Research Article doi: 10.12973/ijem.6.2.237

International Journal of Educational Methodology
Volume 6, Issue 2, 237 - 257.
ISSN: 2469-9632
http://www.ijem.com/

Generalized Discrimination Index

Jari Metsämuuronen*
Finnish Education Evaluation Centre,
FINLAND

NLA University College,
NORWAY

Received: February 17, 2020 ▪ Revised: April 5 2020 ▪ Accepted: April 21, 2020

Abstract: Kelley's Discrimination Index (DI) is a simple and robust, classical non-parametric short-cut to estimate the item discrimination power (IDP) in the practical educational settings. Unlike item–total correlation, DI can reach the ultimate values of +1 and −1, and it is stable against the outliers. Because of the computational easiness, DI is specifically suitable for the rough estimation where the sophisticated tools for item analysis such as IRT modelling are not available as is usual, for example, in the classroom testing. Unlike most of the other traditional indices for IDP, DI uses only the extreme cases of the ordered dataset in the estimation. One deficiency of DI is that it suits only for dichotomous datasets. This article generalizes DI to allow polytomous dataset and flexible cut-offs for selecting the extreme cases. A new algorithm based on the concept of the characteristic vector of the item is introduced to compute the generalized DI (GDI). A new visual method for item analysis, the cut-off curve, is introduced based on the procedure called exhaustive splitting.

Keywords: *Kelley's discrimination index, item parameter, item–total correlation, item analysis, classical test theory.*

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

Introduction

Item discrimination power as a phenomenon and the underlying statistical model

In the general sense, the item discriminating power (IDP) is a loose term for the characteristic of a test item to reflect how accurately or efficiently an item can discriminate between the test-takers with the higher item response from those with the lower item response (see ETS, 2020; Liu, 2008; Lord & Novick, 1968; MacDonald & Paunonen, 2002). In achievement testing with multiple-choice questions resulting binary items, as an example, we ask how accurately the item can make a difference between those test-takers who gave the correct answer and those who gave an incorrect answer. Within test theory and test construction, item discrimination power is one of the two to five item parameters that characterize the test items (e.g., Lord & Novick 1968). Two other parameters most commonly used are item difficulty and pseudo-change score level (or guessing), while the fourth and fifth parameters are rarely in the practical use (see the discussion in Balov & Marchenko, 2016; Barton & Lord, 1981; Loken & Rulison, 2010; Metsämuuronen, 2017).

From the theoretical viewpoint, item discrimination power is usually discussed within the measurement modelling settings with certain statistical models. In these, two ordinal observed variables g (item) and X (score) that have r and s distinctive categories, respectively, are assumed. Latent to the observed variables, we have two continuous variables ξ and η referring to the latent "true" trait (e.g., "true" achievement level) in the test item and in the score. For the latter part of the article dealing with polytomous items, let us assume that the item has 5 categories (r = 0 to 4) and the score includes 40 categories (s = 0 to 39). The threshold values for ξ for each category in g are denoted by γ_i and for η for each category in X by τ_j. Then, the variable g is related to ξ and X to η so that

* **Correspondence:**
Jari Metsämuuronen, Finnish Education Evaluation Centre, P.O. Box 28, FI-00101 Helsinki, Finland. ✉ jari.metsamuuronen@gmail.com

$$g = g_i, \text{ if } \gamma_{i-1} \leq \xi < \gamma_i, i = 1,2,..., R$$

and

$$X = x_j, \text{ if } \tau_{j-1} \leq \eta < \tau_j, j = 1, 2, ..., S.$$

For convenience, we assume that $\gamma_0 = \tau_0 = -\infty$ and $\gamma_R = \tau_S = +\infty$, and $g_1 < g_i < g_r$ and $x_1 < x_j < .x_s$ as illustrated in Figure 1.

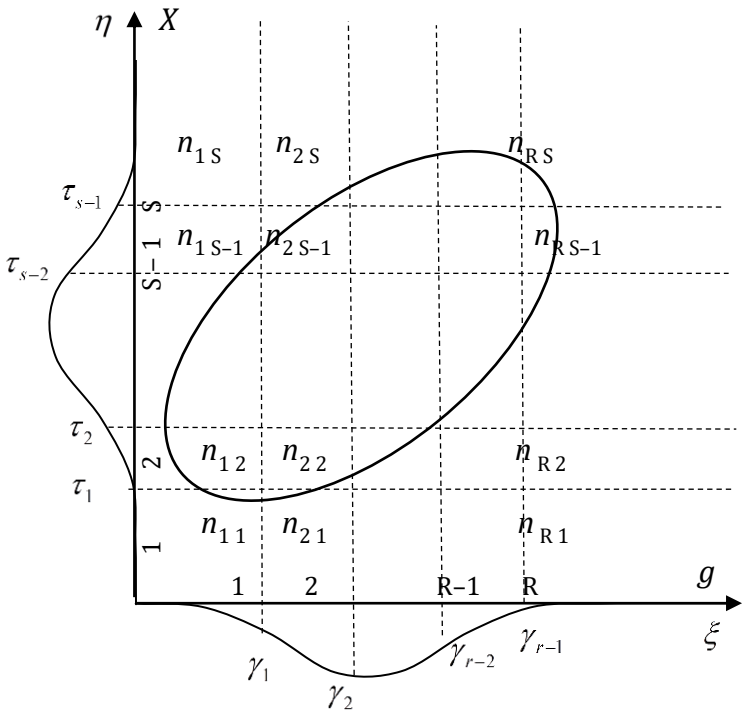

Figure 1. Example of two latent variables ξ and η categorized into ordinal scales; n_{gX} denotes the number of times the observation (g, X) is obtained in the sample.

From the correlation viewpoint related to the item analysis with polytomous items, the inferred correlation between two latent variables is polychoric correlation ($\rho_{\xi\eta}$), the inferred correlation between the latent ξ and observed *X* is polyserial correlation ($\rho_{\xi X}$), the observed correlation between the interval-scaled variables *g* and *X* is point-polyserial correlation, that is, the traditional item–total correlation (ρ_{gX}), and the observed correlation between the ordinal-scaled variables *g* and *X* is rank-polyserial correlation (ρ_{RP}). From Kelley's discrimination index viewpoint focused in the article, the middle categories in the score are omitted in the analysis.

Selected indices of item discrimination power

The indices of IDP reflect the relationship between an item and a trait of interest (Moses, 2017). During the history of test theory, many indices of IDP have been created and developed (see the comparisons by Cureton, 1966a, 1966b; ETS, 1960; Liu, 2008; Metsämuuronen, 2020; Oosterhof 1976; Wolf, 1967). Recently, Metsämuuronen (2020), as an example, studied the efficiency of nine frequently discussed classical indices. Though the contemporary *theoretical* literary of item parameters has, mainly, concentrated on the different aspects of the modern test theory, that is, on item response theory (IRT) modeling (e.g., Balov & Marchenko 2016; Cechova, Neubauer, & Sedlacik, 2014) and the large-scale assessments (e.g., PISA, TIMSS, PIRLS, PIAAC) use mainly IRT modeling in the analysis (e.g., Aslan & Aybek, 2019; Esendemir & Bindak, 2019), many of the classical indices are still in wide use in the *practical* item analysis settings.

Two widely used classical indices for IDP are item–total correlation (*Rit*) (ρ_{gX} ; based on Pearson, 1896) and item–rest correlation (*Rir,* proposed by Henrysson, 1963 and supported by Cureton, 1966b) also known as "corrected item–total correlation" (e.g., in the outputs of IBM SPSS and STATA software packages). These are defaults in widely used general

statistical software packages such as IBM SPSS (e.g.,IBM, 2017), Stata (e.g.,Stata corp., 2018), and SAS (e.g.,Yi-Hsin & Li, 2015). Here, we note the mechanical connection of *Rit* and reliability; because *Rit* is embedded to coefficient alpha (see Eq. 1), and because alpha is the most widely used estimator for the test reliability in the practical settings (see the worry of its too wide use by Dunn, Baguley, & Brunsden, 2013; Graham, 2006; Green & Young, 2009; Hogan, Benjamin, & Brezinski, 2000; Trizano-Hermosilla & Alvarado, 2016; Yang & Green, 2011), *Rit* may be the widest used of all the indices of IDP—though not always consciously.

Both *Rit* and *Rir* are, essentially, Pearson product-moment correlation coefficients—the first between the item and the total score and the latter between the item and the score where the interesting item is omitted. Both embed challenges in the measurement modelling settings. Metsämuuronen (2016; see also 2020) showed that the item–total correlation always underestimates the IDP when the scales of the item and the score differ from each other, and this underestimation may be grave when the difficulty level of the item is extreme. Metsämuuronen (2017) showed that, paradoxically, the "corrected" item–total correlation underestimates IDP even more than the "uncorrected" item–total correlation; this is obvious because the magnitudes of the estimates by *Rir* are always lower than those by *Rit*. The relevant question is what could be a better index for *Rit* and *Rir* within the classical toolbox? After comparing nine indices (*Rit*, *Rir*, bi- and polyserial correlation, polychoric correlation Goodman–Kruskal *Lambda* and *Tau*, Pearson *Eta*, and Somers' *D*), Metsämuuronen (2020) suggests Somers' *D* (Somers, 1962) as one of the "superior alternatives" to *Rit* and *Rir* in the binary dataset.

Another kind of possible "superior alternative" to *Rit* worth of studying is Kelley's Discrimination Index (*DI*; Kelley, 1939) suggested to classroom teachers during the years, among others, by Ebel (1954a, 1954b), Educational Testing Service (ETS, 1960), Wiersma & Jurs (1990), Mehrens & Lehmann (1991), and Metsämuuronen (2017). *DI* is one of the short-cut methods for the practical testing settings because of its simplicity related to estimation without sophisticated statistical tools (see Cureton, 1966a). *DI* and it's generalized version *GDI* are in focus of this article.

Why is item discrimination power important and why is Kelley's DI interesting?

Of the item parameters, discriminating power is an interesting characteristic of the test item because it has a strict connection to the test reliability. We remember that Lord and Novick (1968, p. 344) introduced a modification of the alpha coefficient (α) in which the index of IDP (*Rit* = ρ_{gX}) is embedded:

$$\alpha_{\mathrm{LN}} = \frac{k}{k-1}\left(1 - \frac{\sum_{g=1}^{k}\sigma_{\mathrm{g}}^{2}}{\left(\sum_{g=1}^{k}\sigma_{\mathrm{g}}\rho_{gX}\right)^{2}}\right) \qquad (1)$$

where σ_g^2 refers to the item variances and *k* is the number of items. This coefficient is algebraically identical with the classical formula of the coefficient alpha published in Gulliksen (1950) and Cronbach (1951) based on works by Kuder and Richardson (1937), Flanagan (1937), Rulon (1939), and Guttman (1945). Further, if we take the derivation of Kuder and Richardson seriously —where we assume parallelism of the items (see the critic of using the classical forms by, e.g., Tarkkonen, 1987 and Vehkalahti, 2000)—the less used classic estimator of reliability by Kuder and Richardson (1937), *KR*21, gives us a rough estimate for reliability with minimal factors: $\alpha_{KR21} = \frac{k}{k-1}\left(1 - \frac{1}{k}\left(\sum_{g=1}^{k}\rho_{gX}\right)^{-2}\right)$.

The simplified formula means that the only factor determining the magnitude of estimate of reliability, except the number of items (*k*), would be the magnitude of item discrimination (*Rit* = ρ_{gX}). From this perspective, Ebel (1967) based on Stanley (1964), provided us another kind of estimator of reliability that combines the alpha type of estimator and Kelley's *DI*:

$$\rho_{\text{EBEL}} = \frac{k}{k-1}\left(1 - \frac{6\sum_{g=1}^{k}\sigma_{\text{g}}^2}{\left(\sum_{g=1}^{k} DI\right)^2}\right) \quad (2)$$

where *DI* refers to Kelly's Discrimination Index. The value of Ebel's estimator seems to be, in many cases, lower than that of coefficient alpha. Hence, when knowing that the coefficient alpha underestimates always the real reliability, Ebel's estimator seems to underestimate reliability even more. Maybe this is the reason why the latter formula is not in a practical use.

All in all, to give a rough estimate of reliability of the test, the only things needed—in addition to the number of the items and item variances—are the estimates for item discrimination. In the practical settings of compiling the test from single items, the more items with the high discriminating power we select to the test, the more discriminating the test would be and, contrastingly, the items with very low item discrimination power are usually omitted from the final compilation to raise the reliability. Therefore, the indicators of IDP and the estimates they produce are interesting from the general viewpoint of reflecting the accuracy of the whole test.

DI has some positive characteristics over *Rit* and *Rit* why it turns to be an interesting coefficient to study more: (1) It can detect the deterministic patterns and reaches the value *DI* = ± 1 correctly while *Rit* and *Rir* cannot reach the ultimate value and, hence, they underestimate the association between the item and the score always. Hence, (2) *DI* does not necessarily obviously and mechanically underestimate the association between the item and score as *Rit* and *Rir* do in the case when the scales of the item and the score are not equal. (3) *DI* is easy to calculate in the practical settings related to item analysis, and (4) unlike the coefficients based on Pearson product moment correlation coefficient (*Rit* and *Rir*), as being based on the *order* of the order of the score, it is robust against changes in the dataset and outlier values. Hence, when *DI* may be studied as a "superior alternative" to *Rit* in the binary case, generalized *DI* could be used in the polytomous cases.

The main limitation in *DI* is that it is *restricted only to the binary datasets*. Also, the traditional way of using DI limits its use to fixed cut-offs of the extreme values in the analysis. This article generalizes *DI* to the polytomous dataset with limitless cut-offs.

Research questions

This article discusses the possibilities of Kelley's *DI* as a useful tool in the practical educational testing settings for item analysis. The first part of the article asks how stable the estimates by *DI* are in comparison with those by *Rit*. This is illustrated by using a simple example with deterministically discriminating Guttman-patterned items (Guttman, 1950). The example illustrates also the basic difference in the estimates by *Rit* and *DI*: while *Rit* cannot reach the reach the ultimate value *Rit* = ± 1 in the real life testing settings, *DI can* reach the value *DI* = ± 1 .

The latter part of the article derives the generalized version of *DI* that can be used with binary and polytomous items with multiple cut-offs. This part answers the following research questions: (1) What the characteristics of generalized *DI* (*GDI*) are; (2) How *DI* and *GDI* can be calculated by using new computational algorithms; (3) How *GDI* can be used in the practical educational testing settings; and (4) How a new graphical method of visualizing the item discrimination power could be used in the practical testing settings in locating latent difficulty level, non-logical test behavior, as well as stableness of the estimate of the item discrimination power.

Methodology

The treatment in the article is mainly theoretical and conceptual. Hence, specific methodological tools are not in use. Within the course of study, a new kind methodology is developed for the practical testing settings to be utilized in analyzing both the dichotomous and polytomous items.

The course of the study starts by introducing the original Kelley's *DI*. In this section, the peculiarity of *DI* of not using all the test-takers in the analysis with the rationale of using different cut-offs is discussed. Here, also, *DI* is compared with *Rit* by way of example to illustrate the extreme values in *DI* and how small changes do not change the value in *DI* while they always do in *Rit*.

Generalized *DI* is derived in the next main section. This requires new operationalization of the traditional notation related to *DI*. Some numerical examples of using *GDI* are discussed and new computational algorithms are provided to calculate *DI* and *GDI*.

Finally, a new visual method, cut-of curve, related to *GDI* for illustrating the item discrimination power and item difficulty is introduced as further elaboration of *GDI*. The new method is based on exhaustive splitting procedure (ESP) of the dataset. Some practical hints are given to the practical users how to use the tool.

Kelley's *DI* and it's stable character in comparison with *Rit*

Kelley's DI

Though *DI* is quite an old innovation—originally created for validation of items (chronologically, Long & Sandiford, 1935; Kelley, 1939; Johnston, 1951)—it is still in use in the practical item analysis settings especially in the educational settings (see some examples in Appendix). In some rare works, *DI* has been connected to Rasch- and IRT modelling (e.g., Bazaldua, Lee, Keller, & Fellers, 2017; Kelley, Ebel, & Linacre, 2002; Tristan, 1998) and Bayesian inference (e.g., Batanero, 2007). However, in general, *DI* is not widely handled in the contemporary theoretical writings. Nevertheless, *DI* may be in a semi-wide practical use because it is specifically suggested for teachers by leading authors during the years, and because it is very easy to use in in such environments where sophisticated software packages for items analysis are not in use as discussed above.

In comparison with other indices of IDP, the calculation of *DI* embeds peculiarity that it uses only the extreme cases in the estimation. Though the rationale of not using all the cases in the estimation is not obvious, by using *DI*, we ask the same essential question as with the other indices: how well the test item can discriminate between the lower- and higher-performing test-takers. Because it is difficult to discriminate the medium-range cases from each other and, hence, they may confuse the possible interpretations, the logic behind *DI* to compare only the extreme cases seems reasonable. Because of the mechanism of selecting the extreme cases to the analysis, the different cut-offs for the extreme groups have been widely discussed during the years: in the early phase by, for example, Long and Sandiford (1935), Kelley (1939), and Forlano and Pinter (1941) and later by, for example, Cureton (1966a), D'Agostino and Cureton (1975), Ebel (1967), Feldt (1963), Ross and Lumsden (1964), and Ross and Weitzman (1964).

Kelley's *DI* is traditionally calculated by using the following procedure. Assume a test with *N* test-takers ordered by the score (*X*). The test-takers are divided into two groups consisting, traditionally, only the highest and lowest 25% (e.g.,D'Agostino & Cureton, 1975; Mehrens & Lehman, 1995; Metsämuuronen, 2017) or 27% (e.g.,Ebel, 1967; Feldt, 1963; Kelley, 1939; Ross & Weitzman, 1964) of the test-takers. These cut-offs are denoted by the upper fourth (*U*) consisting the highest scoring test-takers and lower fourth (*L*) consisting the lowest scoring test-takers. By using this notation, assuming a binary item, *DI* can be expressed as follows:

$$DI = \frac{R^U - R^L}{\frac{1}{2}T} = 2\left(p^U - p^L\right) = 2\left(p - 2p^L\right) \quad (3)$$

(e.g., Metsämuuronen, 2017, p. 125) where R^U and R^L refer to the number of correct answers in the upper and lower fourth of the ordered dataset, and T refers to the total number of observations in the two parts together. Consequently, p^U and p^L refer to the proportions of correct answers in the upper and lower part of the reduced dataset, and p is the proportion of the correct answers in the reduced dataset.

Possible cut-offs and a more general notation

The discussion of different cut-offs (see the literature above) is relevant from the viewpoint of the generalized discrimination index introduced in the latter part of the article. The reason for the 25% or 27% cut-offs is that in a normal distribution, the 27% cut-off maximizes the difference in the population and, hence, is considered statistically better than the 25% cut-off (Kelley, 1939, Wiersma & Jurs, 1990). In the original derivation, Kelly assumed normal distribution of the score and 50% of passes for the entire item (i.e., p = 0.50). If the difficulty level of the items would differ from p = 0.50, Kelly wrote: "*The proportions undoubtedly would not be twenty-seven per cent from the extremes*" (p. 70). Forlano and Pinter (1941)—after studying the cut-offs of upper and lower 50%, 33%, 27%, 16%, and 7%—concluded that no method occupies the first rank. However, they preferred 27% because it is a simple and rapid, rough and ready method. Also, Feldt (1963)—after showing that 27% yielded the most precise estimate of the tetrachoric coefficient only when the population correlation was close to zero—suggested no change in the traditional 27% because it yield highly efficient estimate.

All in all, during the history, many different cut-offs have been discussed. Hence, Metsämuuronen (2017) proposed a general notation of *DI*: $DI_i = 2\left(p_i - 2p_i^L\right)$, where p_i refers to the proportion of correct answers in a specific cut-off *i*, and p_i^L refers to the proportion of correct answers in the lower (*L*) part of the cut-off *i*. This kind of notation seems relevant in procedures where many if not all the possible cut-offs are used in the analysis.

Stableness of the estimates by DI in comparison with those by Rit

Because of being robust statistics based on the order of the test-takers, *DI* seems to produce quite stable estimates for IDP. This is discussed by a theoretical example related to so-called Guttman-patterned items (Guttman, 1950; Linacre & Wright, 1996) and minor stochastic error illustrated in Tables 1a and 1b. The ultimately Guttman pattern is a theoretical structure of a dataset where the deterministically discriminating test items form a triangle type of dataset with different difficulty levels with a string of 0s followed by a string of 1s when the cases are ordered consecutively by their total score. Here, the items are called Guttman-patterned when the response pattern is formed with a string of 0s followed by a string of 1s when the cases are ordered consecutively by their latent trait even though the dataset would not be triangle-formed.

Assume a dataset of a hypothetical test with n = 15 test-takers and k = 4 binary items with the deterministically discriminating nature (Guttman pattern) with as in Table 1a. Typical for this theoretical form is that the items discriminate the (hypothetical) test-takers with a higher score in a deterministic manner from those with a lower score. Then, the score explains the behavior in the item in the deterministic manner, and we would expect to see the perfect explaining power ($\rho^2 = 1$) and, consecutively, perfect correlation ($\rho = 1$).

Table 1. Hypothetical data of four Guttman-patterned items

1a) Guttman pattern

Test-taker	**Item1**	**Item2**	**Item3**	**Item4**	**Score**
1	0	0	0	0	**0**
2	0	0	0	0	**0**
3	0	0	0	1	**1**
4	0	0	1	1	**2**
5	0	1	1	1	**3**
6	1	1	1	1	**4**
7	1	1	1	1	**4**
8	1	1	1	1	**4**
9	1	1	1	1	**4**
10	1	1	1	1	**4**
11	1	1	1	1	**4**
12	1	1	1	1	**4**
13	1	1	1	1	**4**
14	1	1	1	1	**4**
15	1	1	1	1	**4**
p	0.667	0.733	0.800	0.867	
Rit	0.891	0.943	0.922	0.812	
$DI_{27\%}$	1.000	1.000	0.750	0.500	

1b) minor stochastic error

Test-taker	**Item1**	**Item2**	**Item3**	**Item4**	**Score**
1	0	0	0	0	**0**
2	**1**	0	0	0	**1**
3	0	0	0	1	**1**
4	0	0	1	1	**2**
5	0	1	1	1	**3**
6	**0**	1	1	1	**3**
7	1	1	1	1	**4**
8	1	1	1	1	**4**
9	1	1	1	1	**4**
10	1	1	1	1	**4**
11	1	1	1	1	**4**
12	1	1	1	1	**4**
13	1	1	1	1	**4**
14	1	1	1	1	**4**
15	1	1	1	1	**4**
p	0.667	0.733	0.800	0.867	
Rit	0.668	0.930	0.896	0.751	
$DI_{27\%}$	0.750	1.000	0.750	0.500	

From Table 1a we note that the item–total correlations (*Rit* = 0.81–0.94) are reasonably high. The latter indicates high item discrimination though the estimates do not reach the perfect 1; the algebraic reason for this underestimation is formalized in Metsämuuronen (2016). The corresponding values of *DI* vary from 0.5 to 1. It is worth noting that in two out of four items, *DI* detects the deterministic pattern in the items. Hence, *DI can* reach the ultimate value of +1 (as well as −1) correctly while *Rit* always underestimates the IDP in the real-life testing settings.

Let us assume that *two* of the test-takers were marked incorrectly (or they, unexpectedly given their ability level, gave correct and incorrect answer) in *item*1 as in Table 1b. Although the difficulty level of the items (proportion of correct answers, *p*) did not change in the process, the magnitude of the estimates by *Rit* decreased in *all* items (up to 0.22 units of correlation) even though there were no changes in items 2, 3, or 4. In contrast, with *DI*, though the magnitude of the estimate for *item*1 has reduced from 1.00 to 0.75 (0.25 units of correlation) the magnitude of the estimates in items 2, 3, or 4 did not change because the order of the test-takers did not change in the process.

The stable character of *DI* is caused by three reasons. First, because the middle-range observations are not used during the calculation of *DI,* the changes in those observations do not change the estimate of IDP. Second, because the correct and incorrect responses can be in any order within the cut-off, *DI* is more robust for the changes of the item structure than *Rit*. Third, because *DI* uses the score only to order the test-takers, the changes in the actual score do not necessarily affect the value of *DI* of the remaining items if the *order* of the test-takers did not change radically.

Though *DI* produces stable estimates, and it can reach the ultimate values correctly, it has two main deficiencies. First, when using the traditional cut-offs of 27% or 25% , IDP may be underestimated radically when the item difficulty is extreme (0.20 > *p* > 0.80; see also Tristan, 1998) as seen with *item*4 (*DI* = 0.50 while *Rit* = 0.812). With items with the

extreme difficulty level, it would be wise to use either another index for IDP or to use another cut-off than the traditional 27% or 25%. All in all, though *DI can* detect the deterministic patterns, it is good to note that there would be better options than *DI* to detect the deterministic patterns. One of these is the nonparametric and directed coefficient of correlation Somers' *D* (Somers, 1962; Metsämuuronen, 2020). In the case of Table 1a, Somers' *D* would detect the deterministic pattern for *all* the items. Second, maybe more crucially, *DI is developed only for the binary items*. The next section generalizes *DI* to allow polytomous responses and several cut-offs.

Generalized DI

Two things are worth pointing from the previous discussion concerning *DI*: first, the classical form of *DI* can be used only with dichotomous items and, second, the cut-offs for *DI* are not deterministically fixed. Hence, generalized *DI* to allow polytomous responses and several cut-offs is discussed and derived in this section. Here the suggestion is given based on the original *DI*. We may note that Brennan (1972) introduced another kind of generalized upper-lower item discrimination index based on Kelly's *DI*. Brennan's *B* is generalized in the sense that the cut-off need not be symmetric. However, Brennan's *B* is still restricted to dichotomous items and uses a fixed cut-off. Harris and Wilcox (1980) showed that Brennan's *B* equals algebraically to Peirce's Theta discussed by Goodman and Kruskal (1959).

General notation for Generalized DI

To formalize the generalized *DI,* a slightly modified notation and radically different operationalization of the symbols are suggested. In a general case, in a specific cut-off *a, DI* can be written as follows:

$$DI_a = \frac{R_a^U - R_a^L}{\frac{1}{2}T_a} = 2\left(p_a - 2p_a^L\right) \quad (4)$$

(cf. Metsämuuronen, 2017, p. 125, that uses the symbol *i* instead of *a*) where the subscript *a* refers to the symmetric cut-off used in the estimation. Assume that we have 20 test-takers and we use 25% cut-off in the calculation of *GDI*. Then, *a* refer to (1) the percentage of the cut-off, such as *a* = 25%, (2) the number of cases in the specific cut-off, such as *a* = 5, and (3) to the actual rank-order of the cut-off, such as *a* = 5 = 5th case, where 5 refers to the rank of the very test-taker in the ordered dataset that benchmarks the 25% cut-off.

At this point, it is good to raise a potential challenge in the calculation of *GDI* (as well as *DI*). Forming the order for *DI* and *GDI* is based on the idea that the test-takers are ranked in a uniquely unambiguous manner, that is, each test-taker has a separate rank order. However, this is not possible when there are ties in the score because, within the tied cases, we do not know the actual order of the test-takers without some relevant rationale. To solve this challenge, an option, relevant within the achievement testing, is suggested to be considered. To acquire the unambiguous order for the cases, the test-takers are double-ordered, first, by the score and, second, by the items—or by other relevant information such as the time used in the task. In the latter ordering, the more difficult items (or less used time in task) are given more weights. The rationale behind the suggestion is that the same score, except in the case of identical profile of answers, is an outcome of a compilation of *different items*. We may think that the test-taker who was able to solve more demanding tasks (or by using less time in task) showed slightly higher achievement than the test-taker with the same score but by solving less-demanding tasks (or by using more time in task). This rationale is embedded to the estimation of IRT modeling; however, its routines are not fully used even in those settings. Another option to solve the challenge, a practical one though less accurate, is to trust the randomization in the ordering: all cases are given their own, unambiguous rank order based on the score, but the tied cases are in a random order. Third option is to include all the test-takers with the same score into the same bin (either L or U) and to change the cut-off dynamically. This requires, however, wide scale in the score to make it possible to keep the symmetricity in the number of test-takers in the cut-offs. Finally, one option is to develop a new coefficient based on non-symmetric cut-offs with polytomous items (cl. Brennan, 1972).

New operationalization of the concepts related to GDI

In order to generalize *DI* to the polytomous scales, new operationalizations of the concepts of R^U, R^L and T are needed. Some new symbols are also used. In what follows, the observed values (O) of the test-takers *i* in the item in the ordered data will be of interest. O_i^U refers to the observed value in the item of the *i*th highest test-taker in the upper half (U) of the ordered data, and O_i^L refers to the observed value in the item of the *i*th lowest test-taker in the lower half (L) of the ordered data. For example, O_2^U is the value for the second highest-scoring test-taker in the item and O_2^L is the value for the second lowest-scoring test-taker in the item.

The first main note to make is that, in the general notation for *DI* in Eq. (4), R_a^U is not the *number* of correct answers in the upper half of the cut-off of 25% or 27% of the highest test-takers. In the general case, R_a^U is the *sum of the observed values* of the test-takers from the highest to the a^{th} highest case

$$R_a^U = \sum_{i=1}^{a} O_i^U \tag{5}$$

and, parallel, R_a^L is the sum of the observed values of the test-takers from the lowest to the a^{th} lowest case

$$R_a^L = \sum_{i=1}^{a} O_i^L \ . \tag{6}$$

In the dichotomous cases, these sums equal the number of 1s in the upper and lower halves of the ordered and reduced data. For the general case, however, this re-operationalization is essential.

Second, in the traditional formula for *DI* (see Eq. 3 and related discussion), *T* refers to the total *number* of test-takers in the reduced data in the specific cut-off *a*, and therefore $\frac{1}{2}T_a$ in Eq 4 refers to the number of test-takers in half of the cut-off *a*. However, in the general case, *T* does not refer to the number of cases but to the *maximum possible sum minus the minimum possible sum* of the observed values of test-takers in the specific cut-off *a*:

$$\begin{aligned} T_a &= max\left[\sum_{i=1}^{a}(O_i^L + O_i^U)\right] - min\left[\sum_{i=1}^{a}(O_i^L + O_i^U)\right] \\ &= max\left[\sum_{i=1}^{a} O_i^L\right] + max\left[\sum_{i=1}^{a} O_i^U\right] - min\left[\sum_{i=1}^{a} O_i^L\right] - min\left[\sum_{i=1}^{a} O_i^U\right]. \end{aligned} \tag{7}$$

where *max[.]* refers to the maximum possible value and *min[.]* refers to the minimum possible value. This definition is obvious when generalizing *DI* to the polytomous items where the minimum value of the item scale is something else than zero, such as in the Likert scale anchored to the values 1 to 5. In the general case, the maximum possible value is the same for all test-takers, and it is the maximum value in the item *g*:

$$max(O_i) = max(O_j) = max(g). \tag{8}$$

Parallel, the minimum possible value is the same for all individual test-takers, and it is the minimum value in the item:

$$min(O_i) = min(O_j) = min(g). \tag{9}$$

Because of (8) and (9), the elements $max\left[\sum_{i=1}^{a}\left(O_i^U\right)\right]$ and $min\left[\sum_{i=1}^{a}\left(O_i^L\right)\right]$ in Eq. (7) can be manipulated as follows:

$$\begin{aligned} max\left[\sum_{i=1}^{a}\left(O_i^U\right)\right] &= max(O_1) + max(O_2) + \ldots + max(O_a) \\ &= max(g) + max(g) + \ldots + max(g) \\ &= a \times max(g) \end{aligned} \tag{10}$$

where *a* refers to the number of cases in half of the reduced dataset. Parallel,

$$min\left[\sum_{i=1}^{a}\left(O_i^U\right)\right] = a \times min(g). \tag{11}$$

Thus, because of (7), (10), and (11)

$$\tfrac{1}{2}T_a = a\left[max(g) - min(g)\right] \tag{12}$$

where *a* refers to the number of observations in the half of the specific cut-off *a* and $max(g) - min(g)$ is the *range* of the values in the scale of the item.

Because of (4), (5), (6), and (12), *GDI* in a specific cut-off *a* is:

$$GDI_a = \frac{R_a^U - R_a^L}{\frac{1}{2}T_a} = \frac{\sum_{i=1}^{a} O_i^U - \sum_{i=1}^{a} O_i^L}{a\left[max(g) - min(g)\right]}. \quad (13)$$

Because the terms *max*(*g*) and *min*(*g*) in (13) are constant for all cut-offs, Eq. (13) can be rewritten:

$$GDI_a = \frac{R_a^U - R_a^L}{\frac{1}{2}T_a} = \frac{\sum_{i=1}^{a} O_i^U - \sum_{i=1}^{a} O_i^L}{a \times C} \quad (14)$$

where the constant *C* is the range of the values in the scale of the item.

Numerical example of calculating GDI with polytomous items

As a numerical example of calculating *GDI*, assume a polytomous dataset with *N* = 25 cases as in Table 2a. The dataset is from Cox (1974, p. 177) and Drasgow (1986, p. 70) without a connection with item analysis. However, let us assume that the dataset would relate with item *g* and the score *X*.

Table 2a . Hypothetic dataset ordered by the score (Cox, 1974; Drasgow, 1986)

g	X	g	X	g	X	g	X	g	X
1	69	1	87	0	88	2	96	1	104
0	72	1	80	1	88	2	96	1	104
1	72	1	81	1	92	1	99	2	104
1	77	2	85	1	92	1	101	1	108
1	78	1	86	1	93	2	103	0	112

Used by permisson of Biometric Society

Table 2b. Statistics for estimating item discrimination by GDI

cut-off (a)	cut-off (%) *N* = 25	cut-off (%) *N* = 24	$\frac{1}{2}T_a = a \times C = a \times 2$	$R_a^U = \sum_{i=1}^{a} O_i^U$	$R_a^L = \sum_{i=1}^{a} O_i^L$	GDI_a
1	4	4.2	2	0	1	-0.500
2	8	8.3	4	1	1	0.000
3	12	12.5	6	3	2	0.167
4	16	16.7	8	4	3	0.125
5	20	20.8	10	5	4	0.100
6	**24**	25	**12**	**7**	**5**	**0.167**
7	**28**	29.2	**14**	**8**	**6**	**0.143**
8	32	33.3	16	9	8	0.063
9	36	37.5	18	11	9	0.111
10	40	41.7	20	13	10	0.150
11	44	45.8	22	14	10	0.182
12	48	50	24	15	11	0.167

Table 2b shows the statistics of all the possible cut-offs related to Table 2a. We note that, because of having an odd number of test-takers, the median case is omitted in the process. Hence, the cut-off percentage may end up 48 or 50 depending on whether 24 or 25 cases are considered. If 25 is considered as the sample size, we have two option for the estimation of item discrimination: 24% cut-off or 28% cut-off. The 24% cutoff gives the estimate

$$GDI_{24\%} = GDI_6 = \frac{R_{24\%}^U - R_{24\%}^L}{\frac{1}{2}T_{24\%}} = \frac{\sum_{i=1}^{6} O_i^U - \sum_{i=1}^{6} O_i^L}{6 \times (2-0)} = \frac{7-5}{12} = 0.167 \quad (15)$$

and the 28% cutoff gives the estimate

$$GDI_{28\%} = GDI_7 = \frac{\sum_{i=1}^{7} O_i^U - \sum_{i=1}^{7} O_i^L}{7 \times (2-0)} = \frac{8-6}{14} = 0.143 \tag{16}$$

As benchmarks, some other estimators are referred here to in relation with Table 2a. The estimates of the observed association between the item and score, based on the mechanics of Pearson's product-moment correlation are by item–total correlation coefficient *Rit* = 0.185 and, after corrected for the inflation, by item–rest correlation coefficient *Rir* = 0.139. The estimate of the inferred association by polyserial correlation coefficient is ρ_{PS} = 0.216 and the corresponding estimate by the polychoric correlation coefficient is ρ_{PC} = 0.123 though the last value depends of the estimation method in some extent. Somers' *D* would give an estimate of *D*(*g*|*X*) = 0.219. With the dataset in Table 2a, the estimates by *GDI*—both 0.167 and 0.143—are at the same range as those by *Rit*, *Rir*, and ρ_{PC}. In any case, the discrimination power of the item is low: it cannot efficiently differentiate between the higher and lower scoring test-takers. Traditionally, this item would be considered as one of those to be omitted from the final compilation of the items.

An alternative way of computing GDI

It may be an obvious fact, though worth formalizing that, in each cut-off *a*+1 after the previous cut-off *a,* the next value of *GDI* is determined by the value of the *next pair* of individual test-takers in the upper half (O_{a+1}^U) and in the lower half (O_{a+1}^L). Namely, Because of (5) and (6),

$$\begin{aligned} R_a^U - R_a^L &= \sum_{i=1}^{a} O_i^U - \sum_{i=1}^{a} O_i^L \\ &= (O_1^U + O_2^U + \ldots + O_i^U + \ldots + O_a^U) - (O_1^L + O_2^L + \ldots + O_i^L + \ldots + O_a^L) \\ &= (O_1^U - O_1^L) + (O_2^U - O_2^L) + \ldots + (O_i^U - O_i^L) + \ldots + (O_a^U - O_a^L) \end{aligned} \tag{17}$$

Then, $R_{a+1}^U - R_{a+1}^L = \sum_{i=1}^{a} O_i^U - \sum_{i=1}^{a} O_i^L + (O_{a+1}^U - O_{a+1}^L) = R_a^U - R_a^L + (O_{a+1}^U - O_{a+1}^L)$.

A new concept, *characteristic vector of the item* ***D*** is introduced. When the values of the i^{th} test-takers are subtracted, the difference is symbolized by D_i:

$$D_i = O_i^U - O_i^L \tag{18}$$

The vector **D** consists of these differences in *all* cut-offs *i*. The total number of these cut-offs is ½*N;* if *N* is odd, the median case is omitted in the analysis. Then, obviously, the number of cut-offs is ½(*N*–1). For the simplicity reasons, let us assume that **D** has ½*N* elements

$$\mathbf{D} = \left(D_1, D_2, \ldots, D_a, \ldots, D_{\frac{1}{2}N}\right). \tag{19}$$

The later computational form of *GDI* uses the *sum* of elements in **D** from the extreme cut-off (*i* = 1) to the particularly interesting cut-off *i* = *a*:

$$S_a = \sum_{i=1}^{a} D_i \tag{20}$$

Because of (17), (18), (19), and (20),

$$R_a^U - R_a^L = \sum_{i=1}^{a} O_i^U - \sum_{i=1}^{a} O_i^L = D_1 + D_2 + \ldots + D_a = \sum_{i=1}^{a} D_i = S_a \tag{21}$$

Eq. (21) can be used in forming an alternative way of computing *GDI* at any cut-off *a*. Namely, according to (13), (21), and (14), the value of the *GDI* in the specific cut-off *a* is

$$GDI_a = \frac{R_a^U - R_a^L}{\frac{1}{2}T_a} = \frac{\sum_{i=1}^{a} D_i}{\frac{1}{2}T_a} = \frac{S_a}{\frac{1}{2}T_a} = \frac{S_a}{a\left[\max(g) - \min(g)\right]} = \frac{S_a}{a \times C} \tag{22}$$

where *a* is the number of test-takers in the half of the cut-off and *C* is a constant referring to the range in the scale of the item.

Numerical examples of using the alternative routine

Let us assume the same dataset as in Table 2a. The alternative way of calculating *GDI* is illustrated in Table 3.

Table 3. Statistics for the alternative way of estimating item discrimination by GDI

cut-off (a)	**cut-off (%) *N* = 25**	**cut-off (%) *N* = 24**	O_i^U	O_i^L	$D = O_i^U - O_i^L$	$GDI_a = \frac{1}{a \times 2}\sum_{i=1}^{a} D_i$
1	4	4.2	0	1	-1	-0.500
2	8	8.3	1	0	1	0.000
3	12	12.5	2	1	1	0.167
4	16	16.7	1	1	0	0.125
5	20	20.8	1	1	0	0.100
6	**24**	**25**	**2**	**1**	**1**	**0.167**
7	**28**	**29.2**	**1**	**1**	**0**	**0.143**
8	32	33.3	1	1	-1	0.063
9	36	37.5	2	2	1	0.111
10	40	41.7	2	1	1	0.150
11	44	45.8	1	0	1	0.182
12	48	50	1	1	0	0.167

The 24% cut-off gives the estimate

$$GDI_{24\%} = \frac{\sum_{i=1}^{6} D_i}{6 \times \left[\max(g) - \min(g)\right]} = \frac{-1+1+1+0+0+1}{6 \times (2-0)} = \frac{2}{12} = 0.167 \qquad (23)$$

and the 28% cutoff gives the estimate

$$GDI_{28\%} = \frac{\sum_{i=1}^{7} D_i}{7 \times \left[\max(g) - \min(g)\right]} = \frac{-1+1+1+0+0+1+0}{7 \times (2-0)} = \frac{2}{14} = 0.143. \qquad (24)$$

Obviously, the estimates are the same as in Eqs. (15) and (16). Again, we have two options for the estimate of item discrimination: 0.167 related to 24% cut-off or 0.143 related to 28% cut-off. Either way, the discrimination power of the item is low: it cannot differentiate between between the higher and lower scoring test-takers. The interpretation of the value of *GDI* is the same as that of the traditional *DI*; the same benchmarks for the low, mediocre, or high item discrimination can be used with both indices.

Further elaboration of GDI

Some further elaborations of *GDI* are discussed in what follows. This includes the procedure of exhaustive splitting (PES) and a new way of illustrating the behavior of the item called the cut-off curve (COC).

Exhaustive Splitting Procedure

A new tool for the further elaboration of the *GDI* is the procedure for exhaustive splitting already employed in Tables 2b and 3. PES is not a necessity in the actual calculation of *GDI* though it may offer a possibility to perform more effective computation and more refined items analysis. In the manual calculation, the formulae can be used without exhaustive splitting.

In PES, instead of using only one fixed cut-off (25% or 27%), *all* possible cut-offs can be used in the item analysis. PES is as follows:

1. Take the ultimately highest and the ultimately lowest observation from the sorted data and calculate *GDI*.
2. Save the discrimination result from this calculation.
3. Take the two highest and the two lowest observations from the sorted data and calculate *GDI* (as in 1). Save the results.
4. Repeat steps 1 and 2, increasing the number of observations and gradually building up to *½N =50%* of the observations at both extremes. When there are odd number of cases, the median case is left outside of the procedure.

A table or graph of the results can be made, and this may be helpful in visualizing the characteristics of the items. In what follows, some relevant graphs are introduced as the discussion turns to the characteristics of the *GDI*. It is worth noting that PES is not restricted to *DI* or *GDI*; Metsämuuronen (2017), for example, used the same idea in illustrating the differences between underestimation in item–test correlation and item-rest correlation in comparison with *DI*.

Cut-off curve

Though visualization is not necessity in understanding the concept of *GDI*, the approach to *GDI* hereafter is easier to adopt with the assistance of graphical demonstrations. The concepts of the 'cut-off curve' (COC) are therefore introduced. Though the concept is not restricted to dichotomous datasets, for the sake of simplicity, dichotomous items are given as examples.

Based on PES, we can form a graphical illustration of the values. This graph is called the cut-off curve (COC) though no actual "curve" exists because the distribution of the estimates is not continuous one. As a preliminary introduction to the graphical item analysis with COC, we assume a Guttman-patterned easy item of $N = 18$ test-takers ordered by the (unseen) score. From the lowest to the highest test-taker, the string is as follows: 000011111 | 111111111—the middle point is marked by a bar. Using the procedure of exhaustive splitting, there are $½N = 9$ possible symmetric extreme cut-offs as shown in Table 4 and Figure 2.

Table 4. All symmetric cut-offs and relevant indicators related to GDI

T_a	$a = ½T_a$	$R_a^L = \sum_{i=1}^{a}(O_i^L)$	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	$R_a^U = \sum_{i=1}^{a}(O_i^U)$	$GDI_a = (R_a^U - R_a^L)/a$
2	1	0	0																	1	1	1.00
4	2	0	0	0															1	1	2	1.00
6	3	0	0	0	0													1	1	1	3	1.00
8	4	0	0	0	0	0											1	1	1	1	4	1.00
10	5	1	0	0	0	0	1									1	1	1	1	1	5	0.80
12	6	2	0	0	0	0	1	1							1	1	1	1	1	1	6	0.67
14	7	3	0	0	0	0	1	1	1					1	1	1	1	1	1	1	7	0.57
16	8	4	0	0	0	0	1	1	1	1			1	1	1	1	1	1	1	1	8	0.50
18	9	5	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	9	0.44

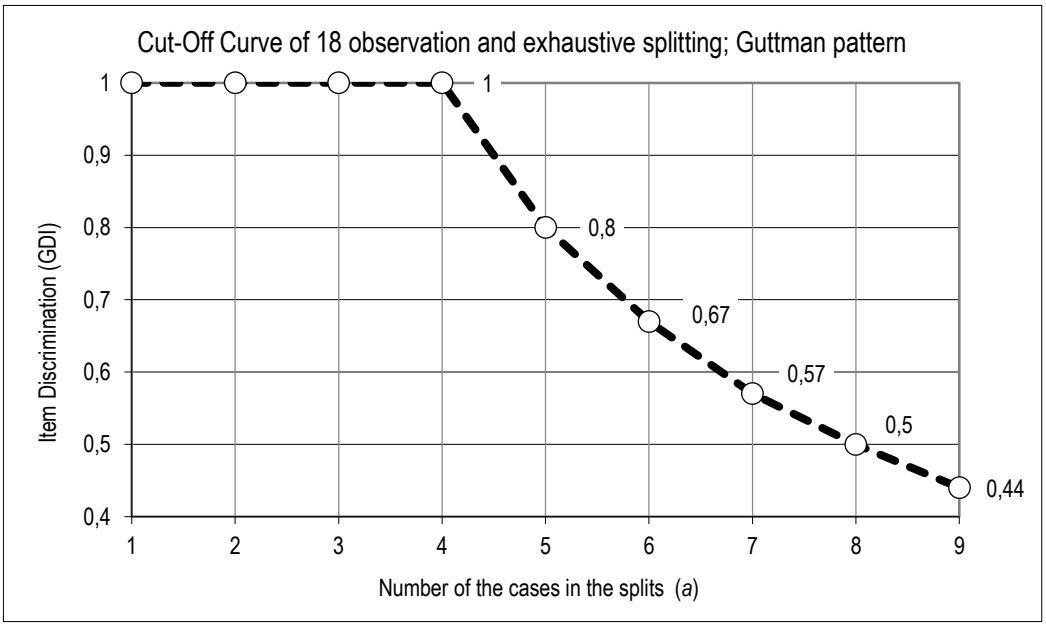

Figure 2. Cut-point curve for a Guttman-patterned item with N = 18 and 4 zeros as a function of a

In COC, the threshold point of a Guttman-patterned item is seen as the point at where the ultimate discrimination (*GDI* = 1) changes dramatically and becomes smaller. In the Guttman-patterned case, this equals with the shorter of the ultimate strings of 0s and 1s; in Figure 2, $a = 4$ is the threshold for the curve. It may be worth noting again that even though it is possible to connect the discrete points together, no actual continuous curve exists in Figure 2. However, connecting the points visualizes the concept of the COC as a "curve".

Another simple example of COC, with a non-Guttman type pattern, illustrates the pattern of COC that leads us to a more practical question of how the value of *GDI* is determined and dependent on the underlying Guttman pattern. Assume a dataset as in Table 5 with five Guttman patterned items with different difficulty level (items 1 to 5) and one non-Guttman-patterned item (item 6). The string of the non-Guttman-patterned item is 00100|01111 after ordered the test-takers by the (unseen) score. After the exhaustive splitting, we get 5 cut-offs for each item. In item 6, two extreme cut-offs show perfect discrimination after which the magnitude of *GDI* starts to get lower though not as logically as in Figure 2. The COCs are illustrated in Figure 3. The moves of the COC of the item 6 are elaborated in what follows.

Table 5. Guttman-pattern and non-Guttman-pattern

test-taker	Item 1 Guttman	Item 2 Guttman	Item 3 Guttman	Item 4 Guttman	Item 5 Guttman	Item 6 Non-Guttman
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	1
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	1	0
7	0	0	0	1	1	1
8	0	0	1	1	1	1
9	0	1	1	1	1	1
10	1	1	1	1	1	1
p	0.1	0.2	0.3	0.4	0.5	0.5

Item discrimination in the cut-off (*GDI*)						
Cut-off (a)	**item1**	**item2**	**item3**	**item4**	**Item5**	**Item6**
1	1	1	1	1	1	1
2	0.5	1	1	1	1	1
3	0.33	0.67	1	1	1	0.67
4	0.25	0.50	0.75	1	1	0.75
5	0.20	0.40	0.60	0.08	1	0.60

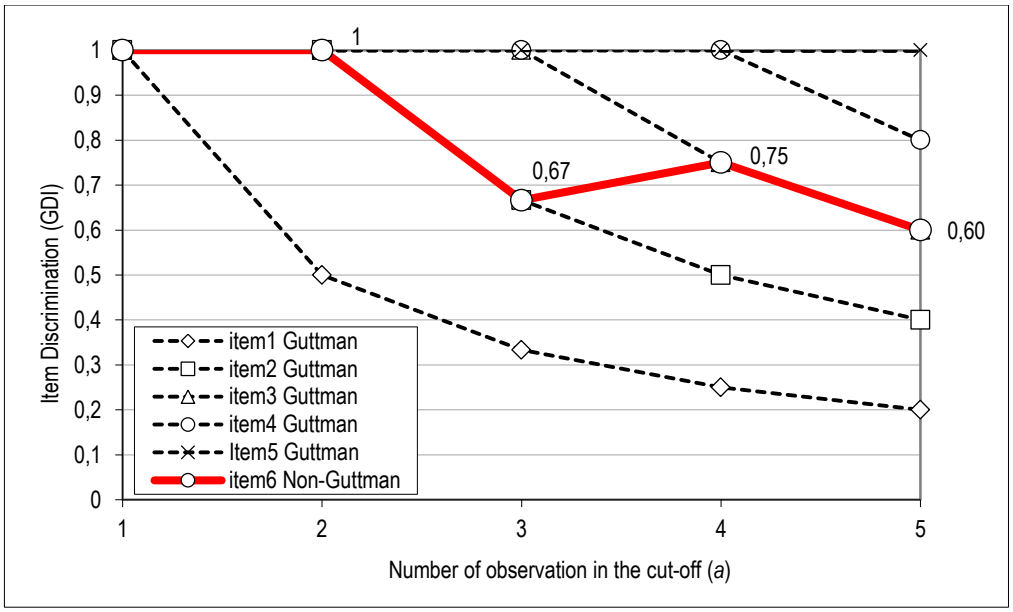

Figure 3. Guttman-pattern and non-Guttman-pattern in the COC

The first non-obvious note to make in Figure 3 is that COC of a non-Guttman-patterned (real-world) item strictly follows the underlying grid formed by the (underlying) Guttman-patterned items. Another non-obvious note to make is that COC of a single item detects the deviations in the non-Guttman pattern in the dataset as a shift in COC of the

underlying Guttman-patterned items in the consecutive cut-offs *a* and *a* + 1. In each point *a,* the COC have only limited options to go because, in each cut-off *a* + 1, the values of *GDI* are determined by the values in the previous cut-off *a.* This determination can be visualized by using COC. These matters are formalized in what follows.

Determination of the value of GDI and the moves in COC

From the practical viewpoint related to binary items, when $D_{a+1} = 1$, the path in COC moves forward to the *next* underlying curve of a Guttman-patterned item (Figure 4). If the result is $D_{a+1} = 0$, the next step will be on the *same* underlying curve as the previous point D_a (*i.e.*, no change in the underlying curve of a Guttman-patterned item). When the result is (theoretically pathological) $D_{a+1} = -1$, the path leads to the *previous* underlying curve (i.e. the path goes to the next cut-off but backwards to the curve of the previous Guttman-patterned item).

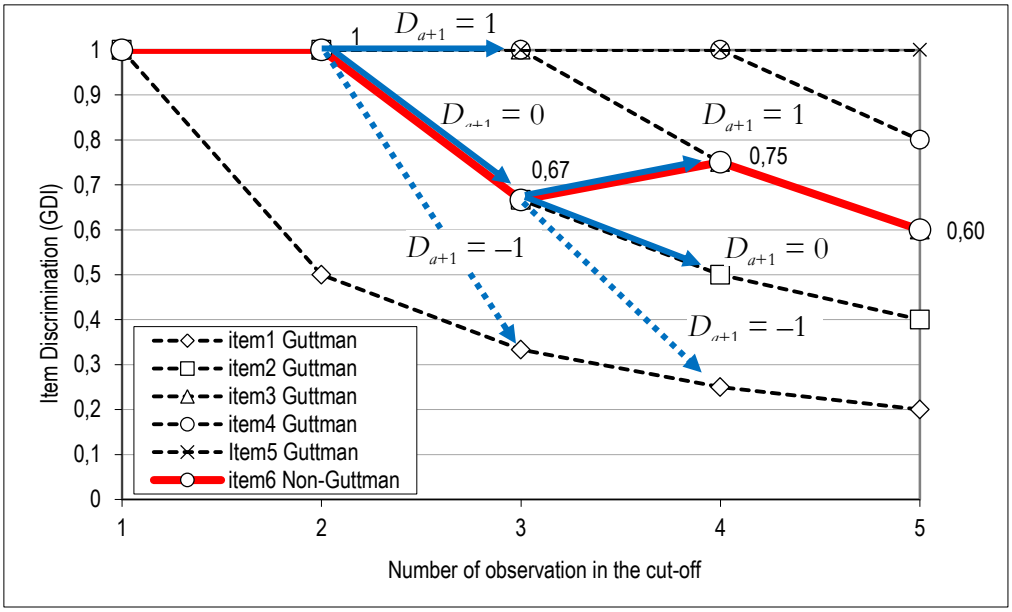

Figure 4. Determination of the value of GDI and COC by the term D_{a+1}

From Eqs. (17), (18), and (19), it is known that the value of *GDI* in the next cut-off *a* + 1 depends on the element D_{a+1}, that is, on the next pair of observations O_{a+1}^U in the upper part and O_{a+1}^L in the lower part. Hence, because the possible values of the element D_{a+1} are limited, the value of GDI_{a+1} and the next move in COC is also strictly limited. In the dichotomous case, at every cut-off *a* + 1 that follows the previous one, the term D_{a+1} can have (only) one of four possible outcomes:

$$D_{a+1} = \begin{cases} 1, when\ O_{a+1}^U = 1\ and\ O_{a+1}^L = 0 \\ 0, when\ O_{a+1}^U = 1\ and\ O_{a+1}^L = 1 \\ 0, when\ O_{a+1}^U = 0\ and\ O_{a+1}^L = 0 \\ -1, when\ O_{a+1}^U = 0\ and\ O_{a+1}^L = 1 \end{cases} \quad (25)$$

Of the three actual outcomes of the term D_{a+1} in (25), the option –1 reflects an outcome of a pathological situation where the lower-scoring test-taker shows a higher response in the item than the higher-scoring counterpart. In the achievement testing this means that a lower-scoring test-taker gives, by a mistake or by a lucky guessing, a correct answer while a higher scoring test-taker gives, by carelessness or sleepiness, unexpectedly given the ability level, an incorrect answer (see the verbal descriptions in Linacre & Wright, 1994) . The option +1 refers to the expected output that the higher-scoring test-takers would give a correct answer while the lower-scoring test-takers would give an incorrect answer. The option 0 comes when the test-takers in both halves give the same value—either correct or incorrect one. All in all, in the dichotomous case, the value of GDI_{a+1} have one of the three fixed options:

$$GDI_{a+1} = \frac{S_a + 1}{(a+1)\cdot C}, \frac{S_a}{(a+1)\cdot C}, \text{or } \frac{S_a - 1}{(a+1)\cdot C}.$$

In the general case, the term D_{a+1} has $2\left[max\left(g\right)-min(g)\right]+1$ options. As an example, in the case of Likert-type of item anchored to values 1–5, D_{a+1} has $2\left[max\left(g\right)-min(g)\right]+1=2\times\left(5-1\right)+1=9$ possible options:

$$D_{a+1}=\begin{cases}+4, when\ O_{a+1}^{U}=5\ and\ O_{a+1}^{L}=1\\ +3, when\ \ O_{a+1}^{L}=O_{a+1}^{U}-3\\ +2, when\ O_{a+1}^{L}=O_{a+1}^{U}-2\\ +1, when\ O_{a+1}^{L}=O_{a+1}^{U}-1\\ 0, when\ O_{a+1}^{U}=O_{a+1}^{L}\\ -1, when\ O_{a+1}^{L}=O_{a+1}^{U}+1\\ -2, when\ O_{a+1}^{L}=O_{a+1}^{U}+2\\ -3, when\ O_{a+1}^{L}=O_{a+1}^{U}+3\\ -4, when\ O_{a+1}^{U}=1\ and\ O_{a+1}^{L}=5\end{cases} \quad (26)$$

The negative values represent options reflecting illogical and pathological behavior in the dataset that may lead to negative item discrimination if being many within one item. All in all, the negative values in the characteristic vector ***D*** in Eq. (19) are strictly indicative for the pathological cases. This matter is elaborated in the next section.

Pathological patterns in the visual item analysis with COC

The pathological cases characterized by negative elements ($D_i = -1$) in the characteristic vector ***D*** can be detected easily by using PES, and those can be seen in COCs. The frequency of the elements $D_i = -1$ in the characteristic vector ***D*** directly indicates the number of pathological pairs of test-takers in an item as discussed above. When the number of these negative pairs is higher than the positive pairs, the value of *GDI* turns to be (pathologically) negative: higher-scoring test-takers appear to give the wrong answer in the items more likely in comparison with the lower-scoring test-takers.

The examples given above were based on rather small and theoretical datasets; it is easy to illustrate the graphs when the number of cases is small. However, the exhaustive splitting procedure and cut-off curves are not restricted to small datasets. As an example of a larger dataset, a random sample of 200 real-world test-takers of a test of the national assessment of mathematics in Finland (FINEEC, 2018) is used as a basis for the illustration. Figure 5 illustrates this kind of COCs of two items with 100 cut-offs. The underlying Guttman-patterned items are shown as lighter lines. The pathological cases within the process are the rare cases where the COC moves to the previous Guttman-patterned latent curve (cf. Figure 4).

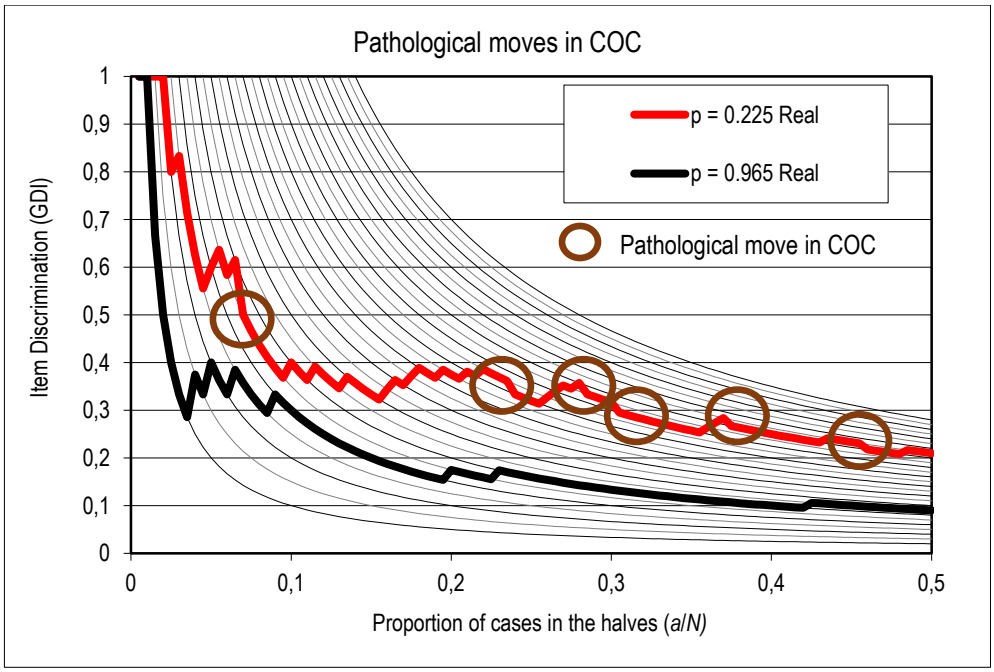

Figure 5. Pathological moves in COC in two real-life items with non-Guttman-pattern (n = 200)

Using PES and COC in the detection of plausible and stable values for GDI

Figure 5 and the underlying exhaustive splitting procedure raise a natural question of how stable and plausible is our point-estimate of IDP if we used only a single cut-off? In practical terms, if the estimate of IDP at the point of a = 25% would be $GDI_{25\%}$ = 0.167 and at the point of a = 27% $GDI_{27\%}$ = 0.143, which of those would be the most credible estimate and why? Could we find a better or more credible estimate? Or an estimate of variance or standard deviation for the point estimate? Would a confidence interval of the estimate enrich our decision making in the analysis of the item behaviour by using GDI? Some initial ideas are discussed here though no final conclusion is reached.

We remember that the idea in item discrimination is to answer how well the item can discriminate the higher scoring and lower scoring test-takers from each other. Traditional *DI* compares the item behaviour of the test-takers in the highest quartile to the lowest quartile and gives an estimate of the general behaviour of the item based on this estimate. On the other hand, by using PES, we would know all the possible estimates related to the same dataset. How could we utilize this kind of information in assessing the discrimination power of the item?

Let us draw COC based on Tables 2b and 3 (Figure 6). By using the PES and COC, we could assess how stable the estimate of 25% or 27% is—or whether there would be some other cut-off that could show evidence of more credible estimate of IDP than what is signalled in the cut-offs of 25% or 27%. Graphical diagnosis of Figure 5 tells that we would not find remarkably credible higher estimates for the item discrimination by *GDI* in any of the cut-off in comparison with the traditional cut-offs. The estimate is very stable between 12% cut-off and 48% cut-off, and the magnitude remains below 0.20 in all cut-offs.

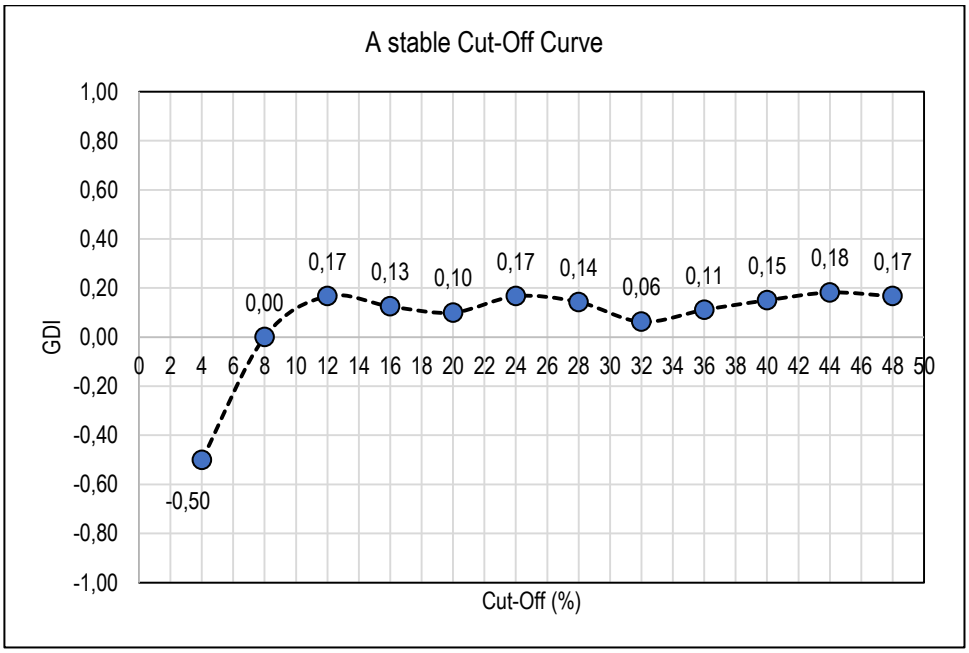

Figure 6. Stability of the estimates by GDI related to Table 2a

Another sample of the use of graphical diagnostics with COC, based on larger sample (Figure 7), gives a clue that, when using the 27% cut-off with the difficult item (p = 0.225), the estimate for IDP is $GDI_{27\%}$ = 0.35 and this seems to be fairly stable estimate. Just by using the graphical possibilities and intuitive heuristics, we may conclude that the value seems quite stable between the cut-offs 10% to 30% ranging from 0.32 to 0.40. The other item in Figure 7, the 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.

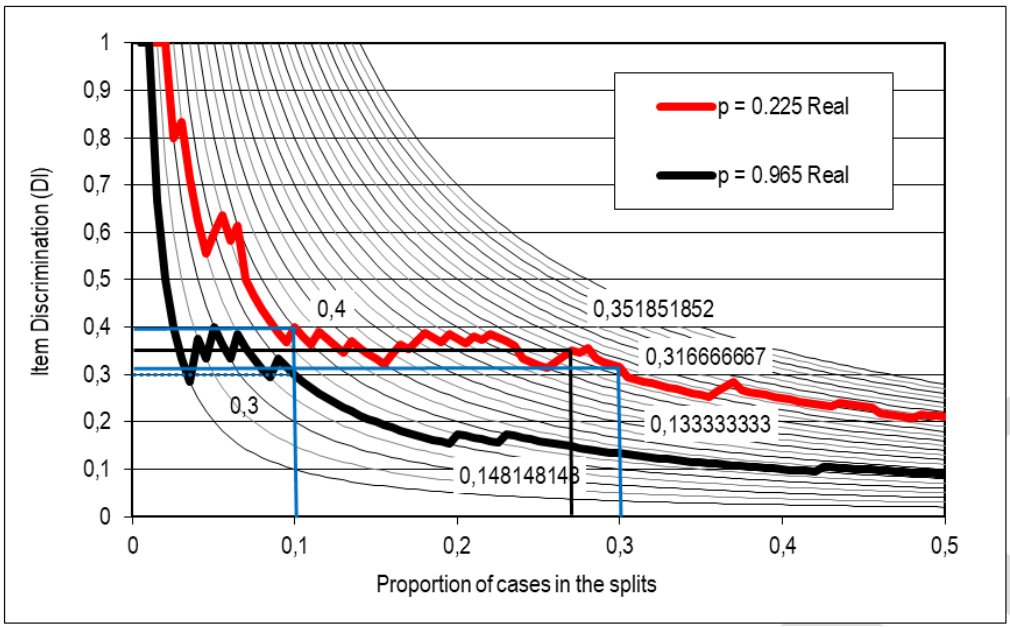

Figure 7. Stability of the estimate of two real life items with non-Guttman-pattern (n = 200)

By using the values of estimates by PES, we could easily compute the average values of the estimates and, hence, the variance of the estimates and consequently, standard errors and the confidence intervals for the estimates. However, this matter is not formalized in this article, or any boundaries are suggested, though we note that PES gives possibilities to develop such tools for the IDP.

Conclusions and Discussion

Results in a nutshell and related discussion

This article has discussed Kelley's DI as a simple nonparametric alternative short-cut method in estimating the item discrimination power in the practical testing settings in a rough manner. Unlike Pearson correlation, DI can reach the ultimate values ± 1 accurately when needed if the item difficulty is of medium level (roughly $0.25 < p < 0.75$), and it is a more stable index in comparison with Rit. Although there would be better options from both underestimation and instability viewpoints—such as Somers' D that can reach the ultimate value also with the extremely easy or difficult items—the advantage of DI is its computational simplicity that makes is applicable in the practical testing settings in schools and other applied areas when the sophisticated tools for item analysis are not in use. This article generalizes DI to allow polytomous item scales in analysis. The generalized DI (GDI) can be applied in a wider sense also in analyzing, for example, the attitude scales or graded items in the achievement tests.

The generalization required new operationalizations of the traditional elements of *DI*, R^U, R^L and *T*. These allow us not only to use polytomous item scales but also varying cut-offs in the item analysis. Hence, the name "generalized *DI*", *GDI*, seems relevant. A new computational method based on the concept of characteristic vector of the item is initiated for the computer-based analysis; the classical way of calculating the value of *GDI* is still valid for the manual calculation. Additionally, a new method of visualizing the item analysis results, the cut-off curves (COC) with the related exhaustive splitting procedure (PES) are initiated in the article. The former can be used in a graphical analysis of the item, detecting the pathological cases in the dataset as well as in assessing the plausibility of the obtained point estimate of *GDI*. The latter application is not, however, elaborated or formalized in this article.

All in all, *GDI* has some positive characteristics: (1) it's easy to calculate even manually, (2) unlike the coefficients based on Pearson product moment correlation coefficient (*Rit* and *Rir*), it is robust against small changes in the dataset, and (3) it can detect the deterministic patterns and reaches the value *GDI* = ± 1 correctly. Hence, (4) *GDI* does not necessarily underestimate the association between the item and score in an obvious and mechanical manner as *Rit* and *Rir* do in the case when the scales of the item and the score are not equal. An underlying discussion relevant to Kelley's *DI* and the *GDI* is what should be the fixed cut-off for index—or should there be any? The standard way of using the fixed cut-off for *DI* (usually 25% or 27% of extreme test-takers) can, in most cases, be quite a good approximation for the real item discrimination even though, in practice, it appears to underestimate item discrimination in items with an extreme difficulty level. On the other hand, PES encourages us to consider whether some other than the traditional fixed cut-off should be chosen. In the theoretical Guttman-patterned case of a deterministically discriminating data structure, it would be an economical option to choose the cut-off that indicates the threshold point of the item, *i.e.*, the cut-off indicated by the shorter of the extreme strings of 1s or 0s in the ordered data. This also leads a pathway toward

the possibility of identifying a unique optimal cut-off in all real-world items as it is in theoretical Guttman-patterned items.

Limitations of the coefficients and the study and further suggestions

The main deficiency related to both *DI* and *GDI* is that they tend to underestimate the item discrimination power if the item difficulty is extreme and the cut-off is selected rigidly. The practical users of *DI* and *GDI* should be aware of this ill-behavior with the items with extreme difficulty level. However, the coefficients may serve as useful short-cut methods in the practical testing settings to evaluate the overall discrimination power of the items or whether some item should be omitted in the compilation as non-discriminative one. Wider simulations of this character of *GDI* and *DI* would benefit us.

We may also note a practical challenge in both *DI* and *GDI* related to the tied cases which was not handled in the article: The exhaustive splitting procedure and the idea of vector *D* is based on the idea that the test-takers could be rank in a uniquely unambiguous manner. However, this is not possible when there are ties in the score because, within the tied cases, we do not know the actual order of the test-takers without some rationale. Some options were discussed within the text: (1) to double-order the test-takers, first, by the score and, second, by the items, (2) to give unambiguous and unique rank order by trusting the randomization in the ordering, (3) to include all the test-takers with the same score in a bin and to use dynamic cut-offs, and (4) to develop a new coefficient based on non-symmetric cut-offs with polytomous items (cl. Brennan, 1972).

The procedure and results presented in this article raises several questions and ideas for further studies. These include comparison of traditional classical indices for item discrimination with the exhaustive splitting procedure, the possibility of locating latent threshold points in real-world items by employing Guttman-patterned items within the classical test theory, possible models for continuous cut-off functions that would lead the *GDI* to allow continuous variables, and asymmetric versions of the cut-off approach for the polytomous variables (cl. Brennan, 1972). Further simulations concerning statistical properties of *GDI* may benefit us—some analyses of *DI* have been administered already (see Bazaldua, et al. 2017; Tristan, 1998; Kelley, et al. 2002). Specifically, the comparison with Somers' *D* or Goodman-Kruskal gamma may be of interest because these all are based on the order of the test-takers rather than the covariance between the item and score.

References

Aslan, S., & Aybek, B. (2020). Testing the effectiveness of interdisciplinary curriculum-based multicultural education on tolerance and critical thinking skill. *International Journal of Educational Methodology, 6*(1), 43–55. https://doi.org/10.12973/ijem.6.1.43.

Balov, N., & Marchenko, Y. (2016). In the spotlight: Bayesian IRT–4PL model. *Stata News*, *31*(1), (2016 quarter 1). https://www.stata.com/stata-news/news31-1/bayesian-irt/.

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parametric logistic item-response model*. ETS Research ReportRR-81-20. Educational Testing Service.

Batanero, C. D. (2007). *Suitability of teaching Bayesian inference in data analysis courses directed to psychologists.* [Unpublished doctoral dissertation]. University of Granada Spain.

Bazaldua, D. A. L., Lee, Y.-S., Keller, B., & Fellers, L. (2017). Assessing the performance of classical test theory item discrimination estimators in Monte Carlo simulations. *Asia Pacific Education Review*, *18*(4), 585–598. https://doi.org/10.1007/s12564-017-9507-4.

Brennan, R. L. (1972). A generalized upper-lower item discrimination index. *Educational and Psychological Measurement, 32*(2), 289–303. https://doi.org/10.1177/001316447203200206.

Cechova, I., Neubauer, J., & Sedlacik, M. (2014). Computer-adaptive testing: Item analysis and statistics for effective testing. In R. Ørngreen & K. Tweddell Levinsen, Proceedings of the 13th European conference on e-learning ECEL-2014 (pp. 106–112). Aalborg University Copenhagen, Denmark 30-31 October 2014. Academic Conferences and Publishing International Limited.

Cox, N. R. (1974). Estimation of the correlation between a continuous and a discrete variable. *Biometrics, 30*(1), 171–178. https://doi.org/10.2307/2529626.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. http://dx.doi.org/10.1007/BF02310555.

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. https://doi.org/10.1007/BF02289461.

D'Agostino, R. B., & Cureton, E. E. (1975). The 27 percent rule revisited. *Educational and Psychological Measurement, 19*(1), 47–50. https://doi.org/10.1177/001316447503500105.

Drasgow, F. (1986). Polychoric and polyserial correlations. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences. Vol 7* (pp. 68–74). John Wiley.

Dunn, T. J., Baguley, T., & Brunsden, V. (2013). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046.

Ebel, R. L. (1954a). How an examination service helps college teachers to give better tests. *Proceedings of the 1953 invitational conference on testing problems.* Educational Testing Service.

Ebel, R. L. (1954b). Procedures for the analysis of classroom tests. *Educational and Psychological Measurement 14*(2), 352–353. https://doi.org/10.1177/001316445401400215.

Ebel, R. L. (1967). The relation of item discrimination to test reliability. *Journal of Educational Measurement 4*(3), 125–128. https://doi.org/10.1111/j.1745-3984.1967.tb00579.x.

Esendemir, O., & Bindak, R. (2019). Adaptation of the test developed to measure mathematical knowledge of teaching geometry in Turkey. *International Journal of Educational Methodology, 5*(4), 547–565. https://doi.org/10.12973/ijem.5.4.547.

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

ETS (2020). Glossary of standardized testing terms. Educational Testing Service. https://www.ets.org/understanding_testing/glossary/.

Feldt, L. S. (1963). Note on use of extreme criterion groups in item discrimination analysis. *Psychometrika*, *28*(1), 97–104. https://doi.org/10.1007/BF02289553.

FINEEC (2018). *National assessment of learning outcomes in mathematics at grade 9 in 2004*. Unpublished dataset opened for the re-analysis 18.2.2018. Finnish National Education Evaluation Centre.

Flanagan, J. C. (1937). A proposed procedure for increasing the efficiency of objective tests. *Journal of Educational Psychology*, *28*(1), 17–21. http://dx.doi.org/10.1037/h0057430.

Forlano, G., & Pinter, R. (1941). Selection of upper and lower groups for item validation. *Journal of Educational Psychology, 32*(7), 544–549. http://dx.doi.org/10.1037/h0058501.

Goodman, L. S., & Kruskal, W. H. (1959). Measures of association for cross classification. II: Further discussion and references. *Journal of the American Statistical Association, 54*, 123–163. https:/doi.org/10.2307/2282143.

Graham, J M. (2006). Congeneric and (essentially) tau-equivalent estimates for score reliability. What they are and how to use them. *Educational and Psychological Measurement, 66*(6), 930–944. http://dx.doi.org/10.1177/0013164406288165.

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika, 74*(1), 121–135. http://dx.doi.org/10.1007/s11336-008-9098-4.

Gulliksen, H. (1950). *Theory of mental tests*. Lawrence Erlbaum Associates Publishers.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282. https://doi.org/10.1007/BF02288892.

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.* The American Soldier, Vol IV. Wiley.

Harris, C. W., & Wilcox, R. R. (1980). Brennan's B is Peirce's Theta. *Educational and Psychological Measurement*, *40*(2), 307–311. https://doi.org/10.1177/001316448004000204.

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

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement*, *60*(4), 523–531. https://doi.org/10.1177/00131640021970691.

IBM. (2017). IBM SPSS Statistics 25 Algorithms. IBM. ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/25.0/en/client/Manuals/IBM_SPSS_Statistics_Algorithms.pdf

Johnston, A. P. (1951). Notes on a suggested index of item validity: The U-L index. *Journal of Educational Psychology*, *42*(8), 499–504. https://doi.org/10.1037/h0060855.

Kelley, T., Ebel, R., & Linacre, J. M. (2002). Item discrimination indices. *Rasch Measurement Transactions, 16*(3), 883–884. https://www.rasch.org/rmt/rmt163a.htm.

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.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151–160. http://dx.doi.org/10.1007/BF02288391.

Linacre, J. M., & Wright, B. D. (1994). Chi-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.

Liu, F. (2008). Comparison of several popular discrimination indices based on different criteria and their application in item analysis. University of Georgia. https://getd.libs.uga.edu/pdfs/liu_fu_200808_ma.pdf.

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology 63*(3), 509–525. https://doi.org/10.1348/000711009X474502.

Long, J. A., & Sandiford, P. (1935). *The validation of test items (Bulletin No. 3).* Department of Educational Research University of Toronto.

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

Macdonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement, 62*(6), 921–943. https://doi.org/10.1177/0013164402238082.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). 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/fileview/November_2016_1478701072__159.pdf.

Metsämuuronen, J. (2017). *Essentials of research methods in human sciences. Vol 1: Elementary basics.* SAGE Publications.

Metsämuuronen, J. (2020). 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.

Moses, T. (2017). A review of developments and applications in item analysis. In R. Bennett & M. von Davier (Eds.), *Advancing human assessment*. *The methodological, psychological and policy contributions of ETS* (pp. 19–46). Springer Open. https://doi.org/10.1007/978-3-319-58689-2_2.

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

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

Ross, J., & Lumsden, J. (1964). Comment on Feldt's "use of extreme groups". *Psychometrika*, *29*(2), 207–209. http://doi.org/10.1007/BF02289701.

Ross, J., & Weitzman, R. A. (1964). The twenty-seven per cent rule. *Annals of Mathematical Statistics*, *35*(1), 214–221. http://doi.org/10.1214/aoms/1177703745.

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*, *9*, 99–103.

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

Stanley, J. T. (1964). *Measurement in today's schools* (4th ed). Prentice-Hall.

Stata corp. (2018). *Stata manual*. Stata. https://www.stata.com/manuals13/mvalpha.pdf

Tarkkonen, L. (1987). *On reliability of composite scales*. *An essay on the measurement and the properties of the coefficients of reliability—unified approach*. Statistical Research Reports 7. Finnish Statistical Society.

Tristan, L. A. (1998). The item discrimination index: does it work? *Rasch Measurement Transactions*, *12*(1), 626. https://www.rasch.org/rmt/rmt121r.htm#Disc.

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to Cronbach's Alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology*, *7*, 769. https://doi.org/10.3389/fpsyg.2016.00769.

Vehkalahti, K. (2000). *Reliability of measurement scales.* Statistical Research Reports 17. Finnish Statistical Society. https://helda.helsinki.fi/bitstream/handle/10138/21251/reliabil.pdf? sequence= 1&isAllowed=y.

Wiersma, W., & Jurs, S. G. (1990). *Educational measurement and testing* (2nd ed.). Allyn and Bacon.

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.

Yang, Y., & Green, S.B. (2011). Coefficient alpha: A reliability coefficient for the 21st Century? *Journal of Psychoeducational Assessment*, *29*(4), 377–392. http://dx.doi.org/10.1177/0734282911406668.

Yi-Hsin, C., & Li, I. (2015). IA_CTT: A SAS® macro for conducting item analysis based on classical test theory. Paper CC184. https://analytics.ncsu.edu/sesug/2015/CC-184.pdf