anales de psicología / annals of psychology
2023, vol. 39, nº 3 (october), 517-527
https://doi.org/10.6018/analesps.535781

ISSN print: 0212-9728. ISSN online: 1695-2294.
https://revistas.um.es/analesps

Is it quality, is it redundancy, or is model inadequacy? Some strategies for judging the appropriateness of high-discrimination items

Pere J. Ferrando and Fabia Morales-Vives*

Universitat Rovira i Virgili, Departament de Psicologia, Research Center for Behavioral Assessment (CRAMC), Tarragona (Spain)

Título: ¿Es calidad, redundancia o inadecuación del modelo? Algunas estrategias para determinar la idoneidad de los ítems altamente discriminativos.
Resumen: Cuando se desarrollan nuevos cuestionarios, tradicionalmente se asume que los ítems deben ser lo más discriminativos posible, como si esto fuera siempre indicativo de su calidad. Pero en algunos casos estas discriminaciones elevadas pueden estar ocultando algunos problemas como redundancias, residuales compartidos, distribuciones sesgadas o limitaciones del modelo que pueden contribuir a inflar las estimaciones de la discriminación. Por lo tanto, la inspección de estos índices puede llevar a decisiones erróneas sobre qué ítems mantener o eliminar. Para ilustrar este problema, se describen dos escenarios diferentes con datos reales. El primero se centra en un cuestionario que contiene un ítem aparentemente muy discriminante, pero redundante. El segundo se centra en un cuestionario clínico administrado a una muestra comunitaria, lo que da lugar a distribuciones de respuesta de los ítems muy sesgadas y a índices de discriminación inflados, a pesar de que los ítems no discriminan bien entre la mayoría de los sujetos. Proponemos algunas estrategias y comprobaciones para identificar estas situaciones, para facilitar la identificación y eliminación de los ítems inapropiados. Por lo tanto, este artículo pretende promover una actitud crítica, que puede implicar ir en contra de los principios rutinarios establecidos cuando no son apropiados.
Palabras clave: Discriminación de los ítems. Análisis factorial. Análisis de ítems. Teoría de Respuesta al Ítem. Redundancia. Evaluación clínica.

Abstract: When developing new questionnaires, it is traditionally assumed that the items should be as discriminative as possible, as if this was always indicative of their quality. However, in some cases these high discriminations may be masking some problems such as redundancies, shared residuals, biased distributions, or model limitations which may contribute to inflate the discrimination estimates. Therefore, the inspection of these indices may lead to erroneous decisions about which items to keep or eliminate. To illustrate this problem, two different scenarios with real data are described. The first focuses on a questionnaire that contains an item apparently highly discriminant, but redundant. The second focuses on a clinical questionnaire administered to a community sample, which gives place to highly right-skewed item response distributions and inflated discriminant indices, despite the items do not discriminate well among the majority of participants. We propose some strategies and checks to identify these situations, so that the items that are inappropriate may be identified and removed. Therefore, this article seeks to promote a critical attitude, which may involve going against routine stablished principles when they are not appropriate.
Keywords: Item discrimination. Factor analysis. Item analysis. Item Response Theory. Item Redundancy. Clinical Measurement.

Introduction

Items from personality and attitude scales intended to measure a single trait or construct (Briggs & Cheek, 1986) can be calibrated by using different modeling frameworks. However, under the most usual choices, items are characterized by two main features: (a) location or extremeness, and (b) discrimination or discriminating power (Ferrando et al. 2022, Henryson, 1971). In the calibration process, indices that estimate these features are computed, and, based on their values, rules often derived from conventional wisdom (e.g., Bollen & Lennox, 1991) are followed for selecting the most "appropriate" items. But an uncritical following of these rules can lead to problems and result in a measure which has far less than optimal properties. In particular, this article focuses on the problems that can arise when items with very high discrimination estimates are uncritically selected.

A preliminary limitation of discrimination indices is that they jointly refer to three interrelated concepts (see Ferrando, 2012). Thus, a large discrimination value can be interpreted as that the item (a) has high quality as indicator of the construct (Lord & Novick, 1968, McDonald, 1999); (b) is highly effective in differentiating respondents by their levels on the construct (Masters, 1988), and (c) is strongly related to the remaining items that measure the construct (i.e. highly consistent; see Bollen & Lennox, 1991, or Nunnally, 1978).

The standard, uncritical view regarding the three facets above, can be summarized in "the more the better" position (Bollen & Lennox, 1991, Masters, 1988): Given that all the three properties are regarded as desirable, the main aim is explicitly to select the most discriminating items (e.g. Nunnally, 1978). Of course, there have also been dissenting positions, summarized by the heading "sometimes, more is worse" (e.g., Masters, 1988).

Although it contains some new methodological contributions, this article has mainly didactic purposes and is aimed at the practitioner. We first discuss the problems related to highly discriminating items, and then propose a strategy for deciding whether a high discrimination estimate may be caused by unwanted determinants. Finally, we illustrate it in two different scenarios.

A non-technical background and interpretations based on a correct model

We shall first discuss a scenario in which (a) the design and sampling processes and (b) the psychometric model and its assumptions, are essentially correct. So, under this initial scenario the discrimination estimates are undistorted or unbiased.

*** Correspondence address [Dirección para correspondencia]:**
Fabia Morales-Vives. Universitat Rovira i Virgili, Departament de Psicologia, Research Center for Behavioral Assessment (CRAMC), Tarragona (Spain).
E-mail: fabia.morales@urv.cat
(Article received: 09-08-2022; revised: 23-10-2022; accepted: 11-11-2022)

The models we consider are unidimensional dominance-based, models in which the expected item score increases with trait level. They are: (a) Classical Item (test) Theory (CIT, e.g. Lord & Novick, 1968), (b) the unidimensional (Spearman's) item factor analysis FA model as fitted to the inter-item correlation matrix (e.g. Lord & Novick, 1968) and (c) the item response theory (IRT) models that can be parameterized as non-linear FA models (McDonald, 1999). Our position is that the FA model in a wide sense is the most general and encompasses the three just referred.

The most common indices of discrimination in the modellings above are: (a) the item total or item-rest correlations in CIT (Henryson, 1971), the standardized factor loading estimates in FA (Ferrando, 2012) and the slope estimate in IRT models (Lord & Novick, 1968). In the correct-model scenario here, the relations between these indices are (almost) one-to one (see Ferrando, 2012). Therefore, provided that the basis model is essentially correct, the interpretation of the relative amount of estimated item discriminating power **will be the same** with any of the three models.

The different parameterizations above, however, help to clarify the "faceted" interpretation of the estimates. Thus, as the square of the item loading is the proportion of item variance that is determined by the construct, the loading can be interpreted as the degree of quality the item has as indicator (first facet). Next, because the IRT index is the slope of the item characteristic curve (ICC), this index directly measures the effectiveness of the item in differentiating respondents according to their trait level (second facet). As for discrimination-internal-consistency, the main result of interest here is that, as the number of items that measure the construct increases without bound, the square of the item-total correlation approaches the average correlation this item has with the remaining items that measure the same construct (Nunnally, 1978). So, when the value of **any** of the discrimination measures considered here increases, the average correlation of the item with the remaining items of the scale also increases, and the relation is (almost) one to one.

In principle, neither the quality nor the effectiveness facets seem to convey potential negative effects. So, the "the more the better" principle would be difficult to refute for them. The internal-consistency facet, however, is more complex, and several authors affine to Cattell's school consider that high internal consistency generally indicates redundancy, narrowness of the scale, and poor content representativity (e.g., Boyle, 1991; Kline, 1987). So, scales with highly discriminating items are highly suspect of having been obtained by writing items that are virtually repetitions of each other, and which, in a FA, are expected to appear in the form of a "bloated specific" factor. Thus, there would be an optimal range of (possibly moderate) discrimination-consistency values that avoids redundancy by, at the same time, ensuring a minimal cohesion that allows us to assume that all items measure the same thing (Briggs & Cheek, 1986, Kline, 1987).

The position above rightly points out that the dominant routine of maximizing internal consistency at any cost is sometimes an unwise strategy. However, we believe it must be qualified. If the FA model holds, none of the items would share specificity beyond the common construct they measure. So, items cannot be accused to be redundant (or virtual paraphrases one of the other), because, if this was so, local dependencies should emerge. Indeed, (a) is very difficult to achieve uniformly high discrimination estimates in an instrument intended to measure a broad-bandwidth construct, and (b) the estimated values can be increased by "narrowing" the manifestations of the construct. However, this narrowing is different from redundancy. On the other hand, high discriminations that can be properly interpreted as reflecting item quality, can be achieved without falling neither into redundancies nor into loss of representativity, when an already medium or narrow bandwidth trait is to be measured.

Upwardly biased discrimination estimates not accounted for by the psychometric model: Three main sources

We shall now discuss the most relevant sources of "inflated" item discrimination estimates when the "ideal" conditions above do not hold. They are: (a) design and sampling inadequacies, (b) item redundancies, and (c) clinical items that measure unipolar traits.

First source: Design and sampling inadequacies

In the calibration process as considered here, the FA model is fitted to a sample correlation matrix **R** which is an estimate of some population correlation matrix **Σ**. Assuming that the sample is representative of the population for which the test is intended, the key issue for **R** to be an appropriate estimate is sample size. As the sample size increases, each element of **R** will increasingly approach its corresponding element in **Σ**. So, unusually high correlation estimates due to sampling fluctuation are less and less likely to appear. In small samples, however, an implausibly high value of a sample correlation is not so unlikely. And, if it occurs, it is expected to give rise to a Heywood or a quasi-Heywood case (estimated communalities near one or even greater than one) at least for one of the variables implied, and, almost surely, the discrimination estimates for these variables will be upwardly biased. This problem is potentially much more relevant when **R** contains tetrachoric/polychoric correlations, where a sample size of at least 200 is always advisable (Ferrando et al., 2022).

Regarding design issues, there are two main potential sources of problems. First, correlations based on different numbers of cases (i.e. obtained under pairwise deletion).

Second, linear dependencies among the item scores (i.e. certain inter-item correlations are unit or near unit, or certain item scores are linear composites of the remaining scores). Again, the main problem here is that of some implausibly high correlations, which, in turn, cause implausibly high discrimination estimates for the involved items. Apart from the general consequences these inflated estimates produce (discussed in the next section), a specific consequence here is that **R** is likely to be not positive definite, and, when this occurs, several problems of estimation, testing, and interpretation of the results are likely to appear (see Lorenzo-Seva & Ferrando, 2021). Given that the sources of biased discriminations in this section are also sources of non-positive definiteness, when offending discriminations appear it should always be checked that **R** is positive definite.

Second source: Item redundancy

The substantive term "item redundancy" refers more operatively to the technical terms "correlated residuals" in FA and "local dependencies" in IRT. The basic idea is that certain items in the pool share specific, non-content related variance beyond the common trait they measure. So, once the influence of the trait is partialled-out, these items continue to be correlated due to non-content reasons, of which, the most relevant are: (a) repeated presentation of the same items, (b) wording or content similarities, (c) similarities in the evoked situation, and (d) context effects (Bandalos, 2021, Edwards et al., 2018).

At its most molecular level, we have a bivariate residual between a pair of items, which is known as doublet (Lorenzo-Seva & Ferrando, 2021). Higher-order local dependencies (triplets, quadruplets and above) might indeed exist (Edwards et al., 2018), but they will also manifest at the bivariate level, which is the level where we shall study them here. The basic mechanism is as follows.

Consider a pair of items that are direct indicators of the common factor, so that the loadings of both are positive. Further, consider that they share specific variance, which is modeled as an additional correlation between their residual terms. If this shared variance is due to the sources discussed above, this residual correlation is also expected to be positive. So, the model expected correlation between these two items is higher than that which would be expected solely on the basis of the common factor they measure. This last expected correlation would be simply the product of their loadings.

The aim of any FA estimation procedure is to keep the residual correlations once the factor has been extracted as close to zero as possible. A way to achieve this when the correlation is higher than it should be, is to over-estimate the loadings of the pair of implied items, which means that the reproduced correlation (the product of the loadings) would be closer to that observed, and the residual correlation closer to zero. If this occurs, the discrimination estimates of these items will be over-estimated, but their residual correlation will remain unsuspectedly low. So, the researcher will find it very difficult to appraise that the high discriminations do not reflect quality but redundancy.

When the problem is more massive than a very small number of doublets, the misspecification can no longer be compensated only by over-estimating the loadings and will also manifest in the fitted residual matrix. If so, the misspecification can be (partly) detected by inspecting the fitted residuals, many of which will depart substantially from zero.

We turn now to the practical consequences of over-estimated discriminations. As for calibration, in the best scenario of very few doublets, the problem will be a loss of information: the item pool would contain less information than the psychometric model would predict. Thus, the few items with very high discrimination estimates would not provide as much accuracy and information as they appear to, because the information they provide would also be almost totally accounted for by the remaining items. The discrimination estimates of the remaining items, however, might be not quite distorted, and the model might fit well.

As the number of non-negligible doublets increase, problems go worse. First, it will be very difficult to select the most appropriate items. This is because the discrimination estimates **for most** of the items will be likely biased, as the estimated loadings for the items implied in the doublets will be generally inflated at the cost of deflated estimates for the remaining items (Chen & Thissen, 1997). Second, the model is likely to fit badly. In our pessimistic experience, when this occurs, a better-fitting multiple correlated-factors solution will be presto fitted to the data. Now, if the most appropriate model is unidimensional but contains additional residual correlations that do not reflect content (as we are assuming here), fitting a multiple correlated-factors model is expected to give rise to artifactual, bloated-specific, and weak factors of little if any substantive interest.

Regarding score-related problems, in the least bad scenario above, the loss of information due to redundancy will translate to score estimates that are less accurate than the reliability or conditional accuracy indices predict. This loss of accuracy, in turn, will affect "internal" processes such as individual assessment, comparison of individuals, cut-off values etc. As for "external" processes the 'true' validity evidence will be also smaller than expected.

As redundancy increases, the scoring problems will get worse. Not only are now the scores less informative than expected, but might be also contaminated by the specificities items share beyond the content factor, thus reflecting a mixture of content and artifactual effects (see Distefano et al., 2022).

Third source: Unipolar traits in clinical measurement

Several authors (e.g., Morales-Vives et al., 2022; Reise & Waller, 2009) have noticed that implausibly high

discrimination values tend to appear in many clinical items, especially when administered in community samples. In our view, this outcome results from the impact of different sources. To start with, many clinical instruments aim to measure narrow-bandwidth constructs and tend to be based on highly repetitive items (Reise & Waller, 2009), which implies that the sources so far discussed are likely to operate. However, there is more than this.

When a normal-range test is administered in a representative sample, the measured trait can be plausibly modeled as a bipolar dimension, which is equally meaningful at both ends, and that has a two-tailed, unimodal, and approximately symmetrical distribution. In the case of a clinical measure administered in a community sample, however, it is more plausible to assume that the trait is unipolar: i.e. it has most (or only) meaning at its upper end, and has a rightly skewed distribution with most cases (the asymptomatic) concentrated at the lower end. If this is so, the distribution of the item scores is also expected to be (strongly) right-skewed, which is the usual result in this case (see Morales-Vives et al., 2022).

The bivariate surfaces of pairs of items of the type above would contain most of the cases piled up at the lowest score in both items. This is a case of lower-bound censoring, which, in turn, is expected to produce upwardly biased (i.e. expansion bias; see Rigobon & Stoker, 2009) correlation estimates. These inflated correlations are in turn expected to translate to inflated estimated loadings.

In terms of IRT parameterization, the items we are discussing have threshold estimates that only spread over a narrow range of trait values, and so, only provide effective measurement in this range, which, generally, is located well above the trait mean. In agreement to this result, the ICCs of these items are virtually flat at the lower end of the trait range (which means null discrimination at this range). Next, starting from a point generally above the mean, the ICC slope increases sharply, which means that the item will be highly discriminating for the minority of individuals which are located at the upper end of the trait range. In other words, the high discrimination estimates do not really reflect a high overall discriminating power (most individuals would in fact remain undifferentiated) but rather that all this power is concentrated in a narrow range.

In conclusion, the problems in the previous sections are also relevant here, but new specific problems that are not due to basic flaws in sampling and design but to fitting a model which was not initially designed for dealing with unipolar traits, are also expected to appear. Now, although more specific models for this scenario exist (Morales-Vives et al., 2022), our position is that the non-linear FA with IRT parameterization continues to be a useful option but, at the same time, great care should be taken when interpreting the results.

What should I do? A proposed diagnosis strategy

The strategy proposed in this section should be seen only as a starting blueprint, and it does not exempt the researcher from the need to think critically. So, we have that when examining the solution, the practitioner observes that certain item discriminations are very high and suspects that they might mask some measurement problems rather than indicating quality. The steps for assessing the discrimination of the items, described below, are also depicted in Figure 1.

For the examination above to be operative, some guidelines should be first provided about the interpretation of discrimination values, and we shall do so using both the correlation/loading metric and the slope-IRT metric. Although the literature is quite consistent, we should warn that the references provided are only indicative and not intended to be used as rigid cut-off values.

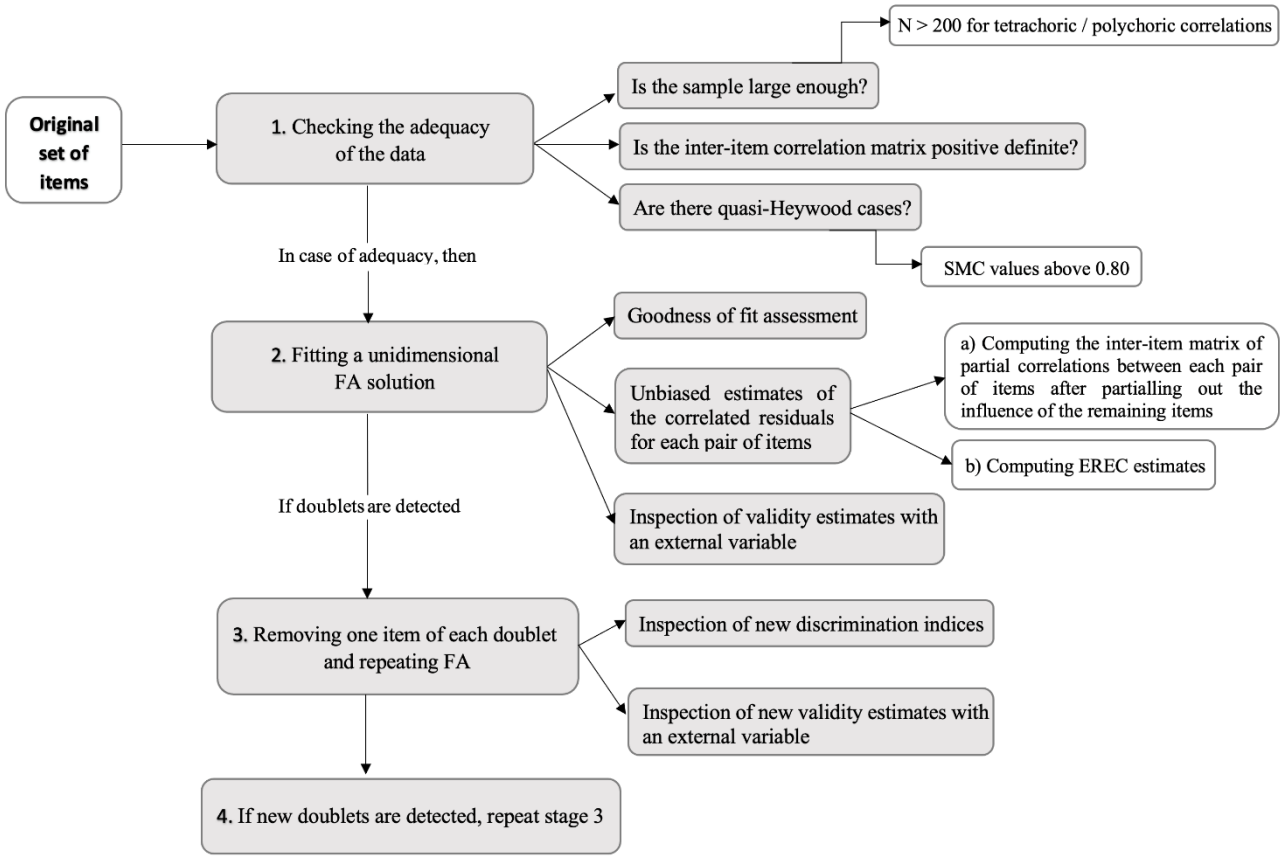
With this proviso, the normal range discrimination values in personality and attitude is about between .3 and .7 in loading metric, which translates to .3 to 1.00 in slope metric. Values between .70 and .85 (loading) or 1.00 and 1.70 (slope) are high but not unusually high, and start to deserve further inspection. Loadings above .90 or slopes above 2 are unusually high and should always be inspected. Furthermore, these values might give rise to estimation problems (mainly Heywood cases). Finally, we note that some clinical studies have reported slopes above 4.0 (loadings above .97) which indicates, with all certainty, that some of the problems discussed here are operating (Masters, 1988, Reise & Waller, 2009).

Let's now start the checks. For any type of test (normal-range or clinical), check first that the sample is large enough to avoid extreme correlation estimates to appear. Second, verify that the inter-item correlation matrix is positive definite. Third, inspect the prior communality estimates to discard the occurrence of Heywood or quasi-Heywood cases. The prior communality estimates we recommend are squared multiple correlations (SMC; see Lorenzo-Seva & Ferrando, 2021). And, as for references, values above, say, .80, can be considered as quasi-Heywood, and point out either to the problem of sampling instability or to item redundancy. Overall, if the prior data-adequacy checks indicate problems, it would be convenient to do an item cleaning or to increase sample size before continuing. If not, we can examine goodness of model-data fit, and turn to the second group of procedures.

If model-data fit is considered acceptable, good. However, a good fit does not rule out a problem due to doublets, although, if it exists, it would imply a very few at most. We recommend to obtain estimates of the correlated residuals for each pair of items, and we provide two alternatives. First, computing the inter-item matrix of partial correlations between each pair after partialling out the influence of the remaining items. This approach is the simplest but produces estimates that are slightly upwardly biased, particularly when there are few items. Second, to use

an estimate called "Expected Residual Correlation Direct Change" (EREC; Ferrando et al., 2022). The EREC estimate is obtained using a sectioning method in which each possible pair of items is sequentially excluded from the core analysis and their loadings are separately estimated using extension analysis, which limits the 'propagation' effect discussed above and leads to a less biased estimate of the residual correlation.

Figure 1
Steps for assessing the discrimination of the items.



If doublets are identified with the procedure above, and they agree with the items with very high discrimination estimates, two more final checks can be envisaged. First is internal: for each doublet, remove one of the offending items and repeat the FA without it. Chances are that the discrimination of the orphan item would now deflate. Indeed, this check can be repeated for each one of the items in the pair. We should stress the importance of inspecting the whole results when removing each item of the doublet, as they may help to decide which of the two is more convenient to remove. In some cases it may be just as good a choice to remove either of both, but in others one of these items may be the source of the problem, because of its relationship with the rest of items or other issues. Therefore, this decision should not be taken at random. Other characteristics of the items, such as their wording, content, etc., should also be taken into account. The example 1 below illustrates these points.

The second type of check is external (Briggs & Cheek, 1986, Kline, 1987), through a relevant variable expected to be related to the construct. We propose two approaches. First, obtain simple sum scores with and without each offending item (one item is removed at a time) and compute the usual validity estimates as the correlations between the sum scores and the criterion. If the removed item was really highly discriminating (i.e. a very good trait indicator), then the validity estimate is expected to suffer when this valuable item is removed. However, if the high discrimination merely reflects redundancy, the validity estimate without this item would hardly vary.

The rationale for the second check is that the items that better measure the construct should be also those more strongly related to the relevant criterion. It is then a simple matter to plot the item-criterion correlations (item validity indices) against the corresponding item discrimination estimates. If the items with the highest estimates are not the most valid, redundancy can be suspected.

We turn now to the bad-fit outcome, a result that might indicate that (a) the data is substantively multidimensional or (b) shared specificities far more massive than a few number of doublets exist. If we focus on the second result, the problem is the same as above. However, being the number of doublets substantial, the lack of fit cannot be longer compensated by inflating the discrimination estimates. The type of inspection so far recommended would possibly entail a considerable labor of trimming before arriving at a locally independent item set that fits well the data and does not lead to inflated discrimination estimates. So, a preliminary screening approach is proposed. It consists of specifying a two-factor solution and leave it unrotated (which means that it is in canonical or principal-axes form). In this type of solution, the first factor is the most general common factor that can be obtained from the data, and the second, orthogonal factor, reflects the residual covariation that the general factor is unable to explain. So, substantial loading values on the second factor (say above .30) point out either to the potential items that share specificities or, alternatively, to groups of items that measure additional contents. Inspection of the item content should then guide the researcher about the next steps to take (detecting doublets or trying to fit multiple solutions). For the reasons given above, the approach is not perfect, as the presence of doublets can inflate the first factor estimates and so maintain lower than should be the second factor estimates. Even so, however, it is submitted to be useful, especially as a first, general screening device.

We shall finally discuss two additional recommendations for clinical measures: (a) try to use far larger samples than in the normal-range case, and (b) fit the non-linear FA model with IRT parameterization. The first recommendation is to ensure that it will be enough variation at the meaningful upper end of the trait. The second is to interpret appropriately the additional relevant information provided by IRT, which, in this case, starts on the spread of item thresholds. In a normal-range item, thresholds are generally spread over the full trait range (between -3 and +3 in normal metric). However, in many clinical items, the lowest threshold might well start above the trait mean. If so, two things can be expected: First, the item would not provide effective measurement at low trait levels. Second, the discrimination estimate of this item is likely to be upwardly biased.

Turning to discrimination, two related pieces of information are highly relevant here. First, for each item obtain the ICC and check the region of the trait at which the item discriminates. Second, obtain the information curve for the test scores. The key issue here is not the raw amount of discrimination, but the region of the trait at which the test provides effective measurement.

The strategy in practice

We shall illustrate how the present proposal works with two different situations: One focused on a questionnaire with highly discriminant but redundant items, and another focused on a questionnaire that assess a clinical unipolar variable.

Example 1: Too good to be true

In this example we used the *Satisfaction With Life Scale* questionnaire (SWLS, Diener et al., 1985), a unidimensional measure, made up of five highly discriminant items, which assesses a narrow-bandwidth construct: the persons' evaluation of the extent to which they are satisfied with their own life. To achieve a large and heterogeneous sample, we administered this questionnaire online, along with the *Connor-Davidson Resilience Scale-10 items* questionnaire (CD-RISC 10; Campbell-Sills et al., 2009), which assesses resilience. A total of 1545 adults (65% women) participated in this study (18-80 years old, M = 41.6, SD = 13.4), which is large enough even for fitting a nonlinear FA solution. Given that (a) the sample was very big, (b) the number of response choices not too large, and (c) some items showed excess kurtosis, this was the chosen solution to be fit.

Regarding the first group of checks, the inter-item polychoric correlation matrix was positive definite, and suitable for FA (KMO = .85). Table 1 shows the SMC communalities for each item. Items 3 and 4 have SMC values of .87 and 1.00 respectively. Therefore, item 3 may be considered as a quasi-Heywood case (above .80) while item 4 is a Heywood case.

The unidimensional solution based on the non-linear model and fitted with Robust Unweighted Least Squares (RULS) as implemented in the FACTOR program (Lorenzo-Seva & Ferrando, 2006) achieved a good fit: GFI = .99, RMSR = .021, CFI = .99, RMSEA = .039. The obtained discrimination estimates are in Table 1. Because the fitted solution can be also parameterized as an IRT solution, Table 1 reports the three discrimination estimates considered in the article, which, as expected, fully agree. Items 2, and 5 have loadings estimates higher than .70 but lower than .85, and slopes estimates higher than 1.00 but lower than 1.70, which can be considered as high but no unusually high. However, the loading estimate of the third item is unusually high, exceeding the .90 value: .95, which agrees with the prior quasi-Heywood qualification based on the SMC. Its item-rest correlation (classical discrimination index) and the slope index are also the highest. In fact, its slope estimation is 3.03, much higher than 2.. In view of this result, we obtained EREC estimates of the correlated residuals for each pair of items. If a rigorous statistical cut-off was used (twice the standard error of a zero population value), results suggests that items 3 and 4 may be considered as a potential doublet although the effect sizes would qualify as small to medium. The EREC index between the pair 3-5 did not reached sig-

nificance, but its value was very similar to the one found for the pair 3 and 4 (.18 for the pair 3-4 and .17 for the pair 3-5), which seems to suggest that item 3 may share higher-order redundancies with both items 4 and 5. This is not surprising, since the wording of item 3 ("Estoy satisfecho con mi vida") seems to define itself the construct that is assessed, which means that this is a "defining" or "prototypical" item, while the remaining items seems to assess what Burisch (1984) named as "correlational" characteristics, being more peripheral. Note also that the SMC Heywood value obtained for item 4 is probably due to the redundancy that it shares with item 3.

If we only took into account the magnitude of the loadings, thinking that the higher the better, together with the fit results, we would conclude that this instrument has remarkably good psychometric properties, despite it includes Heywood and quasi-Heywood cases. In particular, note that the RMSR index (.021) which is directly based on residuals, does not seem to be affected by the redundancies between the item 3 and other items. It seems that in this case the redundancies have particularly affected the size of loadings, *inflating* especially the loading of item 3, which is considerably high.

Table 1
Loading matrix and SMC values obtained from the exploratory factor analysis, skewness and kurtosis of the items in Example 1.

Item	Loading	SMC	Slope	Skewness	Kurtosis	Item-rest correlation
Item 1	.70	.52	.98	- .49	- .50	.60
Item 2	.82	.68	1.42	- .82	.58	.68
Item 3	.95	.87	3.03	- .81	.54	.79
Item 4	.67	1.00	.89	- .30	- .68	.57
Item 5	.73	.52	1.06	- .98	1.26	.59

Considering these results, we firstly decided to remove the item 3. After doing this, all the SMC values were lower than .80 (see Table 2) and not doublets were detected through the EREC index. The loadings of the rest of items remained practically identical (see Table 2), and the fit indices did not vary substantially (GFI = .99, RMSR = .019, CFI = .99, and RMSEA = .037). These results suggest, as expected, that item 3 is redundant with other items, especially with item 4, and that this shared specificity contributed to its high loading although without affecting the remaining loadings.

Table 2
Loading matrix and SMC values when item 3 is removed in Example 1.

Item	Loadings	SMC
Item 1	.70	.73
Item 2	.81	.63
Item 4	.65	.48
Item 5	.75	.55

But another option was to remove item 4 instead of 3. After this, no doublets were detected, and the fit indices were as good as the ones found previously, but the SMC values for items 3 and 5 were higher than .80, and the loading of item 3 was still remarkably high (see Table 3). Therefore, removing item 3 seems to be a better option than removing item 4, since it is item 3 the source of redundancies with other items. So, it seems that this item does not explain additional variance of the construct to that explained by the other items.

We turn to the validity checks by using the resiliency scores as the relevant external variable. Item 3 has the highest correlation with resilience scores (see Table 4), which is not surprising, given its prototypical nature. This may lead to decide that this item should not be removed from the questionnaire, as it seems to be the best predictor. However, removing this item does not substantially affect the correlation between the overall scores of both questionnaires (see Table 4). Therefore, this item does not seem to explain additional variance of resilience to that explained by the rest of items. This is congruent with the SMC value of this item, explained before. If we remove item 4 instead of item 3, the correlation between SWLS and CD-RISC 10 does not change substantially either (see Table 4), as the presence of item 3 compensates for the absence of item 4 and the overall scores do not lose predictive power. However, considering the overall results obtained in the different analyses, it seems that the best option is to remove item 3, since it is the source of redundancy, gives rise to Heywood cases, and its presence is not justified by a substantial gain in predictive power.

Table 3
Loading matrix and SMC values when item 4 is removed in Example 1.

Item	Loadings	SMC
Item 1	.72	.50
Item 2	.83	.69
Item 3	.93	.88
Item 5	.72	.87

Table 4
Correlations between SWLS and CD-RISC 10.

SWLS	CD-RISC 10
Item 1	.31
Item 2	.36
Item 3	.42
Item 4	.34
Item 5	.30
Overall scores	.44
Overall scores without item 3	.43
Overall scores without item 4	.43

Example 2: Bipolarity is highly questionable here

In this example we used the *Beck Depression Inventory* (BDI; Beck et al., 1961). It has 21 items, and each item consists of four statements reflecting increasing levels of depressive symptomatology severity scored from 0 to 3 (0 = absence of symptomatology). To achieve a large and heterogeneous sample, we administered this questionnaire in high schools, along with the *Overall Personality Assessment Scale* (OPERAS, Vigil-Colet et al., 2013), which assesses the Big

Five personality traits. A total of 743 individuals (5.6% women) participated in this study (14-18 years old, M = 15.2, SD = 1.11), which is a sample size sufficiently large, even for fitting a nonlinear FA solution in a community sample.

As shown in Table 5, the distribution of the item scores was strongly right-skewed, with coefficients above 1 for most of the items. All means are lower than 1, which reflects that many participants have scores of 0 in these items.

Table 5
Item descriptives and IRT estimates in Example 2.

Item	Mean	Skewness	Loading	SMC	Slope	Locations		
						b_1	b_2	b_3
Item 1	.53	1.30	.79	.62	1.28	.40	1.39	2.57
Item 2	.59	1.41	.70	.49	.98	.50	1.40	2.20
Item 3	.45	1.69	.81	.71	1.36	.58	1.58	2.39
Item 4	.73	1.18	.66	.48	.88	-.07	1.62	2.25
Item 5	.68	1.23	.70	.51	.98	.01	1.61	2.27
Item 6	.44	2.01	.56	.33	.68	.93	2.48	2.89
Item 7	.60	1.47	.83	.76	1.47	.33	1.32	1.76
Item 8	.75	1.12	.75	.60	1.13	-.08	1.34	1.97
Item 9	.34	2.11	.71	.53	1.02	.81	2.47	2.92
Item 10	.73	1.17	.66	.46	.88	.14	1.38	2.04
Item 11	.66	1.03	.63	.46	.82	-.03	1.79	3.01
Item 12	.53	1.31	.66	.47	.87	.43	1.80	3.13
Item 13	.47	1.62	.71	.52	1.02	.71	1.56	2.56
Item 14	.47	1.81	.81	.68	1.37	.91	1.24	1.72
Item 15	.79	.77	.62	.44	.79	-.37	1.55	3.02
Item 16	.70	1.12	.61	.40	.76	-.11	1.88	2.79
Item 17	.78	1.04	.71	.64	1.01	-.37	1.63	2.28
Item 18	.48	1.55	.65	.42	.85	.72	1.74	2.98
Item 19	.40	1.66	.47	.25	.53	1.50	2.21	4.78
Item 20	.39	1.85	.60	.36	.75	.93	2.31	3.42
Item 21	.22	3.14	.49	.24	.56	2.40	3.08	3.85

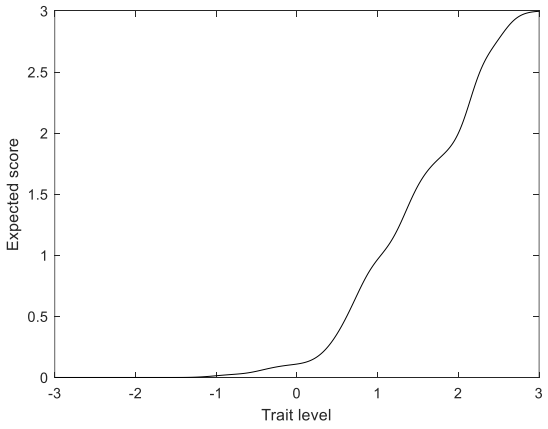
The inter-item correlation matrix was positive definite, and suitable for FA (KMO = .94). Table 5 shows the SMC communalities for each item. None of the items have SMC values higher than .80, which suggests that there are no quasi-Heywood cases.

As recommended above, the data was fitted using the UVA-FA model with IRT parameterization. The unidimensional solution fitted with Robust Unweighted Least Squares (RULS), as implemented in FACTOR, achieved quite a good fit: GFI = .99, RMSR = .050, CFI = .99, RMSEA = .036. Inspection of potential doublets using EREC signaled some doublets as significant, however, the absolute values of the flagged residual correlations were generally quite low. Furthermore, inspection of the item stems did not give rise to assume redundancies. Overall, our interpretation is that these modest residual correlations reflect more the common impact of lower-bound censoring than shared specificities due to wording, situation, etc. For this reason, we decided not to remove any item so far.

Inspection of the IRT estimates in Table 5 shows that all items in general have a narrow range of locations and, some of them, rather high discrimination estimates. In fact, items 3, 7, and 14 had (a) the highest loading/slope estimates, with values higher than .70 but lower than .85, which can be considered as high but not unusually high, and (b) the narrower range of locations, only starting to provide effective measurement well up above the mean. We also note that this triplet was the one that appeared the most in the flagged doublets above. Overall then, we interpret that the higher discrimination estimates of these items partly reflect expansion biases due to the censoring.

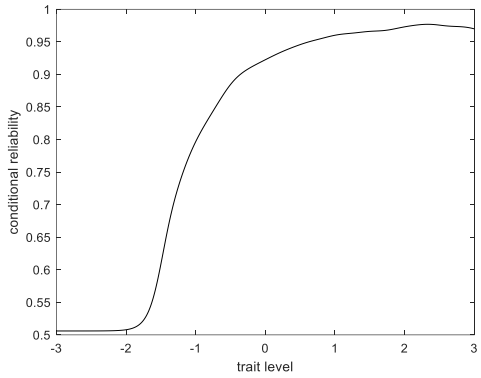
We turn to determinants other than censoring that can partly explain the high discrimination estimates. Figure 2 displays the nonparametric ICC of item 14. The profile is quite clear: Item 14 only starts to provide effective measurement well above the mean. And, although the slope sharply increases from here up to the end, in fact it would only measure accurately a small percentage of individuals: those with strong depression symptoms.

Figure 2
Non-parametric kernel ICC of item 14 of BDI.



At the total-scale level, the trend described for item 14 also applies but not in such an exaggerated way. Figure 3 shows the information curve based on the factor score estimates. So as to use a metric more familiar to the practitioner, the ordinate displays the conditional reliability at the different trait levels. Note that reliability starts to be respectably high above the mean, and continues to be very high until the upper end of the dimension (the meaningful end assuming that depression is unipolar). So, IRT-based BDI scores would measure with high accuracy those individuals that have depression symptoms. The marginal reliability of the scores, however, is only .85; smaller than the curve in Figure 3 suggests at first sight. However, the estimate makes sense, as the accuracy of the scores for measuring the individuals with few or none depressive symptoms is not very high, and these individuals are the majority in a community population.

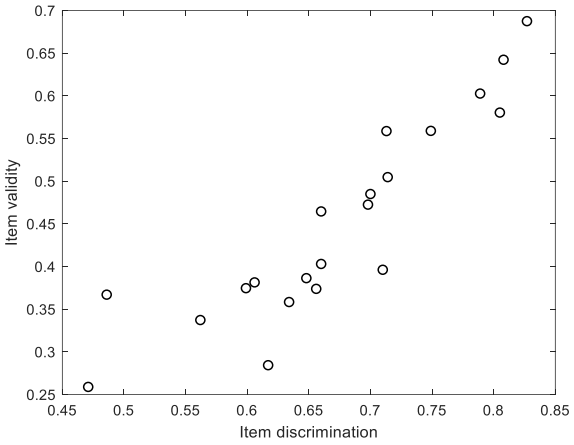
Figure 3
Information curve for the BDI scores.



Regarding the validity checks, we chose emotional stability, assessed with OPERAS, as the external variable. It was strongly related to the BDI scores (the validity coefficient based on the raw BDI scores was r = -.61). For the 21 BDI items, Figure 4 shows the bivariate plot of the item validity estimates (item-external-variable polyserial correlation) against the item discrimination estimates (item loadings in Table 5). For clarity, the validities are positively oriented. The results are impressive, and the correlation between both indices is r = .9. To sum up: in spite of the expansion biases and the reduced-range discriminations, the BDI items with the largest estimated discriminations are also the ones that have the strongest relationships with the external variable.

Overall, BDI can be considered a clinical instrument with good psychometric properties but also with some limitations that require a careful scrutiny of the results. The high estimated discrimination for some of the items reflect partly censoring bias, and the conjectured mechanism is that censoring gives rise to shared specifity between some items that is not explained by their content or wording. This result in apparent doublets, which in turn, lead to inflated discrimination values. Apart from the discrimination biases, the ICCs of these items are only steep in a very narrow range of trait

Figure 4
Scatterplot of the item validity estimates against the discrimination estimates for the BDI items.



levels, and, overall, the score estimates are only highly discriminant at one pole of the dimension, for the minority of individuals with strong depressive symptomatology. Regarding validity, the results show that depressive symptomatology is highly related to emotional stability, as expected, and the items with the highest discriminations are precisely those that contribute the most to this validity. This may be considered as an external indicator of the quality of these items. Because the amount of expansion biases cannot be easily quantified, it is hard to predict what the validity results would have been if these biases could have been controlled. However, despite of this limitation and the items having reduced-range discriminations, and providing effective measurement only well above the trait mean (limitations which are mainly due to the trait unipolar nature), those highly discriminative items still are characterized by their predictive power in relation to the external variable emotional stability. This result shows the robustness of a good questionnaire, able to provide strong validity results (although possibly improvable), even when calibrated and fitted using a non-optimal model that assumes bipolarity.

Discussion

When developing a new questionnaire, or adapting an existing one, it is traditionally assumed that the items should be as discriminative as possible, as if this was always indicative of their quality and the suitability of the questionnaire. In fact, when FAs are carried out, highly discriminative items may result in improved indices (particularly those related to the strength and determinacy of the solution), which reinforces this belief and practice, especially if nothing else is considered. This is illustrated in the example 1 with the item with the highest loading. Although the fit indices of the model are adequate, and this item is apparently the best one, it is precisely the item that it would be advisable to eliminate. Its high discrimination value

and its quasi-Heywood condition reflect that it is a redundant item that does not explain anything of the trait to what is already explained by the other items, and also does not contribute to the prediction of an external variable any more than the other items do. Therefore, when an item is redundant, as in this example, or it has residuals correlated with other items, it may contribute to inflate the discrimination indices, masking the fact that it is not really appropriate.

Example 2 shows another case of inflated discrimination indexes because other causes than content redundancies. First is the distribution of the items, which is very common in the clinical setting: as expected, most of the participants do not present depressive symptomatology, so there are few cases concentrated in the highest scores of the items, resulting in highly right-skewed distributions. This is a case of lower-bound censoring, which produces a bias-expansion effect (Rigobon & Stoker, 2009) that increases the estimated correlations giving rise to inflated estimated loadings. Therefore, the higher loadings obtained in Example 2 are due in part to this effect, and not only to the real discriminating power of the items. Second, the items only provide effective measurement in a range well above the trait mean. In other words, despite the higher loading/ slope indexes, the items do not discriminate well among the majority of participants, which are those concentrated in the lower part of the distribution. In this case, however, external validity does not seem to be affected, since there are precisely the most discriminative items those that best predict emotional stability.

To sum up, it is not advisable to focus only on the discrimination indices of the items, and the favorable appropriateness measures obtained with the whole questionnaire, as these may mask other relevant effects and problems. For this reason, we have proposed here a series of recommendations to collect more information, detect possible redundancies, quasi-Heywood cases, or model limitations, and decide which items should be eliminated and which are really the best. Although we submit that these recommendations make sense, some of them seem to contradict *sacred* stablished principles, mainly maximizing-internal consistency-reliability at any cost. So, we acknowledge they would not be an easy pill to swallow, but we hope the two examples provided will illustrate the usefulness and need for these recommendations. For this reason, we are asking practitioners to be critical and going against routine stablished principles when they are not appropriate. Our recommendations will possibly allow a better test to be obtained but at the cost of loosing apparent accuracy.

The usefulness of an external variable for obtaining further evidence about the discrimination of the items depends indeed on the variable chosen and how it is measured. Ideally this variable should be an objective, non-test variable. However, if a psychometric measure is used as a criterion, it must have adequate psychometric properties, with good items and a clear factor structure, to avoid that the limitations of this questionnaire may difficult the interpretation of the results. Furthermore, this measure should not include items with very similar content and wording to those of the questionnaire which is assessed, as, if this was the case, an additional source of confusion would be added because of the shared item specificity.

In more practical terms, the procedures we propose here can be easily carried out using existing noncommercial programs such as FACTOR (Lorenzo-Seva & Ferrando, 2006) or R packages, such as "psych" (Revelle, 2021) or "lavaan" (Rosseel, 2012), as well as using commercial programs such as the ESEM implementation in Mplus (Asparouhov & Muthén, 2009). Also, at present, our research group is developing a full implementation of the procedures proposed here for detecting correlated residuals in R language.

In closing, we would note that both for maintaining a didactic level and for space limitations, this article has had to put aside methods, indices, and developments that are highly relevant for the problems discussed. Thus, we have proposed to use two specific (and simple) approaches for detecting doublets, but there is a plethora of indices and methods for this purpose worth to be tried. More generally, we have limited ourselves to the standard FA model in which residuals are assumed to be uncorrelated. So, our approach has been based on omitting redundant items in order to avoid biasing effects (among other things), and we believe that this approach is the 'cleanest' specially at earlier stages of item selection. An alternative approach, however, would be to explicitly model correlated residuals within an extended FA model, which, at present, is feasible using the ESEM modeling (Asparouhov & Muthén, 2009). In our view, this approach might be useful at later stages, with a far limited number of items that have been previously selected (as in the first example above). Overall, we believe that it would be interesting for the interested reader not to stay only with the basic proposals made here, but also to explore other alternatives.

Conflict of interest.- The authors of this article declare no conflict of interest.

Financial support.- This research was supported by a grant from the Spanish Ministry of Science and Innovation (PID2020-112894GB-I00) and a grant from the Catalan Ministry of Universities, Research and the Information Society (2021 SGR 00036).

References

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*, 397–438. https://doi.org/1.1080/10705510903008204

Bandalos, D. (2021). Item meaning and order as causes of correlated residuals in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(6), 903-913. https://doi.org/1.1080/10705511.2021.1916395

Beck, A., Ward, C., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring Depression. *Archives of General Psychiatry*, *4*(6), 561-571. http://doi.org/1.1001/archpsyc.1961.01710120031004

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*(2), 305-314. https://doi.org/1.1037/0033-2909.11.2.305

Boyle, G. (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences*, *12*(3), 291-294. http://doi.org/1.1016/0191-8869(91)90115-R

Briggs, S., & Cheek, J. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality*, *54*(1), 106-148. https://doi.org/1.1111/j.1467-6494.1986.tb00391.x

Burisch, M. (1984). Approaches to personality inventory construction: a comparison of merits. *American Psychologist*, *39*(3), 214-227. https://doi.org/1.1037/0003-066X.39.3.214

Campbell-Sills, L., Forde, D., & Stein, M. (2009). Demographic and childhood environmental predictors of resilience in a community sample. *Journal of Psychiatric Research*, *43*(12), 1007-1012. https://doi.org/1.1016/j.jpsychires.2009.01.013

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265-289. https://doi.org/1.2307/1165285

Diener, E., Emmons, R., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, *49*, 71-75. https://doi.org/1.1207/s15327752jpa4901_13

DiStefano. C., Schweizer, K., & Troche, S. (2022) Editorial: controlling psychometric measures for method effects by means of factor analysis. *Frontiers in Psychology, 13*, 1-2. https://doi.org/1.3389/fpsyg.2022.984050

Edwards, M., Houts, C., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods*, *23*(1), 138-149. https://doi.org/1.1037/met0000121

Ferrando, P. J. (2012). Assessing the discriminating power of item and test scores in the linear factor-analysis model. *Psicológica*, *33*(1), 111-134.

Ferrando, P. J., Lorenzo-Seva, U., Hernández-Dorado, A., & Muñiz, J. (2022). Decalogue for the Factor Analysis of test items. *Psicothema*, *34*(1), 7-17. https://doi.org/1.7334/psicothema2021.456

Henryson, S. (1971). Gathering, analyzing and using data on test items. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 130-159). America Council on Education.

Kline, P. (1987). Factor analysis and personality theory. *European Journal of Personality*, *1*(1), 21-36. https://doi.org/1.1002/per.2410010105

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

Lorenzo-Seva, U. & Ferrando, P.J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods*, *38*(1), 88-91. http://doi.org/1.3758/BF03192753

Lorenzo-Seva, U., & Ferrando, P. J. (2021). Not positive definite correlation matrices in exploratory item factor analysis: causes, consequences and a proposed solution. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(1), 138-147. https://doi.org/1.1080/10705511.202.1735393

Masters, G.N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement*, *25*(1), 15-29. http://doi.org/1.1111/j.1745-3984.1988.tb00288.x

McDonald, R.P. (1999). *Test theory: A unified treatment*. LEA

Morales-Vives, F., Ferrando, P.J., & Dueñas, J.M. (2022). Should suicidal ideation be regarded as a dimension, a unipolar trait or a mixture? A model-based analysis at the score level. *Current Psychology*, 1-15. http://doi.org/1.1007/s12144-022-03224-6

Nunnally, J.C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Reise, S.P., & Waller, N.G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27-48. https://doi.org/1.1146/annurev.clinpsy.032408.153553

Revelle, W. (2021). *Psych: Procedures for psychological, psychometric, and personality research*. Software. R package version 2.1.3. https://CRAN.R-project.org/package=psych

Rigobon, R., & Stoker, T. (2009). Bias from censored regressors. *Journal of Business & Economic Statistics*, *27*(3), 340-353. https://doi.org/1.1198/jbes.2009.06119

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/1.18637/jss.v048.i02

Vigil-Colet, A., Morales-Vives, F., Camps, E., Tous, J., & Lorenzo-Seva, U. (2013). Development and validation of the Overall Personality Assessment Scale (OPERAS). *Psicothema*, *25*(1), 100-106. https://doi.org/1.7334/psicothema2011.411