Practical Assessment, Research, and Evaluation

Volume 27 Article 5

2022

Essentials of Visual Diagnosis of Test Items. Logical, Illogical, and Anomalous Patterns in Tests Items to be Detected

Dr. Jari Metsämuuronen
Finnish National Education Evaluation Centre (FINEEC)

Recommended Citation

Metsämuuronen, Dr. Jari (2022) "Essentials of Visual Diagnosis of Test Items. Logical, Illogical, and Anomalous Patterns in Tests Items to be Detected," *Practical Assessment, Research, and Evaluation*: Vol. 27, Article 5.
DOI: https://doi.org/10.7275/n0kf-ah40
Available at: https://scholarworks.umass.edu/pare/vol27/iss1/5

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 27 Number 5, March 2022 ISSN 1531-7714

Essentials of Visual Diagnosis of Test Items. Logical, Illogical, and Anomalous Patterns in Tests Items to be Detected

Jari Metsämuuronen
1) *Finnish Education Evaluation Center* and 2) *Centre for Learning Analytics, University of Turku*

This article discusses visual techniques for detecting test items that would be optimal to be selected to the final compilation on the one hand and, on the other hand, to out-select those items that would lower the quality of the compilation. Some classic visual tools are discussed, first, in a practical manner in diagnosing the logical, illogical, and anomalous patterns in item responses, and second, a new way of illustrating the behavior of the items, cut-off curve, is introduced. All the illustrations in the article are accessible even without specific software packages; they are made by using basic spreadsheet software.

Introduction

In constructing a good measurement instrument including high validity and reliability, selection of good test items is crucial. How to know which items would be the best ones to be selected to the compilation—or, as is in focus in this article, how to detect and out-select those items that would lower the quality of the compilation? Obviously, we are willing to use as wide base of information as possible, "full information" (see Haladyna & Rodrigues, 2021). This article focuses on graphical tools although statistical or analytical tools are discussed too; usually these are used combined. This article concentrates on the technical quality of the items keeping in mind that the content-wise quality of the items, that is, the item validity maybe even more crucial in creating a valid measurement instrument.

Depending on the model used in the item analysis, we have one to five technical parameters describing the behaviour of the item. Their content is somewhat different in the practices within the classical test theory (CTT) and classical item analysis (see Gulliksen, 1950; Lord, Novick, & Birnbaum, 1968) than in the modern test theory including Rasch modelling (Rasch, 1960 onwards), item response theory (IRT) modeling (Lord et al., 1968 onwards), Mokken modelling or non-parametric IRT (NIRT; Mokken, 1971 onwards; see also Sijtsma & Molenaar, 2002; Stout, 2002), multidimensional IRT (MIRT; Mulaik, 1972; Reckase, 1972 onwards; see also Reckase, 2009), and widening the scope also to the confirmatory factor analysis (CFA) or structural equation modeling (SEM; Jöreskog, 1969, 1970 onwards). In practical settings related to item analysis, both the classical and modern test theory are often combined; software packages for IRT are many, and they usually include classical parameters also (see also the theoretical connection of CTT and IRT in, e.g., Bechger et al., 2003; Raykov & Marcoulides, 2016).[1] We also remember that validity of

[1] A useful tool for classical item analysis is, as an example, TIAPLUS software (CITO, 2013). Some widely-used software packages for the IRT modeling include, among others, BICAL (Wright & Stone, 1979), BILOG (Mislevy &

the test construct and, specifically, the construct validity including the effect of individual item's impact in the compilation are usually analysed and illustrated by using SEM software packages such as AMOS (Arbuckle, 2015), EQS (Bentler, 1995; Bentler & Wu, 2008), LISREL (Jöreskog & Sörbom, 1999a; 1999b; Jöreskog, Olsson, & Wallentin, 2016), or MPLUS (Muthén & Muthén, 1998–2017); this area is not discussed in this article even if the content-wise quality of the items, that is, the item validity is a crucial item "parameter" too (see Lord et al., 1968; Metsämuuronen, 2017).

Three generally used technical item parameters are the *discrimination power* (a-parameter in IRT modeling) reflecting the accuracy of the item in separating the test takers from each other, *item difficulty* (b-parameter) reflecting the locations of the item in the ability scale, and *probability to guess the correct answer* (c-parameter), known also as pseudo-change score level (e.g., Verhelst, Glas, & Verstralen, 1995) reflecting the easiness to guess or know the correct answer in the lowest-scoring group in the target group (of the discussion of the parameters, see, e.g. Lord, et al., 1968; Metsämuuronen, 2017). These three parameters are in a common use in one-, two-, and three-parameter (logistic) models within the Rasch- and IRT modelling.

The same three parameters are used within the classical item analysis also. For the item discrimination power, we have several indices of which Oosterhof (1976), as an example, compared 19 of the frequently discussed ones (see also comparisons in Cureton, 1966a, 1966b; Metsämuuronen, 2022a; Wolf, 1967) and several new ones have been suggested during the years (see, e.g., Metsämuuronen, 2020a; 2021a, 2022a; Moses, 2017). As examples of the classical coefficients, two widely-used classical indices based on mechanics of product-moment correlation coefficient (Pearson, 1896 onwards) are the item–score or item–total correlation (*Rit*) and item–rest correlation or corrected item–total correlation (*Rir;* Henrysson, 1963); some alternatives are discussed Section "Connection of item difficulty…". *Rit* and *Rir* are defaults in widely used general statistical software packages such as IBM SPSS and SAS which may explain their wide use. Difficulty level, known also as facility index is usually the proportion of correct answers, that is, the observed sum of the values in the item divided by the maximum possible sum in the item (minus the minimum possible sum if it is not zero) although other options are available (see, e.g., Moses, 2017). Within the classical item analysis, guessing parameter is usually related to the number of alternatives in the multiple-choice questions (MCQ): the classical guessing probability is the inverse of the number of alternatives: $P = ¼ = 0.25$ if having four alternatives and $P = ⅕ = 0.20$ if having five alternatives, as examples.

Writing a good test item is a form of art, and experienced item writer or item evaluator may say without any statistical or visual tool that certain items cannot be good. Many good lists of advice have been given for item writing to avoid the basic flaws (e.g., Ebel & Frisbie 1986; Hopkins 1997; Mehrens & Lehmann 1991; Metsämuuronen, 2017; Miller, Linn, & Gronlund, 2012). These lists include such good advises as "Write the item as clearly as possible", "Avoid giving irrelevant clues to the correct answer", "All distractors (in the multiple-choice items, MCQ) should be plausible and homogenous; no distractor should be automatically eliminated because it is irrelevant or stupid", and "Only one of the alternatives (in MCQ) can be correct or the best" just to mention a few (see more detailed in Metsämuuronen, 2017, pp. 82–85 and,

Bock, 1982) and its later version BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003), BMIRT (Yao, 2003), LOGIST (Wingersky, Barton & Lord, 1982), MULTILOG (Thissen, 1983; 2003), MPLUS (Muthén & Muthén, 2006), NOHARM (Fraser & McDonald, 1988), OPLM (Verhelst, Glas & Verstralen, 1995), PARSCALE (Muraki & Bock, 2003), WINSTEPS (Linacre, 2011), QUEST (Adams & Khoo, 1993) and its later version CONQUEST (Wu, Adams, & Wilson, 1998), TESTFACT (Bock et al., 2003), and a new type of application IRTPRO (Cai, Thissen, & du Toit, 2011) based on exploratory- and confirmatory factor analysis. One can find updated versions of the living software packages, for example, at the site of ssi-central https://ssicentral.com/index.php/products/. In R environments, the parameters can be obtained, for instance, by the packages ltm (https://cran.r-project.org/web/packages/ltm/ltm.pdf), equateIRT (https://cran.r-project.org/web/packages/equateIRT/index.html) or irt (https://cran.r-project.org/web/packages/irt/irt. pdf).

e.g., Mehrens & Lehmann, 1991). Such basic flaws in item writing as the ambiguity in wording, use of extreme words such as "always", "all", "never", and unintended several correct answers, as examples, may lead the test takers confused and, hence, the discrimination power of the item may be reduced because of technical reasons.

Usually, analytical tools, that is, numerical and statistical methods are used in assessing the technical quality or behaviour of the items (see different options in, e.g., CITO, 2013; Haladyna & Rodriges, 2021; Metsämuuronen, 2017; Moses, 2017) and, in many cases, this requires some specific software packages for item analysis; some of these are collected at Footnote 1. However, the visual or graphical diagnosis of the items is often used as a practical additional tool to examine the item behaviour. One of the advances of the IRT modelling over the classical test theory is the effective visualization of the item behavior, that is, the item characteristic curve (ICC; see later Figure 1). Within the classical item analysis, there has been less possibilities of the visualizations. However, using graphical possibilities in item analysis and -diagnostics is usually a good practice and easy way to detect some known specific anomalous, illogical, or maybe even "pathological" patterns (word used by Metsämuuronen, 2017 in this context) embedded to or unrevealed by the numerical analysis. Some of these specific patterns are discussed, for example, by Linacre and Wright (1996), Smith (1996), and Metsämuuronen (2017, pp. 169–172).

This article discusses some alternatives for the visual item analysis without specific IRT or SEM software packages. The potential beneficiary of the article could be a practitioner who does not have access to or knowledge of sophisticated tools for the IRT modelling, maybe, using classical analysis as the main analysis paradigm. These practitioners may come from the realm of practical education settings using classical indices for item discrimination power instead of IRT modelling. Simple tools such as common spreadsheet software packages can be used in the manual calculation in settings where sophisticated software packages for item analysis are not available. In these settings, the graphical options to perform item diagnosis may open wider possibilities to analyze and select the items to the tests—or to out-select items with an illogical or anomalous behavior from the final version. A general software package such as IBM SPSS, SAS, STATA, or different R libraries may be helpful for preparing the graphs; if these are not available, a basic office tool is sufficient for many statistics needed in the graphs. All the visualizations in the article are done by using Microsoft Excel software.

The course of the article starts with the basic visualization used in the Rasch and IRT modelling in the next section. Section "Basic graphical analyses…" shows some traditional ways of illustrating the item behavior within the classical approach. Section "Cut-off curve…" introduces a new type of visualization, the cut-off curve, based on the procedure of exhaustive splitting (PES) proposed by Metsämuuronen (2017; 2020b).

Basic graphical analysis within the Rasch- and IRT modelling approach

Item characteristic curve

One of the strengths of Rasch- and IRT modelling over the classical item analysis is the possibility of visual item analysis provided by the item characteristic function (ICF) and item characteristic curve (ICC). Originally, Rasch (1960) noted that the probability of giving a correct answer, assuming a certain ability level (theta, θ) and the item difficulty (b or beta, β), can be modelled by using the one-parameter logistic (1PL) function:

$$P(X_g = 1 \mid \theta, b) = \frac{1}{1 + e^{-(\theta - b)}} . \quad (1)$$

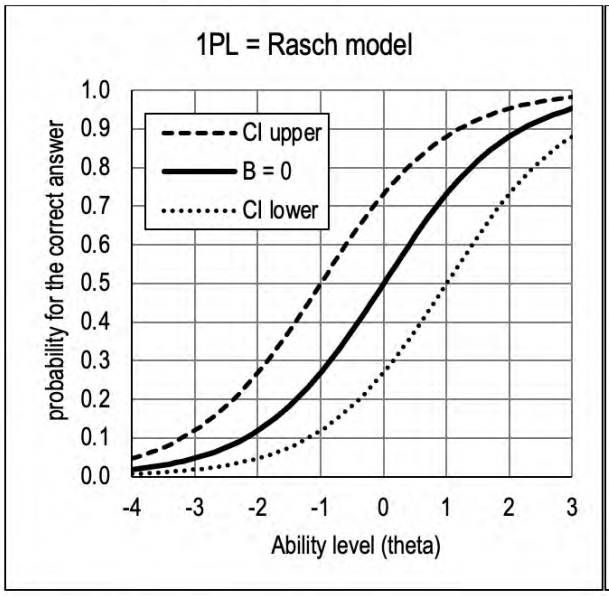
In practical terms, Eq. (1) expresses the probability (P) for the event that the test taker gives the correct answer in the item g ($X_g = 1$) assuming certain (known or estimated) ability (or trait) level θ and item difficulty level b. If we know—usually by using an IRT software—the value of the item difficulty b, we can visualize how the item should behave in the theoretical situation by drawing a graph of the model, that is, in a spreadsheet software, by tabulating the values of θs from –4 to +4 and calculating P with a fixed value in b (see the set of graphs in Figure 1; see also Livingston & Dorans, 2004 or Metsämuuronen, 2017 for examples of outputs by using a IRT software). This graph shows the 'theoretical' ICC (see later the "observed ICC"). As the simplest, the theoretical ICC

of a Rasch model could be as in Figure 1.1, where also the (hypothetic) lower and upper boundary of the confidence interval (CI) are illustrated.

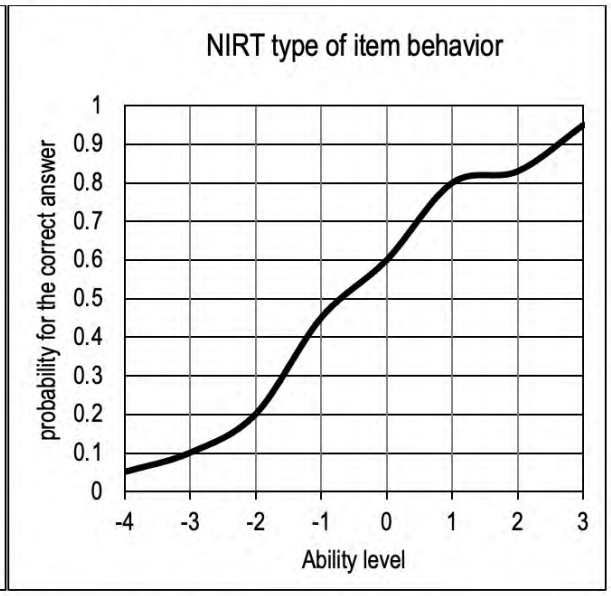
An essential characteristic of the theoretical ICC is that the curve is monotonically growing, that is, in each (higher) ability level, the probability of the correct answer is higher than in the previous (lower) ability levels. In practical terms, the higher-scoring test takers should give the correct answer more probable than the lower-scoring test takers. The more there are test takers the narrower the CI and, parallel, the less test takers the wider is the CI. When the 'observed' ICC, based on the actual dataset, is drawn to the same graph, we can compare how well the observed dataset fits the theoretical curve. If the observed ICC stays within the CI, the inference is that the observed ICC fits the theoretical ICC. The observed ICCs are handled in the next section.

Figure 1. Theoretical ICCs produced by different models

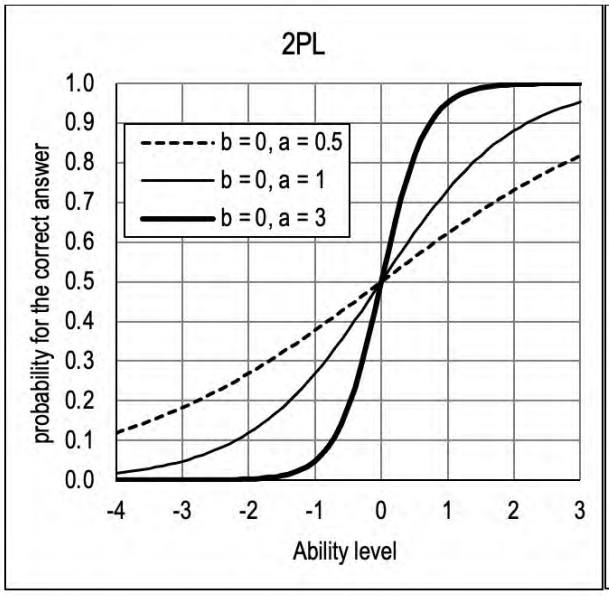
1. ICC of a one-parameter logistic model



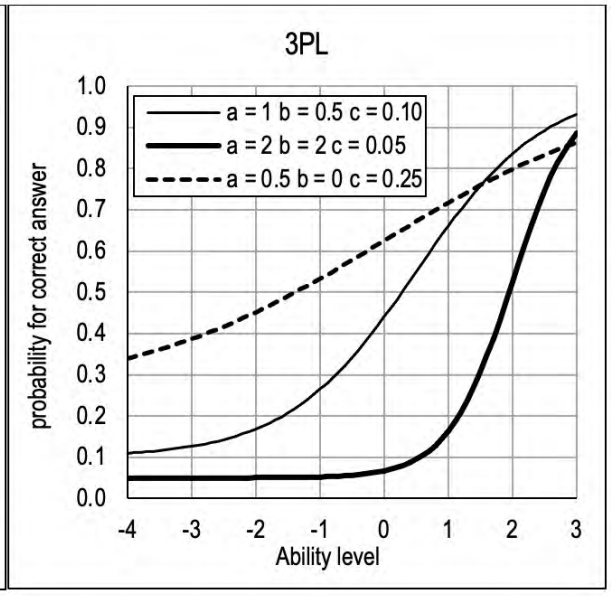
2. ICC of a NIRT model



3. ICCs of the two-parameter logistic models



4. ICCs of the three-parameter logistic models



In many practical settings, the observed dataset does not follow the (parametric) logistic function illustrated in Figure 1.1. Instead, in these cases, a better model could be a nonparametric IRT (NIRT)- or Mokken model visualized in Figure 1.2. An essential characteristic of the observed ICC of a NIRT type of item is that the curve is growing—although not necessarily monotonically. This means that, in each ability level, the probability of the correct answer is higher *or the same* than in the previous ability levels, and that the observed curve need not follow any specific known (parametric) function.

In the two-parameter (logistic) model (2PL), one describes both the difficulty parameter *b* and the item discriminating power *a* (Figure 1.3). The *b*-parameter tells us the location of the information of the item (in terms of the ability level) and the *a*-parameter tells us how accurately the item discriminates among the respondents at that point. The general ICF of the two-parameter model is

$$P(X_g = 1 \mid \theta, b, a) = \frac{1}{1 + e^{-Da(\theta - b)}}, \qquad (2)$$

where θ is a certain value of ability level, b is a certain value of item difficulty, a is a certain value of the item discrimination power, and D is 1 when a logistic model, and 1.7 when a Normal Ogive Model (see, Lord et al., 1968). Three examples of ICCs of two-parameter models with identical values of the *b*-parameter are seen in Figure 1.3 by variating the value of the *a*-parameter to visualize the effect of *a*-parameter. The interpretation of the two-parameter ICCs are straightforward: the steeper the curve is, that is, the higher the value for the *a*-parameter, the more discriminating the item is. The highest information about the item is at location of *b* and near that; the item discriminates the best between the test-takers who are in that range. If the item is an easy one (*b*-parameter is low), and the item is highly discriminating, the item discriminates between the low-scoring and even lower-scoring test-takers; the item cannot discriminate between the high-scoring test takers—they all give a correct answer. Parallel, if the item is a demanding one (*b*-parameter is high), and the item is very discriminating, the item discriminates between the high-scoring and even higher-scoring test-takers.

Finally, in the three-parameter logistic model (3PL), additional to the parameters *a* and *b*, the guessing parameter *c* is also expressed in the ICF and ICCs (see Figure 1.4). The model by Birnbaum (see Lord et al., 1968) is as follows:

$$P(X_g = 1 \mid \theta, b, a, c) = c + (1 - c)\frac{1}{1 + e^{-Da(\theta - b)}}, \qquad (3)$$

where θ is a certain value of ability level, b is a certain value of item difficulty, a is a certain value of the item discrimination power, c is a certain level of guessing, and D is 1 when a logistic model, and 1.7 when a Normal Ogive Model (see Lord et al., 1968). The *c*-parameter tells us how probable it is to obtain the correct answer in the group that performs the lowest in the test. It may be good to note that even if the classical probability for the guessing with MCQs would be 0.20 or 0.25 (see above), in practical settings, the probability for guessing the correct answer may stay much lower, specifically, with semi-difficult items (keeping in mind that, with easy items, even the low-performing test-takers may know the correct answers). This is caused by the fact that the test-takers rarely make random guesses, but they start to use some other strategy in selecting the correct answer. If the distractors are wisely selected, such as in a test of mathematics a certain mistake in the calculation leads to a certain distractor, this leads the low-performing test-takers to select wrong alternatives leading to lower-than-expected probability for guessing the correct answer.

Observed ICCs without the specific IRT software

Though it is possible to produce the theoretical ICCs and their confidence intervals by using common spreadsheet software packages as is done in Figures 1.1–1.4, usually we use specific software packages for these graphs. However, if a specific software package is not available, rough observed ICCs can be drawn either by combining a general statistical software (e.g. SPSS, SAS, STATA, R) with the basic spreadsheet software, or by producing both the basic statistics and illustrations by using the spreadsheet software alone—if not willing to use the general software itself in producing the graphs.

For the basic graphs, one needs the percentage of correct answers in the item (p) and the total score—either a weighted or unweighted one. Then, p is used as an estimate for the probability of correct answers, and the score is used as an estimate for the ability (or trait) level. The ability levels are obtained by dividing

the score into a reasonable number of ability groups; quartiles, quintiles, or deciles, as examples, may serve in this division. Figure 2 illustrates two sets of observed ICCs with different numbers of ability groups. These rough ICCs are composed by dividing the test-takers into four to seven groups (quartiles, quintiles, sextiles, septiles) according to their performance in the test; small sample sizes allow only a small number of groups while, obviously, large sample sizes allow more groups. In every ability group, the percentage of the correct answer is plotted against the ability level.

From the viewpoint of item diagnosis, the items in Figure 2 follow the basic rule for a well-behaving item: the observed ICCs are monotonically increasing functions although the item 1 is easier than the item 2; in the former, reasonably many test takers know the correct answer remarkably more probable than in the latter item even in the lowest-performing group. Next section discusses some known anomalies from this principle.

Basic anomalies in the observed ICCs to be detected

The eloquence in the observed ICCs is that the curves can be used in rough visual diagnostics of item behavior. With a large or large-ish sample size (around 100–200 respondents or more per item), we can assess plausibly how logical the connection between the ability level and probability of the correct answer is. If the dataset is very small, ICC is not necessarily the best option for the graphical item diagnosis. The obvious reason is that, if the number of cases is small, the ability groups are small and unstable—adding just one test taker to the dataset may change the graphs remarkably. However, even a small number of test takers may give us valuable information of the item behavior.

Metsämuuronen (2017, p. 167–172) illustrates six anomalies that can be detected by using the ICCs: illogical guessing, no discrimination, negative discrimination, sleepiness, specific knowledge or imputed outliers, and differential item functioning (DIF). Some of these patterns are discussed also by Linacre and Wright (1994) and Smith (1996). Here, the patterns are illustrated by using rough ICCs with spreadsheet software (Figure 3).

All the patterns seen in Figure 3 show an unwanted behavior in the items to some extent. In some cases, the reason for the anomaly may be caused by a small sample size; with larger sample size, the potential challenge may be rectified. If, however, we detect these patterns with large or large-ish sample sizes in the developmental phase of a test, it would be wise to either rewrite the items or plainly discard those from the final test. Another viewpoint in evaluating the

Figure 2. Observed ICCs based on the score and percentage/proportion of correct answers

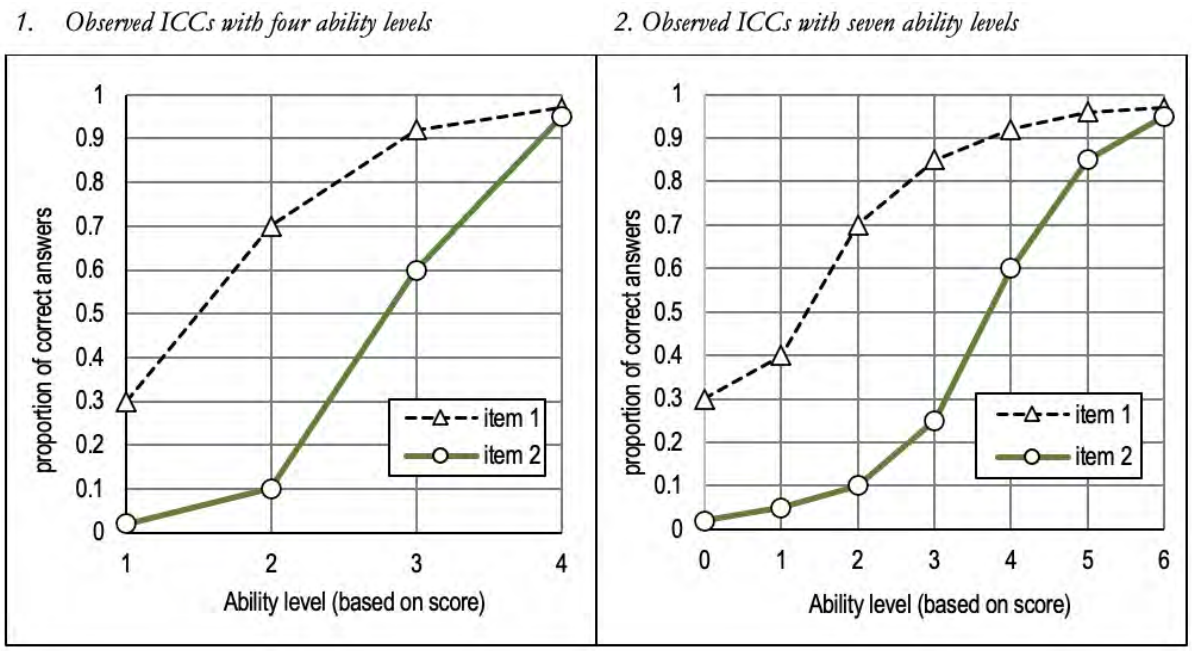
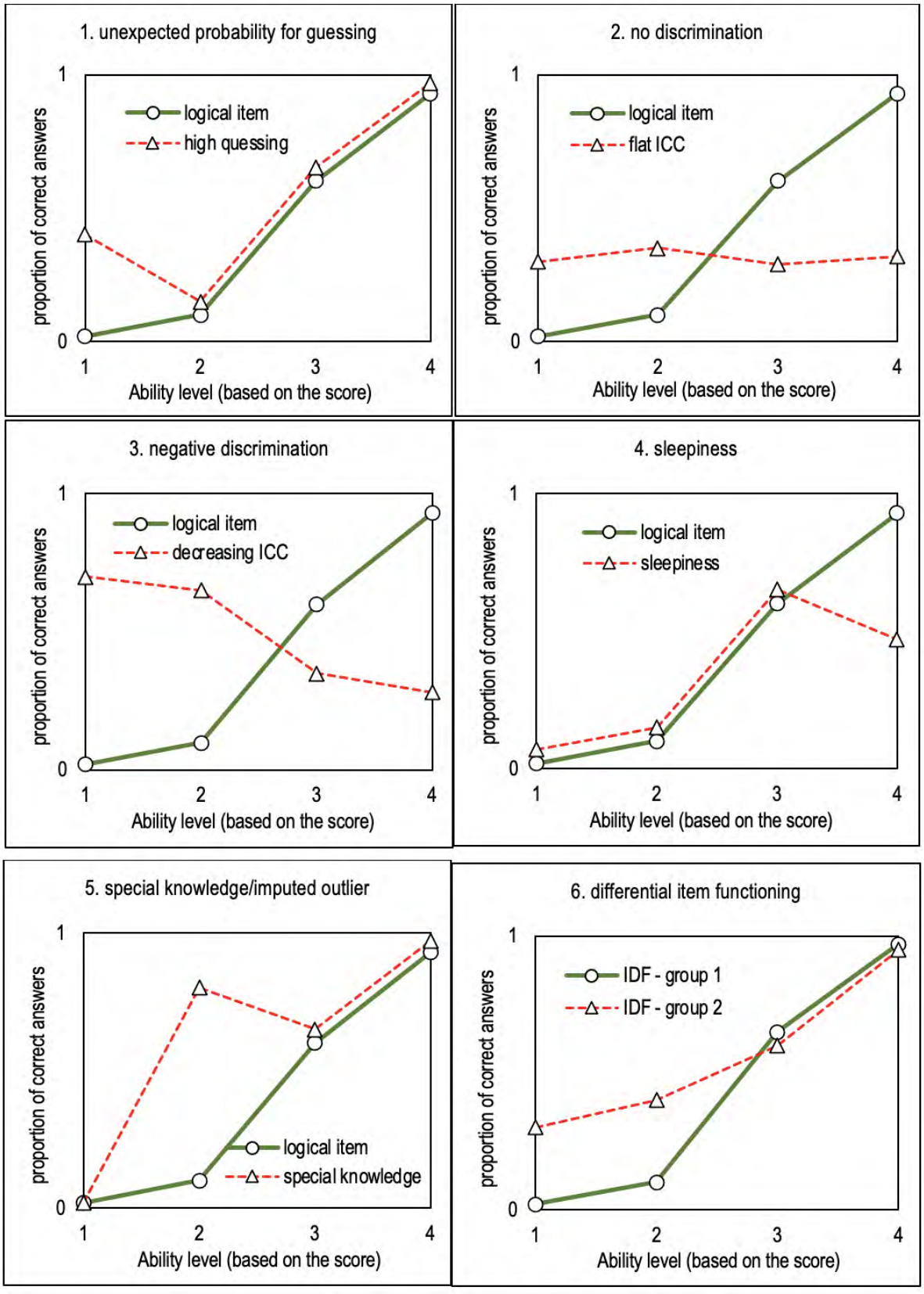


Figure 3. Basic patterns of anomalous observed ICCs based on the score and percentage/proportion of correct answers



possible anomalous pattern is that, always when taking samples, the estimate of the probability to give the correct answer (estimated by the proportion of correct answer) includes random error. This means that the estimate is never exact, and the minor deviances from the logical or theoretical ICC may be explained by random error. Hence, calculating the confidence interval (CI) of the estimate at each ability level may increase our knowledge in evaluating the possible anomaly in ICCs. CI can be plotted to the same graph as the observed ICCs. In Figures 3.1 to 3.6, CI is not illustrated.

Figure 3.1 illustrates the illogical pattern in guessing. We see that the probability to guess the correct answer in the item is too high—the higher achieving test-takers in the ability level 2 give the correct answer *less* probably than those who are in the lowest level group. In the case of MCQs, the challenge is to construct the distractors in a way that they would lure those who do not have enough knowledge to select a wrong alternative. If we detect this pattern, something in the correct option attracts the lowest-levelled test takers to select the correct alternative instead of the distractors even if they did not have any idea of the correct answer. The hint may come from the "irrelevant clue", different length, or phrasing as discussed by Mehrens and Lehmann (1991, p. 134 ff.) and Metsämuuronen (2017, p. 101). Sometimes, the reason may stay unknown although, sometimes—afterwards—we may notice why the correct answer was easily guessed by the test takers belonging to the group of lowest-performing respondents. Usually, rewriting the distractors (or the correct alternative) may enhance the item behavior. Some ideas of distractor analysis are given in Section "Basic graphical analyses…" with the visualization of the classical item analysis.

Figure 3.2 illustrates the patterns that the items cannot discriminate between the higher-performing and lower-performing test takers. From the viewpoint of statistical analysis, this pattern would be detected as the low value for (any of) the index of item discrimination (see options in Section "Connection of item difficulty…"). A classical quick-and-dirty rule suggests that the items with the item-total correlation lower than 0.20 should be considered seriously to be discarded (e.g. Metsämuuronen, 2017, p. 150) although, obviously, no absolute lowest boundary for the *Rit* (except –1) can be set. However, the negative values are anomalies and illogical from the test theory viewpoint. From the item discrimination viewpoint, the items with very low discrimination are useless, and they reduce the test reliability (see the discussion of the deflation in reliability in e.g., Metsämuuronen, 2022a, 2022b, 2022c). Hence, if there are enough items from where the final compilation can be selected, it may be wise to discard the items with very low discrimination power. However, in some cases in an achievement tests these items can be very easy ones and serve as "warming up" type of items in a longer test. Very easy items with low discrimination power also may motivate the lower-achieving test takers to try their best in the test. Then, keeping these in the final test may make sense.

Figure 3.3 illustrates the patterns of negative discrimination; this is even more anomalous or maybe even "pathological" pattern than the low item discrimination (see also Figure 4 in Section "Connection of item difficulty…"). This pattern means that the higher-scoring test takers tend to give an *in*correct answer while the lower-scoring test takers tend to give a correct one. With MCQs the problem is, most probably, in the wrong key: instead of the alternative *A* (that we think is the correct answer) the alternative *C* is the correct one. It is good to note that, sometimes, the best students may identify the correct answer better than the item writer or create a totally new solution for a problem that was not included in the marking scheme. This also may lead to negative item discrimination. If these are not the cases and rectifying the incorrect key does not change the pattern, these items are wise to be discarded from the test.

Figure 3.4 illustrates the patterns where the best students fail to give the correct answer—or at least some of them are behaving differently than what was expected. In some cases, just the best test takers find out that there is (truly) two correct alternatives or they cleverly notice some ambiguity in the wording of the item and mess (unintentionally or intentionally) with two or more options (or they select two alternatives which leads to a technical incorrect answer). In these cases, further analysis of the distractors (see Section "Graphical analysis of (possible)…") may reveal a need for amending the item. Sometimes, this pattern may be caused by "sleepiness" (see Linacre & Wright, 1994):

the item (or the test) is too easy for the best test takers and some of them select an incorrect option just by being lazy, bored, or "sleepy". If the latter can be ruled out, and the reason for the pattern is found to be in the technical quality of the item, it makes sense to try to amend the alternatives of MCQs or just to omit the item.

Figure 3.5 illustrates the patterns that Linacre and Wright (1994) call "specific knowledge" or "imputed outlier". In some ability group, where we would expect lower probability for the correct answer, the probability is unexpectedly high because of some unknown reason. The reason may be the fact that this specific topic has been taught in some specific group that is not, however, very good in the overall scoring (and, hence, the name "specific knowledge"). The same pattern is obtained if imputing the missing values by (obviously) too high value in comparison with the actual ability level (and, hence, the name "imputed outlier"). Logically, this pattern may be also reversed: a group with otherwise higher ability level *lacks* the knowledge of this specific information ("specific ignorance" may be the name for this phenomenon?). This kind of pattern tends to be rectified with larger datasets if it is caused by the "specific knowledge". The "imputed outlier" as a pattern is more difficult to rectify.

Finally, Figure 3.6 illustrates the pattern of differential item functioning (DIF). This phenomenon is usually detected by using statistical tools. However, it can be detected in a less analytical manner by using visual methods. Characteristic to DIF is that the observed ICCs differ notably between two or more populations such as between males and females or between language groups. If one population (such as girls/females) scores higher than the other populations(s) (such as boys/males), it leads, obviously and logically, to a pattern where the *levels* of the curves differ between the groups—this is not DIF in the sense of illogical item behavior. However, when the *forms* of ICCs differ between these groups, the item behaves differently in different populations. In Figure 3.6, it seems that, in population 2, the test takers from the lowest ability levels guess or know the correct answer radically more probably than the test takers in group 1. The reason for this may need some further analysis of the alternatives in MCQs. The relevant question is why the item favors population 2 over the population 1. Did we give an irrelevant clue specifically favoring this population? If no technical reasons (such as a mistake in inputting the data in the deviating populations) for this pattern were found, it would be recommendable to consider omitting these items.

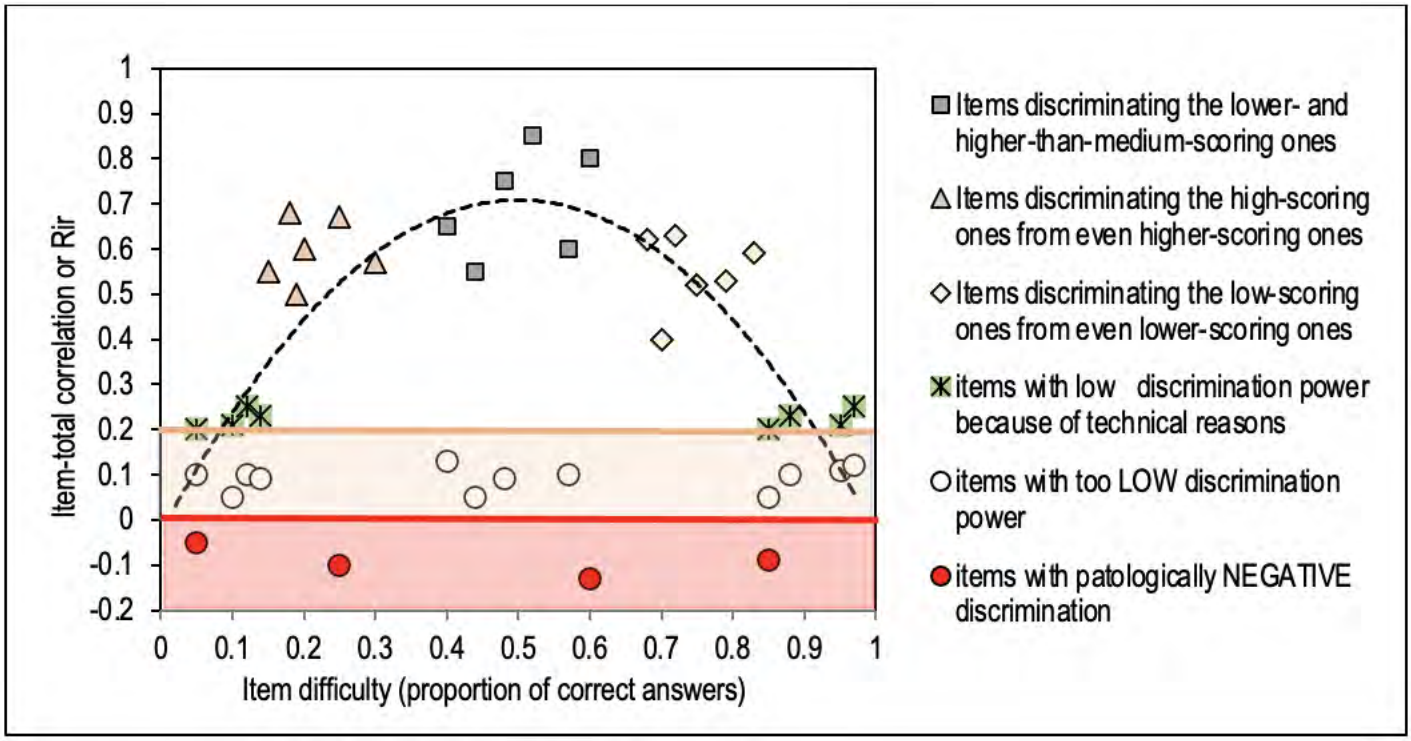
Basic graphical analyses within the classical item analysis approach

Connection of item difficulty and item discrimination

Within the classical test theory, there is no such tradition of illustrating the item characteristics as we can do when assuming the logistic function behind the item in Rasch- and IRT models. We may plot the item discrimination with the item difficulty into a traditional graph as in Figure 4. Although this kind of illustration tells limited information of the item behavior, in some cases, it may be enough for the decision of separating the better-behaving items from the less-well-behaving items.

Figure 4 illustrates mainly, in a schematic way, the behaviour of items from the item discrimination viewpoint (see later real-world examples). The first note to make is that the indices of item discrimination power based on Pearson product-moment correlation coefficient (e.g. *Rit* and *Rir*) cannot reach the perfect value 1 if the number of categories in the scales of the item and the score differ from each other as always is the case in practical testing settings (see algebraic reasons in, e.g., Metsämuuronen, 2016, 2017, and simulation in 2020a, 2021a, 2022a). With binary items, the highest value approximates $Rit = 0.87$, if each test taker would get a different score and the sample size is large enough ($n > 200$; see, e.g., Metsämuuronen, 2020a). The further the difficulty level is from $p = 0.5$, the less is the highest possible value even if the item would discriminate the test takers from each other in a deterministic manner and we would expect to obtain the perfect item discrimination (see simulation in Metsämuuronen, 2021a). This phenomenon is illustrated as a curved trend line in Figure 4. Metsämuuronen (2017, pp. 205–208) discusses algebraic reasons why *Rir* underestimates the discrimination power in items even *more* than *Rit*. The same challenge we face with the classic Kelley's discrimination index (*DI*; Kelley, 1939; see also

Figure 4. Connection of item discrimination and percentage/proportion of correct answers



Metsämuuronen, 2020b) discussed later in the article. The traditional *DI* can reach the ultimate value *DI* = 1 with the mediocre items but it may give a radical underestimation—greater in magnitude than that of *Rit*—with items with extreme difficulty level (0.20 > *p* > 0.80).

Second, among others, Metsämuuronen (2020a, 2020b, 2021a, 2022a; see also Moses, 2017) has studied possible alternatives for *Rit* and *Rir*. From the viewpoint of technical underestimation related to the discrepancy of the scales and item difficulty, polychoric correlation (*RPC*; Pearson, 1900, 1913), Goodman–Kruskal gamma (G; Goodman & Kruskal, 1954), and Somers delta (*D*; Somers, 1962; see the discussion of the correct direction in Metsämuuronen, 2020a), as examples, differ from the previous ones. The latter estimators *can* reach the ultimate value even with extreme patterns either exactly (*G* and *D*) or approximatively (*RPC*). From this viewpoint, they are superior alternatives for *Rit, Rir*, and *DI,* specifically, with the dichotomous items; for the polytomous items, Metsämuuronen (2021a) suggests using dimension-corrected *G* and *D* instead of *G* and *D*. Later, in Section "Graphical analysis of (possible)…", the behaviour of *RPC*, *G*, and *D* is illustrated in comparison with *Rit*.

Third, in practical terms, the curved trend in Figure 4 illustrates the fact that if we would use *Rit* or *Rir* as indicators of item discrimination power and the item is of extreme difficulty level, we would *expect* to obtain relatively low item discriminations because these estimators cannot even reach high values with items of extreme difficulty levels. The classical boundary for an acceptable discrimination has been the value *Rit* = 0.20 as discussed in Section "Basic anomalies…" above with Figure 3 (see also Gulliksen, 1950; Lord et al., 1968; Metsämuuronen, 2017) although this is not, obviously, a strict law. However, below this boundary it may be necessary to argue for selecting the item in the compilation. If there are many items in the pool from where the items can be selected to the compilation—for example, because of pretesting double number of possible items and being able to select only the best ones—there is not much need to select these items with the low discriminating power to the final test. Sometimes, however, these items may be the only ones measuring an important content area; in the case, from the content *validity* viewpoint, we may be interested in keeping the relatively lowly discriminating item in the test.

Fourth, Figure 4 illustrates why the traditional guide (e.g., Lord, 1952) makes sense to select items

with a medium difficulty level when willing to maximize the reliability of the measurement instrument. The estimates of the reliability of the score are strictly related to the item discrimination (see Lord et al., 1968; Metsämuuronen, 2021a, 2022a, 2022b, 2022c), and it is more likely to find highly discriminating items when the item difficulty level is of medium one. Nevertheless, in many practical testing settings related to learning outcomes, we are willing to construct a test by using items with increasing difficulty levels. This may lead to select also such very easy or very demanding items which, on average, are not the best options from the discrimination viewpoint.

Graphical analysis of (possible) underestimation in discrimination power

Above, it was noted that *Rit* and *Rir* underestimate the item discrimination power, and some alternative estimators were given such as *RPC*, *G*, and *D*.[2] Here, two empirical examples are given of this phenomenon and how the (possible) underestimation in the estimators can be detected visually.

First, the estimates by *RPC*, *G*, and *D* need to be calculated. If using a spreadsheet software in calculations, Zaiontz's (2021) procedure and tools can be used for *RPC*. *G* and *D* are easy to calculate even manually (see, e.g., Metsämuuronen, 2017; 2021a, 2021b, 2022a; Siegel & Castellan, 1988) although, for the article, they were calculated by using IBM SPSS (see Table 1; see Appendix A for syntaxes when using such general software package such as IBM SPSS, SAS or R environment). Second, the estimates are plotted in the same graph (see Figure 5).

The examples come from national level assessments of learning outcomes with representative samples. The first is based on a dataset related to a mathematics test of 30 MCQ items with incremental difficulty levels ranging p = 0.913–0.259 (FINEEC, 2018; n = 4,023) from where 10 items with varying difficulty levels are selected for illustration purpose in Figure 5.1. Notably, all the correlations indicating the item discrimination are high or decent (*Rit* = 0.369–0.691) because the low-discriminating items were omitted in the pre-test phase. The other example comes from a more specific, very easy 8-item, 11-point subtest as a part of a larger assessment of learning outcomes reported by Metsämuuronen and Ukkola (2019; n = 7,770) and re-analysed by Metsämuuronen (2022b, 2022c). This subtest measured the preconditions of understanding the instruction language in the main test. It was expected that only students with second language background would make mistakes in the test. Hence, the difficulty levels ranged p = 0.992–0.892 and, of all test takers, 72% were given the full marks. Notably though, all the item–total correlations are acceptable or high (*Rit* = 0.246–0.751) although, in this kind of dataset with extremely easy items, *Rit* tends to underestimate true correlation in an obvious manner. The figures for Figure 5 are collected in Table 1.

Two notes are made. First, the estimates by *RPC*, *G*, and *D* are closer to each other than to *Rit*. This seems to indicate that they reflect the same (latent) correlation even if they estimate different things: while *RPC* estimates the *inferred correlation* of two unobservable continuous variables by their ordinal manifestations, *G* and *D* estimate the *probability* that the test takers are in the same order both in an item and a score (Metsämuuronen, 2021a), and the latter strictly indicate the proportion of logically orders test takers after they are ordered by the score (Metsämuuronen, 2021b). If we assume, as we usually do, that *RPC* do not overestimate the correlation, we conclude that *Rit* underestimates correlation because its values are notably lower than those by *RPC*.

Second, as expected, *Rit* underestimates correlation, specifically, with easy and demanding difficulty levels. In the example of a normal dataset (n = 4,023), the underestimation is 0.349 units of correlation at the highest, that is, even if *Rit* tells that the discrimination power is roughly *Rit* ≈ 0.4 with the extreme items, this is a notably too low value; the proper values would be around $G = D = RPC \approx 0.7$ or higher. In the extreme dataset, the differences are even

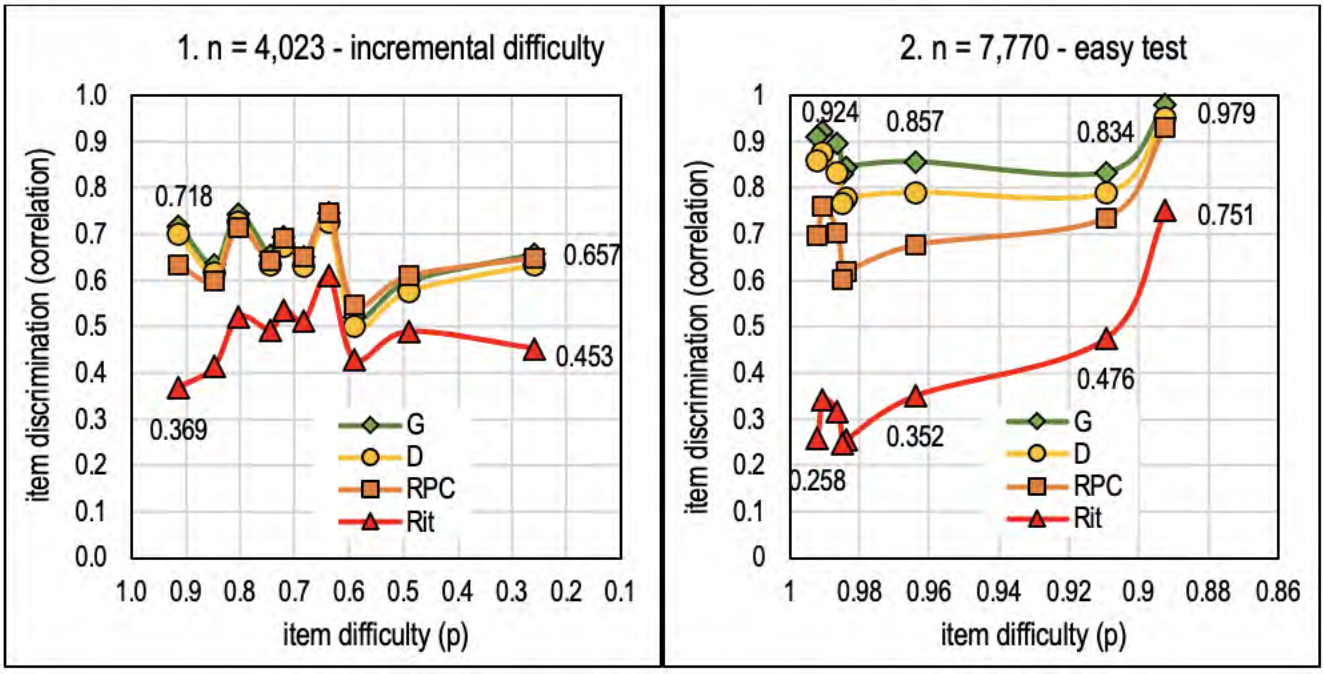
[2] Based on simulations (see Metsämuuronen, 2022a), some other good options for *Rit* and *Rit* are r-bireg and r-polyreg correlation (R_{REG}; Livinstone & Dorans, 2004, Moses, 2017), dimension-corrected *G* (G_2; Metsämuuronen, 2021a), dimension-corrected *D* (D_2; Metsämuuronen, 2020b, 2021a), attenuation-corrected *Rit* (R_{AC}; Metsämuuronen, 2022a), and attenuation-corrected eta (E_{AC}; Metsämuuronen, 2022a).

wider: up to 0.65 units of correlation. In the case, the outcome does not change though; in both extremes the items are suitable for a test. However, in some cases, *Rit* or *Rir* may claim that the item discrimination is too low to be accepted but an alternative estimator may indicate opposite. This kind of illustration may give more evidence in selecting items to the test: which items are poor because *Rit* or *Rir* cannot detect the real discrimination power and which items, really, are those to be omitted.

Table 1. Estimates of item discrimination for Figure 5

FINEEC (2018); n = 4,023										
	g1	g2	g4	g6	g9	g13	g20	g25	g28	g30
p	0.913	0.848	0.804	0.745	0.719	0.683	0.638	0.590	0.490	0.259
Rit	0.369	0.413	0.522	0.493	0.537	0.514	0.610	0.429	0.490	0.453
RPC	0.634	0.599	0.715	0.641	0.691	0.653	0.746	0.548	0.610	0.650
G	0.718	0.634	0.744	0.655	0.695	0.650	0.746	0.521	0.598	0.657
D	0.699	0.616	0.725	0.635	0.675	0.631	0.726	0.502	0.577	0.635
Metsämuuronen & Ukkola (2019); Metsämuuronen (2022b, 2022c); n = 7,770										
	g1	g2	g3	g4	g5	g6	g7	g8		
p	0.964	0.984	0.992	0.909	0.892	0.985	0.986	0.990		
Rit	0.352	0.257	0.258	0.476	0.751	0.246	0.315	0.342		
RPC	0.677	0.618	0.696	0.736	0.931	0.602	0.702	0.760		
G	0.857	0.846	0.911	0.834	0.979	0.831	0.897	0.924		
D	0.791	0.779	0.858	0.789	0.952	0.766	0.832	0.877		

Figure 5. Obvious underestimation when using Rit (or Rir) as an index of item discrimination power



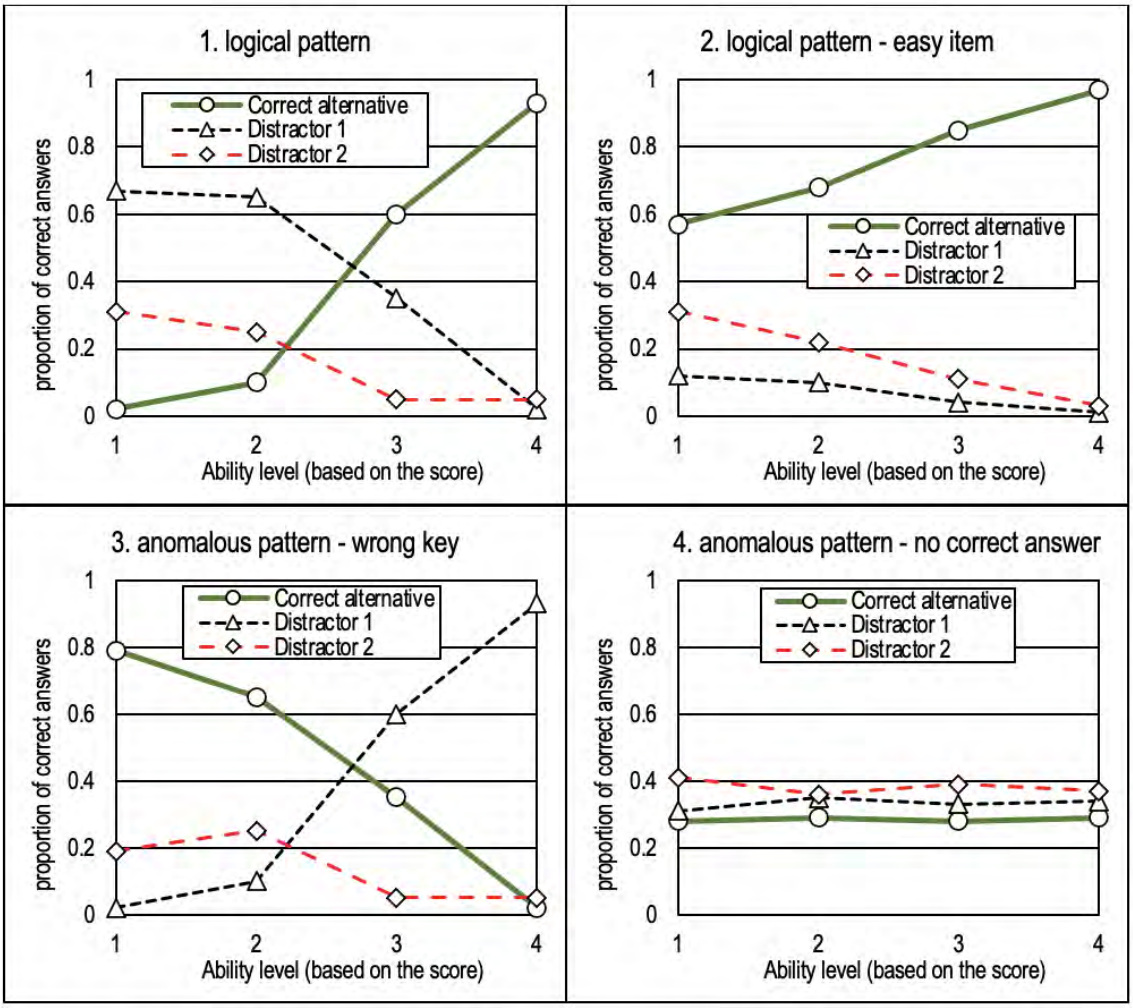
Graphical analysis of distractors

Another kind of graphical tool for the classical item analysis relates with the MCQs with distractors. We may think the distractor analysis as a supplementary analysis for the case that we find something problematic in the item behavior such as a low item discrimination. Distractor analysis may reveal which of the distractors would be worth of amending. Figure 6 illustrates some basic patterns of both the logical and anomalous behavior of the distractors.

In each graph the test takers have been divided into four ability groups (quartiles) and we calculate the proportion of test takers in each alternative in each point of quartile, that is, we reach 100% of the test takers in each quartile. Figures 6.1 and 6.2 show the basic pattern of the distractor in a logically behaving MCQ items: the probability for selecting the correct alternative increases by the ability level. If the item is easy, even the lowest-levelled test takers would know, more probably, the correct answer and, parallel, if the item is difficult, even the best ones would not give, necessarily, the correct answers but, in both cases, the higher-scoring test takers tend to select the correct alternative more probable than the lower-scoring test takers. Two types of illogical patterns are seen in Figures 6.3 and 6.4 illustrating the patterns of negative item discrimination and no discrimination (cf. the illogical patterns in Figures 3.2 and 3.3).

To condense, typical for these anomalous or illogical cases is that the *highest-scoring test takers do not find the correct answer*. If they tend to select totally wrong alternative, it may be wise to consider checking the key. If the selection of the alternatives is random—roughly the same in all ability groups—it may be wise to check whether there is any correct answer at all or whether the item is, in some other way, ambiguous or unclear.

Figure 6. Basic patterns of the logical and illogical behavior of distractors



Cut-off curve and its possibilities in the item diagnosis

Metsämuuronen (2020b; see also 2017) have introduced a new kind of possibility to visualize the item behaviour called cut-off curve (COC). Because COC is, most probably, less known as a concept for the reader, an elementary introduction to the graphs is given here based on Metsämuuronen (2020b). The formulae concerning the approach are not handled except to the extent to make understandable how the values are obtained. If the sample size is large or largish (here, n = 200), the graph may look like in Figure 7. The interpretation and possibilities of COCs in item analysis are discussed in what follows.

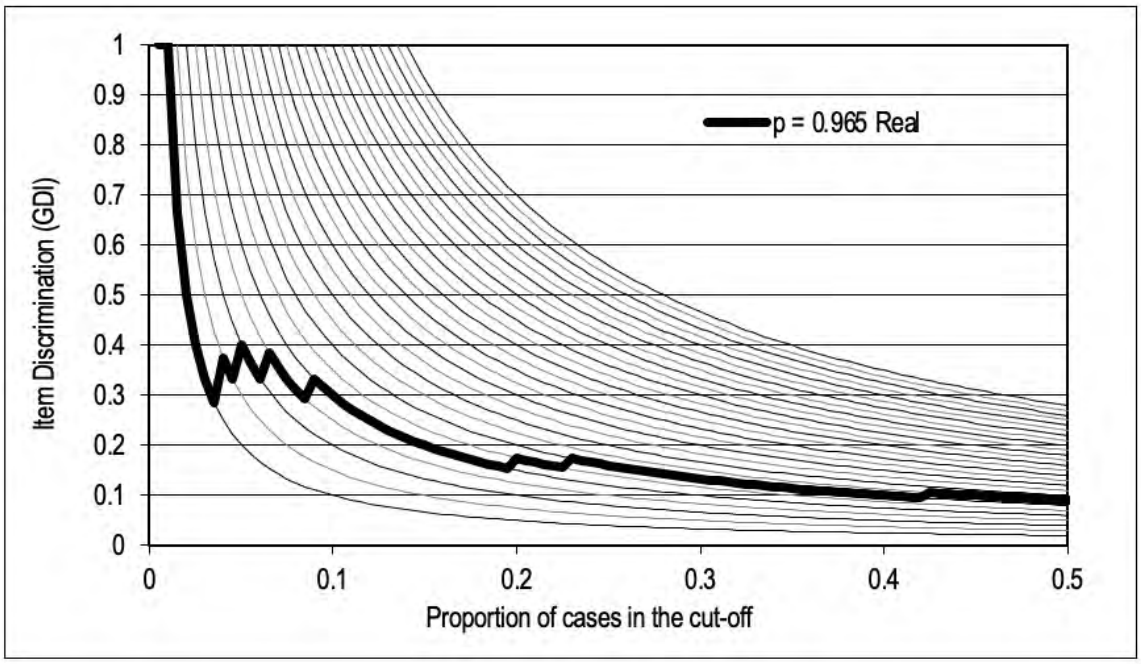
Originally, the concept of "cut-offs" is connected to Kelley's *DI* with which the discussion of different cut-offs is essential. Namely, unlike most of the estimators of item discrimination, *DI* do not use all the cases in the estimation. Instead, only the extreme cases of the ordered dataset are used, and the question is, how many of the extreme cases should be used: either 25% of the highest and lowest scoring respondents of the ordered dataset (e.g. D'Agostino & Cureton, 1975; Mehrens & Lehmann, 1991; Metsämuuronen, 2017) or 27% (e.g. Kelley, 1939; Ross and Weitzman, 1964; Ebel, 1967, Wiersma & Jurs, 1990), or maybe all of the cut-offs as proposed by Metsämuuronen (2020b) for the generalized *DI* (GDI).

Although COCs in Metsämuuronen (2020b) are connected to *DI* or, factually, in *GDI*, the graph type is not restricted to *DI* or *GDI*. Metsämuuronen (2017, p. 209) uses the same idea when comparing the efficiency of four indices of item discrimination power to detect the theoretical Guttman pattern. This theoretical pattern is discussed in Section "Guttman pattern…". The following sections discusses the thinking and practicalities behind the moves in the curves not only in the theoretical Guttman-patterned items but also some other theoretical and real-life patterns. In Section "Possibilities of COC…", some examples are given of using COCs with the real-world datasets. In what follows, the mathematical part of the *GDI* and COCs are kept minimal; an interested reader finds more information of that in Metsämuuronen (2020b; 2020c).

Guttman pattern as an extreme case of item behavior

Zimmerman, Williams, Zumbo, and Ross (2005) highlighted the late Louis Guttman as one of the most neglected theorists within the classical test theory and classical item analysis. Guttman made contributions to reliability theory, factor analysis and scaling theory among other relevant areas of interest. One of

Figure 7. Cut-off curve of a real-life item (n = 200)



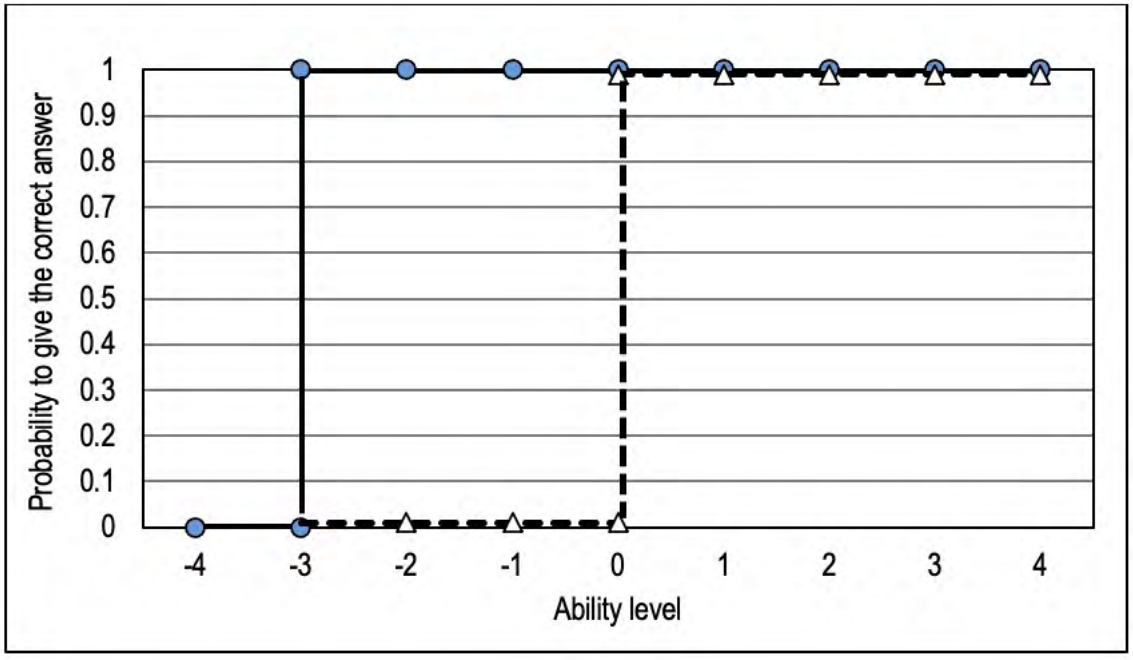
Guttman's innovations was the idea of a scalogram (Guttman, 1950) based on a data structure that follows the so-called Guttman-scalable pattern. Guttman scalability (Linacre, 2000) or Guttman ordering (Linacre, 1992, Roskam & Jansen, 1992) refers to a specific data array where items and respondents arranged in the increasing order by the score and the item difficulty shows a triangular pattern of dichotomous items. Traditionally, the dataset is called Guttman-patterned if all the items are patterned with a string of 0s followed by a string of 1s when the respondents are placed in ascending order by test score even if the data array is not triangular (e.g., Linacre & Wright, 1994). Here, a single item with the structure described above is called Guttman-patterned (GP) even though it is handled without a specific connection to a strictly Guttman-scalable dataset.

The extreme nature of GP items comes from the fact that they can discriminate the higher- and lower-scoring test takers from each other in a deterministic manner. Hence, the pattern is called also a deterministic pattern (Linacre and Wright, 1994). Therefore, there is a fundamental difference between the Rasch models and Guttman models although the Guttman pattern has seen as the ultimate latent form of an item in Rasch modeling (Andrich, 1985; Linacre & Wright, 1996; Linacre, 2000; Linacre, Andrich & Luo, 2003; van Schuur, 2003): the Rasch models are stochastic while the Guttman model is deterministic (Curtis, 2004).

The deterministic nature of the Guttman-patterned items is illustrated in Figure 8 where items discriminate perfectly between those who gave the correct answer in an achievement test, for example, from those who gave the incorrect answer. It may be worth noting the connection between the illustrations in Figures 1.3 and 8: if the item discrimination in the two-parameter IRT model would have been an ultimately perfect one, the ICC would look like the ones in Figure 8. However, IRT models cannot handle this kind of deterministic situation. It may be fair to say that the Guttman pattern is the important underlying theoretical constructs for an ultimately discriminating item (see further discussion in Metsämuuronen, 2020c).

Unluckily, the Guttman pattern is essentially a theoretical situation as noted by Fisher (1992) and Linacre (1992). Usually, in the practical settings, the pattern of zeros and ones include random errors: some lower-scoring respondents give unexpected correct answers or some of the higher-scoring respondents fail to give correct answers. For this reason, whenever this theoretical pattern of an item is found in a real-world data, the phenomenon is not stable: when another sample is taken, the pattern may or may not reoccur.

Figure 8. Hypothetical Guttman-patterned items with the deterministic item discrimination



Kelley's DI, Exhaustive splitting, and Cut-off Curve

Of many classical indices to item discrimination (see lists in Cureton, 1966a, 1966b; ETS, 1960; Henrysson, 1963; Metsämuuronen 2020a, 2021a, 2022a; Moses, 2017; Oosterhof, 1976; Wolf, 1967), Kelley's *DI* shows interesting potential when willing to illustrate the item behavior. Unlike Pearson correlation, *DI can* detect the deterministic pattern and, hence, it can reach the perfect value *DI* = 1, and, in the binary case, it is more stable as an indicator of item discrimination power than Pearson correlation (see discussion and examples in Metsämuuronen, 2020b). Although COCs are not restricted to binary case, here, only the binary case is discussed to simplify the notation. Hence, the simplified, generalized form of *DI*, allowing all cut-offs, can be expressed as

$$GDI_a = \frac{R_a^U - R_a^L}{\frac{1}{2}T_a} \quad (4)$$

(Metsämuuronen, 2020b), where a refers to the number (or proportion or percent) of extreme cases in the half of the cut-off of the ordered dataset, R_a^U and R_a^L refer to the number of correct answers in the upper (U) and lower (L) halves of the cut-off a, and T_a refers to total number of cases in the halves together. Then, for example, if four extreme cases are selected from both extremes, in the binary case, $a = 4$ and $T = 8$. Originally, the cut-off a in Eq. (4) refers to the 25% or 27% of the extreme respondents of the ordered data as discussed above. Following the notation in (4), these cut-offs are special cases of $GDI_{25\%}$ and $GDI_{27\%}$. Equation (4) embeds the idea that there can be *different cut-offs* for estimating the item discrimination. In what follows with the procedure of exhaustive splitting (PES), *all* the possible cut-offs are in use.

Let us use a difficult, hypothetical GP item with 24 respondents and four correct answers as an example of the concept of cut-offs. Ordered from the lowest to the highest test taker based on the (unseen) total score, the string is as follows: 000000000000000000001111. The item can be divided into symmetric cut-offs of extreme cases in many ways. Ultimately, only the most extreme test takers are considered, and we get the string 0 | 1 where the bar just shows the middle point dividing the test takers in to the lower and upper part. Similarly, if four extreme cases are considered, the extreme string of 1s and 0s is 0000 | 1111. In the case, the nearest cut-off to the traditional 25% or 27% is the one with 6 test takers from both extremes: 000000 | 001111. When all the cases are used, 50% of the cases are in each half: 000000000000 | 000000001111.

For the latter parts of the article, we use all the possible symmetric cut-offs. The PES routine is as follows (Metsämuuronen, 2020b): 1) Take the ultimately highest and lowest case from the sorted data and calculate the indices of interest and save the result; 2) Take the *two* highest and the *two* lowest cases from the sorted data and calculate the value for item discrimination (as in 1) and save the results; 3) Repeat phase 2 by increasing the number of cases, gradually building up to *½N = 50%* of the cases at both extremes; and 4) Make a table or graph of the results. This procedure is not bound to any specific statistic. However, the original Kelley's routine is employed here for item discrimination because it appears to be interesting from the item analysis viewpoint.

Let us employ PES with Eq. (4) for the previous item of interest. The statistics for the calculation are collected in Table 2. Notably, for the first four extreme cut-offs, $GDI_{4\%}$ to $GDI_{17\%}$, the item discrimination is perfect indicating that, in these splits, the item can discriminate perfectly between the lowest and the highest respondent. From the fifth cut-off onward, the estimate for item discrimination drops dramatically even though the item remains deterministically discriminating. Finally, when using the whole dataset, $GDI_{50\%} = (4–0)/12 = 0.33$.

Let us prepare a graph of *GDI* as the function of the percentage of respondents in the lower or upper half of the ordered data, an obvious though interesting practical phenomenon is found: COC detects the latent item difficulty level ($\pi_s = p = 0.17$) exactly at the threshold point of the curve in the cut of $a = s$ (Figure 9). More, at the threshold point the item discrimination is perfect (*GDI* = 1) as should be because of the deterministic nature of the item.

It is worth noting the symmetricity embedded in the approach of PES and COC when it comes to GP items with extreme difficulty levels. Namely, the COCs of the patterns 00000000000 | 00000001111 with the proportion of correct answers $p = 4/24 = 0.167$ and 00001111111 | 11111111111 with $p = 20/24 = 0.833$

Table 2. All symmetric cut-offs and values for the item discrimination of an item with the pattern 000000000000000000001111

$\frac{1}{2}T_a$	a (%)	R_a^L		R_a^U	$GDI_a = \frac{R_a^U - R_a^L}{\frac{1}{2}T_a}$
1	4% = 1/24	0	0 1	1	1.00
2	8%	0	0 0 1 1	2	1.00
3	13%	0	0 0 0 1 1 1	3	1.00
4	17%	0	0 0 0 0 1 1 1 1	4	1.00
5	21%	0	0 0 0 0 0 0 1 1 1 1	4	0.80
6	25%	0	0 0 0 0 0 0 0 0 1 1 1 1	4	0.67
7	29%	0	0 0 0 0 0 0 0 0 0 1 1 1 1	4	0.57
8	33%	0	0 0 0 0 0 0 0 0 0 0 1 1 1 1	4	0.50
9	38%	0	0 0 0 0 0 0 0 0 0 0 0 1 1 1 1	4	0.44
10	42%	0	0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1	4	0.40
11	46%	0	0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1	4	0.36
12	50% = 12/24	0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1	4	0.33

would be identical. The essential connection between these patterns is the identical number of the extreme observations producing the perfect item discrimination. Hence, in what follows, the subscript *s* refers to the *shorter of the extreme strings of 0s and 1s*; in the first case, the sorter of the strings is 1111 and, in the latter case, the sorter of the strings is 0000—in both cases 4/24 = 0.167. Hence, when the item is difficult ($p < 0.50$), the latent item difficulty is seen strictly in the graph ($\pi_s = p = 0.167$). When $p > 0.5$, we need a simple transformation of the scale to express the π_s in the p-metric: $p = 1 - \pi_s = 1 - 0.167 = 0.833$.

It may be good to note another Guttman pattern that makes the negative values of item discrimination understandable (see the patterns in Figures 3.3 and 6.3 above). Let's assume a perfectly opposite GP item than the previous example. Out of 24 respondents, the four *lowest* ones give the correct answer and the rest give an incorrect answer. This ultimately anomalous string is as follows: (111100000000000000000000). When performing the PES for the item, *GDI* gives negative values symmetrically in comparison with Figure 9. However, in both cases, the threshold point of the curve is the same (see Figure 10).

Non-Guttman-patterned items and COC

Let us produce some stochastic error to the Guttman pattern. In what follows, in general, this type of real-life item is called a non-Guttman-patterned (NGP) item and two types of NGP items are specified: a common non-Guttman-patterned item with a GP string in the extreme values called later NGP and a specific type of NGP without the GP string called later pseudo-Guttman patterned (PGP) item; the latter is discussed in the next section.

An example of NGP item with minor stochastic error is as follows: (000000000000000000101111). Keeping in mind that the minimum string of Guttman type is of form (0 | 1), that is, the highest in the rank obtains the correct answer and the lowest case gives an incorrect answer, it is worth noting that the reduced data indicated by the string *s* (0000 | 1111) follows the Guttman pattern. Using the PES, the values for *GDI* are as in Table 3 and the corresponding COC is in Figure 11. The underlying theoretical curves of GP items of $p = 0.17$ and $p = 0.21$ are shown in the graph as dashed lines.

It is worth noting that the observed COC (bold line) follows exactly the COCs of the underlying theoretical GP items. The classical item difficulty of the NGP item in the example is $p = 5/24 = 0.21$. What is notable in Figure 11 is that, in comparison with Figure 9 with a strict GP item, the estimator p for item difficulty seems to be biased: the latent item difficulty π seems to be somewhere between $\pi_s = 0.17$ and $\pi_{s+1} = 0.21$ rather than uniquely $p = 0.21$. However, the exact value is not known. This leads to an interesting conclusion: *the proportion of correct answers* p *seems to be*

biased as an estimator for item difficulty when the pattern deviates from the Guttman-pattern (see further discussion in Metsämuuronen, 2020c). This matter is discussed in Section "Possibilities of COC…".

As an example of a nontrivial NGP item, let's assume an item of 24 respondents ordered by the test score from the lowest to the highest with the following structure:

(000100101011 | 111111001111).

Statistics for constructing a COC are calculated (Table 4) and COC is shown in Figure 12. The light curves in the graph are the COCs of the selected underlying theoretical GP items of $p = 3/24$, $p = 4/24$, and $p = 5/24$.

In Figure 12, the first threshold point of the COC is denoted by $\pi_s = 3/24 = 0.125$ (or, factually, $21/24 = 0.875$ because $p > 0.5$). Because of these first Guttman-patterned cut-offs, the COC in Figure 12 starts the same way as in Figures 9 and 11 with $DI = 1$ for three first cut-offs (cf. later Figure 13 of PGP items). Again, it is worth noting that the observed COC of NGP follows exactly the underlying GP items. Later, COC in Figure 12 is further discussed.

Figure 9. Cut-off curve of a hypothetical GP item

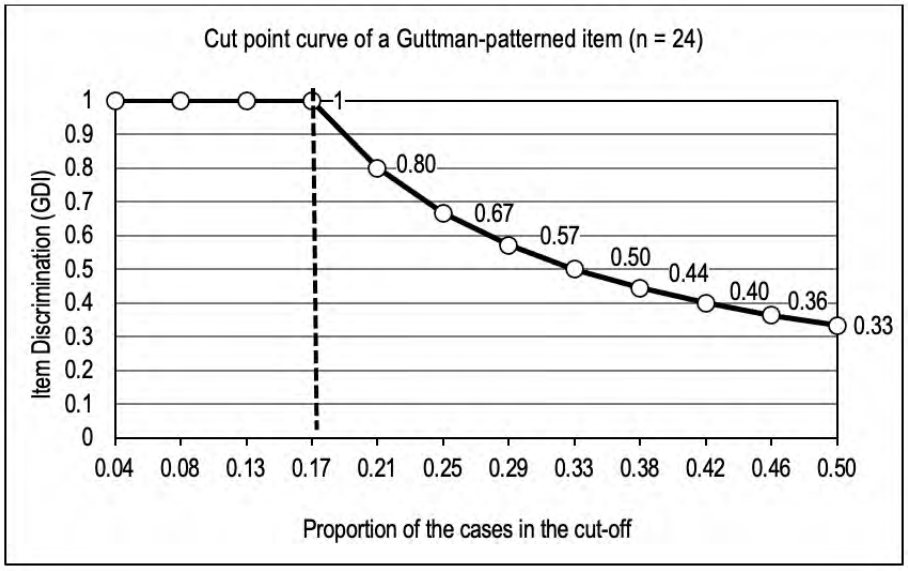


Figure 10. Cut-off curve for positive and negative Guttman pattern

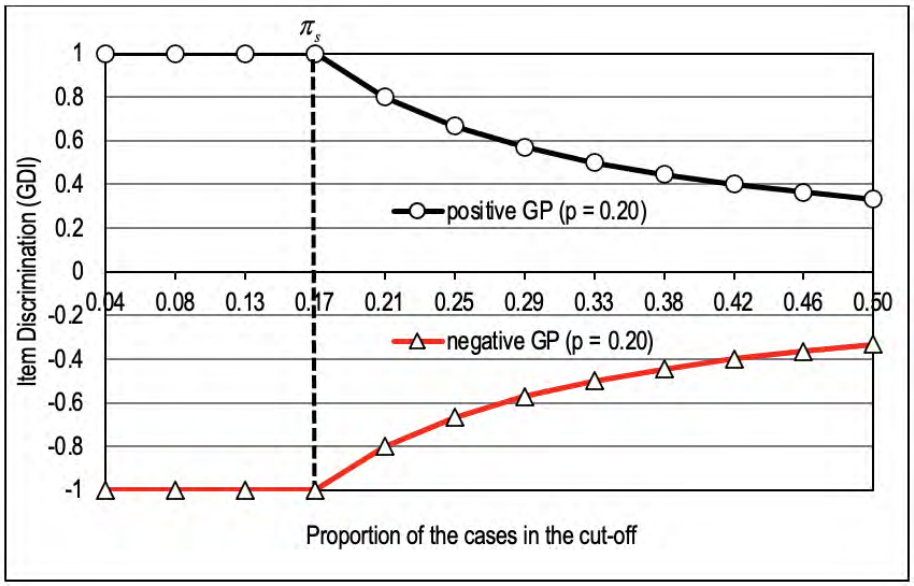
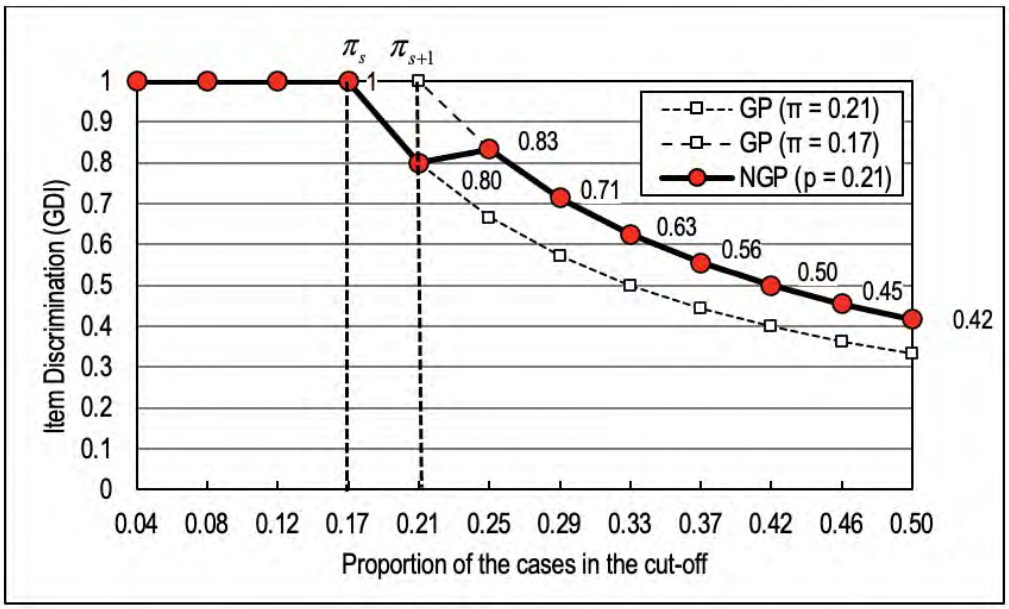


Table 3. All symmetric cut-offs and values for the item discrimination of an NGP item with the pattern 000000000000000000101111

$\frac{1}{2}T_a$	R_a^L		R_a^U	$GDI_a = \frac{R_a^U - R_a^L}{\frac{1}{2}T_a}$
1	0	0 ... 1	1	1.00
2	0	0 0 ... 1 1	2	1.00
3	0	0 0 0 ... 1 1 1	3	1.00
4	0	0 0 0 0 ... 1 1 1 1	4	1.00
5	0	0 0 0 0 0 ... 0 1 1 1 1	4	0.80
6	0	0 0 0 0 0 0 ... 1 0 1 1 1 1	5	0.83
7	0	0 0 0 0 0 0 0 ... 0 1 0 1 1 1 1	5	0.71
8	0	0 0 0 0 0 0 0 0 ... 0 0 1 0 1 1 1 1	5	0.63
9	0	0 0 0 0 0 0 0 0 0 ... 0 0 0 1 0 1 1 1 1	5	0.56
10	0	0 0 0 0 0 0 0 0 0 0 ... 0 0 0 0 1 0 1 1 1 1	5	0.50
11	0	0 0 0 0 0 0 0 0 0 0 0 ... 0 0 0 0 0 1 0 1 1 1 1	5	0.45
12	0	0 0 0 0 0 0 0 0 0 0 0 0 ... 0 0 0 0 0 1 0 1 1 1 1	5	0.42

Figure 11. Cut-off curve of a NGP item with small stochastic error



Pseudo-Guttman-patterned item

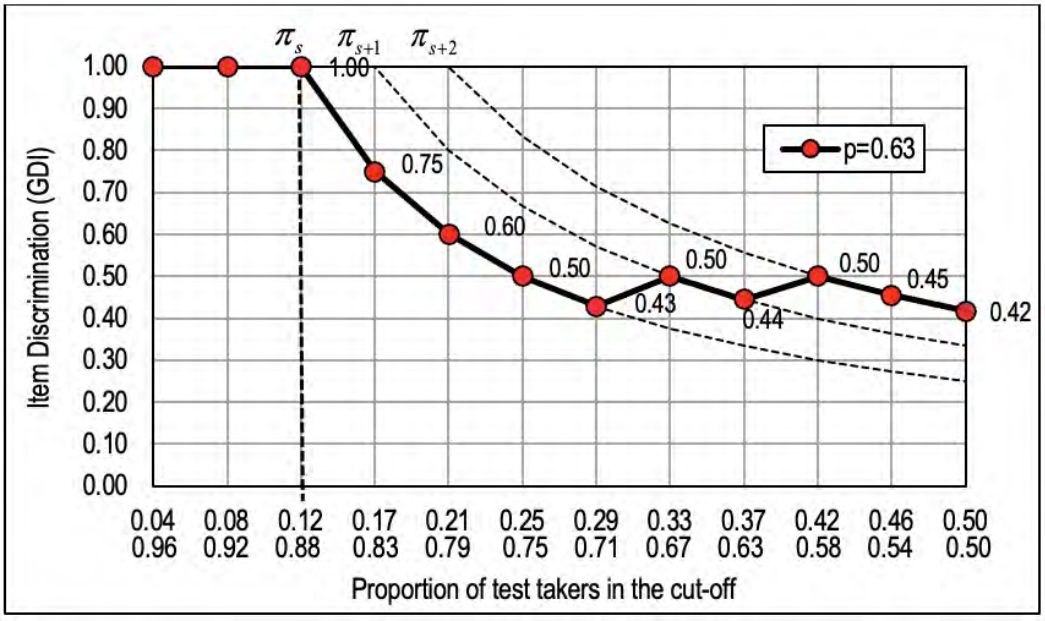
For the sake of completeness, also the PGP is discussed. With the real-world datasets, it may happen that the ultimately extreme respondents give the same value and thus the first pattern is not of Guttman type. One such extreme case is the pattern (0 | 0) when both the lowest and highest respondents give an incorrect answer—this may occur more likely with difficult items. Another extreme case is the pattern (1 | 1) when both the lowest and highest respondents give a correct answer—this may occur more likely with easy items. In comparison with the previous cases, there is no specific string s which would produce the perfect item discrimination though the pattern may otherwise be GP like. Hence, the name 'pseudo-Guttman-patterned'.

Let's assume two PGP items with ordered sequences comparable to Figures 9, 11, and 12 as

Table 4. All symmetric cut-offs and values for the item discrimination of an NGP item with the pattern 000100101011111111001111

$\frac{1}{2}T_a$	R_a^L		R_a^U	$GDI_a = \frac{R_a^U - R_a^L}{\frac{1}{2}T_a}$
1	0	0 ... 1	1	1.00
2	0	0 0 ... 1 1	2	1.00
3	0	0 0 0 ... 1 1 1	3	1.00
4	1	0 0 0 1 ... 1 1 1 1	4	0.75
5	1	0 0 0 1 0 ... 0 1 1 1 1	4	0.60
6	1	0 0 0 1 0 0 ... 0 0 1 1 1 1	4	0.50
7	2	0 0 0 1 0 0 1 ... 1 0 0 1 1 1 1	5	0.43
8	2	0 0 0 1 0 0 1 0 ... 1 1 0 0 1 1 1 1	6	0.50
9	3	0 0 0 1 0 0 1 0 1 ... 1 1 1 0 0 1 1 1 1	7	0.44
10	3	0 0 0 1 0 0 1 0 1 0 ... 1 1 1 1 0 0 1 1 1 1	8	0.50
11	4	0 0 0 1 0 0 1 0 1 0 1 ... 1 1 1 1 1 0 0 1 1 1 1	9	0.45
12	5	0 0 0 1 0 0 1 0 1 0 1 1 1 1 1 1 1 1 0 0 1 1 1 1	10	0.42

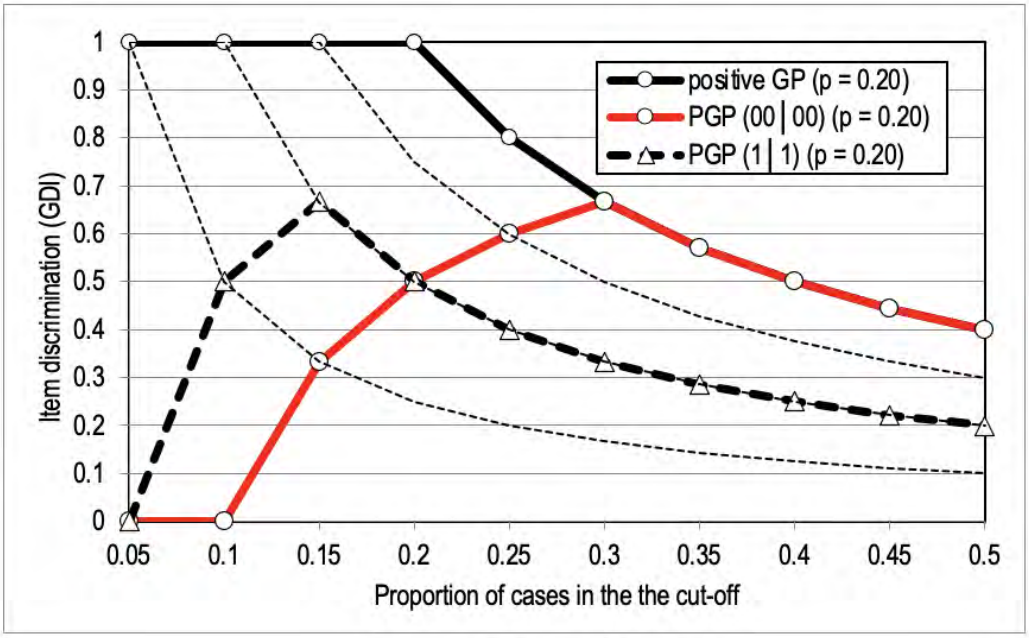
Figure 12. Cut-off curve for a non-trivial NGP item of N=24



follows: (1000000000 | 0000000111) and (0000000000 | 0000111100). The first pattern actualizes when the lowest-ranked case either guesses or unexpectedly knows the correct answer. The latter pattern indicates that two highest-ranked cases were unexpectedly either sleepy, careless, or ignorant. In both examples, p = 0.20. Both the patterns lead to a situation that the item discrimination in the ultimate cut-off(s) of extreme cases equals DI = 0 (Figure 13). The underlying theoretical curves of GP items of p = 1/20 = 0.05, p =2/20 = 0.10, p = 3/20 = 0.15, and p = 4/20 = 0.20 are shown in the graph as lighter lines.

Typical to both the PGP and NGP items is that, first, in the cut-offs *after* the pattern breaking string, COCs follow some of the underlying Guttman patterns and, second, there is not a unique threshold point and thus, there is no unambiguous latent item difficulty as is with GP items.

Figure 13. Cut-off curves for PGP items of n=20



A note of the moves in the COC

In the dichotomous dataset, at every cut-off following a previous one (a + 1), COC can have only one of three possible options because the value of GDI_{a+1} can have only one of the three fixed options:

$$GDI_{a+1} = \left(R_a^U - R_a^L + 1\right)/(a+1), \quad GDI_{a+1} = \left(R_a^U - R_a^L + 0\right)/(a+1),$$

or

$$GDI_{a+1} = \left(R_a^U - R_a^L - 1\right)/(a+1)$$

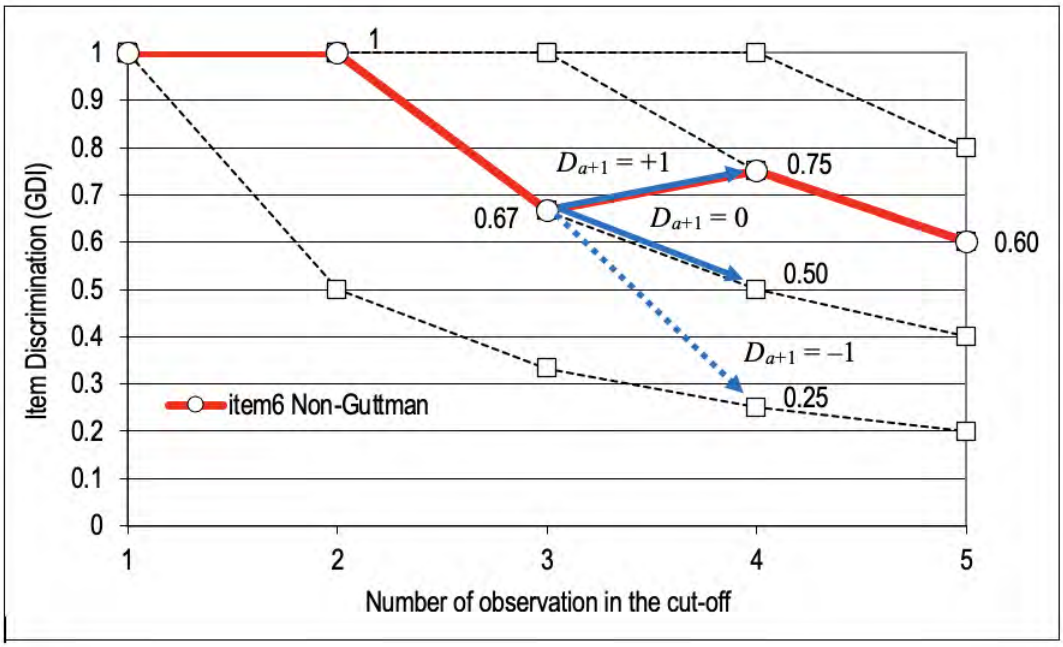
(see in-depth in Metsämuuronen, 2020b). The value +1, 0, or –1 is caused by the difference (D) between the (a + 1)th test taker in the upper and lower half of the cut-off. The value $D_{a+1} = (1 - 0) = +1$ refers to the case that the new test takers in the upper half gives a correct answer but the corresponding new test taker in the lower half gives an incorrect answer and this results the value $D_{a+1} = +1$. Parallel, the value will be resulted $D_{a+1} = 0$ when both test takers give either a correct answer (1 – 1) or an incorrect answer (0 – 0). The value will be $D_{a+1} = -1$, when the (higher achieving) test taker in the upper half gives an incorrect answer but the (lower achieving) test taker in the lower half gives the correct answer (0 – 1).

For the illustrative purposes of the moves in the COCs, a simple item is considered. Let us assume a real-life NGP item with ten observations with the ordered string 0010001111. The consequences of the results obtained from D_{a+1} are illustrated in Figure 14. If the result had been $D_{a+1} = +1$, the next step in the COC would have been on the *next* underlying curve of a Guttman-patterned item, that is, the path would have moved forward to the next underlying curve. If the result had been $D_{a+1} = 0$, the next step will be on the *same* underlying GP curve as the previous one (no change in the path). If the result had been $D_{i+1} = -1$, the path would have led to one step lower in the *previous* underlying GP curve. Hence, by using the approach of COC we can detect the illogical patterns in the dataset easily. In Section "Possibilities of COC…" in Figure 15, we detect some of these kinds of moves in the COCs of the real-life items.

Possibilities of COC with real-life datasets

The examples above were based on rather small datasets; it is easy to illustrate the graphs when the number of cases is small. However, the PES and COCs are not restricted to a specific sample size. As an example of a larger dataset, a real-world random sample of 200 test-takers is used as a basis for the illustration (see Metsämuuronen, 2020b). COCs can be

Figure 14. An example showing options for the next step in the COC



used in detecting illogical and anomalous patterns, stability of the estimates of item discrimination, latent item difficulty, and possible anomalous guessing.

Detecting illogical or anomalous patterns in the dataset by using COC. From the viewpoint of PES, the anomalous patterns are the ones where the COC moves to the previous Guttman-patterned latent curve (see Figure 15). Another non-obvious note of Figure 15 is that, when the proportion of the test takers in the split gets higher and close to 50%, the COCs tends to follow the same Guttman-patterned underlying curve. This is caused by the fact that the probability to find test takers with approximately the same ability level is high at the middle of the ordered dataset; the matter is formalized in Metsämuuronen (2020b).

Detecting plausibility and stability of the value for GDI. We may reasonably ask, how stable and plausible is our estimation of item discrimination power if it is based on only one point estimate? If we would obtain estimates in cut-offs close to each other such as $GDI_{23\%} = 0.55$, $GDI_{25\%} = 0.50$, and, $GDI_{27\%} = 0.47$, which of those would be the most credible estimate and why? COC (and related tabled values) could be used in assessing the variance for the point estimate and that could be used in calculating a confidence interval of the estimate.

Figure 15 illustrates the stability of the estimate of $a = 27\%$. In the example, with the difficult item ($p = 0.225$), the estimate for IDP is $GDI_{27\%} = 0.35$. Just by using the graphical possibilities and intuitional heuristics, we may conclude that the value seems quite stable between the cut-offs 10% to 30% ranging from 0.32 to 0.40. By using the values in COC, specifically, if those are tabled, it could be easy to compute the average value and variance of the estimate and the confidence intervals (see techniques in Metsämuuronen, 2021c). The other item in Figure 16, a very easy one ($p = 0.965$), is less discriminative ($GDI_{27\%} = 0.15$) and, more crucially, the value ranges from 0.13 to 0.30 between the cut-offs 10% to 30% showing two times wider range in comparison with the difficult item ($p = 0.225$).

Detecting latent item difficulty and item discrimination simultaneously by COC. One of the advances of IRT modelling over the classical item analysis is the possibility to estimate item parameters simultaneously (see Section "Item characteristic curve" above). The approach of PES and COC gives possibility to simultaneously estimate two main item parameters, item difficulty and item discrimination within the classical test theory approach. The real-life item on Figure 12 is reinterpreted and reanalyzed here

Figure 15. Cut-off curves for two real life items with non-Guttman-pattern (n = 200)

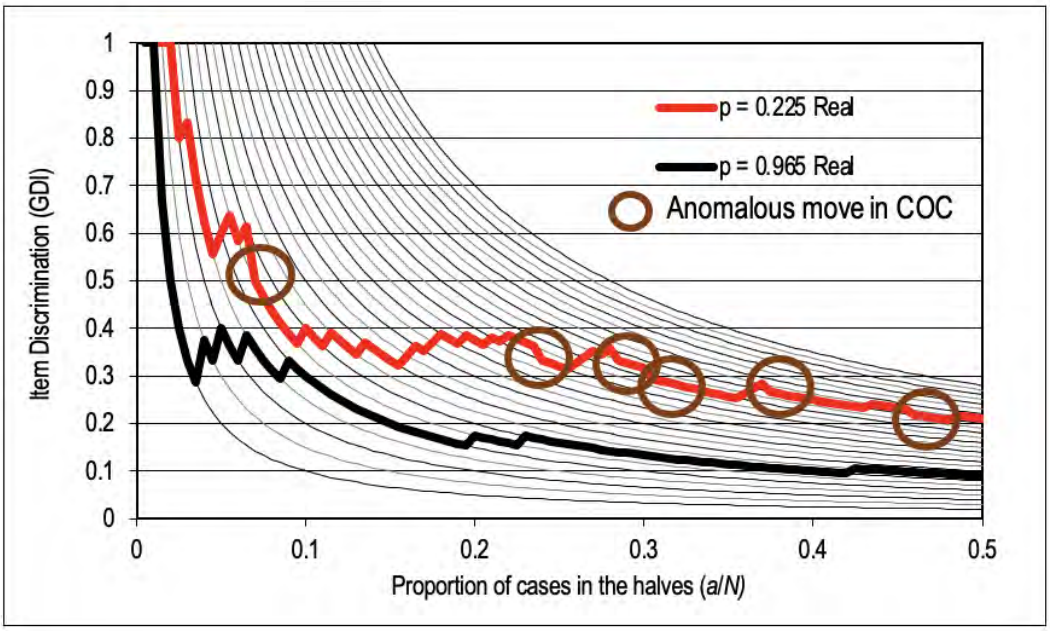
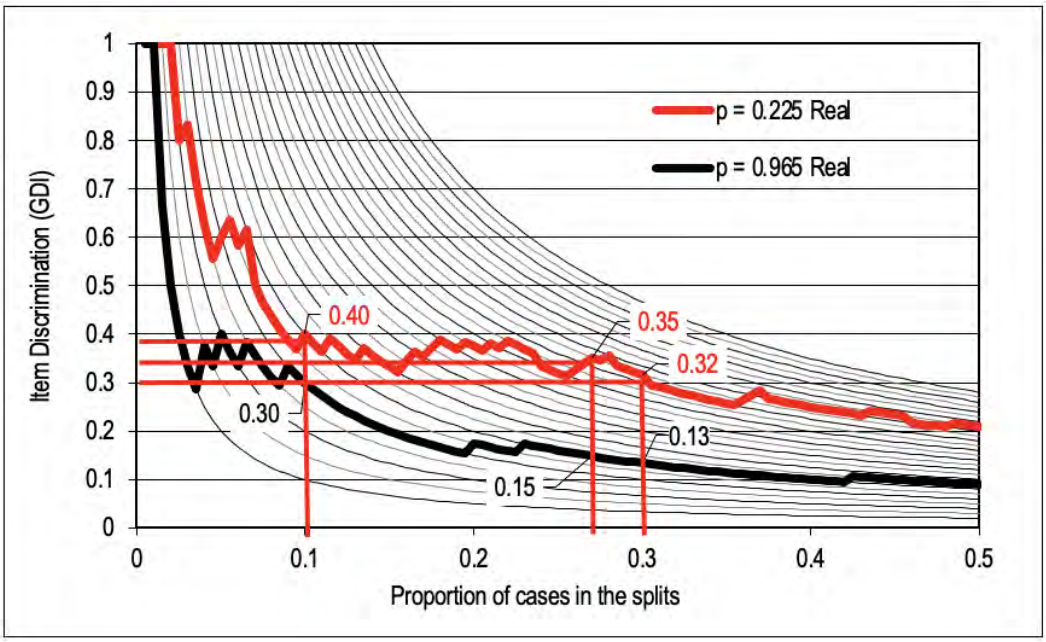


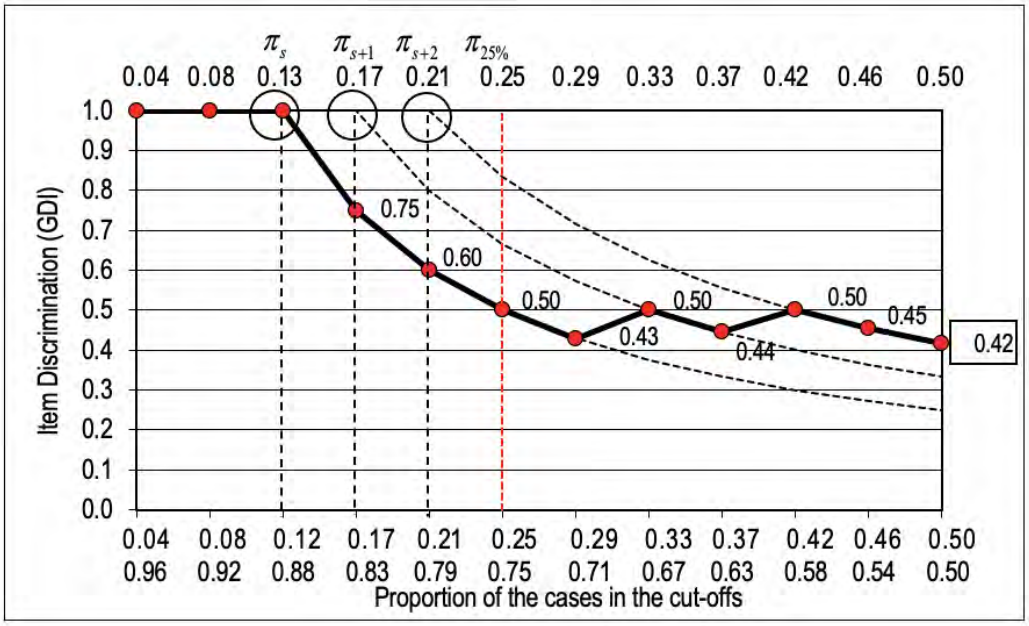
Figure 16. Stability of the estimate of two real life items with non-Guttman-pattern (n = 200)



from the item difficulty and item discrimination viewpoint keeping in mind the note related to Figure 11: The observed proportion of correct answers p seems to be biased as an estimator for item difficulty when the pattern deviates from the Guttman-pattern. The item of 24 respondents ordered by the test score from the lowest to the highest with the following structure is (000100101011 | 111111001111). Out of 24 test takers, 15 gave the correct answer and hence, the observed classical item difficulty is p = 15/24 = 0.625. The task is to find the most credible estimate for the latent item difficulty π.

From the visual point of view, in Figure 17, there seems to be *three* options for the latent difficulty level: $\pi_s = 1 - 0.12 = 0.88$, $\pi_{s+1} = 1 - 0.17 = 0.83$, and $\pi_{s+2} = 1 - 0.21 = 0.79$ in p metrics. A reasonable

Figure 17. COC, item difficulty, and item discrimination power for an NGP item of n=24



approximation of the latent difficulty level of the item can be obtained by weighting the options by the frequency (see options in Metsämuuronen, 2020c). One relevant estimate for the latent difficulty level based on the cut-offs after the specific cut-off s could be

$$\hat{\pi} = (5\times\pi_s + 2\times\pi_{s+1} + 3\times\pi_{s+2})/10$$
$$= (5\times 0.88 + 2\times 0.83 + 3\times 0.79)/10 = 0.838.$$

Simulation by Metsämuuronen (2020c) with real-world datasets suggests, surprisingly, that somewhat better approximation from the bias viewpoint could be obtained if using *all* cut-offs in the estimation. In that case, the estimate for the latent difficulty would be

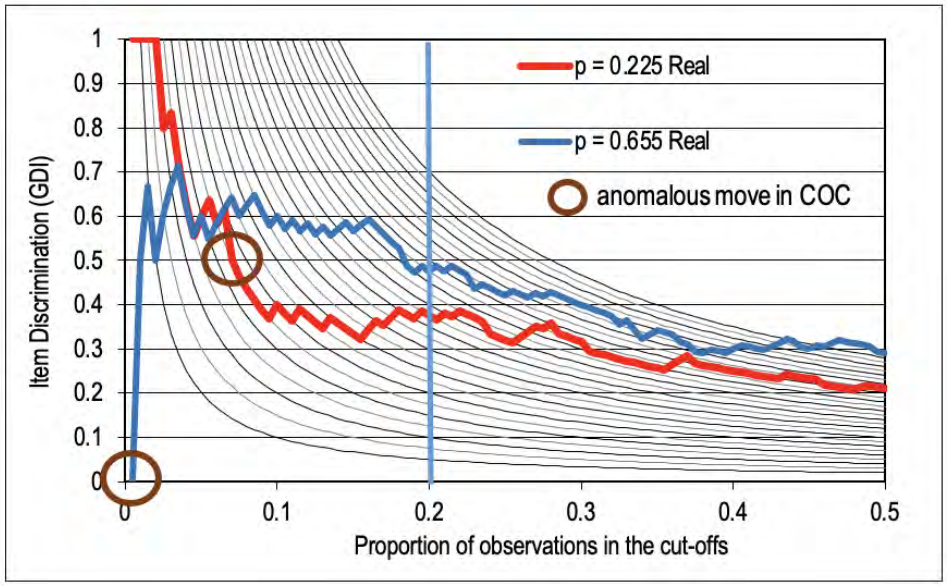
$$\hat{\pi} = (\pi_{s-2} + \pi_{s-1} + 5\times\pi_s + 2\times\pi_{s+1} + 3\times\pi_{s+2})/12$$
$$= (0.96 + 0.92 + 5\times 0.88 + 2\times 0.83 + 3\times 0.79)/12$$
$$= 0.858.$$

Difference between the estimates is not notable. However, it is noteworthy that the difference between the observed item difficulty (p = 0.63) and the latent one ($\hat{\pi}$ = 0.838–0.858) tells us how much the item deviates from the Guttman pattern.

When the item difficulty is estimated, the item discrimination power can be estimated straightforwardly by using the COC (or related tabled figures). The classical estimate for the item discrimination by using *DI* and the cut-off of the 25% cut-off (the cut-off of 27% cannot be used) gives us $GDI_{25\%}$ = 0.50. More credible options could be found when using the estimators of the (latent) item difficult above. This would lead to the same logic as is used in the Rasch and IRT modeling (see Figure 1.3). In the case, the exact cut-off for the estimate $\hat{\pi}$ = 0.838 is not found but a very close is found from the cutoff a = 0.167 leading to $\hat{\pi}$ = 0.833. At this cut-off, the item discrimination is $GDI_{17\%}$ = 0.75. Third option suggested by all cut-offs leads us to intrapolate the item discrimination in-between the cut-offs a = 0.17 and a = 0.13. From the graph, a good approximation could be around $GDI_{17\%}$ = 0.85. The estimations of the item discrimination vary between 0.50–0.85; in all cases, the item discriminates well the test takers from each other.

Detecting possible anomalous guessing by COC. The illogical pairs resulting $D_i = -1$ in the extreme cut-offs of difficult items indicates the possible anomalous cases with guessing patterns. The boundary for "extreme" is debatable; here, the cut-off a = 0.20 is used as an example in Figure 18. Notably, the illogical pairs in the middle range of ability or with easy item may be taken as random errors necessary for the statistical analysis.

Figure 18. Illogical or anomalous moves in COC possibly related to guessing



Let us use Figure 16 as an example of detecting possible pattern of high guessing behavior in an item. One suspicious pair is detected in a difficult item (p =0.225); this do not indicate notable anomalous guessing in general. Another pattern is illustrated by adding a new real-life item (p = 0.655) to the graph. This illustrates how the COC detects the extreme illogical guessing by the lowest-performing test taker at the lower end of the dataset that causes $D_i = 0$, that is, both the ultimately lowest- and highest performing test takers gave the correct answer. The latter leads to pattern of PGP discussed with Figure 13 above.

Concluding remarks

Within the text it has been seen that the visual diagnostic can be done several ways including Rasch modeling and IRT approach, classical test theory as well as by using the cut-off curves. The classical approaches have already established their value and usefulness in the visual item analysis. Adding the PES and COCs approach as a new tool the toolbox may be worth considering for the following reasons:

1. Traditionally, the simultaneous estimation of the item parameters has been possible only within the Rasch modeling. PES with *GDI* allows the simultaneous estimation of both latent item parameters, item difficulty and item discrimination, within the classical test theory.
2. The classical indicator for the item difficulty p seems to be a biased estimator for the latent ability specifically when the item deviates from the pure Guttman pattern. The PES approach gives possibilities to detect these impurities in the datasets. PES and the related statistics can be used in evaluating how much error there are in the observed data structure.
3. In some cases, the classical indicators for item discrimination may doom an item to be undiscriminating. Especially in the case of extremely difficult or easy items the classical indicators as well as IRT models may fail to reach the very essence of GP items' deterministic discrimination. The PES and the resulting graphical applications may increase our knowledge of the item by providing information on whether an item doomed to be undiscriminating by the classical indices can in fact discriminate between respondents. It can then be used as additional information when employing classical or modern methods in item analysis.

4. The procedure for detecting item discrimination with a nonparametric index can be applied in the extreme situations of ultimate discrimination where probabilistic approaches yield indefinite values or are not defined.
5. PES makes it possible to detect and analyze the anomalous, illogical, or even "pathological" cases on the data structure. The undesirable and unexpected possibility of having an incorrect answer among higher respondents while having a correct answer among lower respondents can easily be detected using the approach.
6. In many cases, IRT models need large or moderate sample sizes for the sample free estimation of item parameters. PES is not restricted to any sample size. However, in both cases, larger sample size allows more stable estimates.
7. To understand the technology and theory of Rasch- and IRT models, we usually require advanced knowledge of mathematic and statistic. Although there are a few new symbols, subscripts and superscripts in the text, the logic of the PES is simple, and the approach uses concepts, symbols and mathematical procedures that are mostly well-known. The procedures described here could perhaps be easy to apply without even basic knowledge of more advanced procedures such as Rasch- or IRT modeling.
8. Rasch- and IRT modeling usually requires special computer software. PES is easy to apply in real world contexts where, for example, teachers assess their own classes using a small-scale test by using common spread sheet software.

All in all, the graphical and visual analysis of items are, at the side of analytical and statistical analysis, an integral part of item diagnosis and test construction. Human eyes are quick in detecting anomalous patterns from the graphs when trained to do so. Hence, it is possible to sieve quite a lot of crossing, completing, and cumulating information of an item behavior just by using simple graphs. Screening the items visually makes it also easy to explain to non-technical experts, for example, why some specific item should not be selected to the final test.

References

Adams, R. J., & Khoo, S.-T. (1993). *QUEST. The interactive Test Analysis system.* [Software package]. The Australian Council for Educational Research LTD. ACER.

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. Brandon-Tuma (Ed.), *Sociological methodology* (pp. 33–80). Jossey-Bass.

Arbuckle, J. L. (2015). *IBM® SPSS® Amos™ 24 User's Guide.* IBM Corporation. Amos Development Corporation.

Bechger, T. M., Maris G., Verstralen H. H. F. M., & Béguin A. A. (2003). Using Classical Test Theory in Combination with Item Response Theory. *Applied Psychological Measurement 27*(5), 319–334. https://doi.org/10.1177/0146621603257518

Bentler, P. M. (1995). *EQS: Structural Equations Program Manual.* Multivariate Software Inc.

Bentler, P. M., & Wu, E., J. C. (2008). *EQS 6.1 for Windows. User's Guide.* Multivariate Software Inc. https://mvsoft.com/wp-content/uploads/2021/04/EQS_6_User_Guide_R8.pdf

Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). *TESTFACT 4.0* [Computer software and manual]. Scientific Software International.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO for Windows.* [Software package]. Scientific Software International.

CITO (2013). *TiaPlus.* [Software package + Manual]. Cito. http://tiaplus.cito.nl/TiaPlus_Users_Manual.pdf

Cureton, E. E. (1956). Rank-biserial correlation. *Psychometrika, 21*(3), 287–290. https://doi.org/10.1007%2FBF02289138

Cureton, E. E. (1966a). Simplified formulas for item analysis. *Journal of Educational Measurement, 3*(2), 187–189. https://doi.org/10.1111/j.1745-3984.1966.tb00879.x

Cureton, E. E. (1966b). Corrected item–test correlations. *Psychometrika, 31*(1), 93–96. http://dx.doi.org/10.1007/BF02289461

Curtis, D. D. (2004). Person misfit in attitude surveys: Influences, impacts and implications. *International Education Journal, 5*(2), 125–144. https://pdfs.semanticscholar.org/b9b3/78fbb66a3988bd8f11b41ae1812fd2df0765.pdf

D'Agostino, R. B., & Cureton, E. F. (1975). The 27 percent rule revisited. *Educational and Psychological Measurement, 35*(1), 47–50. http://dx.doi.org/10.1177/001316447503500105

Ebel, R. L. (1967). The relation of item discrimination to test reliability. *Journal of Educational Measurement 4*(3), 125–128. http://www.jstor.org/stable/1434085

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of Educational Measurement*. (Fifth edition). Pear-son.

ETS (1960). *Short-cut statistics for teacher-made tests*. Educational Testing Service.

FINEEC (2018). *National assessment of learning outcomes in mathematics at grade 9 in 2002*. (Unpublished dataset opened for the re-analysis 18.2.2018). Finnish National Education Evaluation Centre (FINEEC).

Fisher, W. P. Jr. (1992). Stochastic resonance and Rasch measurement. *Rasch Measurement Transactions, 5*(4), 186. https://www.rasch.org/rmt/rmt54k.htm

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23*, 267–269. https://doi.org/10.1207/s15327906mbr2302_9

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. Journal of the *American Statistical Association, 49*(268), 732–764. https://doi.org/10.1080/01621459.1954.10501231

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A., Suchman, P. F., Lazarsfield, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction*. Princeton University Press.

Haladyna, T. M., & Rodrigues, M. C. (2021). Using Full-information Item Analysis to Improve Item Quality. *Educational Assessment, 26*(3), 1–14. https://doi.org/10.1080/10627197.2021.1946390

Henrysson, S. (1963). Correction of item-total correlations in item analysis. *Psychometrika, 28*(2), 211–218. https://doi.org/10.1007/BF02289618

Hopkins, K. D. (1997). Educational and Psychological Measurement and Evaluation (Eighth edition). Pearson.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*(2, Pt.1), 183–202. https://doi.org/10.1007/BF02289343

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika, 57*(2), 239–251. https://doi.org/10.2307/2334833

Jöreskog, K. G., Olsson, U. H., & Wallentin, F. Y. (2016). *Multivariate Analysis with LISREL*. Springer. https://doi.org/10.1007/978-3-319-33153-9

Jöreskog, K. G., & Sörbom D. (1999a). *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language*. [Software package]. Scientific Software International.

Jöreskog, K. G., & Sörbom D. (1999b). *LISREL 8 User's Reference Guide*. Scientific Software International.

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30*(1), 17–24. http://dx.doi.org/10.1037/h0057123

Linacre, J. M. (1992). Stochastic Guttman order. *Rasch Measurement Transactions, 5*(4), 189. https://www.rasch.org/rmt/rmt54p.htm

Linacre, J. M. (2000). Guttman coefficients and Rasch data. *Rasch Measurement Transactions, 14*(2), 746–747. https://www.rasch.org/rmt/rmt142e.htm

Linacre, J. M. (2003). What is item response theory, IRT? A tentative taxonomy. *Rasch Measurement Transactions, 17*(2), 926–927. https://www.rasch.org/rmt/rmt172g.htm

Linacre, J. M. (2011). *WINSTEP. Rasch measurement computer program*. [Software package]. Beaverton, OR.

Linacre, J. M., Andrich, D. A., & Luo, G. (2003). Guttman parameterization of rating Scale. *Rasch Measurement Transactions, 17*(3), 944. https://www.rasch.org/rmt/rmt173h.htm

Linacre, J. M. & Wright, B. D. (1994). Dichotomous infit and outfit mean-square fit statistics. *Rasch Measurement Transactions, 8*(2), 350. https://www.rasch.org/rmt/rmt82a.htm

Linacre, J. M. & Wright, B. D. (1996). Guttman-style item location maps. *Rasch Measurement Transactions, 10*(2), 492–493. https://www.rasch.org/rmt/rmt102h.htm

Livingston, S. A., & Dorans, N. J. (2004). *A graphical approach to item analysis*. Research Report No. RR-04-10. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.tb01937.x

Lord, F. M. (1952). The relationship of the reliability of multiple-choice test to the distri-bution of item difficulties. *Psychometrika, 17*(2), 181–194. http://dx.doi.org/10.1007/BF02288781

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Addison-Wesley Publishing Company.

Lorenzo-Seva, U., & Ferrando, P.J. (2015). POLYMAT-C: a comprehensive SPSS program for computing the polychoric correlation matrix. *Behavior Research Methods, 47***,** 884–889. https://doi.org/10.3758/s13428-014-0511-x

Martinson, E. O. & Hamdan, M. A. (1972). Maximum likelihood and some other asymptotical efficient estimators of correlation in two-way contingency tables. *Journal of Statistical Computation and Simulation, 1*(1), 45–54. https://doi.org/10.1080/00949657208810003

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology*. (Fourth edition). Harcourt Brace College Publishers.

Metsämuuronen, J. (2016). Item–total correlation as the cause for the underestimation of the alpha estimate for the reliability of the scale. *GJRA - Global Journal for Research Analysis, 5*(1), 471–477. https://www.worldwidejournals.com/global-journal-for-research-analysis-GJRA/file.php?val=November_2016_1478701072_159.pdf

Metsämuuronen, J. (2017). *Essentials of research methods in human sciences*. SAGE Publications, Inc.

Metsämuuronen, J. (2020a). Somers' D as an alternative for the item–test and item–rest correlation coefficients in the educational measurement settings. *International Journal of Educational Methodology, 6*(1), 207–221. https://doi.org/10.12973/ijem.6.1.207

Metsämuuronen, J. (2020b). Generalized discrimination index. *International Journal of Educational Methodology, 6*(2), 237–257. https://doi.org/10.12973/ijem.6.2.237

Metsämuuronen, J. (2020c). Seeking the true item difficulty. Bias-corrected item difficulty and some consequences in Rasch and IRT modeling. Preprint. https://doi.org/10.13140/RG.2.2.34640.33283/1

Metsämuuronen, J. (2021a). Goodman–Kruskal gamma and dimension-corrected gamma in educational measurement settings. *International Journal of Educational Methodology, 7*(1), 95–118. https://doi.org/10.12973/ijem.7.1.95

Metsämuuronen, J. (2021b). Directional nature of Goodman-Kruskal gamma and some consequences. Identity of Goodman-Kruskal gamma and Somers delta, and their connection to Jonckheere-Terpstra test statistic. *Behaviormetrika, 48*(2). http://dx.doi.org/10.1007/s41237-021-00138-8

Metsämuuronen, J. (2022a). Effect of various simultaneous sources of mechanical error in the estimators of correlation causing deflation in reliability. Seeking the best options of correlation for deflation-corrected reliability. *Behaviormetrika 49***,** 91–130. https://doi.org/10.1007/s41237-022-00158-y; full text is accessible at https://rdcu.be/cGMUt

Metsämuuronen, J. (2022b). Deflation-corrected estimators of reliability. *Frontiers in Psychology, 12*:748672. https://doi.org/10.3389/fpsyg.2021.748672

Metsämuuronen, J. (2022c). Attenuation-corrected estimators of reliability. *Applied Psychological Measurement.* (in printing)

Miller, D., Linn, R., & Gronlund, N. E. (2012). *Measurement and evaluation in teaching*. 11th edition. Pearson.

Mislevy, R., & Bode, R. D. (1982). *BILOG: Maximum likelihood item analysis and test scoring with logistic models*. [Software package] Scientific software.

Mokken, R. J. (1971). A theory and procedure of scale analysis. De Guyter.

Moses, T. (2017). A review of developments and applications in item analysis. In R. Bennett & M. von Davier (eds), *Advancing human assessment. Methodology of educational measurement and assessment*. Springer, Cham. https://doi.org/10.1007/978-3-319-58689-2_2

Mulaik S. A. (1972). *A mathematical investigation of some multidimensional Rasch models for psychological tests*. Paper presented at the annual meeting of the Psychometric Society, Princeton, NJ.

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4: IRT item analysis and test scoring for rating-scale data* [Software package]. Scientific Software International.

Muthén, L. K. & Muthén, B. O. (1998-2017). *Mplus User's Guide*. Eighth Edition. Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2017). *MPLUS*. [Software package]. Muthén & Muthén.

Newson, R. (2008). *Identity of Somers' D and the rank biserial correlation coefficient*. http://www.rogernewsonresources.org.uk/miscdocs/ranksum1.pdf

Oosterhof, A. C. (1976). Similarity of various item discrimination indices. *Journal of Educational Measurement*, 13(2), 145–150. http://dx.doi.org/10.1111/j.1745-3984.1976.tb00005.x

Pearson, K. (1896). VII. Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Philosophical Transactions of the Royal Society of London, 187*, 253–318. https://doi.org/10.1098/rsta.1896.0007

Pearson, K. (1900). I. Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society A. Mathematical, Physical and Engineering Sciences, 195*(262–273), 1–47. https://doi.org/10.1098/rsta.1900.0022

Pearson, K. (1913). On the Measurement of the Influence of "Broad Categories" on Correlation. *Biometrika, 9*(1–2), 116–139. https://doi.org/10.1093/biomet/9.1-2.116

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Danish Institute for Educational Research.

Raykov, T., & Marcoulides, G. A. (2016). On the Relationship Between Classical Test Theory and Item Response Theory: From One to the Other and Back. *Educational and Psychological Measurement, 76*(2), 325-338. http://dx.doi.org/10.1177/0013164415576958

Reckase, M. D. (1972). *Development and application of a multivariate logistic latent trait model*. Unpublished doctoral dissertation, Syracuse University, Syracuse, NY.

Reckase, M.D. (2009). Multidimensional Item Response Theory: Statistics for Social and Behavioral Sciences. Springer.

Reeve, B. B. (2002). *An introduction to modern measurement theory*. National Cancer Inst.

Roskam, E., & Jansen, P. (1992). Rasch model derived from consistent stochastic Guttman ordering. *Rasch Measurement Transactions, 6*(3), 232. https://www.rasch.org/rmt/rmt63e.htm

Ross, J., & Weitzman, R. A. (1962). The twenty-seven per cent rule. *The Annals of Mathematical Statistics, 35*(1), 214–221. http://dx.doi.org/10.1214/aoms/1177703745

Schumacker, R. E. (2005). *Classical test analysis*. Applied Measurement Associates. http://www.appliedmeasurementassociates.com/ama/assets/File/CLASSICAL_TEST_ANALYSIS.pdf

Siegel, S. & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences*. Second Edition. McGraw-Hill.

Sijtsma, K., & Molenaar, I.W. (2002). *Introduction to nonparametric item response theory*. Measurement methods for the Social Science, 5. SAGE Publications, Inc.

Smith, R. M. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions, 10*(3), 516-517. https://www.rasch.org/rmt/rmt103a.htm

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review, 27*(6), 799–811. http://dx.doi.org/10.2307/2090408

Stout, W. (2002). Psychometrics: From practice to theory and back. 15 Years of non-parametric multidimensional IRT, DIF/test equity, and skills diagnostic assessment. *Psychometrika, 67*(4), 485–518. http://dx.doi.org/10.1007/BF02295128

Thissen, D. M. (1983). *MULTILOG: Item analysis and scoring with multiple category response model*. [Software package] Chicago, IL: International Educational Services.

Thissen, D. M. (2003). MULTILOG 7: Multiple categorical item analysis and test scoring using item response theory [Software package]. Scientific Software International.

van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis*, 11(2), 139–163. https://doi.org/10.1093/pan/mpg002

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). *One-parameter logistic model OPLM*. [Software package + Manual]. Cito.

Wiersma, W., & Jurs, S. G. (1990). *Educational measurement and testing* (Second ed.). Allyn and Bacon.

Wingersley, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide*. Educational Testing Service.

Wolf, R. (1967). Evaluation of several formulae for correction of item-total correlations in item analysis. *Journal of Educational Measurement, 4*(1), 21–26. https://doi.org/10.1111/j.1745-3984.1967.tb00565.x

Wright, B. D., & Stone, M. H. (1979). *Best test design*. MESA Press.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ConQuest: Generalized item response modeling software* [Computer program]. Australian Council for Educational Research.

Yao, L. (2003). *BMIRT: Bayesian multivariate item response theory* [Software package]. CTB/McGraw-Hill.

Zaiontz, C. (2021) *Real Statics Using Excel. Polychoric Correlation using Solver*. http://www.real-statistics.com/correlation/polychoric-correlation/polychoric-correlation-using-solver/

Zimmerman, D. W., Williams, R. H., Zumbo, B. D., & Ross, D. (2005). Louis Guttman's contributions to classical test theory. *International Journal of Testing, 5*(1), 81–95. http://dx.doi.org/10.1207/s15327574ijt0501_7

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items (Version 3)* [Software package]. Scientific Software International.

Citation:

Metsämuuronen, J. (2022). Essentials of Visual Diagnosis of Test Items. Logical, Illogical, and Anomalous Patterns in Tests Items to be Detected. *Practical Assessment, Research & Evaluation*, 27(5). Available online: https://scholarworks.umass.edu/pare/vol27/iss1/5/

Corresponding Author:

Jari Metsämuuronen
1) Finnish Education Evaluation Center and 2) Centre for Learning Analytics, University of Turku

Email: jari.metsamuuronen [at] gmail.com

Appendix A. Syntaxes for calculating *Rit*, *RPC*, *G* and *D* by using general software packages

Rit

In the article, *Rit* = item–total correlations = Pearson product-moment correlations (PMC) were calculated manually with a common spreadsheet software by syntax =CORRELATION(matrix1;matrix2). In IBM SPSS, the syntax for PMC is CORRELATIONS /VARIABLES=item score MISSING=PAIRWISE. In SAS, the command PROC CORR provides PMC. In R, PMC can be calculated by cor(x, y, method = c("pearson")) (see http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r).

RPC

In the article, RPCs between items and a score variable were calculated manually by using Zaiontz's (2021) procedure of Martinson and Hamdan (1972) two-step estimator. In IBM SPSS, the syntax for *RPC* is not available although some macros are (e.g., Lorenzo-Seva & Ferrando, 2015). In SAS, the command PROC CORR provides *RPC*. With R, R_{PC} can be calculated by *CorPolychor(x, y, ML = FALSE, control =* list*(), std.err = FALSE, maxcor=.9999)## S3 method for class 'CorPolychor'* print*(x, digits =* max*(3,* getOption*("digits") - 3), ...)* (see, https://rdrr.io/cran/DescTools/man/CorPolychor.html).

G

In the article, estimates by *G* between items and the score were calculated by IBM SPSS. In IBM SPSS, the syntax for *G* is CROSSTABS /TABLES=item BY Score /STATISTICS=GAMMA. In SAS, the command PROC FREQ provides *G* by specifying the TEST statement by GAMMA, SMDCR options. With R, *G* is calculated by *GoodmanKruskalGamma(x, y = NULL, conf.level = NA, ...)* (see, https://rdrr.io/cran/DescTools/man/).

D

In the article, estimates by *D* between items and the score were calculated by IBM SPSS. In IBM SPSS, the syntax for *D* is CROSSTABS /TABLES=item BY Score /STATISTICS=D. In SAS, the command PROC FREQ provides *D* by specifying the TEST statement by D, SMDCR options. with R, *D* can be calculated by *SomersDelta(x, y = NULL, direction =* c*("row", "column"), conf.level = NA, ...)* (see, https://rdrr.io/cran/DescTools/man/). The direction "score dependent" is used (see Metsämuuronen, 2020a).