability and constrained minimum trace factor analysis. *Psychometrika, 46,* 201–213.

Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69,* 613–625.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago: Scientific Software.

Appendix

Table 8. *Comparison of the IRT- and the CTT- based item parameters invariance across different groups of participants (results of t-test for grade and gender; ANOVA for country).*

		Grade				Gender				Country				
		d	*df*	*t*	*p*	*d*	*df*	*t*	*p*	df_1	df_2	*F*	*p*	η^2
Item difficulty														
Math Item Bank	CTT	-0.156	38	-16.22	0.000	-0.002	38	-0.38	0.636	5.0	190.0	25.06	0.000	0.397
	IRT	0.830	38	17.68	0.000	0.008	38	0.30	0.762	3.9	146.2	22.84	0.000	0.375
Science Item Bank	CTT	-0.096	36	-13.27	0.000	-0.029	36	-4.18	0.000	3.8	135.6	2.16	0.081	0.057
	IRT	0.752	36	11.88	0.000	0.172	36	3.01	0.005	5.0	180.0	3.56	0.004	0.090
Item discrimination														
Math Item Bank	CTT	-0.019	38	-2.81	0.008	0.009	38	1.82	0.076	3.7	139.4	7.31	0.000	0.161
	IRT	-0.048	38	-3.38	0.002	0.013	38	1.06	0.294	3.5	132.1	5.52	0.001	0.127
Science Item Bank	CTT	0.020	36	2.34	0.025	-0.019	36	-3.32	0.002	5.0	180.0	13.65	0.000	0.275
	IRT	0.015	36	0.73	0.468	-0.032	36	-2.48	0.018	4.1	147.2	6.08	0.000	0.145

Note. d = Cohen's effect size measure; η^2 = proportion of variance explained by factor.

Table 9. *Comparison of IRT- and CTT-based person/item parameters invariance across different item sets (results of paired t-test between the original and the new parameter estimates).*

		New CTT constructed tests				New IRT constructed tests			
		d	*df*	*t*	*p*	*d*	*df*	*t*	*p*
Person parameters									
New Math Tests	CTT	0.093	3947	46.50	0.000	-0.003	3947	-1.56	0.118
	IRT	0.001	3947	0.22	0.824	0.002	3947	0.33	0.774
New Science Tests	CTT	0.090	3950	45.36	0.000	0.055	3950	29.68	0.000
	IRT	-0.001	3950	-0.16	0.873	-0.001	3950	-0.12	0.906
Item difficulty									
New Math Tests	CTT	same	-	-	-	same	-	-	-
	IRT	0.008	11	3.12	0.010	-0.001	11	-0.06	0.957
New Science Tests	CTT	same	-	-	-	same	-	-	-
	IRT	-0.013	11	-0.55	0.597	-0.008	11	-0.29	0.780
Item discrimination									
New Math Tests	CTT	-0.013	11	-2.86	0.016	-0.011	11	-0.99	0.343
	IRT	0.018	11	0.90	0.387	0.051	11	1.03	0.327
New Science Tests	CTT	-0.010	11	-0.76	0.469	-0.012	11	-0.87	0.403
	IRT	0.038	11	0.88	0.397	0.034	11	0.67	0.518

Note. *d* = Cohen's effect size measure.

Prispelo/Received: 30.04.2008
Sprejeto/Accepted: 05.09.2008