American Psychological Association

Psychological Assessment

ISSN: 1040-3590

https://doi.org/10.1037/pas0000938

Estimating Classification Consistency of Screening Measures and Quantifying the Impact of Measurement Bias

Oscar Gonzalez[1], A. R. Georgeson[1], William E. Pelham III[2], and Rachel T. Fouladi[3]

[1] Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill

[2] Department of Psychiatry, University of California, San Diego

[3] Department of Psychology, Simon Fraser University

Screening measures are used in psychology and medicine to identify respondents who are high or low on a construct. Based on the screening, the evaluator assigns respondents to classes corresponding to different courses of action: Make a diagnosis versus reject a diagnosis; provide services versus withhold services; or conduct further assessment versus conclude the assessment process. When measures are used to classify individuals, it is important that the decisions be consistent and equitable across groups. Ideally, if respondents completed the screening measure repeatedly in quick succession, they would be consistently assigned into the same class each time. In addition, the consistency of the classification should be unrelated to the respondents' background characteristics, such as sex, race, or ethnicity (i.e., the measure is free of measurement bias). Reporting estimates of classification consistency is a common practice in educational testing, but there has been limited application of these estimates to screening in psychology and medicine. In this article, we present two procedures based on item response theory that are used (a) to estimate the classification consistency of a screening measure and (b) to evaluate how classification consistency is impacted by measurement bias across respondent groups. We provide *R* functions to conduct the procedures, illustrate the procedures with real data, and use Monte Carlo simulations to guide their appropriate use. Finally, we discuss how estimates of classification consistency can help assessment specialists make more informed decisions on the use of a screening measure with protected groups (e.g., groups defined by gender, race, or ethnicity).

Public Significance Statement

Ideally, measures used to classify individuals in psychology and medicine (e.g., as depressed vs. not depressed) produce consistent decisions that do not depend on extraneous features of the individuals (e.g., sex, race, or ethnicity). We provide methods to estimate how consistent the decisions based on a measure are and study whether the consistency varies across groups (e.g., sex, race, or ethnicity).

Keywords: screening, decision consistency, item response theory, classification consistency, measurement bias

Supplemental materials: https://doi.org/10.1037/pas0000938.supp

Screening measures ("screeners") are used in psychology and medicine to identify individuals who are high or low on the measured construct or to supplement a clinician's decision making. A respondent answers a series of items and an overall score is estimated by aggregating the item responses. The estimated score is compared to a cut point to determine a decision about an individual. For example, the Center for Epidemiologic Studies Depression scale (CES-D; Radloff, 1977) is used to screen for depression in the general population, with a common cut point of ≥16 (Vilagut et al., 2016). Individuals above the cut point might be given a full diagnostic interview, whereas those below the cut point receive no further assessment.

When screening measures are used to make decisions about respondents, it is important that the procedure be accurate, consistent, and unbiased. Accuracy is the extent to which the measure succeeds in discriminating the respondents of interest (e.g., those high on anxiety) from the rest of the respondents. Accuracy is measured by indices such as the classification rate, sensitivity, and specificity (Gonzalez & Pelham, 2021; Youngstrom, 2014). An inaccurate screening measure will result in misallocation of resources, such as when those who do not truly have a disorder receive services, or those who do truly have a disorder are denied services.

Oscar Gonzalez https://orcid.org/0000-0001-7122-8799
A. R. Georgeson https://orcid.org/0000-0002-6426-9258
William E. Pelham III https://orcid.org/0000-0003-1480-570X
Rachel T. Fouladi https://orcid.org/0000-0002-2873-6170

This article is based on part of the data published on the master's thesis completed by Wallis (2013). The authors report no conflicts of interest. This project involves analysis of data from Research Grant SSHRC-SRG 410-2006-0395, awarded to Rachel T. Fouladi. William E. Pelham III received support from the National Institute on Alcohol Abuse and Alcoholism Grant AA026768. Relevant ethical guidelines regulating research involving human participants were followed throughout the project.

Correspondence concerning this article should be addressed to Oscar Gonzalez, 235 E. Cameron Ave., Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, United States. Email: ogonza13@unc.edu

Consistency of screening measures is also important. *Classification consistency* is defined as the probability that a respondent would be assigned to the same class (either above or below the screening cut point) across repeated instances of the screening measure, assuming no change between administrations (American Educational Research Association [AERA] et al., 2014). From the perspective of the respondent, an inconsistent screening measure is frustrating and will yield arbitrary receipt of resources and services. This is especially important in high-stakes assessment situations. For example, suppose a person is being evaluated for intellectual disability to apply for services from the Developmental Disabilities Administration (DDA). One criterion for intellectual disability is marked deficits in adaptive functioning (American Psychiatric Association, 2013). The evaluator uses a measure of adaptive functioning that produces accurate but inconsistent classifications. Due to the measure's inconsistency, an applicant might score below the cut point and be denied DDA services one day, when they might have completed the same measure, scored above the cut point and received services the next day. Thus, inconsistent screening measures pose a threat to equity, especially when screening is tied to valuable and potentially life-altering services.

It is also important that screening measures are unbiased. For an unbiased measure, the respondent's expected score depends only on their underlying value on the construct of interest (e.g., depression)—it does not depend on other features or group memberships such as age, race, or ethnicity (Millsap, 2011). The presence of measurement bias [or, equivalently, differential item functioning (DIF)] on a screening measure creates a health inequity—the screening decision a respondent receives depends in part on an extraneous personal quality, rather than their true indication for each of the potential decisions (Gonzalez & Pelham, 2021).

Recently, the intersection of two of these important qualities of screening measures has been studied: accuracy and unbiasedness. The presence of measurement bias in a measure can produce different classification accuracy across groups (Gonzalez & Pelham, 2021). However, to our knowledge, no published work has examined the intersection of measurement bias and classification consistency. This may be due to the fact that researchers in psychology and medicine often report on classification accuracy when developing screening measures (O'Connor, 2018; Youngstrom, 2014) but continue to only rarely report classification consistency statistics (Abdin et al., 2018; Kryuen et al., 2013). If this underreporting is due to a lack of resources to measure respondents repeatedly, a more viable alternative is to report model-based estimates of classification consistency that require only a single administration of the measure (Lee, 2010). Even if the applied researcher uses the available methods (e.g., Millsap & Kwok, 2004) to verify that measurement bias has little impact on classification accuracy, there may yet remain large and important impacts of measurement bias on classification consistency. This is problematic because it could lead to a false sense of security in applying the measure across different respondent groups, perpetuating inequities. Thus, we believe a thorough evaluation of a measure's performance across groups involves studying the impact of measurement bias on *both* the accuracy and consistency of classifications.

The purpose of this article is to help researchers describe the consistency of the screening process across groups and prevent health inequities by estimating and reporting classification consistency estimates. The structure of the article is the following. First, we provide conceptual and technical background on classification consistency. Second, we describe how classification consistency of a screener might be affected by measurement bias across groups (Gonzalez & Pelham, 2021). Third, we illustrate how to estimate classification consistency using data from two studies in which participants responded to the CES-D (Chan et al., 2004; Wallis, 2013). Finally, we present simulation results to guide the appropriate use of two approaches to estimate classification consistency indices (Gonzalez & Pelham, 2021; Lee, 2010).

Classification Consistency

Classification consistency refers to the extent to which respondents are classified to the same category (e.g., having vs. not having depression) over repeated replications of the same measurement procedure (Lee, 2010). Other terms related to classification consistency include screening consistency, decision consistency, or reliability of screening decisions. Because classification consistency assumes no change between the administration of the procedures (e.g., due to an intervention, maturation, carry-over effect, etc.), measurement error is the only factor that affects scores across administrations. Classification consistency is central to determine the validity of a screener and if respondents need further assessment before making a decision.

Classification Consistency and Validation

Validity refers to the amount of evidence and theoretical justification that support a specific interpretation and use of the score (Messick, 1989). For example, scores from the CES-D are alternatively used to represent the depression construct in a research setting or used to screen respondents in a clinical setting, meaning that there are two applications with different requirements. As such, a specific interpretation or use of a score is what is validated, not the measure itself. There are many ways to obtain validity evidence, such as consulting with subject matter experts about the content of the measure, examining the relations among the items in the measure, and assessing the relation of the measure with similar measures or other criteria (AERA et al., 2014). The amount of validity evidence needed to support the interpretation or use of a score is related to how consequential is the proposed use of the measure (Kane, 2006). More validity evidence is needed to support the use of a score in a high-stakes screening scenario (e.g., the screener will determine if a respondent will receive services) than in a low-stakes screening scenario (e.g., the screener will help determine the prevalence of a medical condition in a county). Consider a high-stakes screening scenario in which decisions are not easily reversed, such as respondents losing eligibility for services for a set period of time. If the classifications based on the measure are not consistent, respondents may experience adverse consequences arbitrarily, which in turn jeopardizes the valid use of the measure for that purpose. Similarly, if the proposed use of the measure does not specify a particular population in which it will be used, it is assumed that the measure would work with any respondent regardless of their cultural background, language of origin, or disability status. If this is not the case, the valid use of the screener with specific respondent groups is compromised, leading to health inequity. Therefore, consistent decisions from the screener are needed to support the proposed use of the measure and in turn support the consequences that will arise from it.

Classification Consistency to Determine Rescreening Respondents

Consider a two-tiered screening procedure in which respondents would receive a second screening measure depending on the decision from the first measure (see Khowaja et al., 2018, for an example of a two-tiered screening procedure for autism spectrum disorders). When classification consistency for a respondent's score is low, we may prefer to rescreen (i.e., complete the measure once more or administer a more precise measure) for high-stakes decisions. In the DDA example described above, the state might adopt a policy of automatic rescreening for applicants who received a standard score of between 65 and 75 on the measure of adaptive functioning (the standard cut point is 70). Decision about these bounds could be empirically guided by classification consistency at various observed scores.

Estimating Classification Consistency

One way to estimate classification consistency is an empirical approach: administer the screening measure twice to the same respondents and check if the same decision was made each time. For example, Teitelbaum and Carey (2000) asked 135 participants to complete alcohol use screening measures twice, 1 week apart, and found that between 82% and 91% of respondents were consistently classified as at risk versus not at risk across the administrations. However, it is rare that researchers administer the same screening measure repeatedly in quick succession, given constraints in time, setting, and resources, therefore limiting the feasibility of the empirical approach. This has motivated the development of approaches to estimate classification consistency using data from a single administration of the measure. Some of these approaches use observed scores (e.g., Livingston & Lewis, 1995), parameters from item response models (Lee, 2010; Rudner, 2005), Bayesian approaches (Wheadon, 2014), and nonparametric methods (Lathrop & Cheng, 2014). In this article, we focus on the model-based approach by Lee (2010), popular in the area of educational measurement (Lathrop & Cheng, 2013), for three reasons. First, just as in screening scenarios, the approach by Lee (2010) estimates classification consistency in reference to an observed cut point. Second, the approach by Lee (2010) makes use of item response theory (IRT), a framework that would also allow us to test for DIF. Third, the approach by Lee (2010) has direct links with methodology that has been previously used to estimate the classification accuracy of screening measures (Gonzalez & Pelham, 2021). In the sections below, we provide technical background and describe the classification consistency indices.

Background on IRT

For our application, IRT provides a framework to estimate classification consistency for screening measures. IRT refers to a family of latent variable models used to develop and refine measures, administer scales, and scale scoring (Reise & Waller, 2009; Thissen & Wainer, 2001). A commonly used item response model for items with Likert-type responses is the graded response model (GRM; Samejima, 1969), formally described as,

$$p(x_i = k|\theta) = \frac{1}{1 + e^{-a_i(\theta - b_{ik})}} - \frac{1}{1 + e^{-a_i(\theta - b_{i(k+1)})}}. \quad (1)$$

The output of the GRM are trace lines, which describe the probabilityof endorsing response k on item i as a function of a parameter related to the respondent, θ or the latent variable score, and parameters related to the item, the a- and b-parameters. In this case, θ is the respondent's standing on the latent construct that the measure intends to assess, the a-parameter is the strength of the relation between the item and the latent construct (analogous to a factor loading), and the b-parameters represent the level of θ at which respondents have a 50% probability of endorsing a specific response category or a higher category. If an item contains k categories, then there would be $k - 1$ b-parameters to estimate. By definition, the probability of endorsing the lowest category or higher is equal to 1 and the probability of endorsing category $k + 1$ is equal to 0. For items with binary responses, the GRM reduces to the two-parameter logistic model (2PLM),

$$p(x_i = 1|\theta) = \frac{1}{1 + e^{-a_i(\theta - b_{ik})}}. \quad (2)$$

Several assumptions must hold for the item response model to provide meaningful estimates (Thissen & Wainer, 2001). First, it is assumed that the correct number of dimensions has been specified in the model. For our application, we assume unidimensionality. Closely related, we assume local independence, which states that the items are unrelated once the construct is accounted for (i.e., the probability of endorsing two items conditional on the latent variable is the product of endorsing each of the items conditional on the latent variable). Finally, it is assumed that the probability of endorsing item categories as a function of θ does not vary by group. In the aggregate, violations of this assumption are manifested when different groups of respondents have different a- and b-parameters (Thissen et al., 1993). In this case, we say that there is measurement bias in the screener or that some items exhibit DIF (for reviews of DIF detection procedures, see Millsap, 2011; Teresi et al., 2006). Measurement bias might be troublesome when measures are used to make decisions, as described further below.

Classification Consistency Indices

Two model-based indices of classification consistency are *conditional classification consistency* (CCC) and *marginal classification consistency* (MCC). Both range from 0.5 to 1 (see Appendix for explanation) and higher values are better. CCC is the probability of making the same screening decision at a specific value of θ (e.g., $\theta = 1$ *SD* above the mean). MCC is the average[1] of the CCC estimates across a distribution of θ and can be used to describe how consistent is the screening process for a group of respondents with different values in θ (for more information, see Appendix). Both conditional and marginal classification consistency are dependent on the number of items and cut point imposed and these indices change accordingly if the number of items or the cut point changes. In addition, longer tests, more extreme cut

[1] Specifically, the integration over the θ distribution, which can be approximated using quadrature points (θ values where conditional classification consistency is probed) and normalized weights from the normal distribution. This approach is referred to as the *D* method (Lee, 2010). Researchers could also estimate marginal classification consistency by taking an unweighted average of the conditional classification consistency for specific values of θ, which is referred to as the *P* method (Lee, 2010).

scores, stronger item discrimination parameters, and more precise measurement lead to increased classification consistency (Emons et al., 2007).

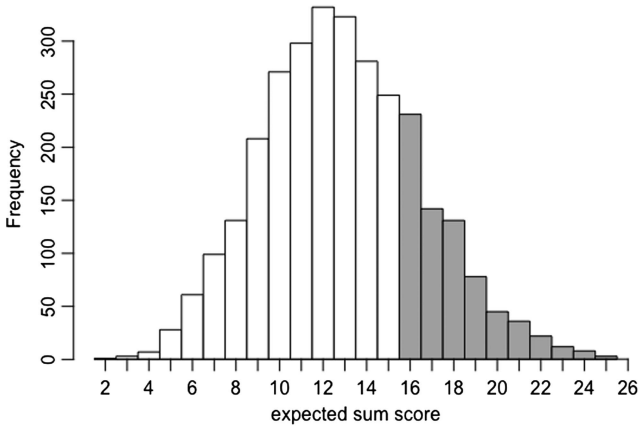
A fundamental step to obtain classification consistency is to estimate $p(\boldsymbol{X}|\theta)$, which is the probability of observing a sum score $\boldsymbol{X}$ on a screener conditional on the latent variable θ.[2] In other words, for every value of θ (i.e., the respondent's standing on the construct assessed by the screener), there would be a distribution of expected sum scores on the measure (see Figure 1). As mentioned above, for classification consistency, we assume that respondents do not change across administration, so the respondent's θ value is fixed, but their observed score $\boldsymbol{X}$ could vary due to measurement error. Because the respondent's θ value is assumed to be fixed, this means that the distribution of observed scores for any repeated administration will be $p(\boldsymbol{X}|\theta)$, and this property is what allows researchers to make inferences about test–retest performance using a single datapoint if assumptions are met. In essence, classification consistency for a θ value is estimated by imposing the screening cut point on the distribution of expected sum scores on the measure, and then estimating the proportion of the distribution that is above and below the cut point. In this case, we could estimate $p(\boldsymbol{X}|\theta)$ analytically using the approach by Lee (2010), or we could simulate data to approximate $p(\boldsymbol{X}|\theta)$ using the approach by Gonzalez and Pelham (2021). Both approaches assume that the item response model fits the data well and that estimates of the item parameters and the distribution of θ (typically assumed to be normal) have been obtained.

Analytical Procedure

Lee (2010) used parameter estimates from the item response models to estimate $p(\boldsymbol{X}|\theta)$ analytically and compute classification consistency indices. The steps to estimate classification consistency indices are the following:

1. Using the estimated item parameters from the GRM (for polytomous items) or 2PLM (for binary items) and a set of θ values (also referred to as quadrature points), estimate the probability of endorsing each item category for each item.

Figure 1
Empirical Distribution of Expected Sum Scores at $\theta = 0$



2. Use a recursive algorithm (Kolen & Brennan, 2004; Lord & Wingersky, 1984; Thissen et al., 1995) to estimate the probability of specific sum score $\boldsymbol{X}$ for each value of θ (further described in Appendix). Note that different patterns of item endorsing can yield the same sum score. For example, with three binary items scored 0/1, there are three ways to obtain a sum score of 2—endorsing item 1 and 2, endorsing 2 and 3, and endorsing 1 and 3. The probability of endorsing each of these response patterns is estimated and then summed.

3. At each value of θ and for a predetermined screener cut point, estimate the proportion of the expected sum score distribution at or above the cut point, p_1, and below the cut point, p_2.

4. Estimate the CCC at a specific θ value by adding $p_1^2 + p_2^2$.

5. Estimate the marginal classification consistency by taking a (un)weighted average of the conditional classification consistencies, weighted by the quadrature weights.

Simulation-Based Procedure

Gonzalez and Pelham (2021) used a simulation-based procedure to examine $p(\boldsymbol{X}|\theta)$ and estimate classification accuracy of screeners. In this case, we extend this procedure to estimate classification consistency. Note that Hambleton and Han also had used simulation-based methods to estimate classification consistency (Bourque et al., 2004; Deng, 2011). However, the two approaches are different. Whereas Hambleton and Han (Bourque et al., 2004) used simulation methods to mimic the administration of two parallel measures, Gonzalez and Pelham (2021) approximate the conditional distribution of the expected summed scores for each value of θ, and then follow steps that largely resemble the approach by Lee (2010). The proposed steps to estimate classification consistency using the simulation-based approach by Gonzalez and Pelham (2021) are the following:

1. Assuming a standard normal distribution for the latent variable, chose quadrature points along the range of the latent variable for which classification consistency will be probed (e.g., 21 equally spaced points in the range of -2.0 and 2.0, as in probing at -2.0, -1.9, -1.8, . . . , 1.9, 2.0). Repeat this vector a large number of times (e.g., 1,000 times).

2. Using the estimated item parameters from the GRM (for polytomous items) or 2PLM (for binary items) and the θ values from step 1, simulate item responses.

3. Sum item responses to estimate $\boldsymbol{X}$ and plot the relation between $\boldsymbol{X}$ and θ.

[2] For researchers familiar with IRT, we could obtain a location-specific standard error of measurement (SEM) for $\boldsymbol{X}$ from the standard deviation of $p(\boldsymbol{X}|\theta)$. If $p(\boldsymbol{X}|\theta)$ were normally distributed, we could use $E(\boldsymbol{X}|\theta)$ and the SEM to estimate classification consistency analytically. Thissen (2000) indicates that $p(\boldsymbol{X}|\theta)$ might not be normally distributed at extreme values of θ, so our approach approximates $p(\boldsymbol{X}|\theta)$ rather than assume its distribution.

4. At each value of θ and for a predetermined screener cut point, estimate the proportion of respondents who are at or above the cut point, p_1, and below the cut point, p_2.
5. Estimate the CCC at a specific θ value by adding $p_1^2 + p_2^2$.
6. Estimate the MCC by taking a (un)weighted average of the CCC estimates across θ values in which they were probed.

Quantifying the Effect of Measurement Bias on Classification Consistency

In most screening applications, screening decisions are made based on observed scores $\boldsymbol{X}$ estimated by summing item responses x. When observed scores across g groups are compared to each other or to a common reference (such as a common cut point in screening procedures), measurement invariance is assumed (Millsap, 2011).[3] Measurement invariance, which is the *absence of* DIF, can be formally expressed as,

$$p(x|\theta) = p(x|\theta, g), \tag{3}$$

where the probability of observing item response x on an item, conditional on the latent variable θ, does not depend on the group g to which respondents belong. If measurement invariance holds, then the measure is free of measurement bias (i.e., free of DIF). In practice, measurement invariance might not always hold. For example, previous research suggests that at the same level of the latent variable, women tend to report greater levels of physical and emotional distress than men, Hispanics tend to use more extreme response styles than non-Hispanics, and older people tend to report more positive self-views than younger people (McHorney & Fleishman, 2006; Reise & Waller, 2009). Although it is important to consider measurement bias on each individual item, it is also important to consider the overall pattern of bias across items. If different items are biased against different groups, then the biases may cancel each other out and produce a score that is unbiased when the item responses are aggregated (Chalmers et al., 2016). In other words, item-level bias might not always affect the decisions made from sum scores.

In screening scenarios, the presence of measurement bias could manifest in three ways. Compared to other participants at the same value on the latent construct, respondents from protected groups could be (a) more likely to be above the screener's cut point, (b) less likely to be above the screener's cut point, or (c) equally likely to be above the screener's cut point, due to the cancellation of bias across items (Gonzalez & Pelham, 2021). Given the consequences of screening decisions (e.g., more assessment, referrals, or estimation of prevalence rates), it is imperative to assess how item-level bias might affect screening decisions when working with protected groups, and specifically, the consistency of those decisions.

Recently, procedures have been developed to describe the effect of measurement bias as changes in sensitivity and specificity for screening measures (Gonzalez & Pelham, 2021; Gonzalez et al., 2020), or selection procedures, in general (Lai et al., 2017; Millsap & Kwok, 2004). For example, after carrying out these procedures, researchers might find that if measurement bias in the measure is ignored, sensitivity to screen Hispanic respondents for anxiety might drop by 5%. We believe that describing the effect of measurement bias in terms familiar to assessment specialists could empower them to make more informed decisions about the use of screening measures on protected groups. With similar motivations, one could use changes in classification consistency to examine how measurement bias might affect the reliability of the screening decisions. Consider a situation in which bias in a screener could lead respondents from protected groups (e.g., ethnic or racial minorities) to receive more or less consistent decisions from an assessment. Consequently, certain groups of respondents might be more likely to receive arbitrary assignment to treatment or referrals than members from a majority group (McHorney & Fleishman, 2006). Changes in classification consistency could indicate whether the individual would still be flagged by the screener after repeated administration in situations in which measurement bias is ignored or accounted for.

To estimate how measurement bias affects classification consistency, the analytical procedure by Lee (2010) and the simulation-based procedure by Gonzalez and Pelham (2021) must be extended to handle multiple groups. Both extensions assume that researchers were able to accurately flag items that exhibit DIF.

Extending the Analytical Approach

The approach by Lee (2010) would be carried out two times just as described above, each time using the same θ estimates. First, a multiple-group IRT model in which DIF-free items have invariant item parameters and items exhibiting DIF have group-specific item parameters is fit to the data set. Then, classification consistency indices are estimated for each group using their respective item parameters. In this situation, DIF is accounted for by allowing items with DIF to have group-specific item parameters. Then, a multiple-group IRT model in which all item parameters are invariant across groups is fit to the data set, and classification consistency indices are estimated. In this situation, DIF is ignored because all the item parameters are the same across groups. Finally, the conditional and marginal classification consistency estimates per group are compared to the corresponding estimate when all the items are treated as invariant. Differences in classification consistency estimates would demonstrate how measurement bias affects the reliability of the screening decisions.

Extending the Simulation-Based Approach

Just as in the extension to the analytical approach, the simulation-based approach by Gonzalez and Pelham (2021) would be carried out two times—(a) a multiple-group IRT model in which only the items that exhibit DIF have group-specific item parameters and (b) a multiple-group IRT model in which all items are treated as invariant. Note that if the groups differ on the θ distribution, then θ values would have to be simulated with respect to a mixed distribution (in this case, by sampling θ values from group-specific distributions). Similarly, the effect of measurement bias on classification consistency is captured by the differences in classification consistency

[3] In this article, we assume that scores from each groups (e.g., males and females) are compared to the same cut point. However, some published measures use group-specific cut points, which may or may not reduce measurement bias. Researchers are encouraged to report classification consistency per group using the group-specific cut points.

indices between group-specific estimates and estimates from the fully invariant model.

Illustration

We illustrate the proposed methods with two examples. In the first example, we conduct secondary data analysis of an existing data set that involved repeated, back-to-back administration of the CES-D. We use this data to compare (a) the model-based classification consistency estimates based on a single administration of a scale to (b) the *empirical* classification consistency estimate using two administrations of a scale. In the second example (found in the supplementary materials), we provide a completely reproducible illustration of estimating the classification consistency of a measures using published item parameters (Chan et al., 2004).

Data

This study was conducted in compliance with (Simon Fraser University) ethics guidelines and Human Subjects Approval from the Institutional Review Board. The data set for this illustration comes from a larger project on response patterns in self-administered questionnaires (see Author note), and on which P. Wallis's Master's thesis on the effect of repeated assessment, questionnaire format, item features, as well as respondent characteristics on participants' responses to the CES-D was based (Wallis, 2013).[4] This project is of interest and a source of data for our illustration because it involves the repeated administration of select questionnaires, including the CES-D (Radloff, 1977), in a single session. With the repeated measures data on the CES-D and data on a focal covariate (e.g., sex), we can contrast classification from a single administration with classification from more than one administration (specifically, two administrations). The three goals of our illustration are to (a) estimate the classification consistency of the CES-D using a single administration (e.g., administration 1), (b) evaluate how measurement bias affects classification consistency, and (c) compare the classification consistency estimate with one administration of the scale to the empirical classification consistency estimate using two administrations of the scale (e.g., administration 1 and 2). For our illustration, we focus on data for a subset of the study participants considered in the study by Wallis (2013).[5]

Participants

The subset of the sample we used consists of 436 undergraduates from a medium-sized university in a metropolitan area in western Canada. Respondents completed the CES-D in English, along with a larger battery of measures via the computer, in a one hour session.

All participants were administered the CES-D in different webpage layouts of the CES-D. The focal sequential administrations for this illustration are (a) when all respondents received all of the items in a single html page, with radio-button response options placed beside the item prompts in a grid- or matrix-like layout and (b) when all respondents received all of the items in a single html page, with radio-button response options placed below the item prompts in typical vertical multiple-choice layout. For simplicity of illustration, we dropped respondents who had missing item responses for any of these two sequential focal administrations of the CES-D, which we are calling administration 1 and administration 2. As such, the final overall sample size for our illustration analyses was $N = 371$ in which the mean age was 19.777 ($SD = 2.192$), 65.0% self-identified as female, 48.1% indicated that English was their first language, and 53.4% self-identified as Asian, 30.5% as Caucasian, and 16.1% in other categories.

Focal Measure

The CES-D consists of 20 items rated with four response categories, ranging from 0 to 3 (Radloff, 1977). Responses from the CES-D are typically aggregated into a single sum score, and a common cut point for screening is ≥16 (Vilagut et al., 2016). It is important to note that prior studies suggest that the factor structure of the CES-D may be complex (e.g., Carleton et al., 2013; Edwards et al., 2010). However, we retain a unidimensional model to match how CES-D scoring is done in practice—Edwards et al. (2010) reviewed articles in major assessment journals from 2000 to 2010 and found that 107 out of 114 studies used only a total score on the CES-D. Our unidimensional model matches the theoretical model that applied researchers assume when they use the CES-D sum score, but we anticipate that we will not meet conventional cutoffs for model fit statistics.

Focal Covariate

For this illustration, we tested whether the CES-D items exhibited DIF as a function of self-identified sex, and whether ignoring any measurement bias adversely impacts the screening process (possibly leading to health inequities).

Preliminary Analyses

Below, we divide our discussion of the preliminary analyses by first addressing the analyses of test–retest data from administration 1 and 2, and then the analyses to estimate classification consistency using administration 1 only.

[4] We recognize that there are important limitations in using this example, and we discuss those limitations accordingly. However, we do not believe that the limitations detract from the illustration of the methods, so we ask readers to focus on the general applications. We do not intend for this article to make a theoretical contribution about the use of the CES-D or comment on the impact of different questionnaire format/layouts.

[5] For the whole session, participants received the CES-D three times in a short time frame, each time in a different format/layout. The respondents that were the focus of Wallis' (2013) study, received (a) one item per page with response options below the single item prompt and (b) all items in one page in which item prompt and response options in grid-/matrix-layout in counterbalanced order; data on these two formats/layouts were the focus of Wallis's thesis. These respondents all received (c) all items in one page with response options below the item prompt in multiple choice format as the third administration format. To simplify the illustration, we use the second and third administrations of the scale because the administration can be considered similar insofar as all items were in a single page and because prior research found that there were time effects between the first and the second timepoint (Wallis, 2013) that are minimally present between second and third timepoints (Wallis & Fouladi, n.d.). Two consequences of this decision are that there may be carry-over effects which could inflate the empirical classification consistency between the second and the third administration of the scale, and that prior exposure to the CES-D items may affect the empirical classification consistency and classification consistency estimates from one administration.

Multiple Administrations (i.e., Test–Retest Data)

Reliability estimate using Cronbach's alpha of the CES-D scores for administration 1 was $\alpha = .88$ (95% confidence interval (CI): .86, .90), and for administration 2 was $\alpha = .89$ (95% CI: .86, .90). For the overall sample, the Pearson correlation between the summed CES-D scores for administration 1 and 2 was $r = .977$ (standard Fisher transform based 95% CI: .97, .98, cf., Fouladi & Steiger, 2008). The mean estimate of the CES-D score for administration 1 was 14.154 ($SD = 9.030$) and for administration 2 was 14.072 ($SD = 9.313$). The mean difference of .081 ($SD = 1.983$; 95% CI: −.122, .283) was not statistically significant, $t(370) = .785$, $p = .433$. The empirical classification consistency across administrations for respondents who had CES-D scores ≥16 was .930 (95% CI: 904, .955). In other words, 93% of respondents received the same classification at administration 1 and 2. In our analysis split by the covariate of interest, the correlation between the summed CES-D scores at administration 1 and 2 for males was $r = .981$ (95% CI: .97, .99) and for females was $r = .975$ (95% CI = .97, .98). For males, the mean CES-D score at administration 1 was 12.908 ($SD = 8.438$), and at administration 2 was 12.916 ($SD = 8.766$). The mean difference of −.008 ($SD = 1.730$; 95% CI: −.307, .291) for males across administrations was not statistically significant, $t(130) = -.051$, $p = .960$. On the other hand, the mean CES-D score for females at administration 1 was 14.833 ($SD = 9.284$), and at administration 2 was 14.704 ($SD = 9.557$). The mean difference of .129 ($SD = 2.111$; 95% CI: −.139, .398) was not statistically significant, $t(239) = 0.948$, $p = .344$. The CES-D scores between males and females were significantly different at administration 1, ($t(369) = -1.970$, $p = .050$; 95% CI: −3.845, −.003; $\eta^2 = .010$), but not at administration 2 ($t(369) = -1.773$, $p = .077$, 95% CI: −3.772, .195). Finally, the empirical classification consistency across administrations for males was .939 (95% CI: 898, .981) and for females was .925 (95% CI: .892, .958).

Single Administration

A unidimensional GRM was fit to the CES-D item responses at administration 1 using the mirt R package (Chalmers, 2012), and the model fit was $C_2(170) = 977.006$, $p < .001$, Comparative Fit Index (CFI) = .882, Root Mean Square Error of Approximation (RMSEA) = .113, Standardized Root Mean Square Residual (SRMR) = .088 (Cai & Monro, 2014). As mentioned above, poor model fit was expected, but we decided to examine classification consistency using this model because using a single CES-D summed score in screening applications assumes unidimensionality of its item responses. Then, we fit a multiple-group GRM[6] and carried out a DIF detection procedure based on Likelihood Ratios (LR) (IRT–LR–DIF procedure; Thissen et al., 1993) in mirt to identify items that were likely to exhibit DIF as a function of sex. The fit of the multiple-group GRM[7] was $C_2(340) = 1{,}150.717$, $p < .001$, CFI = .880, RMSEA = .080, SRMR for males = .108, SRMR for females = .092. The IRT–LR–DIF procedure, applying the Benjamini-Hochberg (1995) adjustment for multiple testing, identified three items that exhibited DIF (see supplement for item parameters). Finally, we fit a multiple-group GRM in which item parameters of DIF-free items were constrained to equality across groups, and items that exhibited DIF had group-specific item parameters. The fit of this model was $C_2(408) = 1{,}208.791$, $p < .001$, CFI = .882, RMSEA = .073, SRMR for males = .116, SRMR for females = .093. Males were the reference group in the multiple-group model (mean = 0, $SD = 1$), and results suggested that females had a higher latent mean (mean = .142), and a similar latent standard deviation ($SD = 1.010$).

Procedure to Estimate Classification Consistency

Classification consistency was estimated using the analytic approach by Lee (2010) and the simulation-based approach by Gonzalez and Pelham (2021) as described above. First, we estimate classification consistency assuming measurement invariance (i.e., ignoring potential DIF) in which both males and females have the same item parameters, but latent mean values and variances were fixed to the estimates from the DIF model. As such, we needed three pieces of information: item parameters, θ values (i.e., the quadrature points) to probe classification consistency, and an observed cut point. Then, we compared classification consistency in a model that accommodates items with DIF in which items that exhibit DIF across sex have group-specific item parameters and DIF-free items have the same item parameters. As such, we needed five pieces of information: the item parameters for males, the item parameters for females, θ values to probe classification consistency for males, θ values to probe classification consistency for females, and the common cut point in the observed score. The differences in classification consistency from estimates that assume no DIF and the estimates that account for DIF provide insight into the impact of measurement bias on the consistency of screening decisions.

Results

For both approaches, we used 41 quadrature points for θ (ranging −2 to 2) to probe classification consistency. Because the simulation and the analytic approaches largely gave the same solution, only the simulation-based results are presented. At a cut point of ≥16, the weighted marginal classification consistency (MCCd) estimate when invariance was assumed was MCCd = 0.864, and the unweighted estimate (MCCp) was MCCp = .910. The left panel of Figure 2 shows the CCC estimates across the range of θ. The desired shape of this curve is a narrow U/V shape in which the vertex is at the θ score that corresponds to the observed cut score. According to the test characteristic curve, a cut point of ≥16 corresponds to a θ score of approximately 0.50. As such, Figure 2 shows that respondents who are outside the range [−0.10, 0.95] on θ have a classification consistence above .90. However, inside the range [−0.10, 0.95], classification consistency drops sharply. As expected, there will be less certainty in the classification of those who are close to the cut point.

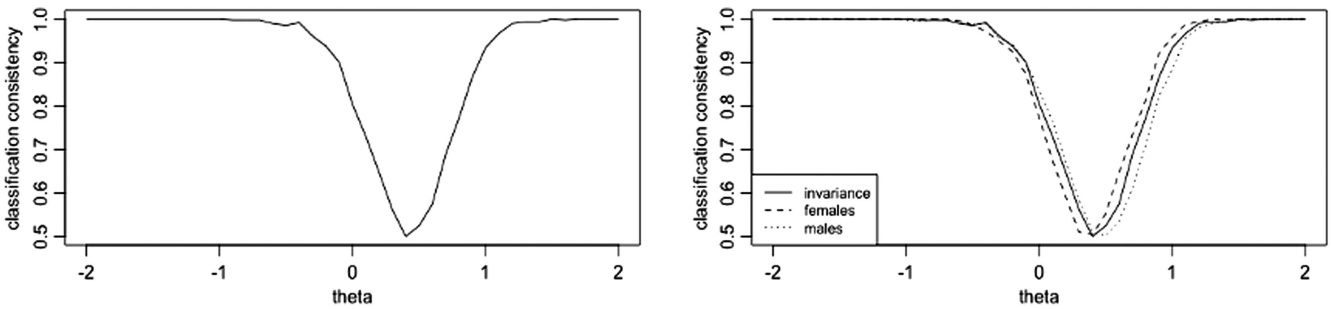
Furthermore, the right panel of Figure 2 shows the CCC estimates across the range of θ for males and females, along with estimates assuming measurement invariance. Marginal classification consistency estimates were MCCd = 0.860 and MCCp = .905 for males and MCCd = 0.859 and MCCp = .911 for females. Given that the

[6] We collapsed the most extreme response category for item 17 because males did not endorse that response. Collapsing did not change the classification consistency estimates in the test–retest analyses.

[7] Goodness of fit indices are all based on single sample calculations (Steiger & Lind, 1980). Multisample goodness of fit indices are available for multiple-group confirmatory factor analyses (CFAs) by adjusting the fit estimates by the number of groups (Dudgeon, 2004; Steiger, 1998), and corresponding confidence intervals could be derived (Steiger & Fouladi, 1997). However, guidance for the fit of item response models has not been investigated. For guidance on multiple-group CFA, readers may be interested in a recent doctoral dissertation by Brace (2020).

Figure 2

Left: Classification Consistency as a Function of θ for the Invariant Condition. Right: Classification Consistency as a Function of θ for Males, Females, and the Invariant Condition



Note. The empirical conditional classification consistency curves are expected to be jagged because they are probed at discrete values on the θ range.

procedure is based on simulations, we could carry out the procedure multiple times to estimate a Monte Carlo confidence interval for the difference of MCCd across respondent groups. With 100 Monte Carlo draws, the mean difference between the MCCd of males and females was .004 with Monte Carlo standard error of .003 so the 95% confidence interval was [−.002, .010]. Also, the CCC estimates were different at specific values of θ. For example, at θ = 0, CCC estimates were CCC = 0.804 for the invariance condition, CCC = 0.835 for males, and CCC = 0.772 for females. The right panel of Figure 2 shows that for males, classification is more consistent below the cut point and less consistent above the cut point, compared to when invariance is assumed. The opposite is true for females: Classification is less consistent below the cut point and more consistent above the cut point compared to when invariance is assumed. Therefore, the results suggest that measurement bias could affect the consistency of the CES-D to screen respondents for depression at specific values of θ. Conservatively, the classification consistency of respondents outside the range [−0.15, 1.00] on θ is not affected by the DIF due to sex (i.e., outside the range [−0.15, 1.00] on θ, classification consistency is above .90).

Estimates Based on Single Versus Multiple Administrations

Finally, we compared the empirical classification consistency across administration 1 and 2 with the classification consistency estimate from administration 1. The empirical classification consistency of .925–.939 was close to the unweighted MCC estimates of .905–.911, and was higher than the weighted MCC estimates of .859–.864. There are several potential reasons why a difference between these estimates was observed, and some of which might be related to study design (see Footnote 5). First, because the CES-D was readministered in close succession, the inflated empirical classification consistency could have been due to carry-over effects across administration. Second, the poor model fit of the unidimensional model could have led to underestimated classification consistency estimates. Third, perhaps, the empirical classification consistency and classification consistency estimates were affected due to prior exposure of the CES-D items before administration 1 (as noted in Footnote 5). Future research clarifying the relation between empirical classification consistency from two administrations (or timepoints) and the classification consistency estimate from a single timepoint would be beneficial, and we encourage researchers who only have a single datapoint to estimate classification consistency to describe the reliability of screening decisions.

Simulation Study

In this section, we present the results from our simulation study on the performance of the simulation-based approach by Gonzalez and Pelham (2021) and the analytic approach by Lee (2010). The main goal of the simulation is to provide guidance on the general use of each approach and, as observed in our illustration, examine if both approaches yield the same estimate across conditions. The two approaches are slightly different—although the approach by Lee (2010) is popular in educational testing and specific to IRT models, the simulation-based approach provides a more general framework to estimate classification consistency from other models [see Gonzalez et al. (under review), for an application of the simulation-based approach to estimate classification consistency for machine learning models]. Also, another purpose of the simulation study is to evaluate the number of simulation iterations per θ value needed to obtain stable estimates and determine its appropriate use. It is expected that the estimates by the simulation-based approach would approximate the analytical estimates of classification consistency as the number of item categories and iterations per θ value increase.

Data Generation

Data were generated from a multiple-group, standardized categorical factor model using the delta parameterization, similar to the simulation by Gonzalez and Pelham (2021). Three different screening scenarios were studied. Scenario 1 had items parameters and θ distributions ($\mu_g = 0$ and $\sigma_g^2 = 1$) invariant across two groups. Scenario 2 had invariant parameters across groups, but the θ distribution for group 1 was $\mu_1 = 0$ and $\sigma_1^2 = 1$ and the θ distribution for group 2 was $\mu_2 = 1$ and $\sigma_2^2 = 0.64$. Finally, scenario 3 had the same θ distribution across groups ($\mu_g = 0$ and $\sigma_g^2 = 1$), but two items exhibited DIF—the values of the two largest factor loadings for group 2 were half of the size of group 1, and the thresholds of those items were 0.50 higher. We expected that group differences in the *a*- and *b*-parameters would lead to different $p(\boldsymbol{X}|\theta)$ across groups, which in turn would affect the classification consistency per group.

In the simulation, we varied the number of items (from 5 to 15), the number of response categories (2 or 5), and the number of θ values sampled (50, 125, and 250) for each group per screening scenario. Note that the number of cases for group 1 and group 2 were balanced, so the total θ values sampled were 100, 250, and 500 per replication. Items were generated so that their scale was unity. The factor loadings were equally spaced between 0.30 and 0.70 (e.g., for conditions with five items, the factor loadings were 0.30, 0.40, 0.50, 0.60, and 0.70), and the residual variance was 1 *minus* the factor loading squared for each corresponding item. The thresholds were set at 0 for conditions with two-category items and at −1.5, −.50, 0.50, and 1.5 for conditions with five-category items. Overall, there were 198 conditions examined in this simulation (11 number of items × 2 number of response categories × 3 number of θ values sampled × 3 screening scenarios), with 500 replications per condition.

General Procedure

First, the loadings and thresholds were transformed from the factor-analysis metric to the IRT metric using the following relations (Wirth & Edwards, 2007):

$$a_i = \frac{D\lambda_i}{\sqrt{1-\lambda_i^2}}; b_{ik} = \frac{\tau_{ik}}{\lambda_i}. \quad (4)$$

In this case, a_i is the *a*-parameter for item *i*, b_{ik} is the *k*th *b*-parameter for item *i*, *D* is a scaling constant of 1.7 to convert IRT logistic estimates to normal-ogive estimates, τ_{ik} is the *k*th threshold in the factor-analysis metric, and λ_i is the factor loading in the factor-analysis metric. Then, we used the functions presented in the supplement to estimate MCC and CCC estimates using the simulation-based approach. In this case, we used 41 equally spaced points that were ±2.0 standard deviations from the μ_g (e.g., if $\mu_g = 0$ and ${\sigma_g}^2 = 1$, then classification consistency was probed from $\theta = -2.0$ to $\theta = 2.0$). Finally, the analytical approach by Lee (2010) was carried out with the cacIRT R-package (Lathrop, 2015). We used a cut score of ≥3 for conditions with binary items and a cut score of ≥12 for conditions with polytomous items. As mentioned above, we are not comparing classification consistency estimates across conditions because we would have to adjust for the number of items and cut point changing, so we only examined if the simulation-based estimates are accurate within each condition.

The main simulation outcomes were the relative difference across estimates and the variability of the simulation-based estimate. Relative difference was estimated by subtracting the classification consistency estimate of the simulation-based approach from the estimate of the analytical approach, divided by the estimate of the analytical approach, averaged across replications. Variability of the simulation-based estimate was described by the standard deviation of the empirical estimates in each condition. Similarity of the analytical and simulation-based estimates was deemed appropriate if the relative difference is below 0.05 and the empirical standard deviation estimate is around 0.025 (i.e., simulation-based estimates differ by ±0.05).

Results

Similar conclusions were observed across the three screening scenarios studied, so we present the results for the first scenario below, and results for the second and third scenarios in the supplement. Groups are fully invariant in the first scenario, so the results are presented as if only one group was examined. Results are presented by group for the second and third scenarios.

Table 1 shows the relative difference and empirical standard deviation of the MCC estimate across the number of items, item categories, and iterations of the θ values. Across conditions, the estimates of MCC showed small relative differences and precise estimates. As expected, the relative difference and the empirical standard deviation decreased as the number of item categories and iterations per θ value increased. Therefore, the results suggest that the simulation-based approach is largely similar to the analytical estimates in the fully invariant screening scenario with as few as 100 iterations (i.e., 50 iterations per group) of the θ values. Furthermore, CCC estimates had small relative differences across conditions, and the standard deviation of the estimate was the lowest in conditions with 500 iterations of the θ values (see supplement for tables at selected θ values). Similarly, Figure 3 shows the CCC estimates by the analytical approach, along with ±1 *SD* of the simulation-based estimates, for conditions with 7 and 11 items. Figure 3 suggests that the simulation-based estimates have a similar trend in the relation between the CCC and θ as the analytical estimates. Overall, the results suggest that the simulation-based approach led to similar CCC estimates as the analytical approach in the fully invariant scenario with at least 500 iterations of θ values (i.e., 250 iterations per group).

General Discussion

In applied settings, assessment specialists desire that the screening process is accurate, consistent, and fair. In previous applications, researchers have studied how accuracy and measurement bias affect the screening process, but reporting classification consistency has been largely ignored. This lack of reporting may be due to insufficient resources to administer the screening measure multiple times, meaning that model-based estimates of classification consistency that only require a single administration may be useful. If researchers do not establish decision consistency across respondents or respondent groups (e.g., protected groups), then the use of scores for screening might not be valid, the allocation of resources may be arbitrary, and health disparities might increase (Manly, 2006). Therefore, researchers could report classification consistency indices per respondent group to determine if classification consistency is the same, thus guide measure use and improve the transparency and fairness of the decision process.

In this article, we discussed the importance of classification consistency in the validation of screening measures and its role on tiered screening, reviewed technical aspects on the estimation of classification consistency, and illustrated the methods with two examples (one in text and one in the supplement). The proposed method yields several outputs that researchers may find useful. For instance, the CCC curves (as shown in Figure 2) describe the range of latent variable scores, θ, where decisions based on the screener are consistent (e.g., decisions might be consistent outside a specific θ range), so researchers could use them to identify the individuals for whom repeated administration is unnecessary as it would likely yield the same decision. Assessment specialists could alternatively use the CCC curves to recommend individuals who would most benefit from rescreening. For example, a horizontal line could be drawn on the left panel of Figure 2 to determine the minimum CCC estimate for which assessment specialists would feel comfortable making a decision (e.g., a horizontal line is drawn at 0.90 because rescreening is

Table 1
Relative Difference and Empirical Standard Deviation [in Brackets] of the Marginal Consistency Classification Estimate by the Simulation-Based Approach

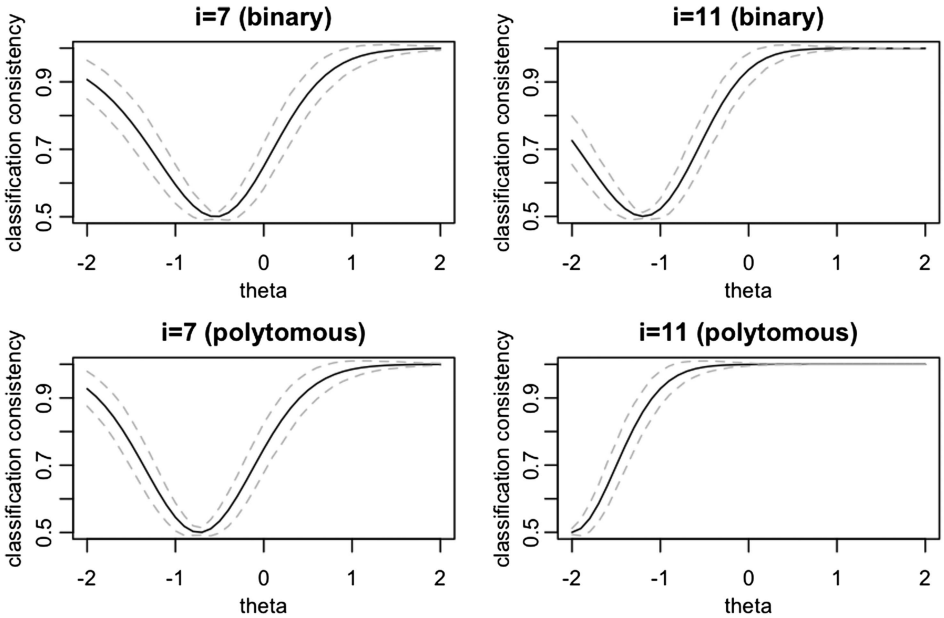
	Iterations per θ value					
	Two-category items			Five-category items		
Items	100	250	500	100	250	500
5	.003 [.006]	.001 [.003]	.001 [.003]	.003 [.006]	.001 [.004]	.001 [.002]
6	.003 [.006]	.001 [.003]	.001 [.003]	.003 [.006]	.001 [.003]	.001 [.002]
7	.002 [.005]	.001 [.002]	.000 [.002]	.002 [.005]	.001 [.003]	.000 [.002]
8	.002 [.005]	.001 [.002]	.000 [.002]	.002 [.005]	.001 [.003]	.001 [.002]
9	.002 [.005]	.001 [.002]	.000 [.002]	.002 [.004]	.001 [.003]	.000 [.002]
10	.002 [.005]	.001 [.002]	.000 [.002]	.001 [.003]	.001 [.002]	.000 [.002]
11	.002 [.004]	.001 [.002]	.000 [.002]	.001 [.004]	.000 [.002]	.000 [.002]
12	.002 [.004]	.001 [.002]	.000 [.002]	.001 [.003]	.001 [.002]	.000 [.001]
13	.002 [.004]	.001 [.002]	.000 [.002]	.000 [.003]	.000 [.002]	.000 [.001]
14	.002 [.004]	.001 [.002]	.000 [.002]	.000 [.003]	.000 [.002]	.000 [.001]
15	.002 [.004]	.001 [.002]	.000 [.002]	.000 [.002]	.000 [.001]	.000 [.001]

Note. Group 1 and group 2 would have the same estimate of classification consistency. *Items* refers to the number of items. The cut point for conditions with binary items was ≥3 and the cut point for conditions with polytomous items was ≥12. Relative difference is the estimate of the simulation-based procedure minus the estimate of the Lee (2010) procedure, divided by the estimate of Lee (2010) procedure.

recommended for individuals in the θ range in which CCC is below 0.90). The points for which the horizontal line crosses the curve are noted, and the corresponding points on the horizontal axes would be the θ range in which participants receive further assessment.

In addition, we discussed how ignoring measurement bias (i.e., DIF) affects the consistency of screening decisions across groups. In our illustration, we found that there is measurement bias in the CES-D as a function of sex, but it did not translate to practical differences in classification consistency across males and females (i.e., mean difference was small and the confidence interval contained zero). In a hypothetical scenario in which measurement bias across sex was to materialize into lower classification consistency estimates for females, then females would be more likely to receive an arbitrary classification than males, which could affect who gets access to services and lead to overall health inequities. Empirically, our illustration suggests that ignoring measurement bias may widen

Figure 3
Conditional Classification Consistency Estimates by the Analytical and the Simulation-Based Approach Across Number of Iterations per θ Value



Note. Solid lines are the analytical estimates and dashed lines are the simulation-based estimates ± 2 empirical standard deviations. *i* is item, and polytomous conditions have five response-categories. The cut score in conditions with binary items was ≥3 and in conditions with polytomous items was ≥12.

the θ range in which the screener makes inconsistent decisions. From a DIF perspective, the difference on the conditional or the marginal classification consistency estimates when DIF is accommodated versus ignored could provide an interpretable estimate of how measurement bias affects screening (Millsap & Kwok, 2004). That is, researchers could determine that the reliability of the decisions in a specific θ range is not enough for the desired purpose, or that classification consistency is too low for a protected group and might place them at a disadvantage for resources. Similarly, assessment specialists could determine the θ range in which DIF is not differentially affecting the classification consistency across groups. Most DIF detection procedures do not evaluate whether item bias materially affects screening decisions, so translating the effects of measurement bias to terms familiar to assessment specialists (e.g., changes in classification consistency) empowers them to make more informed decisions about the tools that they use.

Finally, we presented simulation results to guide the use of two approaches to estimate classification consistency: an analytical approach by Lee (2010) and a simulation-based approach by Gonzalez and Pelham (2021). Results suggest that the estimates by the simulation-based procedure and the analytical procedure were similar when 500 iterations per θ value were used to estimate marginal or CCC. Thus, assessment specialists could use either of these two procedures to estimate classification consistency in their own data (see supplement for R code to carry out these two procedures). Analytical estimates are preferable to simulation-based estimates when available, but preliminary research suggests that the simulation-based procedure could be used to estimate consistency of other models used for screening, such as machine learning methods (Gonzalez et al., under review). However, these extensions of the simulation-based procedure require a more investigation. Overall, these two approaches could be complementary and applicable to a variety of screening scenarios.

Limitations and Future Directions

There are several limitations of the methodology discussed in this article that, if addressed, could further generalize the utility of the methods discussed. It would be valuable to continue to investigate the relationship between classification consistency estimates from a single administration to the empirical classification consistency in test–retest data (MacKinnon et al., 2018). In our empirical example, we observed that our methodology for estimation of classification consistency underestimated the empirical classification consistency of the test–retest data. Although there were several conditions in characteristics of the example that could explain this discrepancy (e.g., poor model fit, carry-over effects in test–retest data, or prior exposure of respondents to CES-D items), a future direction would be to determine if this pattern appeared in other empirical applications of the CES-D. Also, both the Lee (2010) and Gonzalez and Pelham (2021) methods to estimate classification consistency assume that the item response model fits the data well, but as shown in our empirical example, this might not always be the case. Future studies will investigate how model misspecification impacts the estimation of classification consistency (MacCallum et al., 2001) and to evaluate whether nonparametric IRT approaches are more robust to model misspecification (Lathrop & Cheng, 2014). Similarly, the estimation of classification consistency indices assumes that we have true item parameters, but the item parameters are estimates that are paired with standard errors. A Bayesian approach to estimate classification consistency, along with credibility intervals, may better account for the uncertainty in item parameter estimation (Levy & Mislevy, 2016; Wheadon, 2014).

Another interesting future direction will be to investigate if decisions based on two screeners that claim to measure a similar construct (i.e., two measures that are not parallel) yield the same conclusion. For example, Marrie et al. (2018) administered a series of depression and anxiety screeners (e.g., the Patient Health Questionnaire, PHQ-9 and the PROMIS Depression short-form) to respondents with multiple sclerosis. Using methods from this article, one could predict whether a respondent flagged with the PHQ-9 is also likely to be flagged by the PROMIS Depression short-form. Although these procedures would assume that the measures are exchangeable for the same purpose and populations, decision similarity could serve as validity evidence in the development of a new screening measure. Finally, classification consistency could be used for the selection of cut points. When cut points change, it is likely that classification consistency will also change. Thus, future work could examine methods similar to receiver operating characteristic (ROC) curves to find cut points that maximize both classification accuracy and consistency (Youngstrom, 2014).

Overall, we encourage assessment specialists to report estimates of classification consistency when using a screening measure to make decisions about individuals. Reporting classification consistency separately by respondent group makes the screening process more transparent and equitable across protected groups. By simultaneously evaluating classification accuracy, classification consistency, measurement bias, and the impact of measurement bias on accuracy and consistency, assessment specialists can place a screening measure on strong footing for useful and equitable deployment in real-world settings.

References

Abdin, E., Sagayadevan, V., Vaingankar, J. A., Picco, L., Chong, S. A., & Subramaniam, M. (2018). A non-parametric item response theory evaluation of the CAGE instrument among older adults. *Substance Use & Misuse*, *53*, 391–399. https://doi.org/10.1080/10826084.2017.1332645

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed.).

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B. Methodological*, *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bourque, M. L., Goodman, D., Hambleton, R. K., & Han, N. (2004). *Reliability estimates for the ABTE tests in elementary education, professional teaching knowledge, secondary mathematics and English/language arts* [Final Report].

Brace, J. C. (2020). *Relaxed methods for evaluating measurement invariance within a multiple-group confirmatory factor analytic framework*. [Doctoral Dissertation]. University of British Columbia, Vancouver, Canada.

Cai, L. & Monro, S. (2014). *A new statistic for evaluating item response theory models for ordinal data*. [National Center for Research on Evaluation, Standards, & Student Testing. Technical Report].

Carleton, R. N., Thibodeau, M. A., Teale, M. J., Welch, P. G., Abrams, M. P., Robinson, T., & Asmundson, G. J. (2013). The center for epidemiologic

studies depression scale: a review with a theoretical and empirical examination of item content and factor structure. *PLOS ONE*, *8*(3), Article e58067.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1–29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: Improved differential test functioning statistics that account for sampling variability. *Educational and Psychological Measurement*, *76*, 114–140. https://doi.org/10.1177/0013164415584576

Chan, K. S., Orlando, M., Ghosh-Dastidar, B., Duan, N., & Sherbourne, C. D. (2004). The interview mode effect on the Center for Epidemiological Studies Depression (CES-D) scale: An item response theory analysis. *Medical Care*, *42*, 281–289. https://doi.org/10.1097/01.mlr.0000115632.78486.1f

Deng, N. (2011). *Evaluating IRT- and CTT-based methods of estimating classification consistency and accuracy indices from single administrations* [Unpublished doctoral dissertation]. University of Massachusetts, Amherst, MA.

Dudgeon, P. (2004). A note on extending Steiger's (1998) multiple sample RMSEA adjustment to other noncentrality parameter-based statistics. *Structural Equation Modeling*, *11*, 305–319. https://doi.org/10.1207/s15328007sem1103_1

Edwards, M. C., Cheavens, J. S., Heiy, J. E., & Cukrowicz, K. C. (2010). A reexamination of the factor structure of the Center for Epidemiologic Studies Depression Scale: Is a one-factor model plausible?. *Psychological Assessment*, *22*(3), 711–715. https://doi.org/10.1037/a0019917

Emons, W. H., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods*, *12*, 105–120. https://doi.org/10.1037/1082-989X.12.1.105

Fouladi, R. T., & Steiger, J. H. (2008). The Fisher transform of the Pearson Product Moment Correlation coefficient and its square: Cumulants, moments, and applications. *Communications in Statistics. Simulation and Computation*, *37*, 928–944. https://doi.org/10.1080/03610910801943735

Gonzalez, O., Georgeson, A. R., & Pelham, W. E., III (under review). *Estimating classification consistency of machine learning models for screening assessment and tests of individual classification.*

Gonzalez, O., & Pelham, W. E., III. (2021). When does differential item functioning matter for screening? A method for empirical evaluation. *Assessment*, *28*, 446–456. https://doi.org/10.1177/1073191120913618

Gonzalez, O., Pelham, W. E., III, & Georgeson, A. R. (2020). Impact of measurement bias on screening measures. In M. Wiberg, D. Molenaar, J. González, U. Böckenholt, & J. S. Kim (Eds.), *Quantitative psychology. MPS 2019* (pp. 275–284). Springer Proceedings in Mathematics & Statistics, Vol. 322. Springer. https://doi.org/10.1007/978-3-030-43469-4_21

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education and Praeger.

Khowaja, M., Robins, D. L., & Adamson, L. B. (2018). Utilizing two-tiered screening for early detection of autism spectrum disorder. *Autism: The International Journal of Research and Practice*, *22*, 881–890. https://doi.org/10.1177/1362361317712649

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking* (2nd ed.). Assessment Systems Corporation. https://doi.org/10.1007/978-1-4757-4310-4

Kruyen, P. M., Emons, W. H., & Sijtsma, K. (2013). Shortening the S-STAI: Consequences for research and clinical practice. *Journal of Psychosomatic Research*, *75*, 167–172. https://doi.org/10.1016/j.jpsychores.2013.03.013

Lai, M. H., Kwok, O. M., Yoon, M., & Hsiao, Y. Y. (2017). Understanding the impact of partial factorial invariance on selection accuracy: An R script. *Structural Equation Modeling*, *24*, 783–799. https://doi.org/10.1080/10705511.2017.1318703

Lathrop, Q. N. (2015). Practical issues in estimating classification accuracy and consistency with R Package cacIRT. *Practical Assessment, Research & Evaluation*, *20*, Article 18.

Lathrop, Q. N., & Cheng, Y. (2013). Two approaches to estimation of classification accuracy rate under item response theory. *Applied Psychological Measurement*, *37*, 226–241. https://doi.org/10.1177/0146621612471888

Lathrop, Q. N., & Cheng, Y. (2014). A nonparametric approach to estimate classification accuracy and consistency. *Journal of Educational Measurement*, *51*, 318–334. https://doi.org/10.1111/jedm.12048

Lee, W. C. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement*, *47*, 1–17. https://doi.org/10.1111/j.1745-3984.2009.00096.x

Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling*. CRC Press.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179–197. https://doi.org/10.1111/j.1745-3984.1995.tb00462.x

Lord, F. M., & Wingersky, S. M. (1984). Comparison of IRT true-score and equipercentile observed-score "equating". *Applied Psychological Measurement*, *8*, 453–461. https://doi.org/10.1177/014662168400800409

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research*, *36*, 611–637. https://doi.org/10.1207/S15327906MBR3604_06

MacKinnon, D. P., Valente, M. J., & Wurpts, I. C. (2018). Benchmark validation of statistical models: Application to mediation analysis of imagery and memory. *Psychological Methods*, *23*, 654–671. https://doi.org/10.1037/met0000174

Manly, J. J. (2006). Deconstructing race and ethnicity: Implications for measurement of health outcomes. *Medical Care*, *44*(Suppl. 3), S10–S16. https://doi.org/10.1097/01.mlr.0000245427.22788.be

Marrie, R. A., Zhang, L., Lix, L. M., Graff, L. A., Walker, J. R., Fisk, J. D., Patten, S. B., Hitchon, C. A., Bolton, J. M., Sareen, J., El-Gabalawy, R., Marriott, J. J., & Bernstein, C. N. (2018). The validity and reliability of screening measures for depression and anxiety disorders in multiple sclerosis. *Multiple Sclerosis and Related Disorders*, *20*, 9–15. https://doi.org/10.1016/j.msard.2017.12.007

McHorney, C. A., & Fleishman, J. A. (2006). Assessing and understanding measurement equivalence in health outcomes measures: Issues for further quantitative and qualitative inquiry. *Medical Care*, *44*(Suppl. 3), S205–S210. https://doi.org/10.1097/01.mlr.0000245451.67862.57

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). Macmillan.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

Millsap, R. E., & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, *9*, 93–115. https://doi.org/10.1037/1082-989X.9.1.93

O'Connor, B. P. (2018). An illustration of the effects of fluctuations in test information on measurement error, the attenuation of effect sizes, and diagnostic reliability. *Psychological Assessment*, *30*, 991–1003. https://doi.org/10.1037/pas0000471

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385–401. https://doi.org/10.1177/014662167700100306

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27–48. https://doi.org/10.1146/annurev.clinpsy.032408.153553

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research & Evaluation*, *10*, 1–4.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*, 1–97. https://doi.org/10.1007/BF03372160

Steiger, J. H. (1998). A note on multiple sample extensions of the RMSEA fit index. *Structural Equation Modeling*, *5*, 411–419. https://doi.org/10.1080/10705519809540115

Steiger, J. H., & Fouladi, R. T. (1997). Non-centrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.) *What if there were no significance tests?* Erlbaum.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of factors* [Paper presentation]. The annual spring meeting of the Psychometric Society, Iowa City, IA.

Teitelbaum, L. M., & Carey, K. B. (2000). Temporal stability of alcohol screening measures in a psychiatric setting. *Psychology of Addictive Behaviors*, *14*, 401–404. https://doi.org/10.1037/0893-164X.14.4.401

Teresi, J. A., Stewart, A. L., Morales, L. S., & Stahl, S. M. (2006). Measurement in a multi-ethnic society: Overview to the special issue. *Medical Care*, *44*(11, Suppl. 3), S3–S4. https://doi.org/10.1097/01.mlr.0000245437.46695.4a

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, *19*, 39–49. https://doi.org/10.1177/014662169501900105

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–111). Erlbaum.

Thissen, D. (2000). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 159–183). Erlbaum.

Thissen, D., & Wainer, H. (2001). *Test scoring*. Erlbaum. https://doi.org/10.4324/9781410604729

Vilagut, G., Forero, C. G., Barbaglia, G., & Alonso, J. (2016). Screening for depression in the general population with the Center for Epidemiologic Studies Depression (CES- D): A systematic review with meta-analysis. *PLOS ONE*, *11*(5). Article e0155431. https://doi.org/10.1371/journal.pone.0155431

Wallis, P. S. (2013). *The impact of screen format and repeated assessment on responses to a measure of depressive symptomology completed twice in a short timeframe* [M.A. Thesis]. Simon Fraser University, Burnaby, Canada. http://ir.lib.sfu.ca/item/13844

Wallis, P. S., & Fouladi, R. T. (n.d.). *Test-retest effects in a single testing session for a self-administered 1-week recall questionnaire of depressive symptomology* [Unpublished manuscript].

Wheadon, C. (2014). Classification accuracy and consistency under item response theory models using the package classify. *Journal of Statistical Software*, *56*, 1–14. https://doi.org/10.18637/jss.v056.i10

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*, 58–79. https://doi.org/10.1037/1082-989X.12.1.58

Youngstrom, E. A. (2014). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology*, *39*(2), 204–221. https://doi.org/10.1093/jpepsy/jst062

Appendix

Recursive Formula to Estimate $p\ (X|\theta)$

Consider a depression measure with four items—each item has three response categories ranging from 1 to 3 (i.e., sum scores range from 4 to 12) and a cut point of 9. To estimate conditional classification consistency using the analytical approach by Lee (2010), the first value needed is the conditional summed-score probability for every possible sum score x that belongs to classification C (i.e., above or below the cut point) and each potential θ,

$$p_\theta(h) = \sum_{x=x_{(h-1)}}^{x_h-1} \Pr(X = x|\theta), \tag{A1}$$

where x_h is the cut score h. For this formulation, we would include the minimum ($h = 0$; $x_0 = 4$) and the maximum summed score value ($h = C; x_2 = \max(x) + 1$), in addition to our cut score ($h = 1; x_1 = 9$). For the example, x_h separates respondents in two classes: h_{neg} meaning that a respondent is not classified as potentially having depression and h_{pos} meaning that a respondent is classified as potentially having depression. Below is the definition of $p_\theta(h_{neg})$ and $p_\theta(h_{pos})$,

$$p_\theta(h_{neg}) = \sum_{x=4}^{8} \Pr(X = x|\theta), \tag{A2}$$

$$p_\theta(h_{pos}) = \sum_{x=9}^{12} \Pr(X = x|\theta). \tag{A3}$$

$\Pr(X = x|\theta)$, the conditional summed score probability, is computed via multiplication of conditional item response probabilities $p_{ijk}(\theta)$, which is the probability of individual i providing response k to item j given θ. We can compute p_{ijk} (θ) for a range of θ using the output (i.e., the trace lines) from the GRM from Equation 1.

To obtain the probability for a score of 5 using our example, the possible combinations are as follows: 1112, 1121, 1211, and 2111.

Table A1
Polytomous Recursive Formula Example, Assuming $\theta = 1$

Decision	Item r	Score x	$f_r(x)$			
	1	1	$f_1(1)$	$=p_{11}$		
		2	$f_1(2)$	$=p_{12}$		
		3	$f_1(3)$	$=p_{13}$		
	2	2	$f_2(2)$	$=f_1(1)p_{21}$		
		3	$f_2(3)$	$=f_1(2)p_{21}$	$+f_1(1)p_{22}$	
		4	$f_2(4)$	$=f_1(3)p_{21}$	$+f_1(2)p_{22}$	$+f_1(1)p_{23}$
		5	$f_2(5)$		$+f_1(3)p_{22}$	$+f_1(2)p_{23}$
		6	$f_2(6)$			$+f_1(3)p_{23}$
	3	3	$f_3(3)$	$=f_2(2)p_{31}$		
		4	$f_3(4)$	$=f_2(3)p_{31}$	$+f_2(2)p_{32}$	
		5	$f_3(5)$	$=f_2(4)p_{31}$	$+f_2(3)p_{32}$	$+f_2(2)p_{33}$
		6	$f_3(6)$	$=f_2(5)p_{31}$	$+f_2(4)p_{32}$	$+f_2(3)p_{33}$
		7	$f_3(7)$	$=f_2(6)p_{31}$	$+f_2(5)p_{32}$	$+f_2(4)p_{33}$
		8	$f_3(8)$		$=f_2(6)p_{32}$	$+f_2(5)p_{33}$
		9	$f_3(9)$			$=f_2(6)p_{33}$
Negative	4	4	$f_4(4)$	$=f_3(3)p_{41}$		
Negative		5	$f_4(5)$	$=f_3(4)p_{41}$	$+f_3(3)p_{42}$	
Negative		6	$f_4(6)$	$=f_3(5)p_{41}$	$+f_3(4)p_{42}$	$+f_3(3)p_{43}$
Negative		7	$f_4(7)$	$=f_3(6)p_{41}$	$+f_3(5)p_{42}$	$+f_3(4)p_{43}$
Negative		8	$f_4(8)$	$=f_3(7)p_{41}$	$+f_3(6)p_{42}$	$+f_3(5)p_{43}$
Positive		9	$f_4(9)$	$=f_3(8)p_{41}$	$+f_3(7)p_{42}$	$+f_3(6)p_{43}$
Positive		10	$f_4(10)$	$=f_3(9)p_{41}$	$+f_3(8)p_{42}$	$+f_3(7)p_{43}$
Positive		11	$f_4(11)$		$=f_3(9)p_{42}$	$+f_3(8)p_{43}$
Positive		12	$f_4(12)$			$=f_3(9)p_{43}$

(*Appendix continues*)

For the response pattern 1112 and $\theta = 1$, one would compute the probability of that specific combination given $\theta = 1$ via $p_{i11}(\theta = 1) * p_{i21}(\theta = 1) * p_{i31}(\theta = 1) * p_{i42}(\theta = 1)$. One would repeat these steps for each possible combination, and then sum these probabilities to obtain $p_\theta(x = 5|\theta = 1)$. This procedure would be repeated for every possible sum score and every possible θ value. This would become very tedious as the number of items, response categories, and θ increased, so a recursive algorithm is used to obtain these values.

If we define $p_{i11}(\theta_i) = f_1(x = W_{11}|\theta_i)$, where W_{jk} denotes response k for item j, as the probability of selecting a 1 for the first item and $p_{i12}(\theta_i) = f_1(x = W_{12}|\theta_i)$ as the probability of selecting a 2 for the first item, m_j as the number of categories, then the recursive algorithm for finding the probability of obtaining score x across r items is

$$f_r(x|\theta_i) = \sum_{k=1}^{m_j} f_{r-1}(x - W_{jk}) p_{ijk}(\theta_i) \tag{A4}$$

This algorithm is shown in Table A1 for the hypothetical example. The decision column indicates which probabilities would be added per Equation A1 for each of the possible decisions.

After $p_\theta(h)$ is obtained for our two possible h's, these values are then used to obtain the conditional classification consistency index as follows:

$$\phi_\theta = \sum_{h=1}^{H} [p_\theta(h)]^2 \tag{A5}$$

If we were estimating classification consistency using the simulation approach, $p_\theta(h)$ would be determined by simulating item responses rather than using the recursive formula. In this case, item responses would be summed, and the proportion of simulated cases in each decision category (e.g., h_{neg} and h_{pos}) given θ is determined, and these values correspond to $p_\theta(h)$.

Conditional classification consistency is a simple extension of joint probability. For a given θ, these are the possible outcomes and their associated probabilities (Table A2).

Table A2
Possible Decisions for Two Administrations and Their Respective Probabilities

		Administration 2	
Administration		h_{neg}	h_{pos}
Administration 1	h_{neg}	$p_\theta(h_{neg})^2$	$p_\theta(h_{neg})^* p_\theta(h_{pos})$
	h_{pos}	$p_\theta(h_{pos})^* p_\theta(h_{neg})$	$p_\theta(h_{pos})^2$

The bolded probabilities are those that refer to *consistent* classification decisions and are therefore summed to obtain the probability.

In the case of a binary decision, $p_\theta(h_{neg}) = 1 - p_\theta(h_{pos})$. The reason that classification consistency ranges from 0.50 to 1 is because in the scenario with the most uncertainty (i.e., either decision is equally likely), $p_\theta(h_{neg}) = p_\theta(h_{pos}) = 0.50$. Therefore, the classification consistency would be 0.50.

For every θ, we have now obtained a conditional classification consistency value. To obtain the marginal classification consistency, we would need to integrate over θ and their associated densities, $g(\theta)$.

$$\phi = \int_{-\infty}^{\infty} \phi_\theta g(\theta) d(\theta) \tag{A6}$$

For θ values that are quadrature points, this is approximated using quadrature weights using the density function for the normal distribution. Note that the density could be obtained from the *R* function dnorm() or from normal density tables. The density values are then normalized (i.e., divide each density by the sum of the density values so the weights add to 1.0) to yield the quadrature weights. For θ values from a sample, this is approximated using an average.

Received May 24, 2020
Revision received December 8, 2020
Accepted January 6, 2021 ■