Scientific Research Publishing

Open Journal of Statistics, 2018, 8, 519-532
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Use of the IRT Model to Validate Test Items from a Technology Assisted Health Coaching Program

Elysia Garcia[1], Subhash Aryal[1], Emily Spence-Almaguer[2], Danielle Rohr[2], Scott T. Walters[2]

[1]Department of Biostatistics and Epidemiology, UNT Health Science Center, Fort Worth, TX, USA
[2]Department of Health Behavior and Health Systems, UNT Health Science Center, Fort Worth, TX, USA
Email: subhash.aryal@unthsc.edu

How to cite this paper: Garcia, E., Aryal, S., Spence-Almaguer, E., Rohr, D. and Walters, S.T. (2018) Use of the IRT Model to Validate Test Items from a Technology Assisted Health Coaching Program. *Open Journal of Statistics*, **8**, 519-532.
https://doi.org/10.4236/ojs.2018.83034

Received: April 17, 2018
Accepted: June 11, 2018
Published: June 14, 2018

Abstract

Item Response Theory (IRT) models have been extensively used in the field of education to identify link between a response to a test item and underlying latent capability of the test taker. We demonstrate the benefit of using IRT model to analyze health data using data from M. chat program such that statisticians can use the method in lieu of traditional methods including Cronbach's alpha, discriminant analysis and factor analysis. M. chat is a technology based health coaching program and the baseline survey from the participants in the program includes response in different but correlated domains of diet, social habits, leisure practices, mental health, substance abuse, self-sufficiency and medication adherence. We analyzed baseline data from 416 subjects using IRT models. Our results indicated that responses pertaining to alcohol and substance abuse were the most discriminating items with an average discrimination estimate of approximately 4.99 whereas the least discriminating items were the diet habits, with an average estimate of −0.476.

Keywords

IRT, M. Chat Program, Health Coaching

1. Introduction

Item Response Theory (IRT) is a psychometric tool originally applied in the field of education and currently used in multiple fields to yield categorical outcome data [1]. We use IRT models to analyze health and wellness data based on questionnaire-like variables with numerous categorical responses. While basic frequentist descriptive analysis presents statistics on variables and categories independently, an IRT model allows researchers to analyze descriptive aspects of these variables by their latent traits as well as their relationship with other variables and the data set as a whole based on individual responses [2]. IRT modeling quantifies the latent traits as three parameters: the difficulty or threshold parameter, the discrimination parameter, and the ability parameter [3]. IRT is not a new theory, nor is it the only tool that can be applied to analyze health assessment data. Common alternative methods include classical test theory (CTT), such as Cronbach's coefficient alpha, and factor analysis [4]. However, IRT has proved to be popular because of its adaptability and its effectiveness in designing and evaluating questionnaires and its use for scoring respondents [4] [5].

The main concept which is the foundation of IRT is that there is a link between item responses and the various characteristics measured by the test [6]. Based on this concept, IRT suggests that underlying respondent performance on a set of items is a set of personal latent characteristics that can be estimated based on the respondents answers to the items and questions [5]. From these estimates, IRT produces a generalized linear model that can be used to perform further analysis.

IRT was first developed for the field of education in order to calibrate and evaluate tests and score students based on their ability and other latent traits [1]. However, IRT has expanded to more fields, from psychometrics to health assessment and clinical research [7]. Many studies have used IRT to create item banks, which comprise of a collection of already IRT-calibrated questions that are shown to be the best in defining a domain within health measurement [8]. One example of this is the patient-reported-outcomes measurement information system (PROMIS), which is part of the NIH Roadmap Initiative. This system applies both IRT and computerized-adaptive testing (CAT) to improve the precision and efficiency of health measurement, both by reducing the number of questions needed and the number of subjects surveyed [9].

IRT has also been used to assess already established health measurement tools. For example, a study by Hartman *et al.* [10] study used IRT to analyze the DSM-IV abuse and dependence criteria amongst 5587 children of ages 11 - 19. Specifically, the study aimed to answer three questions: 1) do the criteria (dependence and abuse) represent two different levels of substance involvement severity? 2) to what degree does the criterion assess cannabis abuse/problems? and 3) do the criteria work similarly across different adolescent groups? Using IRT, the study concluded that dependence and abuse were not separate constructs for cannabis problems, and that the criteria needed refinement to better assess cannabis abuse and dependence. Other studies have also used IRT to refine established health measurement tools. The results of a 1996 study analyzing measurement instruments for community-living individuals with cerebral palsy and spina bifida found that combining certain items from the Functional Independence Measure and instrumental activity measure was useful for disability assessment [11].

IRT method exhibits unique characteristics not found in traditional approaches such as factor analysis or Cronbach's alpha. One of the principal benefits of using IRT over other classical test theory methods is that IRT takes into account the latent and invariant traits of both the item measurement and the respondent [12]. For example, IRT models simultaneously measures the latent proficiency or ability of an individual subject in answering items along with the difficulty of the item being answered [2] [13]. What makes estimates from IRT useful is that the item parameters are not test dependent, and that the item statistics are independent of individual ability level; rather, item statistics and ability are measured on the same scale, thus allowing predictions of an item or group of items for individuals or groups of individuals [5]. IRT also takes into account the dependence of an item on sampled individuals. Thus, these strengths allow results to be both more precise and generalizable [14].

IRT model is also able to detect variability in responses between groups, also known as differential item functioning [4]. This information can suggest whether a test can be applied to different sub-samples or a group. From testing differential item functioning (DIF), researchers can then reduce bias and increase validity of the model [4]. IRT also allows for more flexible and precise score equating [5]. This score equating not only works between items within a test, but also between multiple scales and questionnaires in order to create a sort of conversion table by which to analyze results [4]. Due to the IRT's ability to equate scores, it also has the benefit of improving already existing measures. .For instance, it can provide information in identifying where along a latent trait scale the measurement provides little information and needs improvement [12].

IRT also has the ability to identify clinically significant differences or change over time [15]. Due to the fact that IRT estimates of latent traits have a direct effect on probability of item response and the fact that items and parameters are measured on an equated scale and linked, changes over time and point estimates have clinical meaning [12]. Thus, researchers can use IRT to determine clinically significant thresholds of change in clinical parameters.

There are multiple models which one can apply when performing IRT. Based on the nature of the measured item outcome, such as dichotomous or ordinal, IRT provides alternatives to achieve the best fit for the data and the most representative results. For example, the one-parameter model, also commonly known as the Rasch model, applies to dichotomous item responses as a function of the latent trait and the difficulty of the item, thus allowing items to vary in difficulty but assumes that all items discriminate equally (*i.e.* equal slopes for each item) [2] [12]. Adding further parameters to measure discrimination and the impact of chance allows one to account for more variability in the data, thus increasing validity of results.

When comparing functionality and convenience of IRT to other common alternative methodologies, IRT also presents numerous advantages. Firstly, IRT provides robust estimates and models [16]. IRT also applies multiple tests and functions at once, thus proving a more time-efficient method for researchers.

Other methods such as factor analysis and Cronbach's alpha, only fulfill certain functions. For instance, Cronbach's alpha only tests the validity of model results, and factor analysis only allows researchers to pick important variables but does not provide a model with which to analyze data and draw inference. However, IRT does perform these functions. Another significant distinction between IRT models and the traditional approaches is that IRT model allows researchers to rank individual respondents based on their answers to items, thus indicating individual risk [9] [12].

Despite many strengths of IRT compared to other classical methods of test analysis, this theory does have its own weaknesses. Some of the IRT's limitations lie in its assumptions that must be satisfied: 1) unidimensionality, 2) local independence of items, 3) and item parameter invariance [17]. However, these assumptions may not always be confidently made [17] [18]. Unidimensionality and local independence can be tested using graphs such as screen plots or weighted least squares means and variance estimator for categorical data [19]. However, these assumption and tests are never conclusive as unarguably true, but instead as an approximation [5]. For unidimensionality, the assumption cannot be strictly true because several latent and test-taking factors always affect test performance to some extent [20].

Another limitation of IRT is that the model selection and building is not a straightforward process. When choosing an IRT model, the main objectives are to find a model that fits the data, properly estimates model parameters, and is used correctly [5]. There are multiple modeling schemes to choose from, such as the Rasch or graded response model. Hard consideration and comprehensive knowledge is needed in order to not only perform IRT testing, but also to consider and interpret results [5] [12]. Results of IRT also cannot indicate how to improve or write items, or what items can fill a noticeable gap in the item difficulty range [12].

Using IRT also poses a practical problem. Utilizing IRT is limited to finding a statistical program that will perform the function. Learning and implementing these programs is not easy [12]. One needs extensive knowledge of statistical program coding for such programs as R, SAS, Stata, or Winsteps to name but a few. Sometimes there is not a direct command to perform IRT, thus requiring extensive coding [2].

Despite the limitations, IRT is an efficient and beneficial tool to analyze not only testing data, but also questionnaire, measurement, and multiple other data forms. Next we illustrate the use of IRT models using data from a technology assisted health coaching program, called m. chat.

1.1. M. Chat Program

The m. chat program is funded by a Medicaid 1115 Waiver to the State of Texas. It is geared towards permanent supportive housing residents in the city of Fort-Worth, Texas with the goal of improving key health indicators of the participants by providing in-person health coaching. Subjects in the program are adult residents of permanent supportive housing who were Medicaid-enrolled or low income uninsured and English speaking. In addition, the subjects reported at least one of the following mental health conditions: having been prescribed a medication for emotional or psychological problem, receiving a pension due to psychiatric disability, reporting hallucination, or a scoring greater than 9 on the Patient Health Questionnaire (PHQ-9) depression screener, indicating moderate to severe depression. Participants were surveyed on domains which comprise general health: diet, social habits, leisure practices, mental health, substance abuse, self-sufficiency and medication adherence. Overall, 90 baseline items were included in the analyses. The program has been described in further details by Walters *et al.* [21].

1.2. Item Response Theory Model

In IRT it is assumed that there is link between the item responses and the various characteristics measured by the test [6]. Based on this concept, IRT suggests that underlying respondent performance on a set of items is a set of personal latent characteristics that can be estimated based on the respondents answers to the items and questions.

To explain the parameters and their role in IRT modeling, we will focus on two specific models which we utilized: the Rasch model and the Graded Response Model. The Rasch model, or the one parameter logistic model, is applied to binary data. The Rasch model, compared to various other IRT models, aims for simplicity more than fitness. The model is as follows:

$$P\left(X_{ij}=1\middle|\theta_i,b_j\right)=\frac{1}{1+\exp\left[-\left(\theta_i-b_j\right)\right]} \tag{1}$$

where *i* is an individual subject and j equals a specific category within a question. The model results in a probability of a Bernoulli random variable with θ_i representing the proficiency or ability of an individual subject and b_j being the difficulty of the specific category. In comparison, the Graded Response Model takes is a multi-parameter model and it can accommodate response with more than two categories.

The Graded Response Model applies specifically to ordinal data of more than two categories and builds upon the Rasch model to calculate parameters and probabilities for question *j* by subject *i* for category level *k*. Whereas the Rasch model strived for simplicity, the Graded Response Model tries to fit a model to the data utilizing more descriptive parameters. The primary assumption of the Graded Response Model is that the item discrimination and difficulty is not equal amongst items. The model can be written as follows:

$$P\left(X_{ij}=1\middle|\theta_i,a_j,b_{j1}\right)=\frac{1}{1+\exp\left[-a_j\left(\theta_i-b_{j1}\right)\right]} \tag{2}$$

$$P\left(X_{ij}=k\middle|\theta_i,a_j,b_{jk}\right)=\frac{1}{1+\exp\left[-a_j\left(\theta_i-b_{j(k-1)}\right)\right]}-\frac{1}{1+\exp\left[-a_j\left(\theta_i-b_{jk}\right)\right]} \quad (3)$$

for $k = 1, \ldots, n$. Then for the last value the model finishes with

$$P\left(X_{ij}=n\middle|\theta_i,a_j,b_{jn}\right)=\frac{1}{1+\exp\left[-a_j\left(\theta_i-b_{jn}\right)\right]} \quad (4)$$

Here θ_i represents the subjects ability which remains the same, and $b_{(jk)}$ continues to represent the difficulty parameter. However, this parameter now includes the step size parameter, $d_{(jk)}$, to create the equation $b_{jk}=b_j+d_{j(k-1)}$. The parameter, $d_{(jk)}$, gives us the latent trait location where one category becomes more likely than the one before it. Finally, the Graded Response Model includes a_j, the slope or discrimination parameter for each question.

2. Calibration

A benefit of the IRT analysis IRT analysis is that all items are placed on the same metric. As a result, direct comparison of the items measuring a variety of domains can be compared to each other. The results of our analysis are presented in Tables 1-8 for the eight varying domains. Table 1 shows items from a modified dietary screener questionnaire [22]. Table 2 shows items from the Mea

Table 1. Estimates of item parameters (category thresholds, item locations, and discrimination) for items in the DIET domain.

Item	Threshold 1	Threshold 2	Location	Slope
How many times a week did you eat fast food meals or snacks?	1.9350	−0.3821	0.7765	−0.7967
How many servings of fruit did you eat each day?	−20.8912	−31.8126	−26.3519	−0.0893
How many servings of vegetables did you each day?	−42.1029	−89.8988	−66.0008	−0.0274
How many regular sodas or glasses of sweet tea did you drink each day?	0.9462	−0.9982	−0.0260	−0.5486
How many times a week did you eat chicken, fish, or cooked beans (like black or pinto)?	−3.4216	−0.7556	−2.0886	0.5698
How many times a week did you eat regular fat potato chips, tortilla chips, or corn chips (not low-fat)?	2.5091	0.5533	1.5312	−0.7688
How many times a week did you eat desserts and other sweets (not the low-fat kind)?	0.7080	−0.7051	0.0014	−1.2727
How much margarine, butter, or meat fat do you use to season vegetables or put on potatoes, bread, or corn?	2.2543	−0.1696	1.0424	−0.6143
How many times a week did you eat red or processed meat, like hamburger, regular hot dogs, or canned meat/spam?	1.7060	−0.8614	0.4223	−0.7358

Table 2. Estimates of item parameters (category thresholds, item locations, and discrimination) for items in the MAPA domain.

Item	Threshold 1	Threshold 2	Threshold 3	Threshold 4	Location	Slope
How much do you think your leisure time activities help improve your mental health?	−2.16033	−1.0604	0.12574	1.94837	−0.286655	1.20521
How much do you think your leisure time activities help improve your physical health?	−1.29623	−0.39399	1.10724	2.94502	0.59051	1.06766
How much do you think your leisure time activities help improve your relationships with other people?	−2.53361	−0.92025	0.56001	3.13771	0.060965	0.77608
Overall, how satisfied are you with your leisure time activities?	−1.32445	−0.68862	0.01551	1.02346	−0.243525	3.27267
Overall, how much would you like a change in your leisure time activities?	−1.02092	0.0066	1.39879	2.36968	0.6885375	1.29444
How often do you feel that you are bored, with little to do?	−1.42831	−0.29443	0.83407	1.79413	0.226365	1.28276

Table 3. Estimates of item parameters for items in the ISEL domain.

Item	Threshold 1	Threshold 2	Threshold 3	Threshold 4	Location	Slope
I can easily find someone to help me think through problems.	−1.0025	−0.3886	0.3764	1.3847	0.0925	2.0122
I can easily find someone to help me sort through my finances.	−0.4600	0.0589	0.6849	1.7184	0.5006	1.5409
I can easily find someone to give me advice when I need it.	−1.3065	−0.5942	0.1841	1.2313	−0.1213	1.8875
If I were sick, I could easily find someone to help me with daily chores.	−0.5501	−0.1258	0.4833	1.4806	0.3220	1.2687
I could easily find someone to give me a ride if I needed one.	−0.5555	0.0617	0.8349	2.0197	0.5902	1.4981
I could easily find someone to loan me $10 if I needed it.	−0.1319	0.2873	1.1523	1.9985	0.8265	1.5673
When I want to socialize, I have a group of friends I can spend time with.	−0.5355	−0.1694	0.5420	1.2515	0.2722	2.3947
When I feel lonely, I have people I can talk to.	−1.1555	−0.5833	0.1200	1.0967	−0.1305	2.4588
I have a group of friends who include me in their activities.	−0.4054	−0.0969	0.7284	1.5917	0.4545	2.2396
People in this neighborhood help each other out.	2.7344	−0.1203	−0.7882	−1.7209	0.0262	−0.8014
There are people I can count on in this neighborhood.	2.2561	0.2204	−0.5916	−1.9861	−0.0253	−1.0010
People in this neighborhood can be trusted.	6.5133	2.2688	0.5280	−2.4448	1.7163	−0.4776
This is a close-knit neighborhood.	3.0407	0.8101	0.0288	−1.6319	0.5619	−0.7180

ningful Activity Participation Assessment (MAPA) [23]. **Table 3** shows items from the Interpersonal Support Evaluation List (ISEL) [24]. **Table 4** shows items from the abuse section of the Addiction Severity Index [25]. **Table 5** shows items from the Quality of Life Enjoyment and Satisfaction Questionnaire (Q-LES-Q) [26]. **Table 6** shows items from the Inventory of Drug Use Consequences

Table 4. Estimates of item parameters (category thresholds, item locations, and discrimination) for items in the "abuse" domain.

Item	Threshold 1	Location	Slope
During the past year has your use of drugs or alcohol contributed to difficulty or inability to meet responsibilities at home, school or work? (Drugs)	1.302	1.302	2.81758
During the past year has your use of drugs or alcohol contributed to difficulty or inability to meet responsibilities at home, school or work? (Alcohol)	1.4683	1.4683	1.9874
During the past year have you used drugs or alcohol even when your use could be putting yourself in physical danger (use while driving, participating in sports, operating heavy machinery, etc.) (Drugs)	1.04811	1.04811	2.92365
During the past year have you used drugs or alcohol even when your use could be putting yourself in physical danger (use while driving, participating in sports, operating heavy machinery, etc.) (Alcohol)	1.32538	1.32538	2.4196
During the past year has your drug or alcohol use led to any problems with the legal system such as drunk and disorderly arrests, being pick-up for drug possession, etc.? (Drugs)	2.26345	2.26345	1.57209
During the past year has your drug or alcohol use led to any problems with the legal system such as drunk and disorderly arrests, being pick-up for drug possession, etc.? (Alcohol)	3.32288	3.32288	1.23289
During the past year have you continued to use drugs or alcohol even though this use has contributed to problems with others such as arguments with friends or family, physical fights, etc.? (Drugs)	1.18043	1.18043	5.52588
During the past year have you continued to use drugs or alcohol even though this use has contributed to problems with others such as arguments with friends or family, physical fights, etc.? (Alcohol)	1.32063	1.32063	3.36117

Table 5. Estimates of item parameters (category thresholds, item locations, and discrimination) for items in the "QLESQ" domain.

Item	Threshold 1	Threshold 2	Threshold 3	Threshold 4	Location	Slope
Taking everything into consideration, during the past week how satisfied have you been with your….						
Physical health?	−1.8962	−0.6183	0.9025	2.6269	0.2537	1.5210
Mood?	−1.7943	−0.9556	0.5089	2.1193	−0.0304	2.1327
Work?	−1.8225	−0.5702	1.3576	2.8174	0.4456	0.9154
Household activities?	−2.6259	−1.3228	0.1278	1.9003	−0.4802	1.2271
Social relationships?	−2.3791	−1.1522	0.0730	2.1025	−0.3389	1.4105
Family relationships?	−2.8627	−1.4861	−0.3274	1.8038	−0.7181	0.9189
Leisure time activities?	−2.4740	−1.1503	0.2056	1.8248	−0.3985	1.5414
Ability to function in daily life?	−2.4313	−1.0472	0.1299	1.9498	−0.3497	1.8071

Continued

Sexual drive, interest, and/or performance?	−1.6501	−0.1195	1.3116	2.9603	0.6256	0.9062
Economic status?	−1.5947	0.1684	1.5346	3.1780	0.8216	1.2982
Living/housing situation?	−4.2376	−2.6664	−1.0159	1.3406	−1.6448	0.9716
Ability to get around physically without feeling dizzy or unsteady or falling?	−2.8601	−1.2415	−0.0848	1.7191	−0.6168	1.2819
Your vision in terms of ability to do work or hobbies?	−3.9600	−1.7660	0.2008	3.4038	−0.5304	0.7792
Overall sense of well-being?	−2.0888	−1.1289	0.0072	1.4336	−0.4442	2.7425
Medication?	−3.8732	−3.0353	−2.1968	−0.5696	−2.4187	2.4815
How would you rate your overall life satisfaction and contentment during the past week?	−1.8051	−1.0765	0.0523	1.4295	−0.3500	2.9012

Table 6. Estimates of item parameters (category thresholds, item locations, and discrimination) for items in the "INDUC" domain.

Item	Threshold 1	Threshold 2	Threshold 3	Location	Slope
My physical health was harmed by my drinking or drug use.	1.09333	1.63281	1.75952	1.49522	3.9704
My physical appearance was harmed by my drinking or drug use.	1.07942	1.64745	1.80245	1.50977	4.51737
I lost weight or didn't eat properly because of drinking or drug use.	0.90064	1.36513	1.61924	1.29500	5.50611
My family was hurt by my drinking or drug use.	1.24036	1.53135	1.71154	1.49442	5.26364
A friendship or close relationship was damaged by my drinking or drug use.	1.21567	1.6832	1.96942	1.62276	4.27188
My drinking or drug use damaged my social life, popularity, or reputation.	1.08413	1.5486	1.66999	1.43424	6.41567
I felt guilty or ashamed because of my drinking or drug use.	0.62646	1.21575	1.36901	1.070407	7.01266
I was unhappy because of my drinking or drug use.	0.52608	1.16081	1.24219	0.97636	9.33647
Drinking or drug use got in the way of my growth as a person.	0.69124	1.16589	1.28388	1.04700	8.89731
I took foolish risks while drinking or using drugs.	0.87429	1.46649	1.69443	1.34507	5.28434
While under the influence, I did impulsive things that I regretted later.	0.83679	1.46166	1.65617	1.31821	5.10543
I had an accident while I was under the influence.	1.74088	2.44278		2.09183	3.20134
I spent too much or lost a lot of money because of drinking or drug use.	0.84482	1.35184	1.58515	1.260603	4.54791

Continued

I failed to do what was expected of me because of drinking or drug use.	0.78277	1.50579	1.63649	1.30835	4.69661
I had money problems because of drinking or drug use.	0.83197	1.37446	1.51214	1.23952	6.19477
Because of drinking or drug use, social or legal authorities were involved in my life (Child Protective Services, Probation/Parole, Court).	2.08686	3.53776		2.81231	1.65717
I spent time in jail or prison because of drinking or drug use.	2.1359	3.00233		2.569115	2.13006
I engaged in illegal or unwanted activities because of drinking or drug use.	2.08785	3.18148	3.55125	2.94019	1.92712

Table 7. Estimates of item parameters (category thresholds, item locations, and discrimination) for items in the "PHQ-9" domain.

Item	Threshold 1	Threshold 2	Threshold 3	Location	Slope
Over the last 2 weeks, how often have you been bothered by any of the following problems?					
Little interest or pleasure in doing things	−0.94877	0.12632	0.77932	−0.01438	1.77019
Feeling down, depressed, or hopeless	−1.04202	0.27716	0.84156	0.02557	2.24252
Trouble falling or staying asleep, or sleeping too much	−1.17737	−0.26468	0.22412	−0.40598	1.52408
Feeling tired or having little energy	−1.78219	−0.36309	0.09106	−0.68474	1.4982
Poor appetite or overeating	−0.9949	0.10302	0.61221	−0.09322	1.10812
Feeling bad about yourself-or that you are a failure or have let yourself or your family down	−0.61294	0.24113	0.77316	0.13378	1.82349
Trouble concentrating on things, such as reading the newspaper or watching television	−0.63518	0.24218	0.6528	0.0866	1.58743
Moving or speaking so slowly that other people could have noticed. Or the opposite-being so fidgety or restless that you have been moving around a lot more than usual	−0.23134	0.85536	1.40856	0.67753	0.93619
Thoughts that you would be better off dead, or of hurting yourself	1.40401	2.25962	2.54382	2.06915	1.38794

(INDUC) [27]. Table 7 shows items from the Patient Health Questionnaire (PHQ-9) [28]. Table 8 shows items from the Morisky Medication Adherence Questionnaire (MMAQ) [29]. Most scales had been adapted from the original to

Table 8. Estimates of item parameters (category thresholds, item locations, and discrimination) for items in the "Morisky" domain.

Item	Threshold 1	Location	Slope
In the past month, did you sometimes forget to take your medicine?	0.17711	0.17711	2.06996
People sometimes miss taking their medicines for reasons other than forgetting. Thinking over the past month, were there any days when you did not take your medicine?	−0.56333	−0.56333	0.79305
In the past month, did you ever cut back or stop taking your medicine without telling your doctor because you felt worse when you took it?	−1.75724	−1.75724	1.2284
When you travel or leave home, do you sometimes forget to bring along your medicine?	−0.69707	−0.69707	1.25763
Did you take all your medicines yesterday?	−3.7356	−3.7356	−0.43213
When you feel like your symptoms are under control, do you sometimes stop taking your medicine?	−1.0656	−1.0656	1.21292
Taking medicine every day is a real inconvenience for some people. Do you ever feel hassled about sticking to your treatment plan?	−0.69038	−0.69038	0.8189
How often do you have difficulty remembering to take all your medicine?	−0.0636	−0.0636	2.3663

fit the target population. Overall, 88 items were analyzed. These tables include the parameter estimates from IRT analyses consisting of the threshold, discrimination, and location parameters. The threshold parameter indicates at what point on the ability spectrum does the sample exhibit equal probability of answering a categorical response versus the next subsequent categories. The ability spectrum is modeled using a standard normal distribution, where $\theta = 0$ equals average ability for the sample. Large negative or positive estimates indicate less or greater ability respectively. Calibration of items places all items on the same ability metric, thus allowing comparison across items. Not all questions in a questionnaire have the same scaling and categorization, so attentive interpretation is needed. The discrimination parameter indicates the ability of the item to discriminate groupings within the sample. Finally, the location parameter equates to the average threshold of the item, indicating the difficulty of the question for the sample to be answered "correctly". A lower location parameter indicates that it is easier for the sample to answer the presumed "correct" answer to the behavioral questions, while a higher location parameter indicates more difficulty. The location parameter is equivalent to the item difficulty parameter in dichotomous models.

3. Results

We included data collected from 416 participants at baseline in the analysis. The average age was 50.65 years. The sample consisted of 41.61% White, 51.77%

Black/African American and 6.62% others. The average BMI was 31.41 with 23.40% in the normal category, 22.93% in the overweight category and 53.66 in the obese category. The burden of disease was also significant in the sample with 5.7% reporting diabetes, 26.41% reporting asthma, 4.94% reporting breathing disorders and 88.54% reporting depression, anxiety or emotional disorder. More than 50% of the sample reported multiple chronic health conditions.

Of all items, the most discriminating were those under the domain "INDUC", a series of questions which ask about negative consequences of alcohol and drug use, with an average discrimination estimate of approximately 4.99. The three most discriminating items were: I was unhappy because of my drinking or drug use (Estimate: 9.33); Drinking or drug use got in the way of my growth as a person (Estimate: 8.89) and I felt guilty or ashamed because of my drinking or drug use (Estimate: 7.01) In comparison, the least discriminating items were the diet habits, with an average estimate of −0.476. The three least discriminating items were: "How many times a week did you eat desserts and other sweets (not the low-fat kind)?" (Estimate −1.27); There are people I can count on in this neighborhood (−1.00) and People in this neighborhood help each other out (−0.80). The low discrimination estimates suggest that there is high probability that subjects at any ability level will endorse any level of categorical responses. In other words, no one group of subjects is more or less likely to answer in a certain category. As a result, these questions hold very little information about the sampled individuals and their behavioral habits. In comparison, the most discriminating items can be regarded as holding the most information and have the ability to discriminate subjects into characteristic groupings based on their responses.

The location parameter provides the extremes of subject ability. The two items with the lowest estimate for the location parameters are: How many servings of vegetables did you each day? (Estimate: −66.00) and How many servings of fruit did you eat each day? (Estimate: −26.35). This indicates that very few study subjects were eating the daily recommended servings of vegetables and fruits respectively. Conversely, the item with the highest estimate for the location parameter is: During the past year has your drug or alcohol use led to any problems with the legal system such as drunk and disorderly arrests, being pick-up for drug possession, etc.? The estimate for this item is 3.322 indicating subjects who had experienced legal problems as a result of their substance use had significantly different substance use patterns than people who did not endorse this item.

4. Discussion

IRT methods have important applications in health outcome measurements. For the most part, statisticians are still using traditional methods including factor analysis, principal component analysis, discriminant analysis together with Cronbach's alpha to build test questionnaires, identify highly discriminating items and to evaluate the internal validity of test domains. In this paper, we have illustrated a method which has already found wide-spread applications among researchers in education, behavioral health and psychometrics as an alternative to commonly used multivariate methods. With the availability of a new procedure in SAS (version 9.4) to conduct IRT analysis along with multiple open source software, statisticians involved in health outcome measurement research can benefit from the use of IRT method.

References

[1] An, X.M. and Yung, Y.F. (2014) Item Response Theory: What It Is and How You Can Use the IRT Procedure to Apply It. SAS Institute, North Carolina.

[2] Sheu, C., Chen, C., Su, Y.H. and Wang W. (2005) Using SAS PROC NLMIXED to Fit Item Response Theory Models. *Behavior Research Methods*, **37**, 202-218. https://doi.org/10.3758/BF03192688

[3] Harris, D. (1989) Comparison of 1-, 2-, and 3-Parameter IRT Models. *Educational Measurement*: *Issues and Practice*, **8**, 35-41. https://doi.org/10.1111/j.1745-3992.1989.tb00313.x

[4] Chang, C.H. and Reeve, B.B. (2005) Item Response Theory and Its Applications to Patient-Reported Outcomes Measurement. *Evaluation & the Health Professions*, **28**, 264-282. https://doi.org/10.1177/0163278705278275

[5] Hambleton, R.K. (2000) Emergence of Item Response Modeling in Instrument Development and Data Analysis. *Medical Care*, **39**, S60-S65.

[6] Drasgow, F. and Hulin, C.L. (1990) Item Response Theory. *Handbook of Industrial and Organizational Psychology*, **1**, 577-636.

[7] Embretson, S.E. and Reise, S.P. (2013) Item Response Theory. Psychology Press, NY.

[8] Reeve, B.B., Hays, R.D., Bjorner, J.B., Cook, K.F., Crane, P.K., Teresi, J.A., *et al.* (2007) PROMIS Cooperative Group. *Medical Care*, **45**, S22-S33. https://doi.org/10.1097/01.mlr.0000250483.85507.04

[9] Fries, J.F., Bruce, B. and Cella, D. (2005) The Promise of PROMIS: Using Item Response Theory to Improve Assessment of Patient-Reported Outcomes. *Clinical and Experimental Rheumatology*, **23**, S53-S57.

[10] Hartman, C.A., Gelhorn, H., Crowley, T.J., Sakai, J.T., Stallings, M., Young, S.E. and Hopfer, C.J. (2008) Item Response Theory Analysis of DSM-IV Cannabis Abuse and Dependence Criteria in Adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry*, **47**, 165-173. https://doi.org/10.1097/chi.0b013e31815cd9f2

[11] Grimby, G., Andrén, E., Holmgren, E., Wright, B., Linacre, J.M. and Sundh, V. (1996) Structure of a Combination of Functional Independence Measure and Instrumental Activity Measure Items in Community-Living Persons: A Study of Individuals with Cerebral Palsy and Spina Bifida. *Archives of Physical Medicine and Rehabilitation*, **77**, 90131-90138. https://doi.org/10.1016/S0003-9993(96)90131-8

[12] Hays, R.D., Morales, L.S. and Reise, S.P. (2000) Item Response Theory and Health Outcomes Measurement in the 21st Century. *Medical Care*, **39**, II28-II42.

[13] (2013) A Simple Guide to the Item Response Theory (IRT) and Rasch Modeling. http://www.creative-wisdom.com/computer/sas/IRT.pdf

[14] DeMars, C. (2010) Item Response Theory. Oxford University Press, NY. https://doi.org/10.1093/acprof:oso/9780195377033.001.0001

[15] Testa, M.A. (2000) Interpretation of Quality-of-Life Outcomes: Issues that Affect Magnitude and Meaning. *Medical Care*, **38**, II166-II174. https://doi.org/10.1097/00005650-200009002-00026

[16] Baker, F.B. and Kim, S. (2004) Item Response Theory: Parameter Estimation Techniques. 2nd Edition, CRC Press, New York.

[17] Weiner, I.B. and Graham, J.R. (2003) Handbook of Psychology, Assessment Psychology. John Wiley & Sons, New York.

[18] Knowles, E.S. and Condon, C.A. (2000) Does the Rose Still Smell as Sweet? Item Variability across Test Forms and Revisions. *Psychological Assessment*, **12**, 245-252. https://doi.org/10.1037/1040-3590.12.3.245

[19] Yang, F.M. and Kao, S.T. (2014) Item Response Theory for Measurement Validity. *Shanghai Archives of Psychiatry*, **26**, 171-177.

[20] De Ayala, R.J. (1993) Review of Review of Fundamentals of Item Response Theory. *Journal of Educational Measurement*, **30**, 84-87.

[21] Walters, S.T., Spence-Almaguer, E., Hill, W. and Abraham, S. (2015) Health Coaching and Technology with Vulnerable Clients. *Social Work Today*, **15**, 6.

[22] Dietary Screener Questionnaires (DSQ) in the NHANES 2009-10. https://epi.grants.cancer.gov/nhanes/dietscreen/questionnaires.html

[23] Eakman, A.M., Carlson, M.E. and Clark, F.A. (2010) The Meaningful Activity Participation Assessment: A Measure of Engagement in Personally Valued Activities. *The International Journal of Aging and Human Development*, **70**, 299-317. https://doi.org/10.2190/AG.70.4.b

[24] Heitzmann, C.A. and Kaplan, R.M. (1988) Assessment of Methods for Measuring Social Support. *Health Psychology*, **7**, 75-109. https://doi.org/10.1037/0278-6133.7.1.75

[25] Mclellan, A.T., Kushner, H., Metzger, D., Peters, R., Smith, I., Grissom, G. and Argeriou, M. (1992) Addiction Severity Index. 5th Edition, 199-213.

[26] Endicott, J., Nee, J., Harrison, W. and Blumenthal, R. (1993) Quality of Life Enjoyment and Satisfaction Questionnaire: A New Measure. *Psychopharmacology Bulletin*, **29**, 321-326.

[27] Kiluk, B.D., Dreifuss, J.A., Weiss, R.D., Morgenstern, J. and Carroll, K.M. (2013) The Short Inventory of Problems-Revised (SIP-R): Psychometric Properties within a Large, Diverse Sample of Substance Use Disorder Treatment Seekers. *Psychology of Addictive Behaviors*, **27**, 307-314. https://doi.org/10.1037/0278-6133.7.1.75

[28] Kroenke, K., Spitzer, R.L. and Williams, J.B. (2001) The PHQ-9: Validity of a Brief Depression Severity Measure. *Journal of General Internal Medicine*, **16**, 606-613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

[29] Morisky, D.E., Green, L.W. and Levine, D.M. (1986) Concurrent and Predictive Validity of a Self-Reported Measure of Medication Adherence. *Medical Care*, **24**, 67-74. https://doi.org/10.1097/00005650-198601000-00007