Edukacja 2015, 3(134), 77–97
ISSN 0239-6858

# Item analysis and evaluation using a four-parameter logistic model

Karolina Świst

Educational Research Institute*

The four-parameter logistic model (4PLM) assumes that even high ability examinees can make mistakes (e.g. due to carelessness). This phenomenon was reflected by the non-zero upper asymptote (*d*-parameter) of the IRT logistic curve. Research on 4PLM has been hampered, since the model has been considered conceptually and computationally complicated – and its usefulness has been questioned. After 25 years, following introduction of appropriate software, the psychometric characteristics of 4PLM and the model's usefulness can be assessed more reliably. The aim of this article is to show whether 4PLM can be used to detect item-writing flaws (which introduce construct-irrelevant variance to the measurement). Analysis was conducted in two steps: (a) qualitative – assessment of compliance of items with the chosen item-writing guidelines, (b) quantitative – fitting 4PLM to compare the results with qualitative analysis – to determine whether the same items were detected as flawed. Other IRT models (3PLM and 2PLM) were also fitted to check the validity of results. Flawed items can be detected by the means of qualitative analysis as well as by 4PLM and simpler IRT models. This model is discussed from the perspective of practical use in educational research.

Keywords: construct-irrelevant variance, item format, item-writing guidelines, item-writing flaws, four-parameter logistic model.

## Multiple-choice item quality

Multiple-choice questions (MCQs) are commonly employed in educational assessment (Hohensinn and Kubinger, 2011). A typical single best-answer MCQ consists of a stem (the question) and distractors (options/answers for the question), of which one choice is correct or the best answer (Cizek and O'Day, 1994). Test constructors have grown fond of multiple-choice questions for various reasons: content-valid test score interpretations; high test score reliability (when the number of high-quality MCQ items is sufficient); easy and cheap storing, use and reuse; objective scoring (not biased by the rater effect); obtaining diagnostic subscores for various types of higher-level thinking processes and the possibility of analysing these scores using numerous psychometric theories (item response theory, classical test theory, generalizability theory; Haladyna

The article is an extended version of the presentation "The four-parameter logistic model – a useful tool in educational research or a conceptual humbug?", presented by the authors at the EduMetric 2014 International Seminar on Educational Research and Measurement, Cracow, December 6–8. The seminar was organised by the Jagiellonian University, as part of the system level project "Quality and effectiveness of education – strengthening institutional research capabilities" carried out by the Educational Research Institute and co-financed by the European Social Fund (Human Capital Operational Programme 2007–2013, Priority III High quality of the education system).



* Address: ul. Górczewska 8, 01-180 Warszawa, Poland. E-mail: k.swist @ibe.edu.pl

and Downing, 1989a). What is more, if properly constructed, MCQ can discriminate very well between students of high and low abilities (Schuwirth and Vleuten, 2004).

Otherwise, MCQ format has been often criticised, believed to test little more than students' memory, rather than complex problem solving (Haladyna, Downing and Rodriguez, 2002). Further, creating suitable MCQs is considered difficult. Several guidelines for writing MCQs have been proposed. The best known were formulated by Haladyna and Downing (1989a; 1989b) and the revised version was proposed by Haladyna et al. (2002). Violations of item-writing guidelines are termed item-writing flaws (IWFs). The result of IWFs is that questions can become unexpectedly easy or difficult to answer (compared to the deviser of the test's intentions), and lead to overestimation or underestimation of examinees' ability (Downing, 2002; 2005). Construct--irrelevant variance (CIV, Haladyna and Downing, 2004) can in this way be introduced to tests – as they can reflect constructs other than those intended for measurement. Such a phenomenon can be detrimental to examinees' scores – e.g., failing a test or rejection from a recruitment processes (e.g. to higher education institutions), if the score is used as a criterion. Therefore consequential validity of the test (see Messick, 1989) is lowered.

Research by Tarrant and Ware (2008) showed that when the test was flawed, low-achieving students actually benefitted. The high-achieving students were punished, as a lower proportion of them passed the test when flaws were present. Low-achieving students could have employed test-wiseness (behaviour that allows examinees to guess or deduce correct answers without knowing the right answer, see also Downing, 2002) rather than knowledge-based strategies.

IWFs tend to frequently recur in various fields, as well as at various stages of education (Hansen and Dexter, 1997; Jozefowicz et al., 2002; Tarrant, Knierim, Hayes and Ware, 2006). For example, Downing (2005) analysed the quality of 4 examinations in medical schools and found that 46% of MCQs could be classified as flawed. As a consequence, about 10–15% of students who had failed exams would have otherwise have passed. Hansen and Dexter (1997) discovered that 75% items present in accountancy test banks violated at least one guideline. The problem has also been found in psychology textbooks. About 60% of MCQs in introductory psychology textbooks have been categorised as flawed (Ellsworth, Dunnell and Duell, 1990).

The original list of item-writing guidelines by Haladyna and Downing (1989a; 1989b) comprised 43 guidelines, and its revision by Haladyna et al. (2002) was shortened to 31. It might be anticipated that violation of particular rules can have more severe effects on test characteristics and scores. Rodriguez (1997) in his meta-analysis of the effects of item-writing flaws, analysed 7 rules: (a) general item-writing procedural consideration: avoid the complex multiple-choice (Type K) format, (b) two item-writing guidelines concerning stem: state the stem in question form, and word the stem positively and (c) four distractor-development considerations: use as many functional distractors as possible; avoid, or use sparingly, the phrase "all of the above"; avoid, or use sparingly, the phrase "none of the above"; and keep the length of options fairly consistent. These are the rules, which were most frequently investigated (Haladyna and Downing, 1989b).

In the following sections, results from Rodriguez' (1997) meta-analysis are shown – the impact of violating these guidelines on item difficulty and discrimination, test reliability and sometimes validity, is given as mean standardised effect sizes.

In order to discuss the first rule – avoidance of the complex multiple-choice format (Type K), the format has to be defined and

an example is needed. Albanese (1993, p. 28) defined a complex multiple-choice (Type K) item as "having a stem, a list of potentially correct answers referred to as primary responses, and a list of combinations of the primary responses called secondary choices". An examinee is obliged to select the correct (or the best) answer from the list of secondary choices. Albanese (1993, p. 28) indicated that "the presence of one or more correct primary responses and the availability of the secondary choices to facilitate machine scoring" distinguishes this item format from others. An example of Type K format provided by Albanese (1993) is shown in Table 1.

Such an item format has been frequently used in medical science and various certification exams (Rodriguez, 1997). The format requires heavier cognitive demand from an examinee than other item formats (Huntley and Plake, 1984). Various researchers (Albanese, 1993; Albanese, Kent and Whitney, 1979; Kolstad, Briggs, Bryant and Kolstad, 1983) discovered that Type K items have lower difficulty compared to when the same question is presented in the form of multiple true-false (MTF) items. This effect was probably caused by clues present in Type K items (see also Albanese, Kent and Whitney, 1977). In Rodriguez's (1997) meta-analysis, Type K format rendered items more difficult (by an average of 0.122) and less discriminating (by an average of 0.145). In consequence, Rodriguez (1997) also discouraged using Type K format, in order to prevent lowering of item discrimination.

Wording the stem negatively (especially using double negatives) might cause students confusion and therefore lead to lowered scores. Examinees are required to perform additional mental operations in order to answer such item properly (Cassels and Johnstone, 1984). Sometimes, test constructors have tended to minimise confusion by offering the best answer (i.e. the incorrect) in as possible a form as to be obvious. Questions then became unexpectedly easy and their ability to discriminate between students was diminished (Tarrant et al., 2006). In practice (Casler, 1983), when the negative term in the stem was somehow emphasised (by underlining, emboldening or capitalising all letters), this proved difficult for students with high ability. Results of Rodriguez' (1997) meta-analysis showed that a negative stem slightly increased item difficulty (on average by 0.032), but decreased reliability of the test (on average by 0.166). The results supported stating the stem positively, but owing to some inconsistencies, they still need to be further investigated.

Designing plausible distractors has been reported as challenging, as the majority of test authors have focused on creating properly-working stems (Tarrant, Ware and

Table 1
*An example of Type K item*

Which of the following is/are appropriate use(s) for the item discrimination index?

| Primary responses | Secondary choices |
|---|---|
| (a) as an index of quality of item functioning<br>(b) flagging items for further review<br>(c) as an index of consistency of item and total test performance by examinees<br>(d) to discriminate between examinees who do and do not guess on the item | A. (a), (b) and (c) (correct answer)<br>B. (a) and (c)<br>C. (b) and (d)<br>D. (d) only<br>E. All of the above |

Mohammed, 2009). What is more, quality of distractors influences the discrimination of MCQs (DiBattista and Kurzawa, 2011). Therefore, each option should be carefully designed and based on the "common misconception about the correct item" (Haladyna and Downing, 1993, p. 1000). As DiBattista and Kurzawa (2011; p. 2) stated: "An effective distractor will look plausible to less knowledgeable students and lure them away from the keyed option, but it will not entice students who are well-informed about the topic under consideration". If the distractor is effective, at least some students will choose it. Haladyna and Downing (1993) showed in their review of 477 items, that over 38% distractors should have been eliminated, as fewer than 5% students chose them. Tarrant et al. (2009) investigated the proportion of non-functioning distractors in 7 tests for nursing. They found out that only 52.2% ($n$ = 805) of all distractors were functioning effectively and 10.2% ($n$ = 158) were not chosen at all. As it has been proved in empirical research that the number of non-functional distractors can be considered high, researchers and practitioners have proposed reducing the number of distractors while developing the test. It was discovered that using fewer options did not change psychometric properties of tests, their reliability or validity (e.g., Aamodt and McShane, 1992; Cizek, Robinson and O'Day, 1998). Rodriguez' (1997) meta-analysis showed that items became slightly more difficult and slightly more discriminating when 4 options were used instead of 3. Test reliability increased when 3 options were used instead of 2. Therefore, a decreasing number of options did not change a test's psychometric properties, although all distractors should be effective, i.e. someone has to select them.

Probably the most controversial guidelines concern using answers such as "none of the above" (NOTA) and "all of the above" (AOTA). Haladyna and Downing (1989b) indicated that NOTA increased item difficulty, as well as lowering discrimination and test reliability – and therefore should not be used. Yet, NOTA (if the correct answer) can prevent simple recognition of a correct answer – examinees have to be sure that all the other distractors are incorrect. Therefore motivation for careful examination of all the options can increase. This option can be especially useful in mathematics exams – it can discourage examinees from guessing or choosing an approximate answer without performing the required calculations (Frary, 1991; Rodriguez, 1997). When an item has "the best" answer (and not a "correct" one), it implies that each distractor is to some extent true. Including NOTA as one of the distractors in such situation can be seriously misleading (Rodriguez, 1997). The Knowles and Welch (1992) meta-analysis of item difficulty and discrimination by the NOTA option was contradictory with the former results – the average effect size for discrimination was 0.01 for discrimination and -0.17 for difficulty. Rodriguez' (1997) meta-analysis reported small and non-consistent increase of difficulty (by a factor of 0.035) and decrease (insignificant yet consistent) in discrimination by 0.027. There were no effects on test reliability and some inconsistent (and insignificant) effects on validity (0.073 on average). Therefore the results obtained by Rodriguez (1997) are still inconclusive.

"All of the above" option was discouraged, as providing certain clues for students who did not know the correct answer, since, if the test has four of five options and students can detect at least two alternatives as correct, then they can deduce that "all of the above" option is correct. So the examinee does not have to know whether the remaining distractors are correct. On the other hand, knowing that at least one distractor is wrong can eliminate "all of the above option" (Hansen and Dexter, 1997; Woodford and Bancroft, 2005). On the contrary, AOTA can function well as a distractor (Rodriguez, 1997). Rodriguez (1997)

reported that Mueller's (1975) was the only study to examine inclusion of both NOTA and AOTA options and to report separate outcomes (though items did not have equivalent stems across different formats). According to Mueller (1975) AOTA items were the least difficult (when compared to Type K and NOTA). The weighted mean item difficulty for Type K was 0.64, for NOTA 0.74 and for AOTA was 0.767. Due to the fact, that it was the only study reporting separate effects for AOTA, AOTA effects were not presented in Rodriguez's (1997) meta-analysis.

In Haladyna and Downing's (1989a) review, all authors reported that the length of options should be similar – as some describe the correct answer as being longer than other distractors. A similar result was shown by Rodriguez (1997) – describing the correct option as longer than the others increased the difficulty index by averagely 0.057 across studies. A correct answer which was longer than the other options, decreased validity across the studies on average by 0.259.

Item-writing flaws can introduce various biases which have been reported as detrimental to examinees' performance. Tarrant and Ware (2008) showed that the effect was especially harmful to the high-achieving group, as these students usually relied on their own knowledge rather than test-wiseness (guessing or deduction). In order to interpret test results properly, detecting such items and adjusting scores for the presence of CIV is necessary. In the following section, the four-parameter logistic model (4PLM) is described and its potential application to detection of flawed items.

## The four-parameter logistic model – assumptions and applications

In the one-parameter logistic model (1PLM) and two-parameter logistic models (2PLM), the probability of supplying a correct answer to an item varies between 0 and 1 and ability level can range from $-\infty$ to $\infty$. Yet, the probability of a correct answer can hardly ever approach 0, even for low-ability students (Liao, Ho, Yen and Cheng, 2012). Therefore, the three-parameter logistic model was introduced (Birnbaum, 1968). The model assumes a non-zero lower asymptote (which indicates that even low-ability students can give the correct answer by guessing). The model is useful when we try to estimate the ability level of low-achievers properly, who answered difficult question correctly by chance. On the other hand, it does not capture the fact that some high-achievers can also answer an easy item incorrectly, since they are stressed or careless. Therefore, 4PLM was introduced by Barton and Lord (1981), as an extension of IRT family models. The model was intended to take such behaviour of examinees into account by estimating the upper asymptote of the logistic curve (which is fixed to 1 in simpler IRT models; Magis, 2013). Therefore, the model can be written as:

$$P(u_{ij}=1\,|\,\theta_i,a_j,b_j,c_j,\mathrm{d}_j)=c_j+(d_j-c_j)\frac{e^{1,7a_j(\theta_i-b_j)}}{1+e^{1,7a_j(\theta_i-b_j)}}, \quad (1)$$

where: $a_j$ – discrimination parameter; $b_j$ – difficulty parameter; $c_j$ – pseudo-guessing parameter and $d_j$ – carelessness parameter, and constant – 1.7.

Barton and Lord (1981) concluded their research by arguing that the 4PLM did not systematically improve the estimation of likelihood and it did not change ability level estimates. What is more, due to the mathematical complexity of the model, its estimation was time-consuming. It is worth reporting that the authors did not estimate the *d*-parameter directly, but compared the goodness of fit of models with fixed d values (to 1, 0.99 and 0.98 respectively). This was the result of computational problems and lack of proper software (Waller and Reise,

2010). Hambleton and Swaminathan (1985; pp. 48–49) argued that 4PLM did not have any practical value. Therefore, research on psychometric properties and application of 4PLM stopped due to problems with estimation (even today the estimation of lower asymptote is considered problematic; Loken and Rulison, 2010) and perceived as uselessness.

After 25 years, interest in 4PLM has been revived. First of all, Bayesian computation methods have been introduced. These methods can definitely accelerate and facilitate estimation, although they do not resolve all the conceptual problems of 4PLM (Raiche, Magis, Blais and Brochu, 2013). Interest in 4PLM has been renewed in the field of clinical and personality psychology, as it is crucial to measure latent trait accurately at its extremes (Reise and Waller, 2003; Stark, Chernyshenko, Drasgow and Williams, 2006; Waller and Reise, 2010). A new practical application of 4PLM has been established for computer adaptive testing (CAT). The model has been used in order to reduce the influence of examinees' early mistakes on estimation of their ability level and 4PLM has reduced such influence in a more effective way than 3PLM (Rulison and Loken, 2009; Loken and Rulison, 2010; Liao et al., 2012). Liao, Ho, Yen and Cheng (2012) also showed that ability level was adequately estimated using 4PLM during CAT. 4PLM prevented the initial drop of ability estimates caused by the incorrect answers to the first two items. Magis (2013) indicated that 4PLM allows more robust estimation of ability due to weighting the log-likelihood function (the aberrant item responses are down-weighted and have less impact on the estimation of ability).

What happens when simpler IRT models are fitted to data with a valid *d*-parameter? Loken and Rulison (2010) showed that when 3PLM was fitted to such data, item discrimination parameters were lowered. The difficulty of items was also shifted higher (by about 0.5 *SD*). When 4PLM parameters were compared to 3PLM parameters, *c*-parameters (pseudo-guessing) obtained within 3PLM had relatively higher root-mean-squared error (RMSE) and relatively lower correlation with the true score. When 2PLM was fitted, mean value of discrimination parameters was shifted by 0.5 and the difficulty parameters were shifted to 0. Although the change between parameters was proved, the correlation of thetas obtained with 4PLM, 3PLM and 2PLM was positive and very high ($r = 0.98$). When simpler IRT models were used, the parameters were not systematically biased.

Fitting simpler IRT models to data with a valid *d*-parameter can influence the test information function (TIF) and modeling of standard errors. When the level of ability was low, the information level present in TIF was underestimated with 3PLM. For levels of high ability, information was overestimated by 3PLM. Modeling of standard errors was also influenced by 3PLM – for people of low ability, standard errors were not high enough (the confidence intervals were too wide). When 2PLM was fitted, the overall information level was still higher than with 4PLM. The precision of the model was highest for the mid-levels of latent trait. Information about extreme levels of latent trait was not captured by the model, so the information level was lower. Therefore when a *d*-parameter is available, fitting simpler models might not bias parameter values significantly, yet influence the precision of measurement of the latent trait, especially at its extremes.

4PLM has its limitations – there is no formal proof for its identifiability. San Martín Gonzalez and Tuerlinckx (2014) showed that the 3PL fixed-effects model (after fixing the difficulty parameters) is still unidentified, which means that the parameters do not have empirical interpretation. What is more, there is no formal proof for 3PL random-effect

model identifiability. As 4PLM is more computationally complex, the formal proof of its identification (or lack of it) is even harder to demonstrate.

Moreover, the interpretation of the *d*--parameter can be confusing. Barton and Lord (1981) assumed that the *d*-parameter results from student carelessness or stress. Loken and Rulison (2010) indicated that if this assumption was valid, a person-specific *d*-parameter should be estimated. Therefore the model should be rewritten in the following form, when correct answer is predicted (assuming that the item is dichotomous, a wrong answer gives 0 points and correct gives 1):

$$P(u_{ij} \mid \theta_i; a_j, b_j, c_j, d_i) = (1-d_i)c_j + d_i\left(c_j + (1-c_j)\frac{e^{1.7a_j(\theta_i-b_j)}}{1+e^{1.7a_j(\theta_i-b_j)}}\right) = c_j + d_i(1-c_j)\frac{e^{1.7a_j(\theta_i-b_j)}}{1+e^{1.7a_j(\theta_i-b_j)}}. \quad (2)$$

Still, if the *d*-parameter was estimated as person-specific, the model would not explain why such a phenomenon occurs. Especially in high-stakes testing, everyone should be motivated (and there is no reason why high-achievers should not be) to answer items as well as possible. The reasons for the presence of a *d*-parameter can be varied – besides carelessness or stress, the test can be "speeded" (e.g., Boughton and Yamamoto, 2007; Mroch, Bolt and Wollack, 2005; Linden, 2007), and therefore even high-achievers can answer items incorrectly due to lack of time. Examinees can also respond to questions in a creative or unusual way, not represented by the scoring key and such response is scored incorrect (Karabatsos, 2003). Nevertheless, determining the cause of *d*-parameter prevalence is highly ambiguous and might give implausible results. 4PLM with a person-specific *d*-parameter is probably even more complicated and might be found to be unidentifiable. Therefore, to simplify the following analysis, it is assumed that a *d*-parameter might only indicate occurrence of item-writing flaws, not the cognitive processes specific to high-achievers and which provoke their mistakes.

## The aim of the article and the research questions

If a test contains item-writing flaws, then they may be detrimental to the performance of high-achieving students. This is especially perilous in high-stakes testing, as it introduces CIV and lowers consequential validity. Biases can be detected by experienced experts (test designers) by means of qualitative analysis, but it obviously takes time and human resources. Introducing an automated method to assess bias should improve the quality of testing.

The aim here was to evaluate 4PLM as a tool to detect item-writing flaws. The standardised external examination papers for Polish, which include reading and writing assessment, were used for the analysis. The reading part of the exam, which usually consists of 20 MCQ items, demands critical analysis and interpretation of the material given (e.g., articles, excerpts from books or poems). The content and format of such items has to be constructed very carefully – as wrong interpretation of reading items can influence the decision for choosing an answer and therefore be misleading in estimation of student ability.

The analysis therefore covered:

- a qualitative analysis of items according to guidelines (avoid Type K, negations in stem, "all of the above", "none of the above", non-functioning distractors and different length of the options) chosen by Rodriguez (1997). The aim of this part of analysis was to show how many items were biased by some flaws and therefore could exhibit the upper asymptote;
- analogical quantitative analysis of the same items with 4PLM, showing how many items exhibited an upper asymptote

in order to determine whether qualitative results were confirmed and therefore valid. It was also shown whether analogical conclusions could be drawn when the simpler IRT models, 3PLM and 2PLM, were used instead of 4PLM, and whether it was worth employing the computationally complex 4PLM.

## Data and methodology

Data for analysis was from Polish language exams (only from the dichotomously scored reading part) covering the whole student population were collected by Central Examination Board in Warsaw, 2012–2014. The choice of this time-frame was for two reasons: (a) availability of databases with information about distractors chosen by the examinee and (b) availability of data concerning only reading skills – so only one construct was intended to be measured by the test (earlier editions of standardised external exams at the end of lower secondary schools had consisted of both reading and writing items, as well as items on history and civic education).

The number of students participating in each standardised external exam is given in Table 2. In each test edition, two versions (A and B) were administered with a different sequence of distractors (to prevent cheating).

The analysis was performed using the *mirt* library (Chalmers, 2012), allowing estimation of the four-parameter logistic model, as well as the simpler IRT models. The analysis was performed with the default specification. Although it might be considered interesting to specify various a priori distributions of parameters, the aim of the research was not to check the stability of solutions.

Parameters from 4PLM, 3PLM and 2PLM were interpreted according to the following assumptions:

- the existence of a *d*-parameter deviating strongly from other *d*-parameter values indicated that an item had some kind of writing-flaw or that some problem with misleading content provoked examinees to choose a wrong answer;
- the occurrence of a *c*-parameter might indicate pseudo-guessing, which may also suggest that the item is somehow flawed. On the other hand, as successful pseudo-guessing favours examinees of low ability, the consequences of pseudo-guessing are not further examined;
- incidence of low *a*-parameter values (discrimination) was further analysed, as it properly indicated poor item differentiation between low and high ability examinees. Low values could indicate item-writing flaws;
- existence of extremely high or extremely low *b*-parameters (difficulty), drifting from the other *b*-parameters values might be another indicator for item-writing flaws, as items become unexpectedly difficult or easy.

Table 2
*The number of students sitting the exams in a given year*

| Year | Number of students who sat version A | Number of students who sat version B | Total number of students |
|---|---|---|---|
| 2012 | 197 094 | 196 737 | 393 836 |
| 2013 | 182 741 | 197 011 | 379 752 |
| 2014 | 181 703 | 181 052 | 362 755 |

## Results

### Qualitative analysis

The results of the qualitative analysis of all reading items according to criteria also chosen by Rodriguez (1997) can be seen in the following tables (Table 3 – 2012, Table 4 – 2013 and Table 5 – 2014). Each table indicates the number of MCQs for a given year and details item-writing guidelines violated.

Table 3 shows MCQ items from 2012. Although the length of options was consistent and AOTA and NOTA options were not used at all, some flaws were present. 3 items can be classified as Type K. There were also negations in stem (although the negative word was always underlined) in 3 items. One feature was especially striking – the number of non-functional distractors. They were present in nearly all items (15 out

Table 3
*Qualitative analysis of reading items from 2012*

| Item | Was type K present? | Were there negations in stem? | Were there non-functional (less than 5% of examinees chose them) distractors present? How many? | | Was "all of the above" option used? | Was "none of the above" option used? | Was the length of options consistent? |
|---|---|---|---|---|---|---|---|
| | | | Version A | Version B | | | |
| 1$2012 | NO | NO | 1 | 2 | NO | NO | YES |
| 2$2012 | YES | NO | 3 | 3 | NO | NO | YES |
| 3$2012 | NO | NO | 1 | 1 | NO | NO | YES |
| 4$2012 | YES | NO | NO | NO | NO | NO | YES |
| 5$2012 | NO | NO | 1 | 2 | NO | NO | YES |
| 6$2012 | NO | YES | 1 | 1 | NO | NO | YES |
| 7$2012 | NO | NO | NO | NO | NO | NO | YES |
| 8$2012 | NO | NO | 2 | 1 | NO | NO | YES |
| 9$2012 | NO | NO | NO | NO | NO | NO | YES |
| 10$2012 | NO | NO | 2 | 2 | NO | NO | YES |
| 11$2012 | NO | NO | NO | NO | NO | NO | YES |
| 12$2012 | NO | NO | 3 | 3 | NO | NO | YES |
| 13$2012 | NO | NO | 2 | 2 | NO | NO | YES |
| 14$2012 | NO | YES | 1 | NO | NO | NO | YES |
| 15$2012 | NO | NO | 1 | NO | NO | NO | YES |
| 16$2012 | YES | NO | 3 | 3 | NO | NO | YES |
| 17$2012 | NO | NO | 1 | NO | NO | NO | YES |
| 18$2012 | NO | NO | 2 | 3 | NO | NO | YES |
| 19$2012 | NO | YES | | NO | NO | NO | YES |
| 20$2012 | NO | NO | 1 | 1 | NO | NO | YES |

of 20 items). The number of non-functional distractors was not consistent between versions A and B, which implied that the order of items might suggest the correct response to the examinee (this hypothesis would need empirical justification, e.g. DIF analysis between both test versions). In items 2$2012 and 16$2012 – both Type K items, 3 non-functional distractors were present. In items 6$2012 and 14$2012 – which had negations in their stems, each item had one non-functional distractor. There were also two items – 4$2012 and 19$2012 with flawed format, but all their distractors were functional.

Table 4 shows MCQ items from 2013. Again, the length of the options was consistent, but there were some items which: could be classified as Type K (4 items); used "none of the above" option (1 item) and had

Table 4
*Qualitative analysis of reading items from year 2013*

| Item | Was Type K present? | Were there negations in stem? | Were there non-functional (fewer than 5% examinees chose them) distractors present? How many? | | Was "the all of the above" option used? | Was "none of the above" option used? | Was the length of options consistent? |
|---|---|---|---|---|---|---|---|
| | | | Version A | Version B | | | |
| 1$2013 | NO | NO | 2 | 2 | NO | NO | YES |
| 2$2013 | NO | NO | 2 | 2 | NO | NO | YES |
| 3$2013 | NO | NO | 2 | 1 | NO | NO | YES |
| 4$2013 | YES | NO | NO | NO | YES | NO | YES |
| 5$2013 | YES | NO | 3 | 3 | NO | NO | YES |
| 6$2013 | NO | YES | NO | 1 | NO | NO | YES |
| 7$2013 | NO | YES | NO | NO | NO | NO | YES |
| 8$2013 | NO | NO | 2 | 2 | NO | NO | YES |
| 9$2013 | YES | NO | NO | NO | NO | NO | YES |
| 10$2013 | YES | NO | 1 | 2 | NO | NO | YES |
| 11$2013 | NO | NO | NO | NO | NO | NO | YES |
| 12$2013 | NO | NO | 2 | 2 | NO | NO | YES |
| 13$2013 | NO | NO | NO | NO | NO | NO | YES |
| 14$2013 | NO | NO | NO | NO | NO | NO | YES |
| 15$2013 | NO | NO | 1 | 1 | NO | NO | YES |
| 16$2013 | NO | NO | NO | NO | NO | NO | YES |
| 17$2013 | NO | NO | 1 | NO | NO | NO | YES |
| 18$2013 | NO | NO | NO | NO | NO | NO | YES |
| 19$2013 | NO | NO | NO | NO | NO | NO | YES |
| 20$2013 | NO | NO | NO | NO | NO | NO | YES |

negations in stem (2 items, but the negative word was marked with underlining). Again, the most striking problem was the scale of occurrence of non-functional distractors. However, they numbered fewer than in 2012 (10 out of 20 items). In 3 items, the number of non-functional distractors varied between test versions. Two items – 5$2013 and 10$2013 were Type K with non-functional distractors. Although items 4$2012, 7$2103 and 9$2013 had the non--recommended item format, there were no non-functional distractors in these cases.

Table 5 shows the 2014 MCQ items. There were 3 items of Type K format and 5 items with negations in stem (although negative words were always underlined, as in years 2012 and 2013). Non-functional distractors were present in 16 out of 20 items. Items1$2014 and 12$2014 both had negations

Table 5
*Qualitative analysis of reading items from year 2014*

| Item | Was Type K present? | Were there negations in stem? | Were there non-functional (fewer than 5% examinees chose them) distractors present? How many? | | Was "the all of the above" option used? | Was "none of the above" option used? | Was the length of options consistent? |
|---|---|---|---|---|---|---|---|
| | | | Version A | Version B | | | |
| 1$2014 | NO | YES | 3 | 3 | NO | NO | YES |
| 2$2014 | NO | YES | NO | NO | NO | NO | YES |
| 3$2014 | YES | NO | 3 | 3 | NO | NO | YES |
| 4$2014 | NO | NO | 1 | 1 | NO | NO | YES |
| 5$2014 | NO | NO | 1 | 1 | NO | NO | YES |
| 6$2014 | NO | NO | 2 | 2 | NO | NO | YES |
| 7$2014 | NO | NO | NO | 2 | NO | NO | YES |
| 8$2014 | NO | NO | 2 | 3 | NO | NO | YES |
| 9$2014 | NO | NO | 1 | 1 | NO | NO | YES |
| 10$2014 | NO | NO | 1 | 1 | NO | NO | YES |
| 11$2014 | NO | NO | 1 | 1 | NO | NO | YES |
| 12$2014 | NO | YES | 2 | 1 | NO | NO | YES |
| 13$2014 | NO | NO | 1 | 1 | NO | NO | YES |
| 14$2014 | NO | NO | NO | NO | NO | NO | YES |
| 15$2014 | NO | YES | NO | NO | NO | NO | YES |
| 16$2014 | YES | NO | 3 | 1 | NO | NO | YES |
| 17$2014 | NO | NO | NO | 3 | NO | NO | YES |
| 18$2014 | NO | NO | 1 | 1 | NO | NO | YES |
| 19$2014 | NO | YES | NO | NO | NO | NO | YES |
| 20$2014 | NO | NO | 2 | 2 | NO | NO | YES |

in stem and non-functional distractors. Items 3$2014 and 16$2014 were of Type K and had non-functional distractors. "All of the above" or "none of the above" options were not used in 2014 and the length of the options was consistent for the items analysed.

Although authors of reading examinations did not frequently use "all of the above" (only 1 item from 2013) or "none of the above" options and the length of options was consistent, Type K format and negations in stem (though in nearly all cases negative word was underlined) could be found among items. Nevertheless, the biggest problem was again the large number of non-functional distractors (they were present in half of the analysed items in 2012 and 2014 and in 8 out of 20 items in 2013).

**Quantitative analysis**

Qualitative analysis allowed the discovery of potential item-writing flaws which could influence student results. To determine if such an effect (defined as the presence of upper asymptote in 4PLM) was present, the results were verified during quantitative analysis. Parameters acquired with simpler IRT models were also checked (3PLM

Table 6

*The parameters for reading items from 2012 obtained by 4PLM, 3PLM and 2PLM*

| Model / Item | 4PLM | | | | 3PLM | | | 2PLM | |
|---|---|---|---|---|---|---|---|---|---|
| | *a* | *b* | *c* | *d* | *a* | *b* | *c* | *a* | *b* |
| 1$2012 | 0.983 | -2.660 | 0.080 | 0.994 | 0.906 | -2.749 | 0.079 | 0.924 | -2.809 |
| 2$2012 | 1.657 | -1.086 | 0.085 | 0.751 | 0.652 | -0.679 | 0.005 | 0.657 | -0.691 |
| 3$2012 | 0.865 | -0.642 | 0.229 | 0.992 | 0.840 | -0.636 | 0.223 | 0.713 | -1.324 |
| 4$2012 | 2.187 | -0.222 | 0.243 | 0.999 | 2.153 | -0.223 | 0.237 | 1.462 | -0.695 |
| 5$2012 | 1.720 | 0.040 | 0.331 | 0.998 | 1.632 | 0.009 | 0.316 | 1.003 | -0.788 |
| 6$2012 | 2.411 | 0.321 | 0.448 | 0.996 | 2.215 | 0.306 | 0.437 | 0.878 | -0.917 |
| 7$2012 | 1.901 | -0.726 | 0.265 | 0.959 | 1.320 | -0.919 | 0.139 | 1.213 | -1.187 |
| 8$2012 | 1.222 | -0.435 | 0.313 | 0.986 | 1.128 | -0.459 | 0.291 | 0.870 | -1.245 |
| 9$2012 | 1.582 | -0.940 | 0.170 | 0.988 | 1.411 | -1.020 | 0.120 | 1.321 | -1.231 |
| 10$2012 | 1.263 | -1.018 | 0.114 | 0.997 | 1.219 | -1.072 | 0.086 | 1.174 | -1.231 |
| 11$2012 | 2.288 | 0.410 | 0.191 | 0.980 | 2.119 | 0.438 | 0.184 | 1.197 | 0.061 |
| 12$2012 | 1.945 | -2.143 | 0.029 | 1.000 | 1.900 | -2.157 | 0.053 | 1.931 | -2.151 |
| 13$2012 | 2.261 | -1.947 | 0.034 | 0.963 | 1.143 | -2.366 | 0.010 | 1.167 | -2.335 |
| 14$2012 | 0.881 | -1.470 | 0.095 | 0.955 | 0.721 | -1.568 | 0.033 | 0.719 | -1.644 |
| 15$2012 | 1.678 | -0.380 | 0.179 | 0.987 | 1.549 | -0.395 | 0.157 | 1.291 | -0.722 |
| 16$2012 | 1.683 | -0.726 | 0.061 | 0.960 | 1.369 | -0.752 | 0.009 | 1.381 | -0.773 |
| 17$2012 | 1.571 | -1.042 | 0.133 | 0.984 | 1.340 | -1.184 | 0.046 | 1.332 | -1.253 |
| 18$2012 | 1.072 | -1.982 | 0.329 | 0.997 | 0.978 | -2.368 | 0.169 | 0.971 | -2.611 |
| 19$2012 | 1.713 | -0.084 | 0.199 | 0.935 | 1.338 | -0.027 | 0.156 | 1.063 | -0.403 |
| 20$2012 | 1.982 | -0.873 | 0.278 | 0.998 | 1.907 | -0.906 | 0.259 | 1.567 | -1.336 |

– where pseudo-guessing, discrimination and difficulty parameters were estimated and the *d*-parameter was fixed to 1 and 2PLM – where only discrimination and difficulty parameters were estimated, the *d*--parameter was fixed to 1 and *c*-parameter to 0). Parameters were not compared between models but it was ascertained whether parameters obtained from simpler models could also point to potentially flawed items.

Values of parameters obtained from 4PLM, 3PLM and 2PLM are presented in Table 6. Only one item had an upper asymptote (d) equal to 1. For 18 items, *d*-parameter magnitudes ranged from 0.935 to 0.999, very close to one. The only *d*-parameter value was away from the rest was item 2$2012 – 0.751.

Choosing the threshold value for *d*-parameter to warn of potential problems with IWFs is arbitrary. Therefore, if the value of *d*-parameter drifted strongly from the parameter values' of other items, they were classified as flawed. Item 2$2012 had d value of 0.751 – the remainder had *d*-values ranging from 0.935 to 1, so it was categorised as potentially biased. This item was part of a testlet. An examinee was presented with a dialogue from *Revenge* (*Zemsta)* by Aleksander Fredro. An examinee had to interpret the dialogue and the intention of the male protagonist's words and behaviour towards the female protagonist. The format of the question was potentially quite misleading. The translation of the item is shown in Table 7 (the original wording item is presented in the Appendix).

The correct answer was 1C (an examinee could find the interpretation in the excerpt). Such item format could be classified as Type K, as it listed primary options. The secondary options were not listed directly. Item format assumed that an examinee would choose between all options from the left and all options from the right (which gave the following combination of answers: 1A, 1B, 1C, 2A, 2B, 2C) but the possible distractors were not listed in the booklet. Therefore, the examinee had to memorise the correct combination and mark it on the answer sheet, which could have led to mistakes. What is more, there was a problem with non-functional distractors in this item. Answers 1B, 2A and 2B were chosen by fewer than 5% of examinees – as they were logically rather implausible, even for examinees who did not read the text very carefully. While answering this item, examinees were prone to another mistake – selecting two alternatives instead of one. This item demonstrated the largest number of such mistakes (about 3.3% of total responses) of all the reading MCQs in 2012. The presence of such a specific mistake indicated that the item was somehow confusing. Analysing the other parameters, discrimination values from both 3PLM and 2PLM were the lowest (when compared to the values of other items within consecutive models), which suggested problems with discriminating low- from high-achievers.

The values of parameters obtained by 4PLM with 3PLM and 2PLM parameters for 2013 are presented in Table 8. Two items

Table 7
*The wording of item 2$2012*

How does the male protagonist (Papkin) start the conversation with female protagonist (Podstolina) and what does he want to achieve?

| | | |
|---|---|---|
| 1. He praises her | | He wants to seduce her |
| | because | He wants to annoy her |
| 2. He is careful while complementing | | He wants to keep the promise |

Table 8
*The parameters of reading items from 2013 obtained by 4PLM, 3PLM and 2PLM*

| Model / Item | 4PLM | | | | 3PLM | | | 2PLM | |
|---|---|---|---|---|---|---|---|---|---|
| | *a* | *b* | *c* | *d* | *a* | *b* | *c* | *a* | *b* |
| 1$2013 | 1.667 | -1.473 | 0.219 | 1.000 | 1.659 | -1.475 | 0.221 | 1.542 | -1.768 |
| 2$2013 | 3.092 | 1.328 | 0.498 | 0.961 | 2.318 | 1.445 | 0.493 | 0.315 | -0.730 |
| 3$2013 | 1.338 | -1.254 | 0.304 | 0.999 | 1.336 | -1.229 | 0.316 | 1.162 | -1.822 |
| 4$2013 | 0.848 | 2.989 | 0.395 | 0.470 | 0.028 | 18.844 | 0.049 | 0.024 | 16.732 |
| 5$2013 | 1.372 | -1.246 | 0.047 | 0.968 | 1.144 | -1.283 | 0.016 | 1.156 | -1.304 |
| 6$2013 | 1.685 | -1.28 | 0.363 | 1.000 | 1.676 | -1.281 | 0.364 | 1.422 | -1.864 |
| 7$2013 | 1.450 | -0.258 | 0.335 | 0.997 | 1.409 | 0.270 | 0.327 | 0.989 | -1.091 |
| 8$2013 | 1.477 | -1.708 | 0.087 | 0.980 | 1.190 | -1.868 | 0.029 | 1.209 | -1.880 |
| 9$2013 | 1.617 | 0.187 | 0.291 | 0.996 | 1.571 | 0.183 | 0.284 | 0.970 | -0.545 |
| 10$2013 | 2.137 | -0.246 | 0.233 | 0.995 | 2.041 | -0.252 | 0.223 | 1.463 | -0.693 |
| 11$2013 | 1.790 | -0.269 | 0.210 | 0.934 | 1.273 | -0.302 | 0.128 | 1.108 | -0.597 |
| 12$2013 | 1.609 | -1.756 | 0.039 | 0.956 | 1.062 | -1.937 | 0.015 | 1.087 | -1.923 |
| 13$2013 | 2.103 | 1.082 | 0.219 | 0.942 | 1.863 | 1.179 | 0.212 | 0.721 | 0.855 |
| 14$2013 | 2.141 | 0.171 | 0.074 | 0.996 | 2.117 | 0.180 | 0.073 | 1.715 | 0.035 |
| 15$2013 | 2.369 | -0.380 | 0.126 | 0.950 | 1.721 | -0.399 | 0.063 | 1.595 | -0.527 |
| 16$2013 | 2.470 | -1.217 | 0.349 | 0.588 | 0.195 | -0.784 | 0.026 | 0.194 | -1.047 |
| 17$2013 | 2.049 | -0.811 | 0.289 | 0.999 | 2.014 | -0.818 | 0.284 | 1.606 | -1.288 |
| 18$2013 | 2.130 | 0.229 | 0.189 | 0.995 | 2.059 | 0.231 | 0.182 | 1.315 | -0.153 |
| 19$2013 | 1.792 | -0.580 | 0.195 | 0.967 | 1.414 | -0.672 | 0.118 | 1.287 | -0.906 |
| 20$2013 | 1.446 | -0.422 | 0.445 | 0.998 | 1.419 | 0.430 | 0.440 | 0.954 | -1.575 |

Table 9
*The wording of item 4$2013*

What is the purpose of repeating the question "Is this courage?" in the article? Choose the right answer.

1. It functions as a framing device, which integrates the article.
2. It enables author to define courage.
3. It encourages the reader to analyse the problem stated in the article.
4. It is a proof of disregarding courageous behaviours.

A. All of the above.
B. Answers 1 and 3 are correct, and 2 and 4 incorrect.
C. Only answer 4 is incorrect, the remaining answers are correct.
D. Only answer 2 is correct, the remaining are incorrect.

out of 20 had upper asymptotes equal to one, upper asymptote values of 16 items ranged between 0.934 and 0.999. Upper asymptote magnitudes in the case of two items (4$2013 and 16$2013) were equal to 0.470 and 0.588 respectively. Again, the values for these items drifted the most, so their flaws were probably the most significant. The content and format

Table 10
*The wording of item 16$2013*

Decide whether the following statements concerning the remark by professor Jerzy Bralczyk are true. Choose "T", if the statement is true or "F", if the statement is false.

| | | |
|---|---|---|
| 1. The remark contains the information that the adjective "white" which is a part of various phrases may be understood differently by different cultures | T | F |
| 2. In order to explain the meaning of the phrase in which the name of colour is mentioned, we have to refer to the history of language | T | F |

of both items 4$2013 and 16$2013 was therefore examined.

Item 4$2013 used Type K format and one option was AOTA. The item was a part of the testlet and the question referred to an article about courage. The translation of the item is given below in Table 9 and original wording in the Appendix.

The correct answer was B. Option A (AOTA) was logically implausible, since option B (answers 1 and 3 were correct and 2 and 4 incorrect) and option D (only answer 2 was correct, the remaining incorrect) could not be simultaneously true. What is more, AOTA could refer both to the answers (A, B, C, D) and to the primary responses (1, 2, 3, 4). This ambiguity could potentially cause confusion. The wrong answer could be also provoked by the fact, that the article did not precisely define "courage" yet showed what "courage" was not. In all the analysed exams there was the implicit assumption that students should base their answers only on the information present in text. Basing the answer on own knowledge was not advantageous and therefore discouraged. However, high-achievers might have been especially tempted to use their own knowledge when they chose an answer. An item's discrimination between low-achieving and high-achieving students was close to zero and the difficulty parameter was extremely high in case of 3PLM. A similar conclusion could be drawn during analysis of the 2PLM results. The values of parameters from simpler IRT models suggested that the item was definitely flawed. Both item format and item content might have violated the guidelines concerned.

The other item was 16$2013 – part of a testlet, in which an examinee was presented with a statement by professor Jerzy Bralczyk concerning the cultural meaning of colours. The question is presented in Table 10 and the original wording is shown in the Appendix. It included two statements and examinees had to decide whether they were true or false. This item was scored dichotomously – an examinee had to answer both items correctly in order to score a point.

The correct answer was "True" for the first statement, "False" for the second statement. The item was a multiple true/false (MTF) question. Kreiter and Frisbie (1989) showed that when compared to the MCQs, MTFs yielded higher reliabilities and higher response rates. Therefore this format should be encouraged in test design. Such a value for the *d*-parameter might therefore suggest some problems with item content. When discrimination parameters from the simpler IRT models (3PLM and 2PLM) were examined, their values were very low. Therefore this item poorly distinguished low-achieving from high-achieving students. High-achieving students might have known that the explanation of the meaning of phrases required reference to the history of language...yet again this information was not mentioned in the text. Therefore, as they based the answer on their own knowledge, not on the information presented in the text – they chose a wrong answer.

Table 11
*The parameters of reading items from 2014 obtained by 4PLM, 3PLM and 2PLM*

| Model / Item | 4PLM | | | | 3PLM | | | 2PLM | |
|---|---|---|---|---|---|---|---|---|---|
| | *a* | *b* | *c* | *d* | *a* | *b* | *c* | *a* | *b* |
| 1$2014 | 2.567 | -1.736 | 0.179 | 0.999 | 2.400 | -1.810 | 0.130 | 2.340 | -1.907 |
| 2$2014 | 1.524 | -0.837 | 0.158 | 0.982 | 1.348 | -0.898 | 0.113 | 1.243 | -1.115 |
| 3$2014 | 1.194 | -0.457 | 0.131 | 0.996 | 1.162 | -0.470 | 0.122 | 1.027 | -0.758 |
| 4$2014 | 1.379 | -1.348 | 0.073 | 0.995 | 1.313 | -1.394 | 0.048 | 1.301 | -1.463 |
| 5$2014 | 2.208 | -0.854 | 0.218 | 0.604 | 0.444 | -0.016 | 0.011 | 0.442 | -0.071 |
| 6$2014 | 1.820 | -1.080 | 0.099 | 0.997 | 1.764 | -1.096 | 0.089 | 1.668 | -1.231 |
| 7$2014 | 1.195 | -1.669 | 0.091 | 0.998 | 1.171 | -1.684 | 0.085 | 1.150 | -1.816 |
| 8$2014 | 1.583 | -1.700 | 0.096 | 0.999 | 1.564 | -1.707 | 0.095 | 1.537 | -1.817 |
| 9$2014 | 1.790 | -0.963 | 0.137 | 0.998 | 1.760 | -0.967 | 0.134 | 1.588 | -1.186 |
| 10$2014 | 1.067 | -0.460 | 0.196 | 0.978 | 0.983 | -0.459 | 0.174 | 0.834 | -0.939 |
| 11$2014 | 0.791 | -1.529 | 0.050 | 0.974 | 0.724 | -1.498 | 0.044 | 0.715 | -1.609 |
| 12$2014 | 1.920 | -1.616 | 0.300 | 1.000 | 1.898 | -1.626 | 0.298 | 1.711 | -2.000 |
| 13$2014 | 1.329 | -2.100 | 0.287 | 0.999 | 1.283 | -2.220 | 0.229 | 1.248 | -2.514 |
| 14$2014 | 2.277 | -0.568 | 0.137 | 0.986 | 2.040 | -0.579 | 0.115 | 1.746 | -0.781 |
| 15$2014 | 1.471 | -0.875 | 0.294 | 0.999 | 1.446 | -0.884 | 0.29 | 1.171 | -1.477 |
| 16$2014 | 1.680 | -0.211 | 0.118 | 0.998 | 1.641 | -0.219 | 0.112 | 1.369 | -0.449 |
| 18$2014 | 1.928 | -1.187 | 0.180 | 0.999 | 1.889 | -1.200 | 0.174 | 1.706 | -1.449 |
| 19$2014 | 1.396 | -0.533 | 0.452 | 0.999 | 1.359 | -0.558 | 0.444 | 0.923 | -1.738 |
| 20$2014 | 1.558 | -0.612 | 0.227 | 0.994 | 1.482 | -0.632 | 0.212 | 1.209 | -1.074 |
| 21$2014 | 0.597 | -2.996 | 0.082 | 0.903 | 0.382 | -3.044 | 0.072 | 0.377 | -3.350 |

The values of parameters obtained by 4PLM, 3PLM and 2PLM for 2014 are shown in Table 11. The upper asymptote of one item was 1. Eighteen parameters ranged between 0.903 and 0.999 and the value of 1item (5$2014) was 0.604 – from which it can be inferred that it drifted strongly from the other values.

Item 5$2014 included two true-false statements – to score the point, an examinee had to answer both items correctly. It was a part of a testlet with an excerpt from *The Shadow of the Sun (Heban)* by Ryszard Kapuściński. The wording of the question is shown in Table 12.

The correct answer was TT (True in the first statement, True in the second statement). As mentioned earlier, such item format was encouraged by empirical research. Therefore, there might have been some problems with item content. The question was classified as the least discriminating according to 3PLM and 2PLM – therefore the content had to be examined. The questions posed by the author are "How else can I get to know this city? This continent?"[1]. It is disputable whether the emotions were

[1] Ryszard Kapusciński (2001). *The Shadow of the Sun*. London: The Penguin Press (translated by Klara Glowczewska).

Table 12
*The wording of item 5$2014*

Decide, whether the following statements concerning the role of questions posed in the first paragraph in text are true. Choose "T", if the information is True or "F", if it is False.

| | | |
|---|---|---|
| 1. They reflect the author's belief about choosing the right place to live. | T | F |
| 2. They add emotions to the presented statement. | T | F |

really visible within these two questions and how examinees understood their role (for example the questions could function as rhetorical devices). Using an ambiguous statement and coercing a dichotomous true or false answer could mislead examinees, causing mistakes.

## Discussion of results

Although 4PLM can be used to detect flawed items, computationally it is still time-consuming. What is more, flawed items can be detected both by means of qualitative analysis (flawed items violated usually at least two item-writing guidelines) and by simpler IRT models (flawed items were usually the ones with the lowest discrimination). Therefore, it has to be asked whether it is worth employing 4PLM, when the same results can be obtained in simpler and quicker ways. It is probably not worth replacing simpler IRT models with 4PLM during analysis, especially when it still lacks formal proof of identifiability. But is it worth replacing IRT models with qualitative analysis or is it better to use both methods? This is a contentious issue, as there were several differences observed classifying items as flawed by quantitative and qualitative analysis. More items were classified as flawed by qualitative analysis. On the other hand, parameters of such items (e.g. those which had negations in stem or even Type K) indicated that items should function well. Usually, even when items contained a large number of non-functional distractors, IRT parameters did not drift from the item parameters which were not categorised as flawed. Some items which did not violate item-writing guidelines, emerged to have a distinct upper asymptote. Such phenomena probably occurred due to ambiguous content (as indicated by qualitative analysis), yet there is no certainty. Such ambiguous relations between qualitative and quantitative analysis might have been caused by choosing specific item-writing guidelines as standards for item quality. To the author's best knowledge, no comprehensive meta-analysis has followed Rodriguez' (1997) study, which was performed nearly 20 years ago. Therefore choosing guidelines for good item-writing should be based on meta-analysis reflecting newer studies. Finally, it has to be asked whether some kind of variance connected with reading ability (e.g. deducing a correct answer from the clues given in stem and distractors) is not introduced during examinations. Therefore, discrimination of some items could be increased (although they possessed some flaws), while validity of the whole test was lowered (Masters, 1988).

Therefore, the analysis led to rather tentative conclusions – both qualitative analysis of item-writing/item-content and IRT modeling (not necessarily 4PLM – at least not before proof of identifiability is obtained), should be used in order to detect substantial flaws. Employing two methods for item analysis should add value to the process of item examination and evaluation. What is more, qualitative analysis can identify potential causes for item-writing flaws (although, of course verifying them is another issue), which is still not possible using quantitative methods.

Further research on item-writing flaws should sum up current studies in the form of a meta-analysis, as performed by Rodriguez (1997) and again verify which rules of item-writing are the most important, how their violations influence test and item parameters, and how specific combinations of flaws (e.g. Type K with non-functional distractors) impact examination results.

Further research on applications of 4PLM in educational research should concentrate on simulation, to examine the stability of psychometric properties of this model and last but not least on obtaining formal proof for its identifiability. Research should also concentrate on examining the psychometric properties of 4PLM with the individual-specific *d*-parameter. The model should offer a better approximation of what Barton and Lord (1981) originally intended, the capture by 4PLM of the processes which lead high-achievers to fail on specific items.

## Literature


Aamodt, M. G. and McShane, T. (1992). A meta-analytic investigation of the effect of various test item characteristics on test scores and test completion times. *Public Personnel Management*, *21*(2), 151–160.

Albanese, M. A. (1993). Type K and other complex multiple-choice items: an analysis of research and item properties. *Educational Measurement: Issues & Practice*, *12*(1), 28–33.

Albanese, M. A., Kent, T. and Whitney, D. (1977). A comparison of the difficulty, reliability, and validity of complex multiple-choice, multiple-response, and multiple true-false items. *Proceedings from the Sixteenth Annual Conference on Research in Medical Education* (pp. 105–110). Washington: Association of American Medical Colleges.

Albanese, M. A., Kent, T. H. and Whitney, D. R. (1979). Cluing in multiple-choice test items with combinations of correct responses. *Academic Medicine*, *54*(12), 948–50.

Barton, M. A. and Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model.* Princeton: Educational Testing Service. Retrieved from http://files.eric.ed.gov/fulltext/ED207996.pdf

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (eds.), *Statistical theories of mental test scores* (chapters 17–20). Reading: Addison–Wesley.

Boughton, K. A. and Yamamoto, K. (2007). A hybrid model for test speededness. In M. von Davier and C. H. Carstensen (eds.), *Multivariate and mixture distribution Rasch models* (pp. 147–156). Springer: New York.

Casler, L. (1983). Emphasizing the negative: a note on the not in multiple-choice questions. *Teaching of Psychology*, *10*(1), 51–51.

Cassels, J. R. T. and Johnstone, A. H. (1984). The effect of language on student performance on multiple-choice tests in chemistry. *Journal of Chemical Education*, *61*(7), 613–615.

Chalmers, R. P. (2012). *mirt:* a multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.

Cizek, G. J. and O'Day, D. M. (1994). Further investigation of nonfunctioning options in multiple-choice test items. *Educational and Psychological Measurement*, *54*(4), 861–872.

Cizek, G. J., Robinson, K. L. and O'Day, D. M. (1998). Nonfunctioning options: a closer look. *Educational and psychological measurement*, *58*(4), 605–611.

DiBattista, D. and Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *Canadian Journal for the Scholarship of Teaching & Learning*, *2*(2), article 4. Retrieved from http://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=1061&context=cjsotl_rcacea

Downing, S. M. (2002). Construct-irrelevant variance and flawed test questions: do multiple-choice item-writing principles make any difference? *Academic Medicine*, *77*(10), S103–S104.

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, *10*(2), 133–143.

Ellsworth, R. A., Dunnell, P. and Duell, O. K. (1990). Multiple-choice test items: what are textbook authors telling teachers? *The Journal of Educational Research*, *83*(5), 289–293.

Frary, R. B. (1991). The none-of-the-above option: an empirical study. *Applied Measurement in Education*, *4*(2), 115–124.

Gross, L. J. (1994). Logical versus empirical guidelines for writing test items: the case of "none of the above". *Evaluation & the Health Professions*, *17*(1), 123–126.

Haladyna, T. M. and Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*(1), 37–50.

Haladyna, T. M. and Downing, S. M. (1989b). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*(1), 51–78.

Haladyna, T. M. and Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, *53*(4), 999–1010.

Haladyna, T. M. and Downing, S. M. (2004). Construct--irrelevant variance in high-stakes testing. *Educational Measurement: Issues & Practice*, *23*(1), 17–27.

Haladyna, T. M., Downing, S. M. and Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309–333.

Hambleton, R. K. and Swaminathan, H. (1985). *Item response theory: principles and applications* (vol. 7). New York: Springer.

Hansen, J. D. and Dexter, L. (1997). Quality multiple--choice test questions: Item-writing guidelines and an analysis of auditing testbanks. *Journal of Education for Business*, *73*(2), 94–97.

Hohensinn, C. and Kubinger, K. D. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational & Psychological Measurement*, *71(4),* 732–746.

Huntley, R. M. and Plake, B. S. (1984). *An investigation of multiple-choice-option items: item performance and processing demands.* Paper presented at the meeting of the National Council on Measurement in Education, New Orleans.

Jozefowicz, R.F., Koeppen, B.M., Case, S., Galbraith, R., Swanson, D. and Glew, H. (2002). The quality of in-house medical school examinations. *Academic Medicine*, *77*(2), 156–161.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277–298.

Knowles, S. L. and Welch, C. A. (1992). A meta--analytic review of item discrimination and difficulty in multiple-choice items using "none--of-the-above". *Educational and Psychological Measurement*, *52*(3), 571–577.

Kolstad, R. K., Briggs, L. D., Bryant, B. B. and Kolstad, R. A. (1983). Complex multiple-choice items fail to measure achievement. *Journal of Research & Development in Education*, *17*(1), 7–11.

Liao, W. W., Ho, R. G., Yen, Y. C. and Cheng, H. C. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behavior & Personality: an international journal*, *40*(10), 1679–1694.

Linden, W. J. van der (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308.

Loken, E. and Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 509–525.

Magis, D. (2013). A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement*, *37*(4), 304–315.

Masters, G. N. (1988). Item discrimination: when more is worse. *Journal of Educational Measurement*, *25*(1), 15–29.

Messick, S. (1989). Validity. In R. L. Linn (ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Mroch, A. A., Bolt, D. M. and Wollack, J. A. (2005). *A new multi-class mixture Rasch model for test speededness.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec.

Mueller, D. J. (1975). An assessment of the effectiveness of complex alternatives in multiple choice achievement test items. *Educational and Psychological Measurement*, *35*(1), 135–141.

Raiche, G., Magis, D., Blais, J.-G. and Brochu, P. (2013). Taking atypical response patterns into account: a multidimensional measurement model from item response theory. In M. Simon, K. Ercikan and M. Rousseau (eds.), *Improving large-scale assessment in education.* New York: Routledge.

Reise, S. P. and Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods*, *8*(2), 164–184.

Rodriguez, M. C. (1997). The art & science of item writing: a meta-analysis of multiple-choice item format effects. Paper presented at the Annual meeting of the American Education Research Association, Chicago.

Rulison, K. L. and Loken, E. (2009). I've fallen and i can't get up: can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement*, *33*(2), 83–101.

San Martín, E., González, J. and Tuerlinckx, F. (2014). On the unidentifiability of the fixed-effects 3PL model. *Psychometrika*. doi:10.1007/S11336-014-9404-2

Schuwirth, L. W. and Vleuten, C. P. van der (2004). Different written assessment methods: what can be said about their strengths and weaknesses? *Medical Education*, *38*(9), 974–979.

Stark, S., Chernyshenko, O. S., Drasgow, F. and Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, *91*(1), 25–39.

Tarrant, M. and Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, *42*(2), 198–206.

Tarrant, M., Knierim, A., Hayes, S. K. and Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today*, *26*(8), 662–671.

Tarrant, M., Ware, J. and Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Medical Education*, *9*(1), 40–48.

Waller, N. G. and Reise, S. P. (2010). Measuring psychopathology with nonstandard item response theory models: Fitting the four-parameter model to the Minnesota Multiphasic Personality Inventory. In S. E. Embretson (ed.), *Measuring psychological constructs: advances in model-based approaches* (pp. 147–173). Washington: American Psychological Association.

Woodford, K. and Bancroft, P. (2005). Multiple choice questions not considered harmful. In A. Young and D. Tolhurst (eds.), *Proceedings of the 7th Australasian conference on computing education* (vol. 42, pp. 109–116). Darlinghurst: Australian Computer Society.

## Appendix

*The original wording of item 2$2012*

W jaki sposób Papkin rozpoczyna rozmowę z Podstoliną i co chce przez to osiągnąć?

| | | |
|---|---|---|
| 1. Przesadnie ją komplementuje | ponieważ | chce ja uwieść. |
| 2. Zachowuje ostrożność w komplementowaniu | | chce ją zdenerwować. |
| | | chce wywiązać się ze złożonej obietnicy. |

*The original wording of item 4$2013*

| Jaką funkcję spełnia dwukrotnie postawione w tekście pytanie: Czy to jest odwaga? | Wybierz właściwą odpowiedź spośród podanych. |
|---|---|
| 1. Stanowi klamrę kompozycyjną spajającą wypowiedź.<br>2. Umożliwia autorowi zdefiniowanie odwagi.<br>3. Ma zachęcić czytelnika do przemyślenia postawionego problemu.<br>4. Dowodzi lekceważenia odważnych zachowań | A. Wszystkie odpowiedzi są poprawne.<br>B. Odpowiedzi 1 i 3 są poprawne, a 2 i 4 błędne.<br>C. Tylko odpowiedź 4 jest błędna, pozostałe są poprawne.<br>D. Tylko odpowiedź 2 jest poprawna, pozostałe są błędne. |

*The original wording of item 16$2013*

Oceń, czy poniższe informacje dotyczące wypowiedzi profesora Jerzego Bralczyka są prawdziwe. Wybierz „T", jeśli informacja jest prawdziwa lub „N", jeśli jest fałszywa.

| | | |
|---|---|---|
| 1. Wypowiedź zawiera informację o tym, że przymiotnik biały występujący w związkach frazeologicznych jest rozmaicie kojarzony w różnych kulturach. | T | N |
| 2. Aby wyjaśnić znaczenie związku frazeologicznego, w którym występuje nazwa koloru, należy odwołać się do historii języka. | T | N |

*The original wording of item 5$2014*

Oceń, czy poniższe informacje dotyczące funkcji pytań w pierwszym akapicie tekstu są prawdziwe. Wybierz „T", jeśli informacja jest prawdziwa lub „N", jeśli jest fałszywa.

| | | |
|---|---|---|
| 1. Podkreślają przekonanie autora o słuszności wyboru miejsca zamieszkania. | T | F |
| 2. Nadają wypowiedzi zabarwienie emocjonalne. | T | F |