Research Article

Psychopathology

Psychopathology 2019;52:358–366
DOI: 10.1159/000505075

Received: August 15, 2019
Accepted after revision: November 27, 2019
Published online: January 21, 2020

Effects of Skip-Logic on the Validity of Dimensional Clinical Scores: A Simulation Study

Adon F.G. Rosen Tyler M. Moore Monica E. Calkins Ruben C. Gur
Raquel E. Gur

Department of Psychiatry, Brain Behavior Laboratory, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Keywords
Skip-logic · Item response theory · Missing data · Criterion validity

Abstract
Structured assessment of clinical phenotypes is a burdensome procedure, largely due to the time required. One method to alleviate this is "skip-logic," which allows for portions of an interview to be skipped if initial ("screen") items are not endorsed. The bias that skip-logic introduces to resultant continuous scores is unknown and can be explored using Item Response Theory. Interview response data were simulated while varying 5 characteristics of the measures: number of screen items, difficulty (clinical severity) of the screens, difficulty of non-screen items, shape of the trait distribution, and range of discrimination parameters. The number of simulations and examinees were held constant at 2,000 and 10,000, respectively. A criterion variable correlating 0.80 with the measured trait was also simulated, and the outcome of interest was the difference between the correlations of the criterion variable and the two estimated scores (with and without skip-logic). Effects of the simulation conditions on this outcome were explored using ANOVA. All main effects and interactions were significant. The largest 2-way interaction was between number of screen items and average item discrimination, such that the number of screen items had a large effect on bias only when discrimination parameters were low. This, among other interactions explored here, suggests that skip-logic can bias results using continuous scores; however, the effects of this bias are usually inconsequential. Skip-logic in clinical assessments can introduce bias in continuous sum scores, but this bias can usually be ignored.

Among the most obvious burdens of standardized clinical assessment, whether for research or patient care, is the amount of time it takes to complete. For patient care, these interview durations are inconvenient, and for large-scale research projects, they can be outright prohibitive. Many interview developers have thus included a time-saving feature into their interview designs. This feature, sometimes called "skip-logic," allows the interviewer to skip large sections of the interview if certain gateway ("screen") symptoms are not endorsed. The prevailing standardized diagnostic interviews (SADS [1]; K-SADS [2]; SCID [3]) all use skip-logic to achieve symptom-based diagnostic categorization.

KARGER

karger@karger.com
www.karger.com/psp

Tyler M. Moore
Brain Behavior Laboratory, Perelman School of Medicine, University of Pennsylvania
3400 Spruce St. – 10th Floor, Gates Pavilion
Philadelphia, PA 19104 (USA)
E-Mail tymoore@upenn.edu

Recently, increased realization that symptom-based diagnostic classifications do not adequately capture disorders has generated interest in continuous measures along symptom dimensions rather than diagnostic classifications alone. Use of clinical interview responses to generate continuous (e.g., sum) scores is increasingly common [4] and fits within the Research Domain Criteria (RDoC) framework [5], but the presence of skip-logic can cause more serious problems than if the interview were being used purely for diagnosis. Unfortunately, it is unknown whether and how much bias skip-logic introduces in estimating dimensional measures.

Item Response Theory (IRT) [6, 7] can help estimate the potential bias that skip-logic may introduce in dimensional assessment. IRT is a psychometric method that focuses on various characteristics of individual test or scale items (rather than a test/scale as a whole). One of the most common IRT models is the 2-parameter model described by the following equation:

$$p_i(\theta) = \frac{1}{1 + e^{-a_i(\theta - b_i)}}, \quad (1)$$

where $p_i(\theta)$ is the probability of endorsement (or a correct response, in the case of cognitive testing), a_i is the item discrimination, b_i is the item difficulty, and θ is the trait level of the person (e.g., a clinical dimension such as depressed mood). The discrimination parameter, a_i, determines how precisely the item can place an individual on a trait spectrum; higher discrimination is always better. The difficulty parameter, b_i, determines how high on the latent trait continuum one has to be in order to have a 50% chance of endorsing the item. The higher the difficulty, the higher someone needs to be on the latent trait in order to have a 50% probability of endorsing the item. Thus, the term "difficulty" in IRT is not limited to items that have correct/incorrect answers; rather, in clinical scales/interviews, difficulty indicates how severe the symptom is on a standard metric. The idea of clinical items having "difficulty" parameters is crucial for understanding the rationale behind skip-logic designs. Reise and Waller [8] provide a comprehensive review of IRT-related issues especially relevant to clinical assessment.

The rationale for skip-logic is that if someone is presented with screening ("easy") items (e.g., depressed mood or loss of interest) and does not endorse them, then there is no reason to expect them to endorse associated symptoms ("harder") items (e.g., suicidality). In this framework, clinical interviews would be designed such that any disorder evaluation (e.g., depression) begins with a set of screen items that ask about entry symptoms, and if none of those symptoms is endorsed, the remaining items in that section are skipped and assumed to be not endorsed (symptoms absent). In an IRT framework, this is a reasonable assumption, because it would be *highly* improbable for someone in the upper range of a trait to not endorse easy items. However, the item parameters (discrimination and difficulty) of screen and non-screen items are often not known during construction of the instrument. The design and question sequences of clinical interviews reflect the hierarchical nature of the Diagnostic and Statistical Manual of Mental Disorders (DSM). "Screen" items assess whether essential symptoms (*necessary* for diagnosis) are present at a threshold level, and if not, the remaining symptom questions (such as appetite or sleep disturbance) are skipped even though they may be present to some degree. It is therefore probable that some screen items will end up with higher difficulty than is optimal for skip-logic to work properly in a dimensional framework. In these cases, where screen items are not as "easy" as they should be, the assumption that their non-endorsement implies non-endorsement of the rest of the items is erroneous. This choice of more difficult items leads to items being skipped, when, in fact, they would have provided symptom-level information if administered. This is problematic specifically when a researcher wants to use a dimensional measure of the trait (e.g., a sum score), because it means that items that would have been endorsed are assumed to be non-endorsed (auto-coded as 0), and the symptom domain will therefore be systematically underestimated in the sample.

Skip-logic can cause the above problems in two inextricable ways. First, as noted above, some item responses will erroneously be auto-coded 0 (not endorsed), and this leads to underestimation of the trait. Second, the erroneous non-endorsement of some items will cause the estimated IRT difficulty parameters to be overestimated – i.e., non-screen items will appear more difficult than they actually are. As a simple illustration, consider an item response vector from 8 examinees to be {1, 1, 1, 1, 0, 0, 0, 0}. The proportion endorsed is 4/8 = 50%, which suggests an item of average difficulty (from equation 1, $b_i = 0$). However, if skip-logic is applied when screen items are too difficult, some of those examinees who would have endorsed the item would be auto-coded to non-endorsement when they "skip out" of the section. This practice would change the hypothetical response vector to {1, 1, 0, 0, 0, 0, 0, 0}. This new proportion endorsed (2/8 = 25%) suggests a more difficult item (from equation 1, $b_i > 0$), and this upwardly biased difficulty parameter will subsequently affect all IRT-related applications, such as IRT-

based scoring and creation of item banks for computerized adaptive tests.

The purpose of the present study was to investigate how implementation of skip-logic affects the predictive validity of scores. Given that the data are simulated, we know the "ground truth" (population parameters) of how strongly the measured traits relate to a validity criterion (see below). The primary outcome of interest is the difference between the estimated relationship with and without skip-logic via a difference of correlations.

Methods

All analyses described below were performed using the *psych* package [9] in R [10]. Data were simulated using the sim.irt() function, and items were calibrated using the irt.fa() function. All item parameters are in the logistic metric ($D = 1.0$) (see Embretson and Reise [6]). All codes can be found online in a github repository (https://github.com/adrose/skipLogic).

Simulation Conditions

Simulation conditions used here were chosen based on a review of the literature, as well as item parameters estimated in our own clinical data collected on a large community cohort. Results from our own data (not shown here) can be found in Moore et al. [11]. Other (intentionally diverse) publications used to determine simulation conditions included IRT analyses of substance use disorders [12], suicidality [13], DSM V personality disorders [14], depression [15], psychosis [16], health literacy [17], as well as other IRT simulation studies [18–20].

Simulation conditions were varied in 5 ways, for a total of 48 conditions. The condition types were:

1 Number of screen items. This condition type varied the number of items used to determine whether an examinee should be administered the full scale (caused by endorsement of any single screen item). The number of screen items was set to 2 (10% of total items), 4 (20%), or 6 (30%).
2 Difficulty of screen items. This condition type varied the screen item difficulty thresholds—i.e., how high on the latent trait an examinee has to be to have a 50% probability of endorsement. Difficulties of screen items were drawn randomly from a uniform distribution ranging from [–3 to –1] or [–1 to 1]. Note that screen item difficulties were never selected from a more difficult range (e.g., [1–3]), because highly difficult screen items inevitably cause such an overwhelming loss of information that the simulations often failed for technical reasons. For example, highly difficult screen items will result in *most* examinees (rather than only some) endorsing none of the screens and therefore having response vectors of all 0s (non-endorsements).
3 Difficulty of non-screen items. These are the same as No. 2 above, but for the non-screen items. Difficulties of non-screen items were drawn randomly from a uniform distribution ranging from [–1 to 1] or [1 to 3]. Note that non-screen item difficulties were never selected from a less difficult range (e.g., [–3 to –1]), for the same reason that screen items were not simulated with very high difficulty. Specifically, very easy non-screen items will result in *most* examinees who endorse all screen items also endorsing all non-screen items.
4 Shape of the θ (trait) distribution. Most traits are assumed to be normally distributed, but it is quite common in clinical measurement for this distribution to be positively skewed. We thus varied the shape of the θ distribution to be either normally distributed or positively skewed. To achieve the skewed distribution, a standard normal distribution was generated, squared (creating the skew), and re-standardized to maintain mean = 0 and SD = 1.
5 Range of discrimination parameters for all items. This condition type varied the overall quality of items, as determined by the slope of each item response characteristic curve at its inflection point (a from equation 1). Item discrimination parameters were sampled from a uniform distribution ranging from [0.3 to 1.5] (very low to moderate) or [1.5 to 3.5] (moderate to very high).

Table 1. Simulation conditions

Condition description	Conditions
Number of screen items	2 || 4 || 6
Screen item difficulties (range)	–3 to –1 || –1 to 1
Non-screen item difficulties (range)	–1 to 1 || 1 to 3
θ distribution shape	standard normal || skewed
Item discriminations for all items (range)	0.3 to 1.5 || 1.5 to 3.5

The above 5 conditions are summarized in Table 1. The following conditions were constant across all simulations: number of simulations (n = 2,000), number of simulated examinees (n = 10,000, but see below), and number of items (n = 20). Finally, all simulated data sets included a criterion variable correlating exactly 0.80 with the true trait (θ) values. The criterion could be thought of as any outcome variable that might be used to assess the validity of a dimensional clinical test. The main outcome of interest here is the difference between the score-outcome relationship when skip-logic is used versus not used. For completeness, all simulations in all conditions above were repeated using a much smaller sample size (n = 200) to check for unexpected effects of n and confirm that the results generalize to smaller samples.

Note that a typical approach in a simulation study is to compare estimated parameters to "true" population parameters. For example, if a true (population) item discrimination parameter is 1.0 and that value is estimated to be 1.0 under typical circumstances, one might introduce atypical circumstances via simulation to determine whether the discrimination of 1.0 is still accurately estimated. One might simulate from a non-normal distribution and discover that under these atypical circumstances, the estimated discrimination value is 0.90. The difference between the true and estimated values (1.0–0.9 = 0.1) is called "bias" and is the central focus of most simulation studies. However, here, we are most interested in the difference between the ability of a test score to predict a criterion *with* versus *without* skip-logic. The true (population) value of that relationship (0.80) is less relevant. For example, consider a simulation result in which the score without skip-logic relates 0.40 to the criterion, and the score with skip-

Table 2. ANOVA results predicting bias of sum scores by simulation condition

Condition	η^2	Cohen's *F*
Screens	0.169	1.002
DiscriminationRange	0.090	0.729
Difficulty_of_Screens	0.167	0.996
Difficulty_of_NonScreens	0.113	0.820
Distribution	0.001	0.074
Screens*DiscriminationRange	0.067	0.631
Screens*Difficulty_of_Screens	0.050	0.545
Screens*Difficulty_of_NonScreens	0.043	0.503
Screens*Distribution	0.002	0.121
DiscriminationRange*Difficulty_of_Screens	0.023	0.373
DiscriminationRange*Difficulty_of_NonScreens	0.010	0.241
DiscriminationRange*Distribution	0.001	0.073
Difficulty_of_Screens*Difficulty_of_NonScreens	0.062	0.605
Difficulty_of_Screens*Distribution	0.000	0.023
Difficulty_of_NonScreens*Distribution	0.001	0.072
Screens*DiscriminationRange*Difficulty_of_Screens	0.006	0.194
Screens*DiscriminationRange*Difficulty_of_NonScreens	0.006	0.192
Screens*DiscriminationRange*Distribution	0.000	0.024
Screens*Difficulty_of_Screens*Difficulty_of_NonScreens	0.010	0.250
Screens*Difficulty_of_Screens*Distribution	0.005	0.177
Screens*Difficulty_of_NonScreens*Distribution	0.000	0.035
DiscriminationRange*Difficulty_of_Screens*Difficulty_of_NonScreens	0.000	0.039
DiscriminationRange*Difficulty_of_Screens*Distribution	0.000	0.027
DiscriminationRange*Difficulty_of_NonScreens*Distribution	0.000	0.043
Difficulty_of_Screens*Difficulty_of_NonScreens*Distribution	0.000	0.043
Screens*DiscriminationRange*Difficulty_of_Screens*Difficulty_of_NonScreens	0.001	0.075
Screens*DiscriminationRange*Difficulty_of_Screens*Distribution	0.000	0.054
Screens*DiscriminationRange*Difficulty_of_NonScreens*Distribution	0.000	0.049
Screens*Difficulty_of_Screens*Difficulty_of_NonScreens*Distribution	0.001	0.073
DiscriminationRange*Difficulty_of_Screens*Difficulty_of_NonScreens*Distribution	0.000	0.017
Screens*DiscriminationRange*Difficulty_of_Screens*Difficulty_of_NonScreens*Distribution	0.000	0.016

Screens = number of screens (2, 4, 6); Distribution = θ distribution type (normal, skewed).
All results are significant at the $p < 0.0001$ level.

logic relates 0.39 to the criterion. While it is true and interesting that both scores in this simulation condition do a very poor job of predicting the criterion (0.40 and 0.39 vs. the true value of 0.80), what is of key interest here is the difference between the scores' predictive abilities with and without skip-logic (0.40 vs. 0.39). That is, while the score with skip-logic relates poorly to the criterion, the score without skip-logic does not do better, and the key conclusion from the analysis would therefore be that skip-logic is acceptable in that circumstance.

Results

Table 2 shows the results of an ANOVA relating the simulation conditions (plus all interactions) to the difference between the score-criterion relationship using skip-logic versus not using skip-logic. All results are statistically significant, but note that significance of effects is confounded by the number of simulations. Therefore, meaningful interpretation of the ANOVA results requires

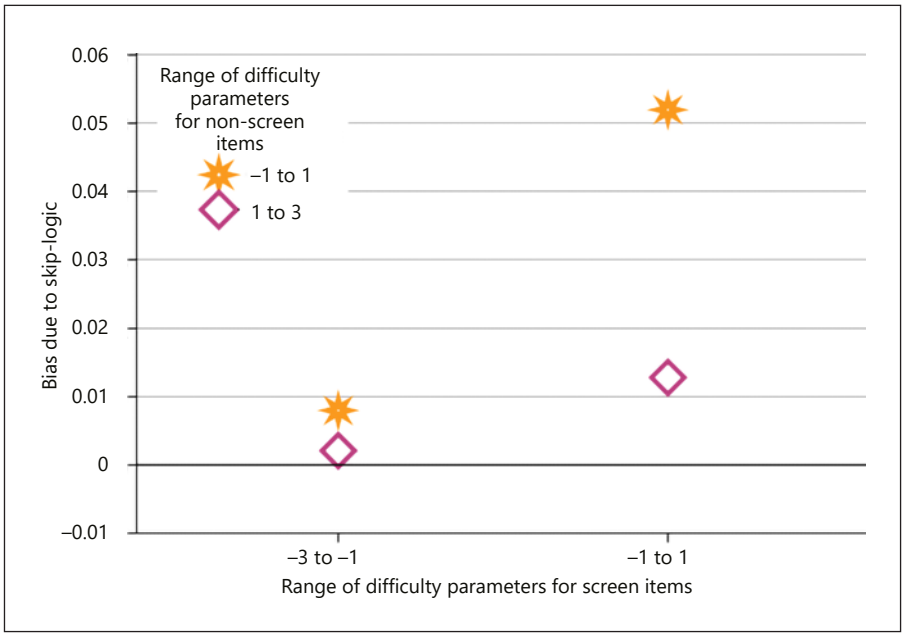


Fig. 1. Relative bias due to difficulty parameters of screen and non-screen items.

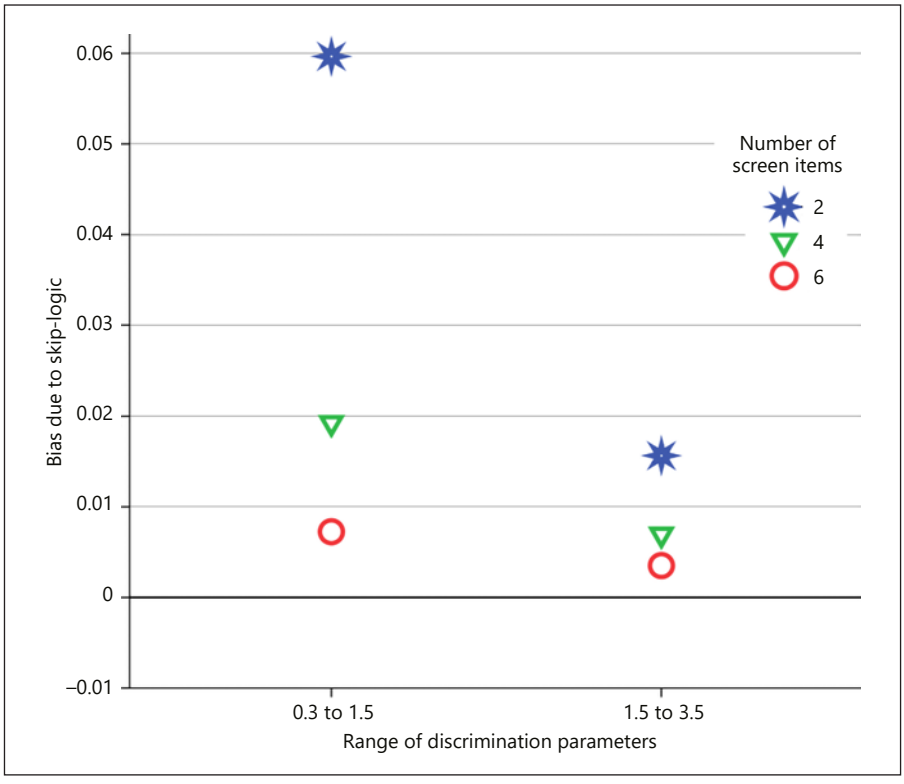


Fig. 2. Relative bias due to range of discrimination parameters and number of screen items.

effect sizes; Table 2 includes η^2 and Cohen's *F*. Of the main effects, the largest is for the number of the screen items ($\eta^2 = 0.169$) and the difficulty of the screen items ($\eta^2 = 0.167$). The smallest was for the shape of the θ distribution ($\eta^2 = 0.001$). Online supplementary Figure 1 (see www.karger.com/doi/10.1159/000505075) shows these main effects graphically.

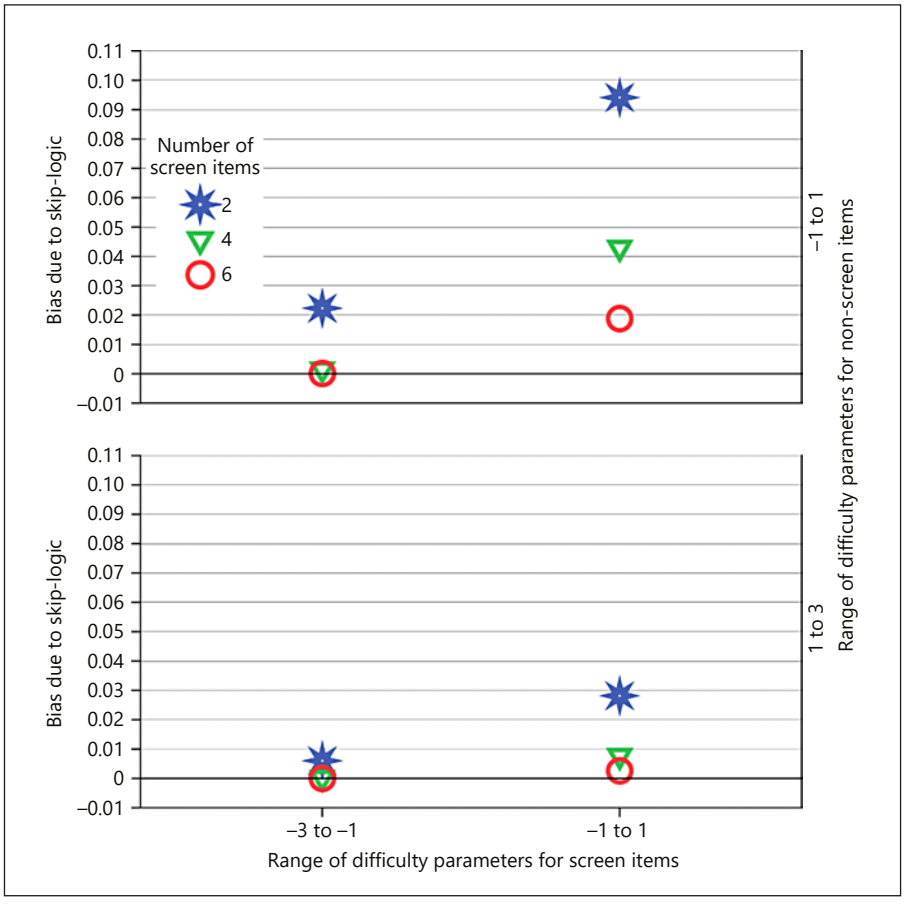
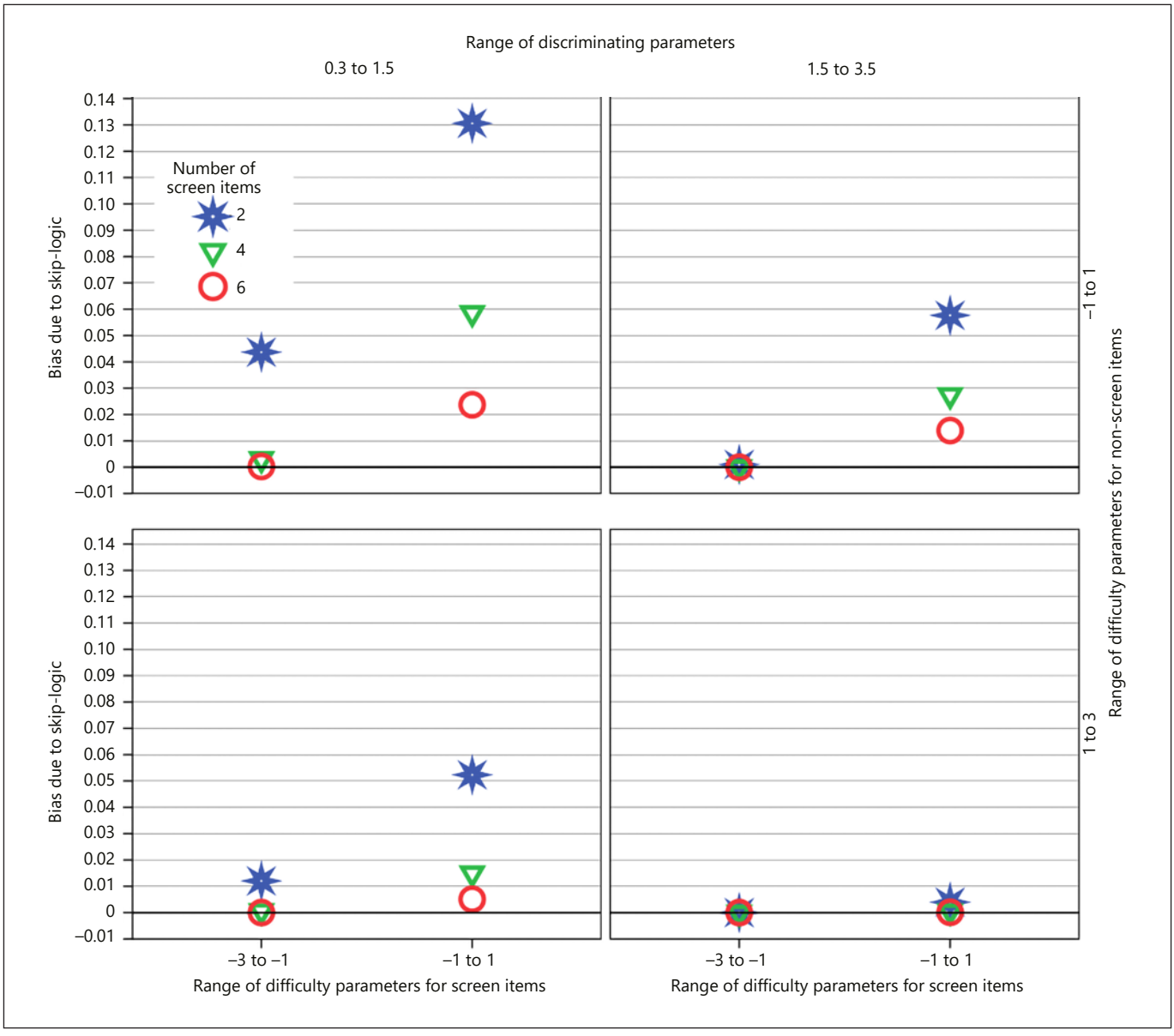


Fig. 3. Relative bias due to difficulty parameters of screen and non-screen items, separated by the number of screens.

Of the 2-way interactions, two stood out: (1) interaction between difficulties of the screen and the difficulties of the non-screen items ($\eta^2 = 0.062$), and (2) interaction between the number of screen items and the range of discrimination parameters ($\eta^2 = 0.067$). Figure 1 shows the first interaction graphically. When the screen items have relatively low difficulty (left side of the graph), bias is minimal regardless of the difficulties of the non-screen items. However, if screen items have moderate difficulty (right side of the graph), bias will be higher, especially in the case where screen and non-screen items have equal difficulty ranges (star in upper right corner of the graph). The second important interaction – between the number of screen items and the range of discrimination parameters – is shown in Figure 2. When discrimination parameters are low (left side of the graph), the number of screens can make a large difference in determining the amount of bias. When discrimination parameters are high (right side of the graph), the number of screens is less consequential, though more screens will always lead to less bias.

Of the 3-way interactions, the largest was among the number of screens, difficulties of screens, and difficulties of non-screens ($\eta^2 = 0.010$). Figure 3 shows this interaction graphically. As in Figure 2, there is a clear tendency for difficult screens and easy non-screens to cause more bias, especially when combined (top right corner of Figure 3). However, when the number of screens is high enough (6; circles in the graph), bias remains low even in the worst case of equally difficult screens and non-screens.

Finally, Figure 4 shows the most significant 4-way interaction, which is a combination of the four conditions already mentioned above (discriminations, screen difficulties, non-screen difficulties, and the number of screens). As in Figure 3, Figure 4 shows that as long as the number of screens is high enough (6; circles in the graph), bias remains low even when discrimination parameters are low, and screen and non-screen items have equal difficulty. However, when there are fewer screens (2; stars in the graph), bias can reach "unacceptable" levels (>0.05) when any 2 of the remaining 3 problematic conditions are



Color version available online

Fig. 4. Relative bias due to difficulty parameters of screen and non-screen items, separated by the number of screens and range of discrimination parameters.

met (low discrimination, difficult screens, or easy non-screens). Notably, when discrimination parameters are high, screen difficulties are low, and non-screen difficulties are high, there is near-zero bias even when only 2 screens are used.

Note that all results described above were identical when a smaller sample size ($n = 200$) was used, but, as expected, variability among simulations was higher with the smaller sample size. The implication is that researchers can expect the above phenomena to occur to the same extent regardless of sample size, but when sample size is small, the effect of skip-logic can be masked by the noise of the small sample.

Discussion

Given the time and effort burdens of clinical interviews, the use of skip-logic is fully justifiable and often desirable, especially in research settings with time-limited

Psychopathology 2019;52:358–366
DOI: 10.1159/000505075

access to participants. However, as with any procedure that results in loss of information and affects some participants more than others, there is potential for skip-logic to introduce systematic bias in the measure when used to estimate dimensional scores. Introduction of "noise" (unexplained variance) to the measure is bad enough (increased type II error), but systematic bias is especially worrisome because it can result in spurious effects (even in the opposite direction of reality). Here, we explored the effects of skip-logic by simulating item response patterns from a latent trait with a known relationship (0.80) to an external criterion. We found that bias introduced by skip-logic will be minimal as long as (a) screen items are relatively easy or non-screen items are relatively hard; (b) there are at least 6 screen items (or 30% of total items are screens); and (c) discrimination parameters of items are generally high (>1.5 on logistic metric).

Remarkably, these three conclusions mostly stand true on their own. For example, in a worst-case scenario such as having moderate-difficulty screens, easy non-screens, and low item discriminations, the effect of skip-logic will still be minimal as long as there are at least 6 screens. Likewise, even if there are only 4 screens, and both screens and non-screens have moderate difficulty (bad combination), bias will still be minimal as long as item discrimination parameters are generally high. Our results suggest that problematic levels of bias occur only with certain "worst case" combinations of item characteristics. The most potentially damaging is the number of screen items, where having few screens (e.g., 10% of scale) could result in severe underestimation of the trait in participants who "skip out." The second most potentially damaging is the difficulty of the screen items, where moderate-to-high difficulty screen items could result in severe underestimation of the trait. An overall conclusion taken from the above findings is that most (probably all) contemporary clinical interviews that utilize skip-logic can safely provide dimensional sum scores (if desired) with minimal or no skip-logic-related bias. However, some questionnaires may benefit by adding additional screeners; for instance, in the GOASSESS, generalized anxiety disorder has only two screen items, whereas (within the same interview) phobias contains eight screeners. Comparisons across interviews also suggest potential for improvement (either by removing or adding screens) – e.g., whereas the SCID includes two screen items for mania, the K-SADS includes 7.

The recommendations above are somewhat abstract and assume that the researcher has information (e.g., difficulty of screen items) that they might not have. Unfortunately, more specific recommendations are unlikely to be useful given the variety of clinical interviews in use (not to mention the diversity of psychopathological phenomena themselves). General steps likely to be useful are as follows:

- Examine probe items for symptoms that are commonly endorsed in the absence of the disorder. For example, Cole et al. [21] found that on the KSADS, the probe depression symptoms of sleep disturbance, feelings of guilt, and concentration difficulties were endorsed more often than were the screener symptoms (depressed mood, anhedonia, and irritability). In this specific example, the KSADS includes more severe symptoms (motor retardation, suicidal ideation) that balance out the less severe probes listed above, but other interviews might not have such a wide range.
- Examine screen items for symptoms that are present only in moderate-to-severe cases of the disorder. For example, the first screen item for schizoid personality disorder on the SCID asks whether the person has *no* desire to make/form close relationships. While this might seem like a good screen item, multiple IRT analyses [22, 23] have found it to be the most difficult (least endorsed) item of all schizoid personality disorder items, meaning it is likely that SCID schizoid personality disorder sum scores are unacceptably biased by skip-logic.
- Count the number of screen items. If there are at least 6, bias due to skip-logic is unlikely to be a problem, even if all other parameters are conducive to bias. The same is mostly true if there are at least 4 screen items. If there are only 2 screen items, there is a higher chance of skip-logic-related bias, and steps No. 1 and 2 above should therefore be taken with extra care.

This study has some notable limitations. First, the dimensional approach to psychopathology conflicts with some established theoretical conceptions thereof, such as the idea of "cardinal" symptoms *necessary* for a latent trait to be validly labeled. For example, by this reasoning, if someone denies the cardinal symptoms of depression (depressed mood and anhedonia), then any other depression symptoms endorsed (e.g., sleep disturbance, appetite change, etc.) cannot be indicative of depression; they must be due to something else. By contrast, the approach used in the present study assumes that items on a scale are conceptually interchangeable (except for item parameter estimates, which will of course differ). Second, an assumption of all analyses here was that skipped symptoms *might have been* endorsed – i.e., there is a >0% probability of someone endorsing a probe item even if

they did not endorse any screen items – but this assumption is false for some disorders. Using posttraumatic stress disorder as an example, after the patient is asked about previous traumatic events, all subsequent items reference those traumatic events; therefore, if there were no traumatic events reported, posttraumatic stress disorder probe items can be skipped with exactly 0 loss of information. Despite the above weaknesses, however, the present study provides evidence that it is generally safe to assume non-administered probe items are "not endorsed" when calculating sum scores, and this evidence is especially compelling when there are at least 4 screen items.

Statement of Ethics

This research did not involve human or animal subjects.

Disclosure Statement

None of the authors has a conflict of interest to declare.

Funding Sources

This work was supported by NIMH grants MH089983, MH019112, MH096891, and MH117014, the Lifespan Brain Institute (LiBI), and the Dowshen Neuroscience Fund.

Author Contributions

T.M.M. and A.F.G.R. contributed the conceptual design of the study and performed the simulations. All authors contributed substantially to the interpretation of results, drafting the manuscript, and revising the manuscript. All authors approved the final version of the manuscript and accept accountability for all aspects of this study and its communication.

References

1 Endicott J, Spitzer RL. A diagnostic interview: the schedule for affective disorders and schizophrenia. Arch Gen Psychiatry. 1978 Jul; 35(7):837–44.
2 Kaufman J, Birmaher B, Brent D, Rao U, Flynn C, Moreci P, et al. Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL): initial reliability and validity data. J Am Acad Child Adolesc Psychiatry. 1997; 36(7):980–8.
3 Glasofer DR, Brown AJ, Riegel M. Structured clinical interview for DSM-IV (SCID). Wade T, editor. Singapore: Springer; 2015. 4 p.
4 van Nierop M, Viechtbauer W, Gunther N, van Zelst C, de Graaf R, ten Have M, et al. Childhood trauma is associated with a specific admixture of affective, anxiety, and psychosis symptoms cutting across traditional diagnostic boundaries. Psychol Med. 2015;45(6): 1277–88.
5 Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am J Psychiatry. 2010 Jul;167(7):748–51.
6 Embretson SE, Reise SP. Item response theory for psychologists. Mahwah: Erlbaum; 2000.
7 Lord FM. Applications of item response theory to practical testing problems. Hillsdale: Erlbaum; 1980.
8 Reise SP, Waller NG. Item response theory and clinical measurement. Annu Rev Clin Psychol. 2009 Apr 27;5(1):27–48.
9 Revelle W. psych: Procedures for Personality and Psychological Research [Internet]. Evanston: Northwestern University; 2018. Available from: https://CRAN.R-project.org/package=psych
10 The R Core Team. R: a language and environment for statistical computing [Internet] Vienna, Austria: R Foundation for Statistical Computing; 2018. Available from: https://www.R-project.org/
11 Moore TM, Calkins ME, Satterthwaite TD, Roalf DR, Rosen AFG, Gur RC, et al. Development of a computerized adaptive screening tool for overall psychopathology ("p"). J Psychiatr Res. 2019 Sep;116:26–33.
12 Kirisci L, Tarter RE, Reynolds M, Vanyukov M. Individual differences in childhood neurobehavior disinhibition predict decision to desist substance use during adolescence and substance use disorder in young adulthood: A prospective study. Addict Behav. 2006;31(4): 686–96.
13 De Beurs DP, de Vries AL, de Groot MH, de Keijser J, Kerkhof AJ. Applying computer adaptive testing to optimize online assessment of suicidal behavior: a simulation study. J Med Internet Res. 2014 Sep;16(9):e207.
14 Suzuki T, Samuel DB, Pahlen S, Krueger RF. DSM-5 alternative personality disorder model traits as maladaptive extreme variants of the five-factor model: an item-response theory analysis. J Abnorm Psychol. 2015 May;124(2): 343–54.
15 Olino TM, Yu L, Klein DN, Rohde P, Seeley JR, Pilkonis PA, et al. Measuring depression using item response theory: an examination of three measures of depressive symptomatology. Int J Methods Psychiatr Res. 2012 Mar; 21(1):76–85.
16 Shevlin M, Adamson G, Vollebergh W, de Graaf R, van Os J. An application of item response mixture modelling to psychosis indicators in two large community samples. Soc Psychiatry Psychiatr Epidemiol. 2007 Oct; 42(10):771–9.
17 Paige SR, Krieger JL, Stellefson M, Alber JM. eHealth literacy in chronic disease patients: An item response theory analysis of the eHealth literacy scale (eHeals). Patient Educ Couns. 2016;100(2):320–6.
18 Kirisci L, Hsu T, Yu L. Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. Appl Psychol Meas. 2001 Jun;25(2):146–62.
19 Reise SP, Yu J. Parameter recovery in the graded response model using MULTILOG. J Educ Meas. 1990 Jul 1;27(2):133–44.
20 Bulut O, Sünbül Ö. R Programlama Dili ile Madde Tepki Kuramında Monte Carlo Simülasyon Çalışmaları. Egit Psikol Olcme Deger Derg. 2017 Sep 30;266–87.
21 Cole DA, Cai L, Martin NC, Findling RL, Youngstrom EA, Garber J, et al. Structure and measurement of depression in youths: applying item response theory to clinical data. Psychol Assess. 2011;23(4):819–33.
22 Cooper LD, Balsis S. When less is more: How fewer diagnostic criteria can indicate greater severity. Psychol Assess. 2009;21(3):285–93.
23 Balsis S, Gleason ME, Woods CM, Oltmanns TF. An item response theory analysis of DSM-IV personality disorder criteria across younger and older age hroups. Psychol Aging. 2007 Mar;22(1):171–85.