Original Research Article

SAGE Open Medicine

Use of item response theory to investigate disability-related questions in the National Health and Nutrition Examination Survey

SAGE Open Medicine
Volume 9: 1–11

DOI: 10.1177/20503121211012253
journals.sagepub.com/home/smo
SAGE

Samuel W Terman[1,2] and James F Burke[1,2]

Abstract

Objectives: Accurately measuring disability is critical toward policy development, economic analyses, and determining individual-level effects of health interventions. Nationally representative population surveys such as the National Health and Nutrition Examination Survey provide key opportunities to measure disability constructs such as activity limitations. However, only very limited work has previously evaluated the item response properties of questions pertaining to limitations in National Health and Nutrition Examination Survey.

Methods: This was a cross-sectional study. We included participants ⩾20 years old for the 2013–2018 National Health and Nutrition Examination Survey cycles. Activity limitations, and a smaller number of body function impairments or participation restrictions, were determined from interview questions. We performed item response theory models (a two-parameter logistic and a graded response model) to characterize discriminating information along the latent continuum of activity limitation.

Results: We included 17,057 participants. Although each particular limitation was somewhat rare (maximally 13%), 7214 (38%) responded having at least one limitation. We found a high amount of discriminating information at 1–2 standard deviations above average limitation, though essentially zero discrimination below that range. Items had substantial overlap in the range at which they provided information distinguishing individuals. The ordinal graded response model including 20 limitations provided greater information than the dichotomous two-parameter logistic model, though further omitting items from the graded response model led to loss of information.

Conclusion: National Health and Nutrition Examination Survey disability-related questions, mostly specifically activity limitations, provided a high degree of information distinguishing individuals with higher than average limitations on the latent continuum, but essentially zero resolution to distinguish individuals with low or average limitations. Future work may focus on developing items which better distinguish individuals at the "lower" end of the limitation spectrum.

Keywords

Epidemiology, public health

Date received: 4 December 2020; accepted: 30 March 2021

Introduction

Approximately 15% of the world's population has any sort of disability,[1] and rates of disability are only increasing due to an aging population. Accordingly, accurately measuring disability is critical toward policy development, economic analyses, and determining individual-level effects of health interventions. The International Classification of Functioning, Disability, and Health (ICF) framework of the World Health Organization (WHO)[2,3] provides the international standard for conceptualizing and measuring disability. In this model, disability is an umbrella term referring to impairments of body functions or structures, activity limitations, and participation restrictions. The ICF emphasizes an integrated biopsychosocial model of human

[1]Department of Neurology, University of Michigan, Ann Arbor, MI, USA
[2]Institute for Healthcare Policy & Innovation, University of Michigan, Ann Arbor, MI, USA

Corresponding author:
Samuel W Terman, Department of Neurology, University of Michigan, Taubman 1st Floor, Reception C, 1500 E. Medical Center Drive, SPC 5316, Ann Arbor, MI 48109, USA.
Email: sterman@umich.edu

functioning that acknowledges the interaction of medical conditions and environmental context.

Nationally representative epidemiological monitoring surveys provide one key opportunity for operationalizing the ICF. The National Health and Nutrition Examination Survey (NHANES) provides one relevant example. NHANES is a nationally representative face-to-face household survey that is carried out every 2 years, measuring a wide range of variables[4] that map onto the ICF framework. These include body functions like mental and sensory functions (i.e. hearing and seeing) and participation in life (i.e. participating in social or leisure activities). However, the greatest number of disability-related questions relates to activity limitations (i.e. walking, lifting, sitting, grasping, etc.). Given the rich array of available data, researchers have used these data extensiv ely.[4–17] For example, early work in NHANES documented the prevalence of a wide variety of activity limitations (i.e. limitations in walking, stopping/crouching, lifting, getting up from a chair) stratified by age, race, and sex.[4] Subsequent work turned attention to assessing associations between particular nutritional, medical, or social factors (i.e. dietary inflammation,[7] socioeconomic status,[13] sleep disorders,[14] hypertension,[16] epilepsy,[18] or mortality risk[8]) and disability (defined as limitations in activities of daily living, instrumental activities of daily living, lower extremity mobility, or participating in social activities).

Despite researchers using NHANES extensively, the properties of individual questions used for disability assessment within this dataset remain incompletely explored. Item response theory (IRT) was initially developed in the field of education to help create tests comprised of a limited collection of individual questions with a range of difficulty to distinguish students across abilities.[19] However, it applies more broadly whenever an investigator begins with a unidimensional concept in mind and a large pool of items which could distinguish individuals along a unidimensional latent continuum. Note that while disability is a multidimensional concept, the items in this study predominantly assessing activity limitations were considered adequately unidimensional. Then, IRT analysis occurs on the larger item pool to select those items that distinguish ability over the relevant range of substantive importance. Advantages of IRT include explicitly modeling the degree to which each particular item and also an entire collection of items can or cannot distinguish individuals across the range of a given latent trait, allowing deeper understanding of each item by calculating a difficulty and discrimination parameter for each item regarding the position (difficulty) and steepness (discrimination) along the latent trait at which an item distinguishes individuals (though different IRT models can capture more or less complexity than this), explicitly modeling measurement error for observed versus expected values, and assessing the degree to which items might be measuring abilities differently between subgroups. Particularly, IRT helps test designers decide on the items to include in a scale and also informs where items should be added if existing items do not adequately cover the range of substantive importance. Therefore, IRT provides a unique opportunity to advance the science of disability measurement.[20–23]

In this study, we applied IRT to all adult NHANES participants in the three contiguous most recent waves of available data (2013–2018). We included items focusing on activity limitations plus several additional available items dealing with body functions (seeing, hearing, mentation) and participation (social, leisure participation). We assessed unidimensionality, and then item response properties of these self-rated items to assess where along the latent continuum of limitations do these items best distinguish individuals. Although we did include several items referring to body functions and participation, for the most part our study evaluated activity limitations; thus, our terminology will mostly refer to each item as a "limitation" for consistency within the ICF framework.

Methods

Study design and dataset

This was a cross-sectional analysis of the ongoing, landmark NHANES using data collected from 2013 to 2018. NHANES is a long-standing semi-annual cross-sectional study run by the Centers for Disease Control and Prevention. Its goal is to understand broad trends in health and nutrition in the United States. NHANES samples approximately 5000 to 10,000 non-institutionalized individuals from 15 counties across the United States each 2-year cycle. Thus, three cycles were more than adequate for our study In order to be nationally representative, NHANES uses complex, stratified, multistage probability cluster sampling, in addition to oversampling certain individuals (over 60 years old, African Americans, Hispanics) selected from the US Census. The design and operation of NHANES are available online (https://wwwn.cdc.gov/nchs/nhanes/default.aspx).

Procedures involving human subjects

Given use of publicly available de-identified datasets, this study was deemed exempt by the University of Michigan Institutional Review Board.

Patient selection

Because NHANES collects limitations data only for those at least 20 years old, our study included all NHANES participants at least 20 years old.

Variables

All questionnaires used for data collection are available online.[24] Data collection procedures undergo extensive

quality control monitoring, interviewers are accompanied by field supervisors to attest to the protocol, at least 10% of every interviewer's work is randomly selected and validated by field supervisors by phone confirmation or repeat visit, and interview recordings are reviewed by supervisors for validation of responses.[25]

Baseline variables to describe our population were selected to map onto domains within the ICF framework, in addition to selecting variables that prior literature has suggested predict disability (i.e. neurological diseases, multimorbidity, socioeconomic status, and depression).[26] Demographics included age, sex, race, and income-to-poverty ratio (a family's income as a ratio of poverty guidelines). We recorded the following medical conditions: asthma, chronic obstructive pulmonary disease, congestive heart failure, coronary disease, hypertension, diabetes, epilepsy, liver disease, stroke, thyroid disease, and malignancy. We calculated the number of comorbidities as the sum of each of these individual conditions. Most conditions were determined by whether participants reported that a healthcare professional had previously diagnosed a particular condition, though see the footnote in Table 1 for exceptions. Other variables included insurance coverage, Patient Health Questionnaire–9 (PHQ9) to evaluate depression severity, body mass index (kg/m^2), a measure of physical activity (whether participants reported at least 10 min of moderate to vigorous activity each week), self-reported health status (rated as excellent, very good, good, fair, or poor), number of prescription medications, and current smoking status.

NHANES asks participants a wide variety of questions regarding body functions, activity limitations, and participation restrictions.[24] Supplemental Table 1 lists the items in detail that we used in this study and their corresponding ICF category. Items included a small number of mental or sensory functions (i.e. concentrating, memory, hearing, seeing), a large list of activity limitations (i.e. walking, grasping, pulling, finances, meals), and a smaller list of participation restrictions (i.e. social or leisure activities). Some variables allowed only binary "yes/no" responses (concentrating, hearing, seeing, working, errands, dressing/bathing). All other variables were rated on an ordinal scale ranging from "no difficulty" to "some difficulty" to "much difficulty" to "unable to do." In the two-parameter logistic (2PL) IRT model described below, ordinal variables were dichotomized according to "no difficulty" versus "at least some difficulty."

Table 1. Population description.

	Mean (SD; *N*) or Raw no./No. non-missing (weighted %)
Age, per decade	47.9 (17.2; 17,057)
Male sex	8207/17,057 (48%)
Race	
Mexican American	2497/17,057 (9%)
Non-Hispanic Black	3673/17,057 (11%)
Non-Hispanic White	6270/17,057 (64%)
Comorbidities	
Number of chronic conditions	1.4 (1.5; 17,057)
Asthma	2556/17,043 (15%)
Cancer	1684/17,050 (11%)
Coronary disease	741/16,997 (4%)
Congestive heart failure	597/17,021 (2%)
Diabetes mellitus[a]	2969/17,057 (13%)
Hypertension[a]	7851/17,057 (41%)
Liver disease	787/17,020 (4%)
PHQ9[b]	3.2 (4.3; 14,848)
Stroke	648/17,037 (3%)
Thyroid disease	1878/17,020 (12%)
Self-rated health	
Excellent	1282/14,979 (10%)
Very good	3735/14,979 (31%)
Good	6216/14,979 (41%)
Fair	3208/14,979 (16%)
Poor	538/14,979 (3%)
Current smoker	3268/17,044 (19%)
Body mass index (mg/kg^2)	29.5 (7.1; 16,101)
Number of prescription medications	2.1 (2.9; 17,057)

SD: standard deviation; PHQ9: Patient Health Questionnaire–9; NHANES: National Health and Nutrition Examination Survey; SBP: systolic blood pressure; DBP: diastolic blood pressure.

[a]Comorbidities were assessed by self-report that a physician has previously diagnosed the patient with each condition. However, for the definition of hypertension and diabetes, we also accepted if participants reported at least one medication treating these conditions, or else NHANES measurements suggested the diagnosis (hypertension: of up to three readings, average SBP $\geq$ 140 mm Hg ($\geq$130 mm Hg if diabetes) or average DBP $\geq$ 90 mm Hg ($\geq$80 mm Hg if diabetes); diabetes: A1c $\geq$ 6.5%)[27] as has been done in prior NHANES studies.[28]

[b]Common interpretation thresholds for depression include 0–4 minimal, 5–9 mild, 10–14 moderate, 15–19 moderately severe, 20–27 severe.[29]

Statistical analysis

To describe our baseline population, we used mean and standard deviation (SD) for continuous variables and number (%) for categorical variables. All analyses were survey-weighted as required by NHANES's complex sampling frame, which allows for extrapolation to the US population as a whole. The weights provided in each biennial cycle's dataset were divided by 3 (the number of included interview cycles: (1) 2013–2014, (2) 2015–2016, (3) 2017–2018), as recommended by NHANES's analytical documentation.[30]

Then, we described the items listed in Supplemental Table 1. We depicted the distribution of each variable as a horizontal 100% stacked bar graph. We reported the frequency of participants reporting at least one limitation. We then described the summed number of limitations (whether binary responses were "yes," or ordinal responses were at least "some") via a histogram.

IRT modeling involves several assumptions. For example, models assume monotonicity (the probability of responding "yes" to a given question increases as the person's latent

trait increases). This is reasonable to assume given increase in the latent trait would only be expected to increase the chance of any particular activity limitation; it would be implausible to consider that an increase in the latent trait would initially increase, then later decrease chance of any particular activity limitation. IRT also assumes unidimensionality, which means that all items contribute to a single underlying latent trait. We tested for unidimensionality of non-binary items through exploratory factor analysis. As employed by others,[23] adequate unidimensionality is demonstrated in IRT analysis[21] if the ratio of the eigenvalues between the first and the second factor is $\geqslant 4$[31] and if factor loadings on the dominant factor are $\geqslant 0.40$.

We performed a 2PL IRT model.[32] We included all 25 items listed in Supplemental Table 1. Because a 2PL model requires dichotomous items, we used all binary items as is, plus we dichotomized ordinal items ("no difficulty" versus "at least some difficulty"). In a 2PL IRT model, the probability of person *j* providing a positive answer to item *i* is calculated by

$$\Pr\left(Y_{ij} = 1 \mid \theta_j\right) = \frac{\exp\left\{a_i\left(\mid \theta_j - b_i\right)\right\}}{1+\exp\left\{a_i\left(\mid \theta_j - b_i\right)\right\}} \Theta_j \sim N(0,1)$$

In the above formula, θ is the estimated latent trait (disability). This is understood to be an individual's relative value of the latent trait compared to the population mean; here, we use a collection of items measuring mainly activity limitations to quantify the latent trait of disability. It is a standardized value, such that a value of 0 corresponds to the population's mean, and values away from 0 refer to the number of SDs above or below the mean a given person falls. The first parameter (α_i) represents the discrimination of item *i*. Discrimination refers to how quickly the probability of responding "yes" rises with an increase in the latent trait; hence, higher discrimination would more sharply distinguish individuals with relatively similar latent traits. Note though that while each item only has a single discrimination parameter, each item has a different discrimination at different points along its distribution, and items are most discriminating at their difficulty parameter where the item characteristic curve is steepest (see next paragraph for the description of graphical methods). The second parameter (b_i) represents the "difficulty" of item *i*. In this sense, "difficulty" is a mathematical term, so while NHANES does phrase items as how much "difficulty" an individual has with a particular task, in this article we reserve "difficulty" for referring to the IRT concept whereas we refer to trouble performing tasks as "limitations" to avoid confusion. The difficulty parameter represents the left–right location along the latent continuum—specifically, the value of θ at which 50% of individuals respond "yes." A high difficulty parameter means that only individuals with a high severity of the underlying latent trait will respond "yes." In other words, discrimination describes the shape or steepness of the information curve that is maximal at the item's difficulty, whereas the difficulty parameter itself describes the left–right location of the curve where such distinguishing information exists. Ultimately, the model estimates each individual's underlying latent variable θ_j (limitations) through the above parametric model, and predicts the probability that each individual item *i* is positive given that estimated θ_j. We displayed difficulty and discrimination of each item, sorted by difficulty, to demonstrate the ranges of the latent variable and the degree to which each question separates individuals with higher versus lower limitation.

We then displayed graphical item information.[32] First, we included the Test Information Function. This is an overall picture of where items enable discrimination along the latent continuum of activity limitation. Regions with high test information and low standard error represent regions of the latent continuum at which the items collectively excel at distinguishing individuals, whereas items poorly distinguish individuals in regions with low information and high standard error. Test Information is akin to the classical test theory concept of reliability. We then showed item characteristic curves for each item, which shows the probability of a "yes" response to each item according to the value of the latent trait, where items shifted right have higher difficulty, and steeper items have higher discrimination. We then displayed Item Information Functions, which are akin to the Test Information Function, except for each individual item. We also repeated the Test Information Function, except dropping certain subsets of items (i.e. either items that were not purely activity limitations, or else every other item in terms of difficulty, or else the least discriminating items), to determine how much information was lost when paring down the items. In order to assess model fit, we displayed test characteristic curves with observed points versus the predicted (expected) curve based on each model. The test characteristic curve (also known as the total characteristic curve) is the sum of all item characteristic curves for a model and thus plots the expected latent trait along the entire continuum for a given model. Observed points are obtained by plotting the predicted latent trait ("*x*") versus the total number of "yes" responses.

Next, in order to retain the more detailed ordering and potentially added power of ordinal responses, albeit at the expense of removing the strictly binary items (generally, body function items), we performed a graded response model including only the 20 ordinal items noted in Supplemental Table 1. Note while the 2PL model included the dichotomous variable "bathing/dressing," the graded response model included the ordinal variable "dressing," as no such variable existed for just bathing. We displayed similar information as we did in the 2PL model. Note in the 2PL model where there are only two possible choices per item (yes/no), there was only one boundary of interest, and thus, each item could be represented by a single item characteristic curve. In

contrast, our ordinal items include four possible Likert-type choices, and thus must be represented by numerous category characteristic curves per item. Estimation of a graded response model is similar to the 2PL model, with the addition of subscripting each difficulty parameter by particular ordinal response.

Data were analyzed using SAS 9.4 (Cary, NC, USA) and Stata 14.2 (College Station, TX, USA).

Data accessibility statement

All datasets are freely available for download at https://wwwn.cdc.gov/nchs/nhanes/default.aspx.

Results

Patient population

Our study included 17,057 participants. Table 1 depicts participant characteristics; 48% were male, 64% were non-Hispanic White, and 19% were current smokers. They had a mean 1.4 (SD = 1.5) chronic conditions.

Item distributions

Figure 1 depicts the distribution of each item. For each binary item (Figure 1(a)), between 4% and 13% of participants endorsed a given limitation. For each ordinal item (Figure 1(b)), between 0.5% and 10% of participants endorsed at least some limitation. While no single limitation was particularly common, 7214/17,057 (38%) responded having at least one of these limitations. When we summed the number of binary limitations plus ordinal limitations with at least "some difficulty," the median number of limitations was 0 (interquartile range = 0–3) and the mean number of limitations was 2.5 (SD = 4.5). The right-skewed distribution is depicted in the histogram in Figure 1(c).

IRT models

Table 2 displays results from exploratory factor analysis. Factors 1 and 2 demonstrated eigenvalues of 7.3 and 0.9, respectively (ratio = 7.9). All variables loaded ⩾0.4 onto Factor 1 or quite close (minimum = 0.39). These findings supported IRT's unidimensionality assumption.

Figures 2 and 3 depict results from the 2PL (Figure 2) and graded response (Figure 3) models. These models included 17,054 out of 17,057 participants.

Figure 2(a) demonstrates the Test Information Function for the 2PL model. This graph demonstrates a high amount of information discriminating participants whose position on the latent construct continuum is within 1–2 SDs above the mean (i.e. this region had high Test Information and low standard errors), but very limited ability to distinguish individuals near or below average number of limitations. Item characteristic curves (Figure 2(b)) and Item Information

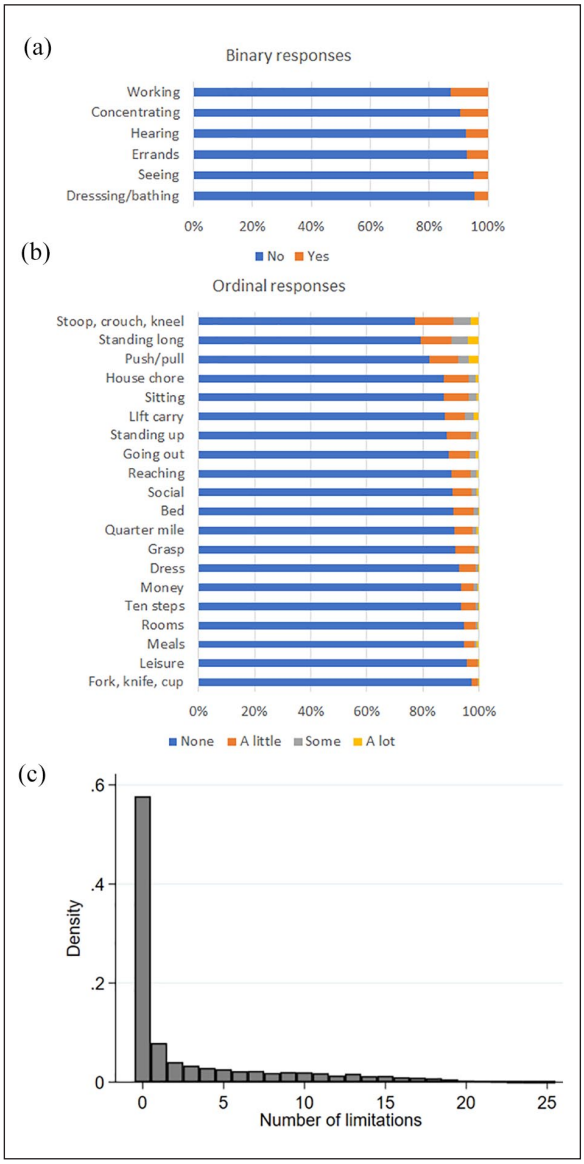

Figure 1. Distribution of limitation variables. (a) 100% stacked horizontal bar chart of binary variables, sorted by descending order of frequency. Note that Supplemental Table 1 contains a fuller description of each item. Each horizontal bar represents the percentage that responded yes versus no to each limitation, (b) 100% stacked horizontal bar chart of ordinal variables, also sorted by descending order of frequency. Each horizontal bar represents the percentage that responded with each of the four possible options for each limitation (none, a little, some, or a lot) and (c) histogram of summed total number of limitations.

Functions (Figure 2(c)) expand upon this observation—the probability of any given item being "yes" was nearly 0 until participants reached approximately 0–1 SDs above the mean, after which point curves rose rapidly and nearly all individuals responded positively to most questions after exceeding approximately 2 SDs above the mean. While no item contributed to distinguishing between individuals at or below average limitations, several items contributed a relatively small

Table 2. Factor analysis to assess for unidimensionality.

Eigenvalue	Factor 1	Factor 2
	7.33 (88% proportion of variance)	0.93 (11% proportion of variance)
Item[a]		
Standing long	0.75	–0.28
Chores	0.74	0.02
Pushing, pulling	0.73	–0.21
Going out	0.70	0.23
Stooping, crouching	0.69	–0.28
Sitting long	0.67	–0.09
Lifting, carrying	0.67	–0.14
Quarter mile	0.67	–0.29
Standing up	0.63	–0.07
Social	0.62	0.30
Bed	0.61	0.03
Ten steps	0.61	–0.23
Reaching	0.58	–0.03
Dressing	0.57	0.19
Meals	0.51	0.35
Grasping	0.51	0.11
Leisure	0.43	0.30
Rooms	0.43	0.12
Money	0.39	0.28
Fork, knife, cup	0.39	0.26

[a]Supplemental Table 1 contains a full description of the limitation items listed here, which contain only brief labels in this table.

amount of information to distinguish individuals at particularly high degrees of limitations (smaller flatter curves in the bottom right of Figure 2(c), that is, high difficulty but less steep ability to differentiate participants who are near each other on the latent continuum). Figure 2(d) lists the difficult parameters in descending order along with their discrimination parameters. For example, items with high difficulty (i.e. items that only those with the greatest latent trait endorsed) and low discrimination (i.e. low ability to distinguish individuals) from Figure 2(c) are made clear in Figure 2(d) such as seeing, hearing, fork/knife/cup, leisure, and money. Items with the lowest difficulty (i.e. items that even those with a lower latent trait may endorse) included stooping/kneeling/crouching, standing for long periods, and push/pulling heavy objects. Figure 2(e) shows the test characteristic curve for the 2PL model. This shows good calibration between observed versus expected datapoints.

We then evaluated peak Test Information when we dropped subsets of items from the 2PL model. When we included only the 19 purely activity limitations (i.e. dropping the six items labeled in Supplemental Table 1 as body function impairments or participation restrictions), the Test Information Function appeared similar to Figure 2(a), except peak information was approximately 85 (instead of approximately 100). When we ordered items by difficulty and included only every other item (13 out of 25 items), the Test Information Function appeared very similar to Figure 2(a), except peak information was approximately 50. When we ordered items by discrimination and included only those half (13 out of 25 items) with the highest discrimination parameters, the Test Information Function again appeared similar to Figure 2(a), except peak information was approximately 90.

Figure 3 depicts results from the graded response model. This included our 20 ordinal items. Again, items contained the most robust amount of information around 1–2 SDs above the mean (Figure 3(a)). Note, though, that Test Information Function peaks at approximately 150 in the graded response model, whereas it peaks at approximately 100 in the 2PL model (Figure 2(a)). As per Figure 3(b), participants tended to respond with "no difficulty" until they reached approximately 0–1 SDs above the mean (the upper left curves that start at 1 and drop to 0); there was a relatively wide spread of latent limitations at which individuals responded "little" or "some" limitation between 1 and 3 SDs above the mean, and participants tended to respond "a lot" of limitation to most or all questions after exceeding 2–3 SDs above the mean (the upper right curves that start at 0 and increase to 1). Figure 3(c) depicts just one such example extracted from Figure 3(b) for illustration—difficulty transferring between rooms. Participants almost universally responded "none" until reaching at least 1 SD above the mean latent limitations, then there was a relatively narrow range in which respondents answered "little" or "some" between 1.5 and 2.5 SDs above the mean, then almost universally responded "a lot" above 2.5 SDs above the mean. Again, there was substantial overlap in the range of latent limitations tapped by most items. Figure 3(d) illustrates Item Information Functions for each item. Results were similar to the 2PL model (Figure 2(c)), for example, that no item distinguished individuals below average limitations. Similar to the 2PL model, managing money, participation in leisure activities, and using a fork/knife/cup had the highest difficulty parameters (Figure 3(e), which demonstrates the difficulty parameter for the most severe response ("a lot")). On the other end of the spectrum, participants were more likely to respond "a lot" at lower ranges of latent limitations (lower "difficulty") to standing for long periods, push/pulling heavy objects, and stooping/kneeling/crouching. Figure 3(f) shows the Test Characteristic Curve for the graded response model, which shows good fit between expected versus observed datapoints.

We then evaluated peak Test Information when we dropped subsets of items from the graded response model. When we included only the 18 purely activity limitations (i.e. dropping going out and social given these are participation restrictions), the Test Information Function appeared similar to Figure 3(a), except peak information was approximately 130 (instead of 150). When we ordered items by difficulty and included only every other item (10 out of 20 items), the Test Information Function appeared very similar

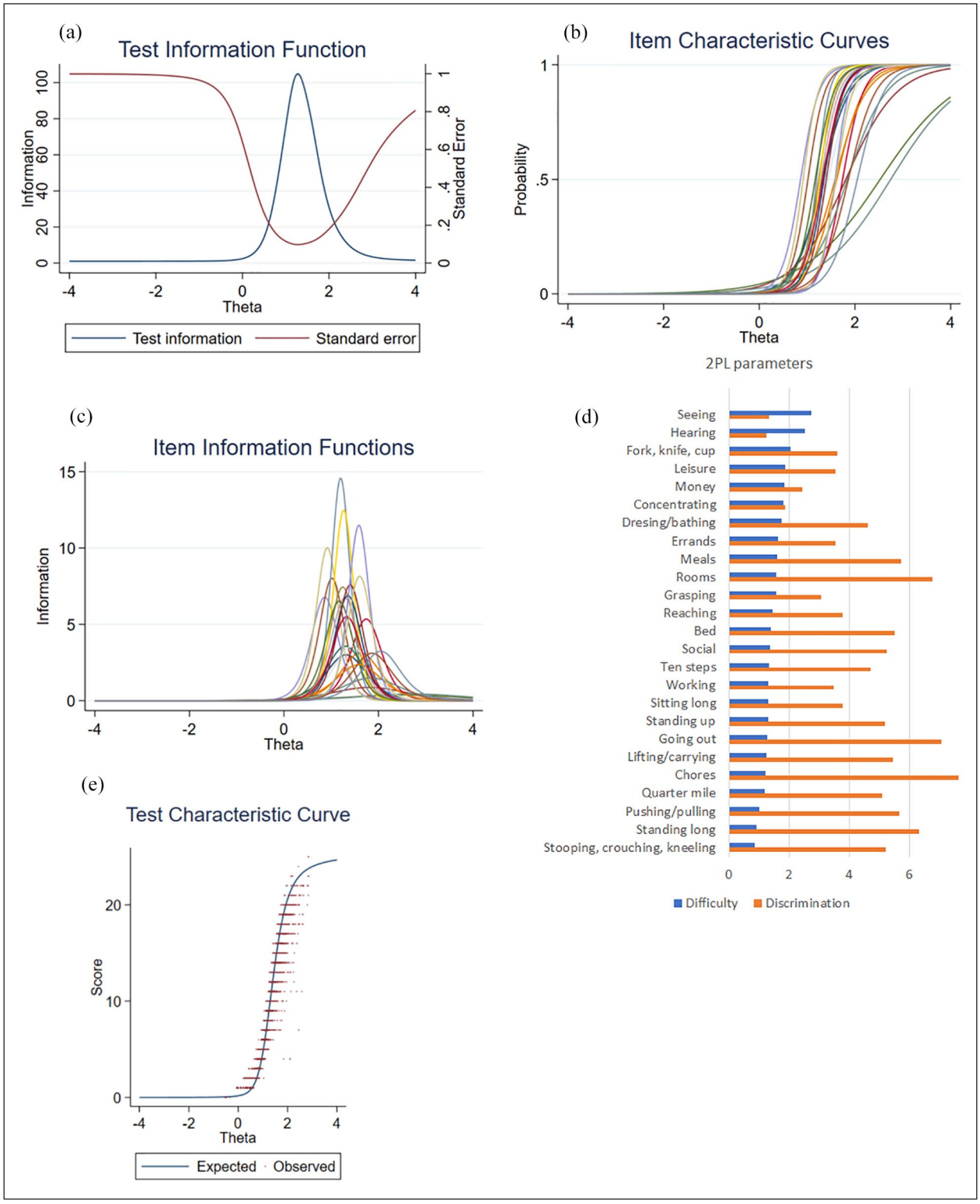

Figure 2. 2PL model graphical results. (a) Test Information Function depicts total amount of information about disability across all items (blue line), which is inverse to the standard error around Test Information (red line). (b) Item characteristic curves for all items depict the probability of responding "yes" to each item across the range of the latent variable θ. (c) Item information function represents the amount of information for each individual item across the range of the latent variable. (d) 2PL model parameters. This bar graph is sorted from highest to lowest difficulty parameters (the value of the latent trait above which 50% of participants indicated a given limitation). Thus, a higher difficulty parameter refers to an item where only those with the greatest degree of the latent trait endorse a given limitation. Discrimination parameters are also displayed. (e) Test characteristic curve demonstrating observed versus predicted calibration.

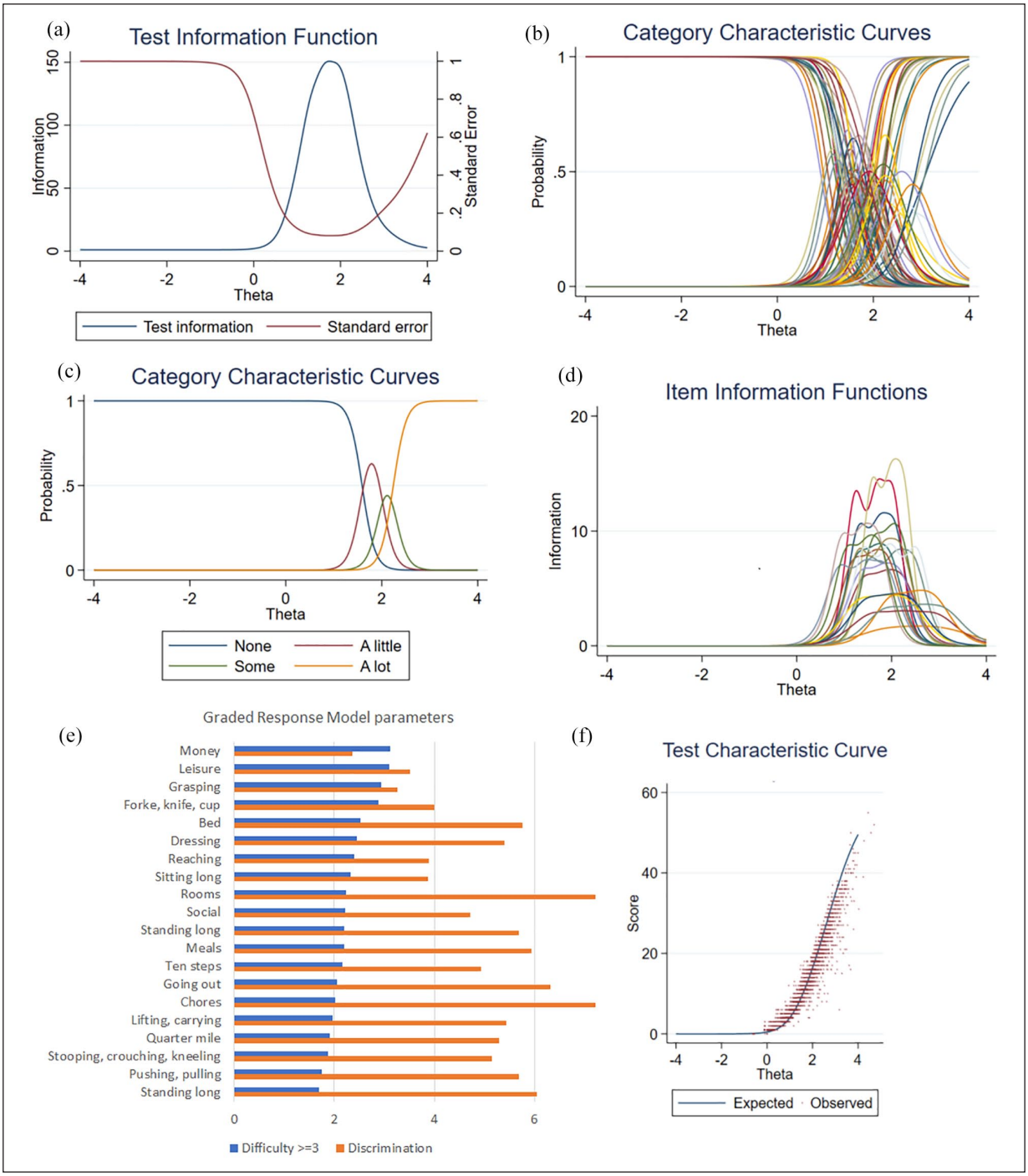

Figure 3. Graded response model. (a) Test Information Function depicts total amount of information about disability across all items (blue line), which is inverse to the standard error around Test Information (red line). (b) All superimposed category characteristic curves. Whereas in the above 2PL model there was just a single item characteristic curve per item representing a response of "yes" (Figure 2(b)), in the graded response model (Figure 3(b)) there are four curves for each item representing each of the possible responses ("none," "a little," "some," "a lot"). Each curve represents the probability of responding with a particular response for each particular item across the range of the latent variable. Hence, curves to the left that start at probability = 1 represent choices of "none" given it is highly likely for a respondent to answer "none" to all questions if they have very low disability. In contrast, the curves to the right that end at probability = 1 represent choices of "a lot" given it is highly likely for a respondent to answer "a lot" to all questions if they have very high disability. (c) Example boundary characteristic curve, a single item (here, walking between rooms), (d) Item Information Functions for each item represent the amount of information provided by each item across the range of the latent trait. (e) Graded response model parameters. This bar graph is sorted from highest to lowest difficulty of responding with "a lot"—for simplicity of data display, other less severe categories are omitted. Discrimination parameters are also displayed. (f) Test characteristic curve demonstrating observed versus predicted calibration.

to Figure 3(a), except peak information was approximately 70. When we ordered items by discrimination and included only those half (10 out of 20 items) with the highest discrimination parameters, the Test Information Function again appeared similar to Figure 3(a), except peak information was approximately 80.

Discussion

In a large nationally representative sample, we applied IRT to evaluate characteristics of individual items (measuring activity limitations, in addition to several body impairments and activity restrictions) and the collective ability of a set of these items to distinguish individuals along the latent continuum of disability. Although any single item was somewhat rare, 38% of the population endorsed at least one of the listed disabilities. We found that in total, the measured items provided high ability to distinguish individuals who were 1–2 SDs above average on the latent trait of activity limitation, and that ordinal items (even though fewer items in total) provided superior test information compared with including a larger array of binary responses. Typically, if individuals endorsed limitations, they endorsed physical activity limitations, whereas only those with the highest degree of limitations endorsed sensory deficits (i.e. seeing, hearing) or participation restrictions (i.e. leisure). A limited number of particular items (e.g. managing money, using a fork/knife/cup, or participant in leisure activities) provided moderate information distinguishing individuals 2–3 SDs above the mean. However, the items provided essentially zero information toward distinguishing individuals with below average limitations and very little information among individuals 0–1 SDs above the mean. We also evaluated whether certain items could be dropped without sacrificing Test Information. We showed that certain dichotomous items could be dropped (i.e. body functions and participation restrictions, or else those least discriminating items) without sacrificing much information, that the ordinal items generally provided greater information than the dichotomous items as expected, but that dropping either items with neighboring difficulty or even the least discriminating ordinal items did in fact meaningfully sacrifice information.

Surely an outcome measure should distinguish individual at higher versus medium disability. However, depending on the research question, distinguishing individuals at low versus medium disability would likely be equally useful at the both the individual and population levels. Lower levels of disability in a large portion of the population may still have a large public health or economic impact, and detecting lower levels of disability could be important for early detection of mild disease when preventive interventions exist. In NHANES, given most individuals responded "no" or "none" to most questions, this collection of items provided essentially zero ability to distinguish an individual with low versus average limitations who all provided nearly identical responses. If even subtle differences in disability are of interest for a given research question, then items would need to be added in future questionnaires with lower "difficulty," that is, questions to which an individual with even subtle limitations would respond "yes."

Prior studies of disability using IRT have been completed in other national datasets such as the Health and Retirement Study,[22] national data outside of the United States,[33] and the National Health and Aging Trends Study (NHATS).[23] For example, one large study in Australia performed Rasch analysis to demonstrate the usefulness of various questions surrounding "support needs" (with self-care, mobility, communication, etc.) for measuring "extent" of disability.[33] A major strength of our study in comparison, though, is that while Rasch modeling by definition involves the strong assumption that all discrimination parameters are constrained to 1, our 2PL and graded response models were more flexible in terms of empirically calculating a unique discrimination parameter for each item for better model fit. In another example study relevant to our own, investigators used IRT to compare discrimination across the disability latent continuum for previously validated self-reported items[34] (akin to those in our study, where participants rate their perceived level of difficulty performing tasks) versus performance-based items (not available in our study, where survey staff objectively measure participants' balance, walking speed, grip strength, etc.) They found that self-reported items best distinguished participants in the average to above average (0–1 SDs) region of limitations, performance-based items distinguished participants across a broader range, and combining self-report and performance-based items actually provided more information than either type of item used alone. In NHANES, the only objectively measured performance-based item was grip strength; thus, we did not perform this comparison between self-report and performance-based items. Still, their findings are consistent with ours, that self-reported limitations best differentiated participants in the 0–2 SD range, with essentially zero ability to discriminate individuals in the <0 SD range. Together with this prior work in NHATS, our results now demonstrate consistency of findings across studies, which had some differences in how questions were asked (e.g. language in NHATS describes trouble "in the last month," whereas NHANES asks about current difficulty performing certain tasks), what exact questions were asked, and what populations were studied (our study was over 20 years old, whereas NHATS was over 65 years old). Disability items in NHATS have been validated in terms of test–retest reliability and convergence validity:[34] while we are not aware of NHANES questions having been subjected to test–retest analysis, here we nonetheless provide evidence for convergence validity and thus do not suspect that questions in NHANES would be any less valid than those of NHATS to explain the results.

Our study has several limitations. Self-reported variables such as medical conditions could misclassify participants. For example, self-reported physical activity may overestimate actual physical activity compared with direct

accelerometer-based measurement.[35] Second, NHANES is a cross-sectional study, and thus itself does not allow for correlation between risk factors and future development of disability in a longitudinal fashion. Third, sample size and power calculations are complex in IRT models. While we did not perform power calculations for this study, literature supports that our sample size ($N = 17{,}057$) was more than adequate; for example, simulation-based work suggests that sample sizes exceeding $N = 500$ achieve adequate precision for certain graded response models,[36] and other authors have suggested N at least 200 to 500 is adequate for 2PL models.[37,38]

Conclusion

NHANES disability-related questions provided a high degree of information at distinguishing individuals with higher than average limitations with favorable convergent validity and consistency with other national data, and good model fit. However, they provided essentially zero resolution to distinguish individuals with low or average limitations. If a given research question requires ability to distinguish between individuals with low or average limitation, then this dataset would not provide adequate resolution. Furthermore, many existing questions provided substantial overlap in the range of the latent limitations trait over which they help distinguish individuals, and our work highlights that certain questions may be omitted for analysis if using all dichotomous items, and that ordinal items provided greater information than dichotomous items though with less possibility for dropping a meaningful number of items without sacrificing information. Finally, we demonstrated how IRT predictions can be used as outcomes themselves when exploring correlations between ICF-related concepts and disability-related traits.

Acknowledgements

The authors thank Dr Vicki Freedman at the University of Michigan who provided expert feedback regarding contextualizing relevant literature, research approach, and data interpretation.

Author contributions

S.W.T. contributed to hypothesis generation, study design, data collection, statistical analysis, and manuscript preparation. J.F.B. contributed to study design, and manuscript preparation and editing.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Dr Samuel W Terman is supported by the University of Michigan Department of Neurology Training Grant 5T32NS007222-38. He has no relevant disclosures. Dr James F Burke is supported by the National Institute of Neurological Disorders and Stroke K08 NS082597 and National Institutes of Health, National Institute on Minority Health and Health Disparities R01 MD008879.

Ethical approval

Ethical approval for this study was waived by the University of Michigan Institutional Review Board.

Informed consent

Written informed consent was obtained from all subjects before the initial National Health and Nutrition Examination Survey (NHANES) study, but not needed for the present analysis as this study was deemed exempt by the University of Michigan Institutional Review Board (Ethics Approval Number: HUM00175050).

Writing assistance

No additional individuals provided writing assistance.

ORCID iD

Samuel W Terman https://orcid.org/0000-0001-6179-9467

Supplemental material

Supplemental material for this article is available online.

References

1. World Health Organization (WHO). Disability and health, 2020, https://www.who.int/news-room/fact-sheets/detail/disability-and-health (accessed 26 October 2020).
2. Guralnik JM and Ferrucci L. The challenge of understanding the disablement process in older persons. J Gerontol A Biol Sci Med Sci 2009; 64(11): 1169–1171; discussion 1175.
3. World Health Organization (WHO). International Classification of Functioning, Disability and Health (ICF), https://www.who.int/classifications/icf/en/ (accessed 29 May 2020).
4. Ostchega Y, Harris TB, Hirsch R, et al. The prevalence of functional limitations and disability in older persons in the US: data from the National Health and Nutrition Examination Survey III. *J Am Geriatr Soc* 2000; 48(9): 1132–1135.
5. Cook CE, Richardson JK, Pietrobon R, et al. Validation of the NHANES ADL scale in a sample of patients with report of cervical pain: factor analysis, item response theory analysis, and line item validity. *Disabil Rehabil* 2006; 28(15): 929–935.
6. Wang T, Wu Y, Wang W, et al. Association between coffee consumption and functional disability in the US older adults. *Br J Nutr* 2020; 125: 1–19.
7. Wang T, Jiang H, Wu Y, et al. The association between Dietary Inflammatory Index and disability in older adults. *Clin Nutr* 2021; 40: 2285–2292.
8. Wu LW, Chen WL, Peng TC, et al. All-cause mortality risk in elderly individuals with disabilities: a retrospective observational study. *BMJ Open* 2016; 6(9): e011164.
9. Xu B, Houston D, Locher JL, et al. The association between Healthy Eating Index-2005 scores and disability among older Americans. *Age Ageing* 2012; 41(3): 365–371.

10. Idler EL, Russell LB and Davis D. Survival, functional limitations, and self-rated health in the NHANES I epidemiologic follow-up study, 1992. *Am J Epidemiol* 2000; 152(9): 874–883.
11. Manns P, Ezeugwu V, Armijo-Olivo S, et al. Accelerometer-derived pattern of sedentary and physical activity time in persons with mobility disability: National Health and Nutrition Examination Survey 2003 to 2006. *J Am Geriatr Soc* 2015; 63(7): 1314–1323.
12. Tjia J, Briesacher BA, Peterson D, et al. Use of medications of questionable benefit in advanced dementia. *JAMA Intern Med* 2014; 174(11): 1763–1771.
13. Plantinga LC, Johansen KL, Schillinger D, et al. Lower socioeconomic status and disability among US adults with chronic kidney disease, 1999-2008. *Prev Chronic Dis* 2012; 9: E12.
14. Puri S, Herrick JE, Collins JP, et al. Physical functioning and risk for sleep disorders in US adults: results from the National Health and Nutrition Examination Survey 2005-2014. *Public Health* 2017; 152: 123–128.
15. Steeves JA, Shiroma EJ, Conger SA, et al. Physical activity patterns and multimorbidity burden of older adults with different levels of functional status: NHANES 2003-2006. *Disabil Health J* 2019; 12(3): 495–502.
16. Stevens A, Courtney-Long E, Gillespie C, et al. Hypertension among US adults by disability status and type, National Health and Nutrition Examination Survey, 2001-2010. *Prev Chronic Dis* 2014; 11(8): E139.
17. Suzuki R. The interaction effects between race and functional disabilities on the prevalence of self-reported periodontal diseases—National Health and Nutrition Examination Survey 2011-2012. *Community Dent Health* 2017; 34(4): 234–240.
18. Terman SW, Hill CE and Burke JF. Disability in people with epilepsy: a nationally representative cross-sectional study. *Epilepsy Behav* 2020; 112: 107429.
19. Van der Linden WJ and Hambleton RK. *Handbook of modern item response theory*. New York: Springer, 1997.
20. Spector WD and Fleishman JA. Combining activities of daily living with instrumental activities of daily living to measure functional disability. *J Gerontol B Psychol Sci Soc Sci* 1998; 53(1): S46–S57.
21. McHorney CA and Cohen AS. Equating health status measures with item response theory: illustrations with functional status items. *Med Care* 2000; 38(9 suppl.): II43–II59.
22. McHorney CA. Use of item response theory to link 3 modules of functional status items from the asset and health dynamics among the oldest old study. *Arch Phys Med Rehabil* 2002; 83(3): 383–394.
23. Kasper JD, Chan KS and Freedman VA. Measuring physical capacity: an assessment of a composite measure using self-report and performance-based items. *J Aging Health* 2017; 29(2): 289–309.
24. National Health and Nutrition Examination Survey. NHANES 2017-2018 questionnaire data, https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire&CycleBeginYear=2017 (accessed 27 January 2021).
25. National Health and Nutrition Examination Survey. NHANES 2017-2018 questionnaire data overview, https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overviewquex.aspx?BeginYear=2017 (accessed 27 January 2021).
26. Moen VP, Drageset J, Eide GE, et al. Dimensions and predictors of disability—a baseline study of patients entering somatic rehabilitation in secondary care. *PLoS One* 2018; 13(3): e0193761.
27. American Diabetes Association. Diagnosis, 2020, https://www.diabetes.org/a1c/diagnosis (accessed 8 September 2020).
28. Muntner P, DeSalvo KB, Wildman RP, et al. Trends in the prevalence, awareness, treatment, and control of cardiovascular disease risk factors among noninstitutionalized patients with a history of myocardial infarction and stroke. *Am J Epidemiol* 2006; 163(10): 913–920.
29. Kroenke K, Spitzer RL, Williams JBW, et al. The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review. *Gen Hosp Psychiatry* 2010; 32(4): 345–359.
30. National Health and Nutrition Examination Survey: analytic guidelines 2011-2014 and 2015-2016. 3.1.1 determining the appropriate sample weight for analysis, 2018, https://wwwn.cdc.gov/nchs/data/nhanes/analyticguidelines/11-16-analytic-guidelines.pdf (accessed 20 April 2021).
31. Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 2007; 45(5 suppl. 1): S22–S31.
32. Stata. *Item response theory reference manual*. College Station, TX: Stata Press, 2019.
33. Anderson P and Madden R. Design and quality of ICF-compatible data items for national disability support services. *Disabil Rehabil* 2011; 33(9): 758–769.
34. Freedman VA, Kasper JD, Cornman JC, et al. Validation of new measures of disability and functioning in the National Health and Aging Trends Study. *J Gerontol A Biol Sci Med Sci* 2011; 66(9): 1013–1021.
35. Prince SA, Adamo KB, Hamel ME, et al. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *Int J Behav Nutr Phys Act* 2008; 5: 56.
36. Jiang S, Wang C and Weiss DJ. Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Front Psychol* 2016; 7: 109.
37. Morizot J, Ainsworth AT and Reise SP. Toward modern psychometrics: application of item response theory models in personality research. In: Robins RW, Fraley RC and Krueger RF (eds) *Handbook of research methods in personality psychology*. New York: Guilford Press, 2007, pp. 407–423.
38. Thorpe GL and Favia A. Data analysis using item response theory methodology: an introduction to selected programs and applications. *Psychology Faculty Scholarship*, 2012, p. 10, https://digitalcommons.library.umaine.edu/cgi/viewcontent.cgi?article=1019&context=psy_facpub (accessed 26 March 2021).