An Item Response Theory Analysis of the SCOFF Eating Disorders Questionnaire in a Seventh Grade Population

Three common eating disorders - anorexia nervosa, bulimia nervosa, and binge eating - are defined in the *Diagnostic and Statistical Manual of Mental Disorders* [DSM-V]. Anorexia nervosa is defined as a condition that primarily affects adolescent girls and young women. It is characterized by distorted body image and excessive dieting that leads to severe weight loss with a pathological fear of becoming fat. Bulimia nervosa is defined as frequent episodes of binge eating followed by inappropriate behaviors such as self-induced vomiting to avoid weight gain. Finally, binge eating disorder is defined as recurring episodes of eating significantly more food in a short period of time than most people would eat under similar circumstances, with episodes marked by feelings of lack of control. (American Psychiatric Association, 2013).

In an extensive review of the prevalence and correlates of eating disorders in adolescents, Swanson et al. (2011) found lifetime prevalence rates of 0.3% for anorexia nervosa, 0.6% for bulimia nervosa, and 1.6% for binge eating disorders. The age of onset for anorexia was 12.3 years, 12.4 years for bulimia, and 12.6 years for binge eating. In addition, studies have reported increasingly higher rates for younger children, boys, and minority groups (Campbell & Peebles, 2014; Kinasz et al., 2016). Finally, there are complicated gender differences in the prevalence of eating disorders (Streigel-Moore et al., 2009), in correlates of extreme dieting behaviors (Brown et al., 2015), in symptom trajectories (Allen et al., 2013), and in the clinical presentation of eating disorders (Kinasz et al., 2016).

Further, eating disorders are associated with substantial medical and psychiatric comorbidities among adolescents, including but not limited to: amenorrhea, endocrine changes, osteopenia, depression, anxiety disorders (particularly obsessive-compulsive disorder), substance abuse, attention-deficit hyperactivity disorder, and personality traits and disorders (Allen et al., 2013; Campbell & Peebles, 2014; Herpetz-Dahlmann, 2009; Swanson et al., 2011). These comorbidities often cause adverse effects on adolescents' physical and social functions, as well as their achievement in adulthood.

Given the serious consequences of eating disorders, early detection is necessary. Since school social workers are at the frontline in dealing with student mental, social, and behavioral health and development (Jarolmen, 2013), detecting and responding to eating disorder risk is clearly in the purview of school social work responsibilities (Early & Drew, 2013). While Rosen and The Committee on Adolescents (2014) noted that the clinical assessment of eating disorders in children and adolescents is complicated by various physiological, psychological, and social characteristics that can contribute to an actual diagnosis of an eating disorder, they suggested that using a screening tool initially in the clinical process is a good practice (p. 1241). Given adequate psychometrics in school settings, the SCOFF questionnaire may have substantive value for school social work practice.

The SCOFF questionnaire

The SCOFF questionnaire is a simple, easily administered eating disorder screening instrument. It was developed for use with adult females in primary care settings in the United Kingdom (Morgan et al., 1999). It was developed through a rigorous process which included the use of experts in the eating disorder field to identify essential elements of eating disorders and a series of studies designed to assess the questionnaire's ability to identify cases of anorexia nervosa and

bulimia nervosa (Hill et al., 2010). Questionnaire designers made the case for the content validity of the instrument and early studies testing the diagnostic validity of the questionnaire indicated it was able to identify true cases of eating disorders (sensitivity, true positives) and true cases of no eating disorder (specificity, true negatives) (Hill et al., 2010).

SCOFF is an acronym for the content of each item in the questionnaire. In the American version (used in this study, see below), 'S ' stands for sick in Item 1, 'C' stands for control in Item 2, 'O' stands for others in Item 4, 'F' stands for fourteen pounds in Item 3, and 'F' stands for food in Item 5 (Parker et al., 2005). As noted, core features of both anorexia nervosa and bulimia nervosa were woven into item content. For example, Item 1 askes about intentional vomiting; Item 2 addresses loss of control over eating; Item 3 is concerned with weight loss; Item 4 addresses body dissatisfaction; and Item 5 is concerned with food intrusive thoughts (Hautala et. al., 2009). The questionnaire was not designed to diagnose a specific eating disorder; rather it was designed to suggest an eating disorder might be present (Morgan et al., 1999).

Although it was developed for use in adult populations, the questionnaire is frequently recommended for use with children and adolescents in the US (Campbell & Peebles, 2014; Rindahl, 2017; Rosen and The Committee on Adolescents, 2014). In addition, the questionnaire has been translated for use with adolescents and young adults in China (Leung et al., 2009), Finland (Hautala et al., 2009), Germany (Herpetz-Dahlman et al., 2015), Italy (Siervo et al., 2005), Spain (Muro-Sans et al., 2008), and Mexico (Sanchez-Armass et al., 2012).

Various studies have examined the SCOFF questionnaire's psychometric properties in children and adolescent populations (Hautala et al., 2009; Leung et al., 2009; Muro-Sans et al., 2008; Rueda et al., 2005; Parker et al., 2005). Results from these studies indicated there was wide variability in the psychometric properties and screening accuracy (diagnostic validity) of the questionnaire. For example, coefficient alphas reported in the studies were low ranging from .44 to .57. Sensitivity coefficients ranged from 53.3% to 81.9% and specificity coefficients ranged from 75.8% to 93.2%.

A recent systematic review and meta-analysis of the performance of the questionnaire (Kutz et al., 2020) also found wide variability in various measures of diagnostic validity with sensitivities ranging from 53.7% to 97.7% and specificities ranging from 21.0% to 97.1%. Although there was substantial diversity in ages, settings, and eating disorder reference standards for the studies reviewed, the authors note that the highly variable diagnostic performance of the SCOFF is a problem issue for applied use. (p.892).

Recently Bean (2019) examined SCOFF questionnaire psychometrics in a high school population using item response theory methods. Results from the study indicated that SCOFF items varied substantively in their statistical relationship with the latent trait of eating disorder risk. Further, there were differences in how male and female students endorsed items. An important conclusion from the study was that practitioners using the SCOFF questionnaire in schools or other youth serving settings should not uncritically use the published scoring rule that a summed score ≥ 2 is an indicator of eating risk. This scoring rule assumes that each item is an equally weighted predictor of risk which was not supported by study findings either overall or within gender groups.

Study goals

This study is an extension of the Bean (2019) high school study from the standpoint that the same IRT methods and framework were used to explore the SCOFF questionnaire in a population of seventh grade students. As noted above, Swanson et al. (2011) determined that the

age of onset for anorexia was 12.3 years, 12.4 years for bulimia, and 12.6 years for binge eating. Students in the seventh grade typically are in the 11-12-year-old range so this study is well-placed in assessing the SCOFF questionnaire as an early eating risk detection instrument. The purpose of the study was to explore, in detail, how SCOFF questionnaire items performed in measuring the construct of eating disorder risk in this age group. In addition, we were interested in testing for gender differences in item responses. As noted, there is evidence that there are gender differences in the course and characteristics of eating disorders (Kinasz et al., 2016) and IRT methods are useful in examining how items (differential item functioning (DIF)) and scales (differential test functioning (DTF)) operate across groups. In a recent article, Nugent (2017) stressed the importance of examining DIF and DTF in social work measurement research. He noted that given the diversity of our populations, we should be wary of assuming that a scale, or items on a scale, function the same for persons in different groups. DIF and DTF have important implications for how an instrument is scored and used.

A brief description of item response theory

Since the language and concepts of IRT might be new to some readers, what follows is a description of essential IRT concepts (see Nugent, 2017 for an accessible introduction to IRT for social work measurement). Briefly, IRT is a statistical process that links assessment, survey, or test item responses to a latent trait (sometimes referred to as a latent variable or a construct). (Baker & Kim, 2017). The process proceeds as follows. An assumption is made that each respondent has an amount of the latent trait which influences the probability that the respondent will endorse an item. In this study, we assumed that each student possessed a level of eating risk ranging from extremely low to extremely high and that the level of eating risk influenced the probability of endorsing (saying "yes" to) a SCOFF item. In IRT modeling it is necessary to establish a measurement scale for the latent trait of interest. In this study, the scale for eating risk was expressed as theta (θ) which is in a standard score form with a mean of 0 and a standard deviation of 1.

A product of the item-linking process is a set of model parameters that characterize the relationship between each item and θ. An item location parameter (also referred to as a difficulty or b-parameter) locates an item on the θ scale. It is interpreted as the point on θ where a respondent has a .5 probability of endorsing that item (Baker & Kim, 2017, p.18). An item slope parameter (also referred to as a discrimination or a-parameter) is interpreted as a measure of an item's ability to discriminate between different levels on the θ scale (Baker & Kim, 2017, p.4). An item slope also is interpreted as measure of the strength of the relationship between that item and the latent trait (similar to a factor loading in factor analysis).

Typically, item parameters are estimated using a marginal maximum likelihood fitting function (Chalmers, 2012). Once the parameters are estimated a variety of indexes are available to assess how well the model fits the data (e.g., root mean square of approximation, standardized root mean square residual, comparative fit index). If the model adequately fits the data, it is then possible to compute various IRT components that provide insights into the item and scale attributes that form the basis for a comprehensive IRT analysis. These attributes include item and scale information, conditional standard errors, conditional reliability, model-based person scores in both the θ metric and transformed estimated true scores, differential item functioning, and differential scale functioning.

Method

Participants

The data used in this study came from 3,298 seventh grade students in eighteen Ohio school districts. The characteristics of the sample of students were as follows: 697% were in suburban schools, 30.3% were in city schools; 48.8% were male, 51.2% were females; 68.9% were White, 15.3% were African-American, 15.8% were Other Race; 57.1% lived with both parents, 28.3% lived with one parent or split time between parents, 14.6% lived with another caretaker. The eighteen districts were in one large urban county. Data were collected following consent procedures prescribed in each district.

Instrument

As noted above, the SCOFF questionnaire is composed of five questions which were presented to respondents as follows:

1. Do you make yourself sick because you feel uncomfortably full?
2. Do you worry you have lost control over how much you eat?
3. Have you recently lost more than fourteen (14) pounds in a three-month period?
4. Do you believe yourself to be fat when others say you are too thin?
5. Would you say that food dominates your life?

The response scale for each item is 'yes' or 'no'. The summary scale score is a count of items that have a 'yes' response: The range of the summary score is 0 to 5. The scoring rule typically applied is that a 'yes' response to two or more questions indicates that the respondent is at risk of having an eating disorder (Hill et al., 2010). A Flesch-Kincaid readability analysis indicated that the SCOFF questionnaire reads at a 6^{th} grade level.

Data Analysis

IRT model testing has some theoretical assumptions— unidimensionality and local independence— that require attention before proceeding to model building. Unidimensionality refers to the assumption that the items which compose a scale measure something in common; that is, they are influenced by a single, underlying latent trait. Local independence is closely related to unidimensionality. It refers to the assumption that items should be uncorrelated after controlling for the effect of the latent trait. We tested the unidimensionality assumption of our items using a parallel analysis based on minimum rank factor analysis (PA-MRFA) as implemented in the FACTOR program (Ferrando & Lorenzo-Seva, 2017). We assessed local independence using a standardized signed phi residual correlation method implemented in the R mirt package (Chalmers, 2012).

For the IRT model fitting analysis, we examined a 2-parameter logistic (2PL) model following suggestions that 2PL models are appropriate for dichotomous clinical assessment items (Reise & Waller, 2009; Thomas, 2011). We conducted the IRT analyses using R statistical computing environment (R Development Core Team, 2019) and the packages ltm: An R Package for Latent Variable Modeling for Item Response Theory Analyses (Rizopoulos, 2006) and mirt: A Multidimensional Item Response Theory Package for the R Environment (Chalmers, 2012). We assessed the fit of our 2PL model fit using a limited information strategy recommended by Maydeu-Olivares & Joe (2014). To assess fit we examined two fit statistics: the root mean square error of approximation (RMSEA) and the standardized root mean square residual (SRMSR).

Finally, as noted above, we explored gender based DIF and DTF using procedures recommended by Meade (2010). Using this approach, it was possible to compute DIF and DTF effect sizes which are helpful in quantifying the extent to which there is DIF and DTF between males and females. For this analysis, we used an Excel-based program called VisualDF to compute various DIF and DTF effect sizes (Meade, 2010).

Results

Item descriptive statistics

Descriptive statistics for each item are shown in Table 1. Item means are measures of the proportion of students who endorsed (responded 'yes') to an item. Item 1 was the least endorsed item (p = .09) while Item 4 was the most endorsed (p = .24). Item-total correlations with the item included ranged from .41 (Item 3) to .71 (Item 2). Item-total correlations with the item excluded ranged from .10 (Item 3) to .41 (Item 2). Cronbach's alpha for the scale was .53.

Table 1
Item Descriptive Statistics

Item	Mean	Item-Total Correlation if Item Included	Item-Total Correlation if Item Excluded	Cronbach's Alpha if Item Excluded
Item 1. Do you make yourself sick because you feel uncomfortably full?	.09	.52	.28	.48
Item 2. Do you worry you have lost control over how much you eat?	.22	.71	.41	.39
Item 3. Have you recently lost more than fourteen pounds in a three-month period?	.13	.41	.10	.58
Item 4. Do you believe yourself to be fat when others say you are too thin?	.24	.68	.34	.44
Item 5. Would you say that food dominates your life?	.10	.60	.37	.43

2PL model parameters and interpretation

After determining that our items met the assumptions of unidimensionality and local independence, we proceeded to fit a 2PL model using a full-information marginal maximum likelihood fitting function to estimate model parameters (Chalmers, 2012; Rizopoulos, 2006). The RMSEA = .04 (95% CI [.03, .06]) and SRMSR = .03 indicated an adequate model fit using the suggested threshold values of < .06 for the RMSEA and <.05 for the SRMSR (Maydeu-Olivares & Joe, 2014).

Parameter estimates and standard errors for the 2PL model are shown in Table 2. Values for the location parameter ranged from .92 for Item 2 to 5.21 for Item 3. Recall that the location parameter represents the point on the θ scale where a respondent has a .5 probability of endorsing that item. Thus, a respondent would have less eating risk (corresponding to a value of .92 on the θ scale) to have a .5 probability of endorsing Item 2 compared to the eating risk it would take to have a .5 probability of endorsing Item 3 (corresponding to a value of 5.21 on the θ scale). Location values for the three remaining items can be similarly interpreted.

The slope parameters shown in Table 2 are interpreted as a measure of an item's ability to discriminate between different levels on the θ scale; steeper slopes are more discriminating. The slope parameters presented in Table 2 ranged from .37 for Item 3 (the least discriminating item) to 2.67 for Item 2 (the most discriminating item). Using a framework for interpreting slope parameters recommended by Baker and Kim (2017, p. 26), we considered estimates for Item 1, Item 2, and Item 5 to be very high in their ability to differentiate respondents with different levels of θ; Items 1 and Item 4 were moderate in their ability to differentiate respondents; and Item 3 was low in its ability to differentiate respondents. (Note: The labels are as follows: 0 = No ability; .01-.04 = Very low; .35-.64 = Low; .65-1.34 = Moderate; 1.35-1.69 = High; >1.70 = Very high).

Table 2

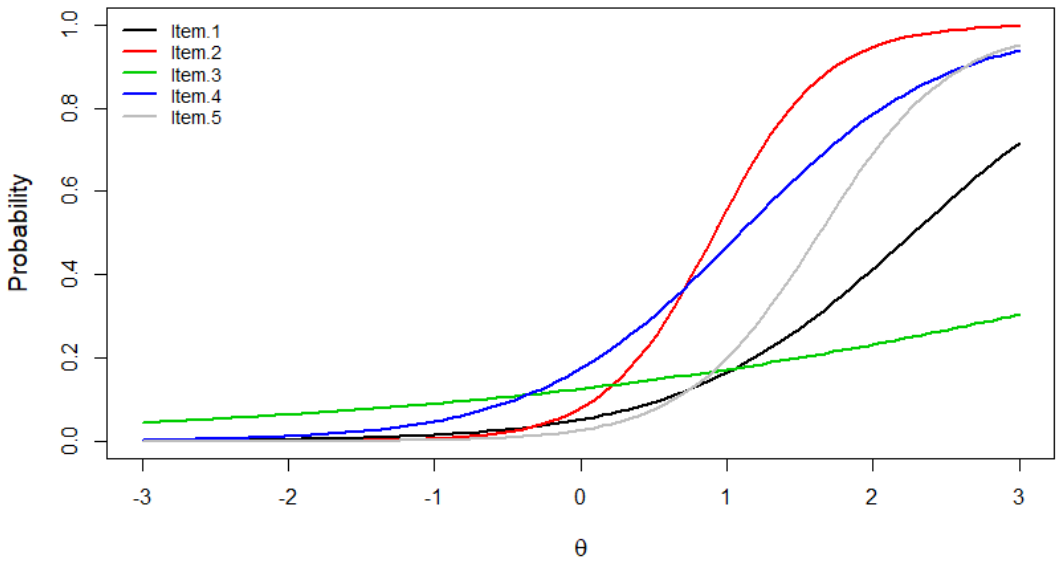
2PL Model Coefficients, Standard Errors, and Item Information

Item	Slope Parameter (SE)	Location Parameter (SE)	% of Total Information
Item 1. Do you make yourself sick because you feel uncomfortably full?	1.28 (.11)	2.27 (.14)	16.2
Item 2. Do you worry you have lost control over how much you eat?	2.67 (.28)	.92 (.04)	33.8
Item 3. Have you recently lost more than fourteen pounds in a three-month period?	.37 (.07)	5.21 (.97)	4.1
Item 4. Do you believe yourself to be fat when others say you are too thin?	1.43 (.10)	1.09 (.06)	18.1
Item 5. Would you say that food dominates your life?	2.20 (.19)	1.63 (.07)	27.8

The relationship between each item and θ is graphically presented by the item characteristic curves (ICCs) shown in Figure 1. An ICC traces the increasing monotonic relationship between θ and the probability of responding to an item using a cumulative logistic function. It is a smooth S-curve which shows the probability of not endorsing an item is near zero at low levels of θ and steadily increasing as θ increases. It is interesting to examine how 2PL model parameters shown in Table 2 define the trace lines shown in item ICCs in Figure 1. For example, the ICC for Item 2 was located lower on θ (θ = .92) than the other items. Further, it had the steepest slope (slope parameter = 2.67). On the other hand, the location parameter for Item 3 (θ = 5.21) placed it higher on θ than the other items. It also had the shallowest slope (slope parameter = .32). Overall, the ICCs indicated that item location parameters and slopes for Item 1, Item 2, Item 4, and Item 5 were concentrated in approximately the $+.5 \leq \theta \leq +2.5$ range. This pattern is not uncommon in scales designed for clinical or screening use (Reise & Waller, 2009) where takes a higher θ level (more eating risk) to endorse an item designed to detect that risk. Our location parameters indicated that this was the case for this item set. On the other hand, the location parameter for Item 3 placed it well outside this θ range—in fact, the extremely high location parameter for this item flagged it as a possible problem (to be discussed later).

Figure 1

Item Characteristic Curves for all items.



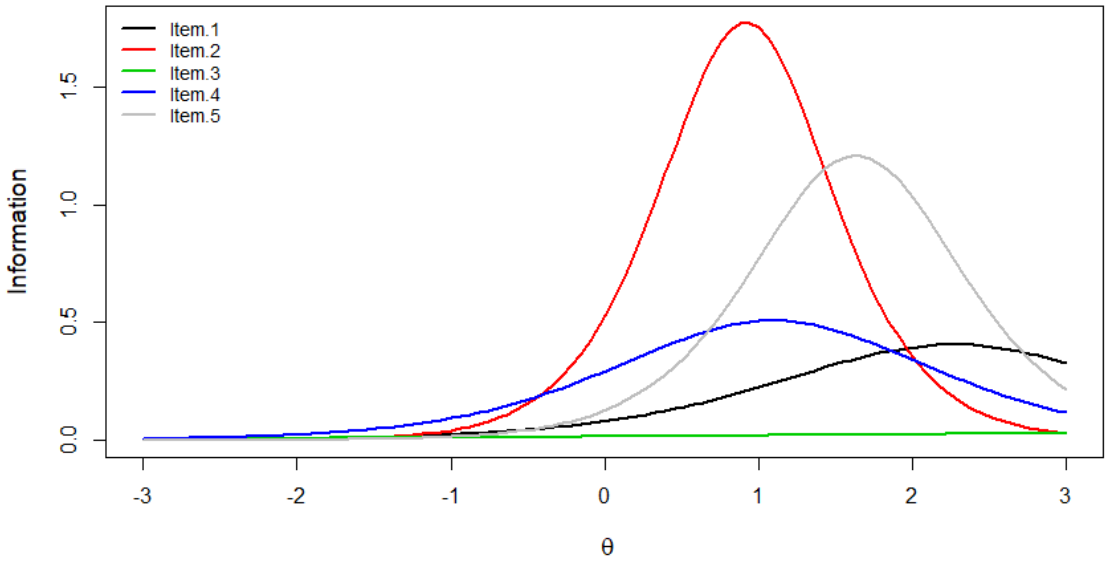
Item information

A primary goal of an IRT analysis is to generate estimated θ scores for respondents which can then be used in making clinical decisions or for use in further statistical analyses. When we generate estimates of scores, we are interested in their precision. In IRT, information is a key statistical concept that refers to the ability of an item to provide precise estimates of scores on θ (Baker & Kim, 2017, p. 89). Item information is largely a function of an item's slope—the larger the slope, the more information that item provides (Nugent, 2017, p. 313). Stated differently, items that have a stronger relationship with θ provide more statistical information than items with weaker relationships. For example, the relationship between slopes and information is illustrated in Table 2. Item 2 provided the most information (33.8% of total information) followed closely by Item 5 (27.8% of total information). Item 3 was the least informative (4.1% of total information).

Item information is visually displayed in the item information curves (IICs) shown in Figure 2. Each IIC curve is mathematically defined as a function of an item's location parameter, discrimination parameter, and θ scores. Thus, there is a direct relationship between item information presented in an item's IIC and the location and discrimination parameters of our 2PL model shown in Table 2. The location of the peak of each curve is approximately defined by that item's location parameter; the height of the curve is a function of that item's slope parameter. For example, the high peak curve for Item 2 graphically shows that it is the most informative item in the scale; the shallow curve for Item 3 shows graphically it is the least informative item in the scale.

Figure 2

Item Information Curves for all items.



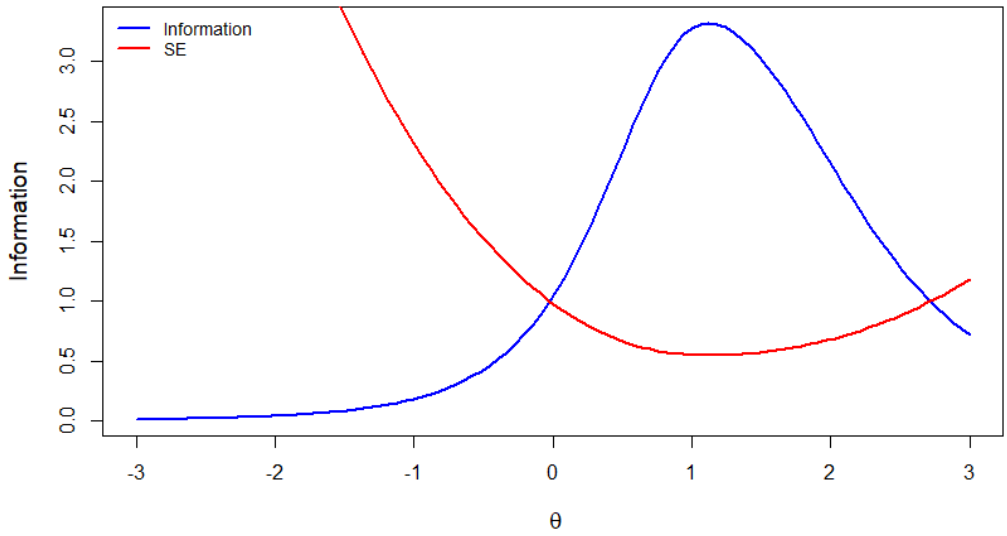
Scale information and conditional standard errors

An important feature of IRT is that information functions for individual items can be summed to form a scale information function (SIF) (Note: scale information functions also are called test information functions). The scale information function is a summary of how well items, overall, provide statistical information. Scale information values can be used to compute conditional standard errors which serve as indicators about where on θ scores are most precisely estimated. The relationship between scale information and conditional standard errors is illustrated in Figure 3. The blue line represents the scale information function. The red line represents the conditional standard errors. Taken together, these trace lines provide a visual reference about how estimate precision varies across θ with larger information values and smaller values of SE values corresponding to better estimate precision (Note: Information and standard errors are mathematically linked via a simple transformation: SE = 1 / √ Information).

These two functions are critical in understanding how a scale operate over θ. By examining the curves—especially the SE curve—it is possible to determine where on the θ scale estimates are most precise. For example, scale information peaks at about θ = +1 and concentrates in the $0 \leq \theta \leq +3$ range. Standard error values are lowest in the $+.5 \leq \theta \leq +2.5$ range; thus, the most precise θ estimates are in the that range. These curves indicate that the SCOFF scale optimally operates in a narrow range of eating disorder risk in the upper levels of θ.

Figure 3

Scale information curve and standard errors



Differential Item and Test Functioning

Our next analysis was concerned with differential item functioning and differential test functioning. Item means and 2PL model parameters for males and females are displayed in Table 3. There were notable differences between genders on some of the items. For example, males were less likely to endorse Item 2 (mean = .15) than females (mean = .29). Also, males were less likely to endorse Item 4 (mean = .13) than females (mean = .35). The female location parameters (range = .67 to 5.23) were more variable than male location parameters (range = 1.31 to 4.53). Gender variability was less evident in the slope parameters where female variability was .38 to 2.74 and male variability was .42 to 2.42.

Table 3

Item Means, 2PL Model Parameters, and Standard Errors for Males and Females,

Item	Male Mean	Male Slope (SE)	Male Location (SE)	Female Mean	Female Slope (SE)	Female Location (SE)
Item 1. Do you make yourself sick because you feel uncomfortably full?	.08	1.37 (.18)	2.24 (.19)	.10	1.30 (.15)	2.18 (.18)
Item 2. Do you worry you have lost control over how much you eat?	.15	2.20 (.30)	1.31 (.08)	.29	2.74 (.42)	.67 (.05)
Item 3. Have you recently lost more than fourteen pounds in a three-month period?	.14	.42 (.11)	4.53 (1.16)	.13	.38 (.10)	5.23 (1.33)
Item 4. Do you believe yourself to be fat when others say you are too thin?	.13	1.55 (.18)	1.67 (.12)	.35	1.25 (.12)	.67 (.07)

Item	Male Mean	Male Slope (SE)	Male Location (SE)	Female Mean	Female Slope (SE)	Female Location (SE)
Item 5. Would you say that food dominates your life?	.08	2.42 (.35)	1.73 (.11)	.12	2.16 (.25)	1.51 (.09)

Results from the DIF and DTF analysis are shown in Table 4. The expected score standardized difference (ESSD) coefficients are measures of the magnitude of DIF and DTF. Meade suggested that because they are like Cohen's *d* coefficients it is possible to interpret ESSDs in the standardized mean difference framework proposed by Cohen (Cohen, 1988). Cohen suggested that a *d* of .20 would be considered a small effect size, .50 would be considered a medium effect size, and .80 would be considered a large effect size. Using that framework, the ESSD for Item 1 (ESSD = .03) was small in magnitude indicating that males and females with the same θ score had similar probabilities of endorsing that item. Item 5 (ESSD = .20) showed a small amount of DIF. This indicated that for a given value of θ, there were modest gender probability differences in endorsing the item. The ESSDs for Item 2 (ESSD = .53) and Item 3 (ESSD = -.39) were considered to have medium DIF. Finally, the ESSD for Item 2 (ESSD = 1.01) is a large DIF indicating that for a given value of θ, there were substantive gender probability differences in endorsing that item. The expected scale score standardized difference (ESSSD) shown at the bottom of Table 4 is a measure of DTF. The value (ESSSD = .55) is medium effect size.

Table 4
Gender DIF and DTF

Item and Scale	Expected Score Standardized Difference
Item 1. Do you make yourself sick because you feel uncomfortably full?	.09
Item 2. Do you worry you have lost control over how much you eat?	.53
Item 3. Have you recently lost more than fourteen pounds in a three-month period?	-.39
Item 4. Do you believe yourself to be fat when others say you are too thin?	1.01
Item 5. Would you say that food dominates your life?	.20
Scale Level	.55

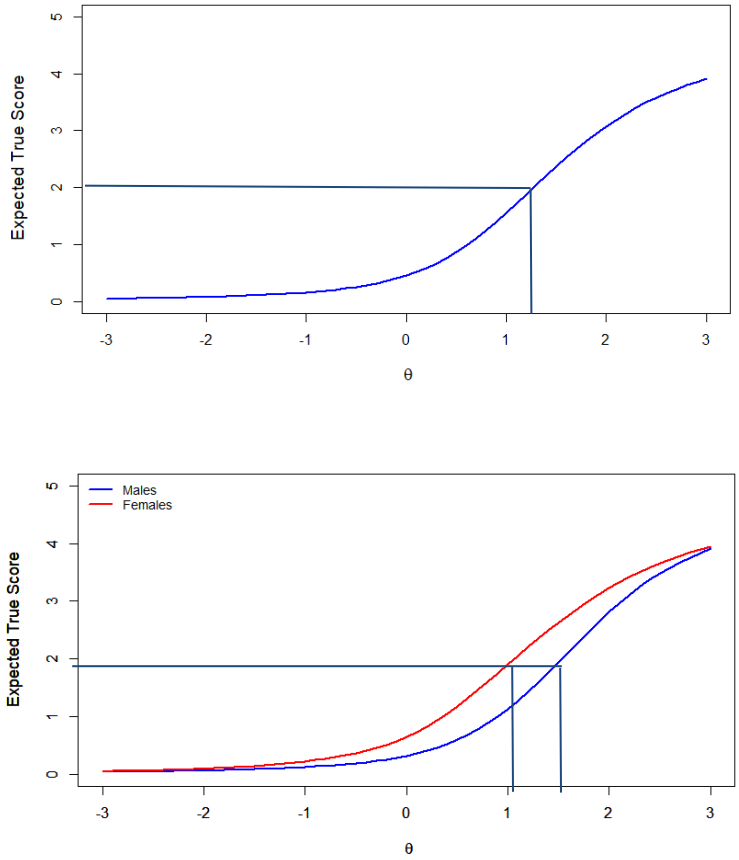
Scoring

Thus far we have been working with θ scores as representations of eating disorder risk. In this section, we address the idea that often it is of interest to transform those θ estimates into the original scale metric (the sum of the number of items endorsed with a range of 0-5). The transformed scores are called expected true scores. Because they are expressed in the original scale metric, they often provide a more familiar frame of reference for score interpretation (Baker & Kim, 2107, pp. 59-60).

A scale characteristic curve (SCC) provides a means for graphically showing how estimated θ scores map to expected true scores. The SCC for our data is shown in the top pane of Figure 5. It has a straightforward use; for any given θ score we can easily find a corresponding expected true score. For example, to have an expected true score at the clinical cut-off of two, a student would need to have a θ = +1.14 value. SCCs also are helpful in examining the impact DTF on scoring. SCCs for males and females are shown in the bottom pane of Figure 5. It takes a higher risk score (θ = 1.27) for males to have an expected true score at the clinical cut-off of two than for females (θ = 1.06). Examining DTF scoring differences has important practical implications. It could be important to use two different scoring models – in this case one for females and one for males – to be more precise in determining eating risk.

Figure 5

Scale characteristic curves for all students and for males and females



Discussion

The primary goal of this study was to examine the SCOFF questionnaire's properties in a seventh grade population using an IRT analytic strategy. As noted above, this study was conducted as a companion to a similar study using IRT methods to analyze SCOFF questionnaire items in a high school population (Bean, 2019). This study mirrors many of the results from that study. What follows is a summary of key results:

- The basic descriptive summary of each item illustrated the 'relatively rare' nature of eating disorder symptoms. Item means ranged from .09 for Item 1 to .24 for Item 4. Various item-total correlations indicated that items had positive correlations with the summed score scale. Item 2 had the highest item-total correlation (both item-included and item-excluded); Item 3 had the lowest correlation. The overall Cronbach's alpha for all items was .53.
- The unidimensionality and local independence analysis indicated the SCOFF items met the assumptions of IRT model-building by measuring a latent construct in common. We conceptualized the latent construct as eating disorder risk. For all the analyses we conducted, eating disorder risk was represented as Theta (θ).
- There was substantial variability in how items linked to the latent trait of eating disorder risk. Location and slope parameters provided important insights into this variability with location parameters indicating how much eating disorder risk it took to endorse an item and item slope parameters indicating how well an item discriminated risk. Item 2 ("Do you worry you have lost control over how much you eat?") was the most discriminating item (slope = 2.67) and took the comparatively least amount of risk to have a .5 probability of endorsement (location = .92). Item 3 ("Have you recently lost more than fourteen pounds") was the least discriminating item (slope = .37) and took the most amount of risk to have a .5 probability of endorsement (location = 5.21).
- We considered the relationship between Item 3 and eating disorder risk so weak that we recommend it not be used in scoring (or be used cautiously). Item 3 information was less than 5 percent of total information and the location parameter placed it in the extremely high range of θ; location parameters this high are considered problematic.
- Model-based estimates of eating disorder risk (θ) were concentrated in the $0 \leq \theta \leq +3$ range with the most precise estimates falling in the $+.5 \leq \theta \leq +2.5$ range. The estimated scores clustered around $\theta = +1$ which is the θ value that corresponds to an expected true score close to the clinical cutoff value of two. It is not uncommon for clinical and screening scale with low item endorsements to measure a narrow band of a construct and to concentrate around a clinical threshold value (Reise & Waller, 2009, p.31).
- We detected gender DIF in four items. Item 5 had a small effect size of .20. Effect sizes for Item 2 (ES = .53) and Item 3 (ES = -.39) were in the medium range. Item 4 (ES = 1.01) had a very large effect size. The DTF effect size (ES = .55) was in the medium range. We considered both the medium and large DIF and DTF values to be clinically meaningful differences.
- We illustrated how to use model-based θ score estimates to compute estimated true scores. Estimated true scores were expressed in the original summed score metric (Range = 0-5). They tended to be more accurate measures of eating disorder risk than the recommended simple summed score, however, because they were computed using model parameters.

Implications for practice

The most important take-away from this study is that our results do not support the generally recommended scoring rule that a summed score of two or greater flags a seventh grade respondent as being at eating disorder risk. This scoring rule assumes that SCOFF items are equally weighted as indicators of risk. Our results suggest that items are not equally weighted either overall or within male and female groups. A more realistic approach in an applied setting would be to use the information in Table 5 to guide screening. This guide assumes a school social worker should use SCOFF items that are most informative (in the statistical sense where informative means more precise in estimating risk). For example, we suggest that a school social worker first look at Item 2 to see if it was endorsed, then next look at Item 5 to see if it was endorsed, and then proceed to examine the other items. Paying attention to the most informative items and then using other items in the decision process makes the most sense to us (per suggestions by Cotton, Ball, & Robinson, 2003). As a case in point, a summed score of two that uses Item 2 and Item 5 (overall and for males and females) would be based on highly informative items and would, therefore, be potentially more clinically relevant than a summed score of two that uses Item 1 and Item 4. The latter score would meet the technical recommended risk threshold, but it would be based on poorly discriminating items.

The gender DIF and DTF we detected is an important clinical and screening issue. Thielemann et al., (2018) noted there is a growing research and clinical interest in gender differences in the age of onset of eating disorders, in how body image relates to eating disorders, in eating disorder symptom patterns, and in various compensatory behaviors. As an example of body image difference, females tend to prefer a thin body whereas males tend to prefer a more muscular body. Relative to compensatory behaviors, while females tend to engage in bulimic behaviors like vomiting and use of laxatives, males tend to use excessive to regulate weight. Since the SCOFF questionnaire was developed as a screening instrument for use with women, some of the issues related to male eating disorders might not be adequately covered.

Table 5

Ranking of Most Informative to Least Informative Items for All Students, Male Students, and Females Students

	All	Males	Females
Most Informative	Item 2 – loss of control over eating (33.8%)	Item 5 – food intrusive thoughts (30.6%)	Item 2 – loss of control over eating (35.2%)
	Item 5 – food dominates life (27.8%)	Item 2 – loss of control over eating (27.8%)	Item 5 – food dominates life (27.8%)
	Item 4 – body dissatisfaction (18.1%)	Item 4 – body dissatisfaction (19.6%)	Item 1 – make yourself sick (16.8%)
	Item 1 – make yourself sick (16.2%)	Item 1 – make yourself sick (17.3%)	Item 4 – body dissatisfaction (16.1%)
Least Informative	Item 3 – weight loss (4.1%)	Item 3 – weight loss (4.8%)	Item 3 – weight loss (4.1%)

In a recent systematic review, Rindahl (2017) concluded that the SCOFF questionnaire was the best eating disorder screening tool available for school nurses. She noted that the "validity and reliability of questionnaire offers a real tide-turning opportunity for the identification of adolescents that are in danger of eating disorders" (p. 4). Our assessment of the SCOFF questionnaire literature about the use of the instrument in children and adolescent populations did not support this sweeping statement. We found that reported Cronbach's coefficient alphas were consistently low and that diagnostic validity was variable. She further suggested that the brevity of the questionnaire is a significant strength in the context of a busy school day. On that point we agree—school nurses and social workers can benefit from short screening scales. However, short scales must be psychometrically sound; brevity is only one consideration in selecting and using clinical and screening scales in schools. Understanding how items and scales operate is the primary consideration when making clinical decisions.

Rosen and the Committee on Adolescence (2010) provided guidelines for screening children and adolescent eating disorders in practice. These guidelines include use of the SCOFF questionnaire, plus consideration about other factors, including deviations from age-appropriate growth, the occurrences of inappropriate dieting and amenorrhea. Furthermore, because eating disorders may be hidden by adolescent patients, specific history and family history, as well as physical symptoms of eating disorders, should be regarded as the important references in screening among adolescents. We cautiously recommend the SCOFF questionnaire for use with children and adolescent populations as suggested by the Rosen and the Committee on Adolescence (2010) guidelines. However, based on our results we strongly urge school social workers and other school clinicians not to rigidly adhere to the recommended scoring rule and to factor in gender differences in the screening process.

It is important to note that the SCOFF questionnaire is one of several eating disorder screening scales available for school social workers. Rindahl (2017) identified eleven eating scales that could be used in school settings. Although there was variability in the focus of these scales (e.g., binge eating, non-specific screening, emotional eating), they do provide options for school social workers who need tools for eating disorder assessments and interventions. For example, the Eating Attitudes Test (EAT-26) is a popular general eating disorder screening tool with good psychometric properties (Garner et al., 1982). It is a comprehensive scale with items that map to anorexia nervosa, bulimia nervosa, and binging; it could be a viable alternative for use by school social workers.

Limitations and future research

As with all studies, our conclusions must be tempered by a few cautions and limitations. The study used a purposive sample of seventh grade students. Some might argue that this limits the generalizability of our findings. An advantage of IRT is that the estimated item parameters (slopes and thresholds) are population invariant which means that, theoretically, item parameters will be the same (or nearly the same) in different populations (DeMars, 2010). To our knowledge, only one other study examined the SCOFF questionnaire using IRT methods in an adolescent population (Bean, 2019). We would like to encourage more SCOFF questionnaire research in children and adolescent populations using IRT to see if various item and scale characteristics operate in a similar fashion in various populations. Further, we think it is especially important to further explore gender DIF and other possible sources of DIF and DTF such as age, weight status (BMI), and mental health status (depression, anxiety, self-harm). We agree with Nugent's (2017) call for attention to DIF and DTF in social work measurement

research. Detecting new sources of DIF and DTF will aid substantially in refining the precision using the SCOFF questionnaire in the eating disorder screening process.

References

Allen, K. L., Crosby, R. D., Oddy, W. H., & Byrne, S. M. (2013). Eating disorder symptom trajectories in adolescence: Effects of time, participant sex, and early adolescent depressive symptoms. *Journal of Eating Disorders (Open Access)* (August), 1:32 http://www.jeatdisord.com/content/1/1/32

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: Author.

Baker, F.B. & Kim, S. (2017). *The basics* of *item response theory using R.* Cham, Switzerland: Springer International Publishing AG.

Bean, G.J. (2019). An Item Response Theory analysis of the SCOFF Questionnaire in a high school population. *Journal of Evidence-Informed Social Work, 16*, 404-422. doi:10.1080/26408066.2019.1617212

Brown, C.S, Kola-Palmer, S., & Ghingra, K. (2015). Gender differences and correlates of extreme dieting behaviours in US adolescents. *Journal of Health Psychology*, *20*, 569-579. doi:10.1177/1359105315573441

Campbell, K. & Peebles, R. (2014). Eating disorders in children and adolescents: State of the art review. *Pediatrics, 134*, 582-592. doi:10.1542/peds.2010-2821

Chalmers, R. Phillip. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*, 1-29. URL http://www.jstatsoft.org/v48/i06/

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Cotton, M. A., Ball, C., & Robinson, P. (2003). Four simple questions can help screen for eating disorders. *Journal of General Internal Medicine, 18*, 53-56. doi:10.1046/j.1525-1497.2003.20374.x

DeMars, C. (2010). *Item response theory*. New York, NY: Oxford University Press.

Early, T. J., & Drew, H. (2013). Effective interventions for students with eating disorders. In C. Franklin, M. B. Harris, & P. Allen-Meares (Eds.). *The school services source Book: A guide for school-based professionals* (pp.179-190). New York, NY: Oxford University Press.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. New York, NY: Psychology Press.

Ferrando, P.J., & Lorenzo-Seva, U. (2017). Program FACTOR at 10: Origins, development and future directions. *Psicothema, 29*, 236-240. doi:10.7334/psicothema2016.304

Hautala, L., Junnila, J., Alin, J., Grönroos, M., Maunula, A. M., Karukivi, M., Saarijärvi, S. (2009). Uncovering hidden eating disorders using the SCOFF questionnaire: Cross-sectional survey of adolescents and comparison with nurse assessments. *International Journal of Nursing Studies*, *46*, 1439-1447. doi:10.1016/j.ijnurstu.2009.04.007

Herpertz-Dahlmann, B. (2009). Adolescent eating disorders: Definitions, symptomatology, epidemiology and comorbidity. *Child and Adolescent Psychiatric Clinics of North America, 18*, 31-47. doi:10.1016/j.chc.2008.07.005

Herpetz-Dahlman, B., Dempfle, A., Konrad, K., Klasen, F., & Ravens-Sieberer, U. (2015). Eating disorder symptoms do not just disappear: the implications of adolescent eating-disordered behaviour for body weight and mental health in young adulthood. *European Child and Adolescent Psychiatry, 24*, 675-684. doi:10.1007/s00787-014-0610-3

Hill, L. S., Reid, F., Morgan, J. F., & Lacey, J. H. (2010). SCOFF, the development of an eating disorder screening questionnaire. *International Journal of Eating Disorders*, 43, 344-351. doi:10.1002/eat.20679

Jarolmen, J. (2013). *School social work: A direct practice guide.* Thousand Oaks, CA: Sage Publications, Inc.

Kinasz, K., Accurso, E.C., Kass, A.E., & LeGrange, D. (2016). Sex differences in the clinical presentation of eating disorders in youth. *Journal of Adolescent Health*, 58, 410-416. doi:10.1016/j.jadohealth.2015.11.005.

Kutz, A.M., Marsh, A.G., Gunderson, C.G., Maguen, S., & Masheb, R.M. (2020). Eating disorder screening: A systematic review and meta-analysis if diagnostic test characteristics of the SCOFF. *Journal of General Internal Medicine, 35*, 885-893. doi:10.1007/s11606-019-05478-6

Leung, S. F., Lee, K. L., Lee, S. M., Leung, S. C., Hung, W. S., Lee, W. L., . . .Wong, Y. N. (2009). Psychometric properties of the SCOFF questionnaire (Chinese version) for screening eating disorders in Hong Kong secondary school students: A cross-sectional study. *International Journal of Nursing Studies, 46*, 239-247. doi:10.1016/j.ijnurstu.2008.09.004

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49*, 305-328. doi:10.1080/00273171.2014.911075

Meade, A.W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95*, 728-743. doi:10.1037/a0018966

Morgan, J. F., Reid, F., & Lacey, J. H. (1999). The SCOFF questionnaire: Assessment of a new screening tool for eating disorders. *British Medical Journal*, *319*, 1467-1468. doi:10.1136/bmj.319.7223.1467

Muro-Sans, P., Amador-Campos, J. A., & Morgan, J. F. (2008). The SCOFF-c: Psychometric properties of the Catalan version in a Spanish adolescent sample. *Journal of Psychosomatic Research, 64*, 81-86. doi:10.1016/j.psychores.2007.06.011

Nugent, W.R. (2017). Understanding DIF and DTF: Description, methods, and implications for social work research. *Journal of the Society for Social Work & Research*, *8*, 305-334. doi:10.1086/691525

Parker, S. C., Lyons, J., & Bonner, J. (2005). Eating disorders in graduate students: Exploring the SCOFF questionnaire as a simple screening tool. *Journal of American College Health, 54*, 103-107. doi:10.3200/jach.54.2.103-107

R Development Core Team. (2019). R: A language and environment for statistical computing, reference index version 3.4.3. [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

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

Rindahl, K. (2017). A systematic review of literature on school screening for eating disorders. *International Journal of Health Sciences, 5*, 1-9. doi:10.15640/ijhs.v5n3a1

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17*, 1-25. URL http://www.jstatsoft.org/v17/i05/

Rosen, D., & The Committee on Adolescence. (2010). Clinical report – Identification and management of eating disorders in children and adolescents. *Pediatrics, 126,* 1240-1253. doi:10.1542/peds.2010-2821

Rueda, G.E, Diaz, L.A., Campo, A., Barros, J.A., Avila, Orostegui, L.T., Osorio, B.C., & Cadena, L.P. (2005). Validation of the SCOFF questionnaire for screening of eating disorders in university women. *Biomedica, 25*, 196-202.

Sanchez-Armass, O., Drumond-Andrade, F. C., Wiley, A. R., Raffaelli, M., & Aradillas-Garcia, C. (2012). Evaluation of the psychometric performance of the SCOFF questionnaire in a Mexican young adult sample. *Salud Publica Mex, 54*, 375-382.

Siervo, M., Boschi, V., Papa, A., Bellini, O., & Falconi, C. (2005). Application of the SCOFF, Eating Attitude Test 26 (EAT 26) and Eating Inventory (TFEQ) Questionnaires in young women seeking diet-therapy. *Eating and Weight Disorders, 10*, 76-82. doi:10.1007/bf03327528

Striegl-Moore, R.H., Rosselli, F., Perrin, N., DeBar, L., Wilson, G. T., May, A., & Kraemer, H. (2009). Gender difference in the prevalence of eating disorder symptoms. *International Journal of Eating Disorders, 42*, 471-474. doi:10.1002/eat.20625

Swanson, S. A., Crow, S. J., Le Grange, D., Swendsen, J., & Merikangas, K. R. (2011). Prevalence and correlates of eating disorders in adolescents. *Archives of General Psychiatry, 68*, 714-723. doi:10.1001/archpenpsychiatry.2011.22

Thielemann, D., Richter, F., Strauss, B., Braehler, E., Altmann, U., & Berger, U. (2018, May 3). Differential item functioning in brief instruments of disordered eating. *European Journal of Psychological Assessment*. Advance online publication. http://dx.doi.org/10.1027/1015-5759/a000472

Thomas, M. L. (2011). The value of item response theory in clinical assessment. *Assessment, 18,* 291-307. doi:10.1177/107319110374797.