BMC Medical Research Methodology

RESEARCH

Open Access

Pre-statistical harmonization of behavioral instruments across eight surveys and trials

Diefei Chen[1*], Eric Jutkowitz[2,3], Skylar L. Iosepovici[2], John C. Lin[2] and Alden L. Gross[1,4]

Abstract

Background Data harmonization is a powerful method to equilibrate items in measures that evaluate the same underlying construct. There are multiple measures to evaluate dementia related behavioral symptoms. Pre-statistical harmonization of behavioral instruments in dementia research is the first step to develop a statistical crosswalk between measures. Studies that conduct pre-statistical harmonization of behavioral instruments rarely document their methods in a structured, reproducible manner. This is a crucial step which entails careful review, documentation and scrutiny of source data to ensure sufficient comparability between items prior to data pooling. Here, we document the pre-statistical harmonization of items measuring behavioral and psychological symptoms among people with dementia. We provide a box of recommended procedure for future studies.

Methods We identified behavioral instruments that are used in clinical practice, a national survey, and randomized trials of dementia care interventions. We rigorously reviewed question content and scoring procedures to establish sufficient comparability across items as well as item quality prior to data pooling. Additionally, we standardized coding to Stata-readable format, which allowed us to automate approaches to identify potential cross-study differences in items and low-quality items. To ensure reasonable model fit for statistical co-calibration, we estimated two-parameter logistic Item Response Theory models within each of the eight studies.

Results We identified 59 items from 11 behavioral instruments across the eight datasets. We found considerable cross-study heterogeneity in administration and coding procedures for items that measure the same attribute. Discrepancies existed in terms of directionality and quantification of behavioral symptoms for even seemingly comparable items. We resolved item response heterogeneity, missingness and skewness, conditional dependency prior to estimation of item response theory models for statistical co-calibration. We used several rigorous data transformation procedures to address these issues, including re-coding and truncation.

Conclusions This study highlights the importance of each aspect involved in the pre-statistical harmonization process of behavioral instruments. We provide guidelines and recommendations for how future research may detect and account for similar issues in pooling behavioral and related instruments.

Keywords ADRD, Dementia, Behavioral symptom management, Epidemiology, Pre-statistical harmonization

*Correspondence:
Diefei Chen
dchen95@jhmi.edu
[1] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States, 2024 E. Monument Street, Baltimore, MD 21205, USA
[2] Health Services, Policy & Practice, Brown School of Public Health, Providence, RI, USA
[3] Center of Innovation in Long Term Services and Supports, Providence VA Medical Center, Providence, RI, USA
[4] Johns Hopkins University Center on Aging and Health, Baltimore, MD, USA

Background

Often there are multiple instruments that evaluate the same underlying construct. Data harmonization is a powerful method that combines data obtained from different items that represent the same underlying construct. For example, in dementia research, behavioral symptoms are collected using different measures such as Neuropsychiatric Inventory-Questionnaire (NPI) and Problem Behavior Checklist. Such instruments can be combined by anchoring on the shared behavioral symptomology. Integrating data from different study populations encourages collaborations, increases sample size and statistical power, improves generalizability, facilitates subgroup analyses and investigation of rare phenotypes [21, 23, 24], ensures reliability of published study results [9], and optimzies existing data and research infrastructures [17]. Data harmonization has been used to advance research in genome-wide association studies [8, 26], neuroimaging [42], and dementia care [20, 25].

There are several approaches to data harmonization. For example, Item Response Theory (IRT) consists of modeling a latent variable based on different sets of items that represent the same underlying construct. Items can be measured within studies by different instruments or across studies [21, 23, 24]. Other commonly used statistical methods include standardization and missing data by design with multiple imputation [18]. Regardless of the statistical harmonization method of choice, pooling of data is complex and requires careful scrutiny of source datasets and items [2]. Most methods require data to have some common items to be used for linking purposes---this necessitates undertaking the qualitative process of pre-statistical harmonization [17]. However, this crucial step for optimizing existing research resources and infrastructures is rarely described in research.

Pre-statistical harmonization is the series of procedures undertaken before data pooling. The goal of pre-statistical harmonization is to identify items that are likely comparable across studies [17]. Pre-statistical harmonization involves selection of participant studies (e.g., careful review of study design, methods and study population), and identification of items to be harmonized [11]. It is crucial to identify items that are measured using comparable instruments. Candidate items for linking can be those measuring a comparable underlying construct, and can be harmonized using a simple transformation algorithm via IRT or other approaches. Several studies have described pre-statistical harmonization from disease areas including substance use [40] and cognitive impairment [6, 17].

In this study, we document the pre-statistical harmonization of dementia behavioral symptom measures captured in National Institutes of Aging funded Alzheimer's Disease Research Centers clinical evaluations, a National Institutes of Aging national survey of cognitive health, and six National Institutes of Aging funded randomized trials of nondrug dementia care interventions that include assessments of dementia related behaviors. This study is the first step in a larger initiative to develop a statistical crosswalk between the different measures of dementia related behaviors administered in clinical practice, national surveys, and randomized trials of dementia care interventions. A major challenge in harmonization of behavioral data is the remarkable variability in questions across instruments and how they are asked. Studies vary in terms of response options, directionality in coding responses (e.g. 0 = No, 1 = Yes vs. 0 = Yes, 1 = No), quantification of behavioral manifestations, and other factors. To ensure the quality and reproducibility of data pooling, careful scrutiny of items to be harmonized is a crucial step to account for these differences. However, this phase of data harmonization is rarely discussed in published research.

In this paper, we aim to describe procedures we undertook during the pre-statistical harmonization process of measures of dementia related behaviors. We do so for the sake of reproducibility and to provide a guide for future studies requiring the consolidation of multiple data sources to evaluate behavioral symptoms of dementia [38, 39]. Specifically, we describe approaches to review question content and scoring procedures in order to establish comparability across items before data pooling. We then summarize our findings on how heterogeneity in items both within and across studies might lead to difficulty in interpretations of statistical models. Finally, we offer guidelines for how future research needs to acknowledge and address similar issues in pooling behavioral instruments.

Methods

Studies

We identified measures of dementia behavioral symptoms that are used in clinical practice, a national survey, and randomized trials of dementia care interventions. Because our ultimate goal is to develop a statistical crosswalk between common measures of behavioral symptoms, we only needed a single record per participant in studies with longitudinal data. National Institute on Aging funded Alzheimer's Disease Research Centers submit longitudinal clinical evaluations to the National Alzheimer's Coordinating Center (NACC) uniform dataset, which includes an assessment of behavioral symptoms. For NACC, we used data from clinical evaluations submitted between September 2005 and May 2020 ($N = 14{,}654$), and we selected a single random visit

for each participant with a dementia diagnosis [1, 4]. The Aging, Demographics and Memory Study (ADAMS) is a US representative survey of cognitive health, in which participants were administered a clinical assessment that includes measures of behaviors [32]. We restricted our analysis to ADAMS Wave A participants with a dementia diagnosis ($N=308$). Care of Older Persons in Their Environment (COPE) ($N=237$), the Tailored Activity Program (TAP) ($N=60$), the Alzheimer's Quality of Life Study (ALZQOL) ($N=88$), the Advancing Caregiver Training (ACT) project ($N=272$), the Resources for Enhancing Alzheimer's Caregiver Health project (REACH) ($N=670$), the Adult Day Services Plus study (ADS PLUS) ($N=194$) are National Institute on Aging funded trials of non-drug dementia care interventions that included measures of dementia related behaviors and are trials in which the study principle investigator was willing to share data. We used baseline data from these trials so that responses would not be confounded by participation in the trial.

Specifically, COPE was a randomized trial to test the effectiveness of a nonpharmacologic intervention that aims to ameliorate physical functioning, quality of life, and behavioral outcomes for people living with dementia [15]. The TAP trial tested a home-based occupational therapy intervention that aimed to reduce behavioral symptoms of people living with dementia [16]. ALZQOL was a randomized trial to assess potentially modifiable risk factors associated with quality of life for persons with dementia and caregivers [12]. ACT was a randomized trial testing a home-based multicomponent intervention targeting environmental and behavioral factors contributing to quality of life of persons with functional disabilities [13, 14]. REACH II was a randomized controlled trial of the effects of a multicomponent intervention on quality of caregiving among caregivers of dementia patients [3]. ADS Plus study was a randomized controlled trial of the effect of a management intervention on quality of caregiving among dementia caregivers, service utilization and institution placement of care recipients [13, 14]. We selected only baseline data from COPE, TAP, ALZQOL, ACT, REACH II, ADS Plus to be merged with other studies.

Procedure

We acquired codebooks, data entry and test forms, and procedural instructions from each study. We then identified common behavioral instruments and items used within and across studies. We reviewed each individual item to identify its respective behavioral attribute, skip patterns (e.g. questions that are conditional on other items), question stems, response options or scoring types, and theoretical score ranges. This step revealed multiple sources of variation across studies.

Upon reviewing available documentation, we created a crosswalk document that links common items from assorted instruments adopted within and across studies that assess behaviors. As implemented here, a crosswalk is a table that maps common elements between different studies. Each row represents an item of interest (e.g., whether a respondent exhibits false beliefs). The relevant individual test item for a study associated with a construct is placed on the corresponding row in the crosswalk. Items judged by experts to be similar across studies were placed on a row together. Additionally, this crosswalk contained relevant information about each item in each study including the name of source dataset, specific section of the survey, name and version of the instrument being used, study-specific name for each item, question stem, and responses options which included the possible score range. The crosswalk was updated throughout the pre-statistical harmonization procedures listed below. For the purpose of data sharing, this crosswalk will be made available upon request to the senior author.

Workflow

Establishing a workflow is a process that encompasses all aspects involved in data management. We followed procedures described in *The Workflow of Data Analysis Using Stata,* by J. Scott Long [33]. In our data analysis project, we used a generalizable file structure sharable via a secured online server that can be accessed by team members from different terminals. This structure facilitates reproducible research.

There are nine common folders: 1) Hold then delete, 2) To file, 3) Admin, 4) Documentation, 5) Mailbox, 6) Posted, 7) Private, 8) Readings, and 9) Work. Hold then delete and To file are folders that temporarily hold files so that we can determine the purpose of these files later, as needed. Admin is a folder for budgeting, correspondence with other investigators, IRB paperwork, and the proposal. Posted is probably the most important folder: it contains sub-folders for analysis, data (both source data and data derived with our analytic code; distinguishing between these is especially crucial for purposes of reproducible research), descriptive statistics, figures and interim analyses. Other folders are self-explanatory by their names. Under the Posted folder, there is sub-folder containing the common set of analytic files. Analytic files contain sections of code pertinent to a specific task during the data cleaning and processing. Descriptions of each analytic file are below:

Data management files:

1. Master.do: sets up working directory and calls all files;
2. Preambling.do: sets local macro and global macro to store study-specific item names;
3. Start-latex.do: produces pdfs for reports;
4. Call-source.do: calls data from source data files and processes the raw data minimally, such as reshaping data into long format and generating record ID;
5. Renamevars.do: generates the rename, recode and labeling commands for each item and store them as global macros;
6. Mergestudies.do: calls on renaming, recoding and labeling global macros and merges studies together;
7. Create-variables.do: performs data cleaning so that items have the same values across datasets;
8. Select-cases-for-analysis.do: Identifies data-specific global macros of items for each attribute;

Data analysis files:

9. Model-fitting.do: This program runs IRT analyses of behavioral items in each study separately.
10. Models.do: This program conducts statistical co-calibration via IRT of behavioral items.
11. Sensitivity-analysis.do: This is an optional syntax file for any sensitivity analyses necessary to probe robustness of results.

Conditional dependency

Responses to some items are conditional on others. For example, answering "yes" to a question about a behavior may prompt, in some datasets, additional questions about frequency or recency of the behavior. These items are inappropriate to be include in statistical harmonization because the items are conditionally dependent on each other. To address this in our datasets, we underwent rigorous efforts to manually review each instrument and items therein. We found that, in this project, items assessing severity or frequency of a behavior were usually conditional on a binary yes/no question assessing presence of a behavior.

Standardization of item coding

A critical stage of pre-statistical harmonization is to ensure common items, and items that can be made comparable via transformation, are comparable across instruments or studies. One way to achieve this is to code response options in the same way across multiple instruments and studies. For example, ADAMS adopted the Neuropsychiatric Inventory (NPI) as an instrument to measure behavioral outcomes. One specific test item measures whether delusion of danger is present in the participant's behavior. The question stem is "Does (NAME) believe that (HE/SHE) is in danger--that others are planning to hurt (HIM/HER)?". Response options include: yes (coded 1), no (coded 5), invalid (coded 6), skipped or not asked because screening symptom not confirmed (coded 96), not assessed/not asked (coded 97), don't know (coded 98), not applicable/not assessed for this item (coded 99). To standardize response options, we edited the crosswalk column and replaced text descriptions of value labels with stata-readable code that revised values so that a resulting variable would be ready for analysis. For the above example, we set values of 6, 96, 97, 98 and 99 to missing, and values of 5 to 0, such that the final item for this question to be used for analysis has values of 0 (not endorsed) and 1 (endorsed).

Comparability of items

In addition to manually reviewing each test item, we leveraged several automated approaches to uncover potential cross-study differences in items. We summarized and visualized data by displaying item values specific to each study. For example, we cross-tabulated items and studies conditioning on that item having different minimum and maximum values across studies. Resulting tables can identify items that have different scoring procedures across studies. Another approach to uncover potential sources of heterogeneity is to estimate correlation matrices. These matrices help identify items which have sizable negative intercorrelations with other items, and thus which may need to be reverse coded to be in the same direction as other items. Within each study, we tabulated the frequency of each item to identify skewness and potential outliers. We cycled through every item within each study, filtering out items which have the same minimum value and maximum value within a specific study (indicating no variability). We filtered out items which have maximum value between 90 to 100, or that had negative values because for our scales, such values indicated missing data codes that should be removed prior to analysis via recoding in our crosswalk. In our preliminary IRT analysis, we leveraged both univariate and bivariate residual analysis to identify items that have mismatched model estimated correlation and empirical correlations. Details on how each approach is adopted for a specific harmonization task are given in following sections.

Missingness and skewness

On top of missingness in responses already coded in original documentation (e.g. In ADAMS, 6 = Invalid, 96 = Skipped, or not asked because screening symptom not confirmed, 97 = Not assessed/Not asked (NPI

not completed), 98. = DK (Don't Know), 99 = Not applicable/not assessed for this item), we paid close attention to other sources of outlying values or skewness. For example, some items represent severe or extremely rare behavioral symptoms (i.e. inappropriate sexual contact), such that the frequency of its being endorsed within a particular study is low. Another possible scenario is when an item is only assessed for a subset of participants, as an artifact of conditional dependency (i.e. an item can only be answered given another item's response). Sometimes there is little to no variability in responses due to small sample sizes. ALZQOL and TAP are both small randomized trials.

Items which have excessive missingness – which is usually explainable – and skewness in responses could create issues when we run statistical models such as IRT analysis later, mainly because they would have poor correlations with other items.

Directionality

Items should be consistent in their polarity or direction to facilitate interpretability of harmonized factor analysis. We decided to code "higher" values as indicative of adverse behaviors. If the presence of a certain behavior is considered "worse", we gave this response option a higher value. The same procedure was undertaken for items that indicate severity or frequency, such that higher values indicate more severe presentations of a behavioral symptom. This step required careful review of question content and response options.

On top of these qualitative review procedures, we adopted an automated approach to identify items in need of reverse coding. Within each study, we ran correlation matrices and flagged potentially problematic negative correlations ($r < -0.2$). If not by chance, a negative correlation tends to indicate an item may be in need of reverse scoring relative to other items within the same study.

Scoring type and scales

Other than ensuring consistent directionality, all studies must have the same response levels for a given item (e.g., 4-point Likert scale, 5-point Likert scale, binary no/yes). Discrepant response levels are present when different instruments were used but judged to have common items. We undertook rigorous efforts to parse out presumably comparable items that were subject to differential scoring across studies.

Model fitting

To check whether that above procedures helped and were not detrimental for statistical co-calibration, in each of the eight studies we estimated two-parameter logistic Item Response Theory (IRT) models. This procedure is akin to testing configural measurement invariance (e.g., [31]). The two-parameter IRT model predicts the probability of an individual endorsing a behavioral symptom or item, as a function of the discrepancy between one's unobserved level on the underlying trait and the item difficulty parameter, modified by the item discrimination parameter. Item difficulty, akin to a threshold in factor analysis, is the level of the underlying trait at which a randomly selected individual from the population has a 50% probability of endorsing the item. Item discrimination, analogous to a factor loading in factor analysis, describes the strength of association between the item and the underlying trait, or how well the item separates individuals having low and high level of the underlying trait [34, 41]. We scrutinized Mplus output for high residuals between empirical and model-estimated covariances (specifically, standardized residuals greater than 0.3), which would suggest a mismatch between model estimated and empirical correlations [37]. High residuals could imply items measure a similar behavior, or a heretofore undetected problem with conditional independence. Estimation of separate models within each individual study helps establish configural invariance by detecting couplet items that may be subjected to multidimensionality [30]. This procedure helps us to examine whether participants from different studies interpret the behavioral measurement items in a conceptually equivalent way [5]. We use three fit statistics to examine our model fit: root mean square error of approximation (RMSEA), comparative fit index (CFI) and standardized root mean square residual (SRMR). By convention, we adopted a cut-off value of RMSEA less than 0.08 to indicate excellent fit, an RMSEA between 0.05 to 0.08 to indicate mediocre fit. As for CFI, a value greater or equal to 0.90 is indicative of goof fit. A SRMR value lower than 0.08 indicates good fit [28, 29, 35]. Fitting these data to such a model is not conclusive of a problem with harmonization because misfit can also be due to model misspecification, however, we used model fitting as an exploratory approach.

Results

Characteristics of study samples

Among the eight samples, NACC has the largest sample size ($n = 14{,}564$). Thus, the baseline characteristics of our pooled sample are largely driven by participants from NACC. TAP ($n = 60$) and ALZQOL ($n = 88$) are both community-based trials and have the smallest sample sizes. Most study cohorts are balanced in terms of sex of the participants, except for ADAMS and COPE which had predominantly female participants. Mean ages are reasonably comparable across studies. ADAMS has the oldest cohort (85.2 years) and ADS Plus has the

Table 1 Characteristics of study participants

Characteristics	Adult Day Services Plus study (ADS PLUS)	Advancing Caregiver Training (ACT)	Aging, Demographics and Memory Study(ADAMS)	Alzheimer's Quality of Life Study (ALZQOL)	Care of Older Persons in Their Environment (COPE)	National Alzheimer's Coordinating Center(NACC)	Resources for Enhancing Alzheimer's Caregiver Health(REACH II)	Tailored Activity Program (TAP)
Sample Size	194	272	308	88	237	14,564	670	60
Age(in years), mean(SD)	[a]67.3(12.8)	82.4(8.44)	85.2(6.91)	81.7(8.02)	82.7(8.66)	74.6(9.82)	[a]79.0(9.2)	79.4(9.40)
Female, n(%)	99(51.0)	147(54.0)	213(69.2)	46(52.3)	165(69.2)	8064(55.4)	[a]390 (58.2)	26(43.3)
Race, n(%)								
White	150(77.3)	193(71.0)	224(72.7)	67(76.1)	170(71.7)	11967(82.2)	321(47.9)	46(76.7)
African American	29(15.0)	73(26.8)	70(22.7)	20(22.7)	62(26.2)	1716(11.8)	214(31.9)	13(21.7)
Others	15(7.73)	6(2.21)	14(4.55)	1(1.14)	5(2.11)	881(6.05)	195(29.1)	1(1.67)
Ethnicity, n(%)								
Hispanic/ Latino	18(9.28)	5(1.84)	23(7.47)	2(2.27)	4(1.69)	1271(8.73)	207(30.9)	2(3.33)
Non-Hispanic/Latino	174(89.7)	267(98.2)	285(92.5)	86(97.7)	233(98.3)	13243(90.9)	436(65.1)	58(96.7)
Unknown	2(1.03)	0	0	0	0	50(0.34)	27(4.03)	0
Education, n(%)								
Less than high school/ GRE	21(10.8)	[a]25(9.19)	179(58.1)	[a]2(2.27)	[a]7(2.95)	1709(11.7)	314(46.9)	12(20.0)
High school graduate	55(28.4)	[a]69(25.4)	70(22.7)	[a]21(23.9)	[a]66(27.9)	3335(22.9)	149(22.2)	19(31.7)
Some college	50(25.8)	[a]83(30.5)	32(10.4)	[a]33(37.5)	[a]74(31.2)	2420(16.6)	56(8.36)	4(6.67)
College degree	39(20.1)	[a]62(22.8)	14(4.55)	[a]10(11.4)	[a]46(19.4)	3147(21.6)	64(9.55)	14(23.3)
Post-graduate study	27(13.9)	[a]33(12.1)	13(4.22)	[a]22(25.0)	[a]44(18.6)	3953(27.1)	37(5.52)	8(13.3)
Unknown	2(1.03)	[a]0	0	[a]0	[a]0	0	50(7.46)	3(5.00)

[a] Note: Demographic information of patient with ADRD is not avaliable because it is a study about care giver not care recipient

youngest (67.3 years). Study cohorts are predominantly White, non-Hispanic origin. Detailed baseline characteristics of study participants of each cohort are available in Table 1.

We identified 59 items from 11 instruments that measure a theoretically similar construct of behavioral symptomatology. Among these items, 29 items are unique to only one study; 4 items are common across 2 studies; 8 items are common across 3 studies; 6 items are common across 4 studies; 4 items are common across 5 studies; 1 item is common across 6 studies; 4 items are common across 7 studies; 3 items are common across 8 studies.

Conditional dependency

A threat to statistical harmonization is conditional dependency among items. For example, the Neuropsychiatric Inventory-Clinician rating scale (NPI-C) first determines if a behavior is present (e.g. verbal aggression). If so, conditional questions regarding different types of verbal aggression would be asked. During our qualitative review of instruments and item descriptions, we identified 13 items that are conditional upon other items in NACC; 136 conditional items in ADAMS; 38 conditional items in COPE; 42 conditional items in TAP; 120 conditional items in ALZQOL; 39 conditional items in ACT; 57 conditional items in REACH; 34 items in ADSPlus. We excluded these conditional items because they are redundant with other items in a given dataset and cannot depend on other items. As with most statistical methods, dependency among items can inappropriately boost reliability and give a false sense of measurement precision in a psychometric model.

Missingness and skewness

In reviewing every item within each study, we found that one item, indicating refusal to cooperate with appropriate help or resisting care with daily activities, had no

variability in ALZQOL. We thus removed this item from ALZQOL.

Directionality

Per our qualitative review of question stems and response options, together with automated statistical analysis on correlation matrices, we reverse-coded items within each study to ensure consistent directionality. Specifically, we reverse-coded 22 items in ADAMS; 2 items in COPE; 4 items in TAP; 24 items in ALZQOL; 2 items in ACT; 4 items in REACH.

One example of items that were flagged during qualitative review is one item for poor concentration in the Dementia Quality of Life Instrument (DEMQOL) from ALZQOL. The question stem is: *In the last week, how worried have you been about poor concentration?* The original response options are: *1=A lot, 2=Quite a bit, 3=A little, 4=Not at all, −5=Can't answer.* Since we decided that for all items, higher values should be indicative of worse or more severe symptomatology, we recoded this item as *3=A lot, 2=Quite a bit, 1=A little, 0=Not at all,* and *Can't answer* was recoded as missing. Another example is one item from ADAMS using the Neuropsychiatric Inventory (NPI) assessing whether the participant has enough energy. The question stem is: *Does (HE/SHE) have enough energy?* The original response options are: *1=Yes, 5=No, 6=Invalid, 96=Skipped, or not asked because screening symptom not confirmed, 97=Not assessed/Not asked (NPI not completed), 98.=DK (Don't Know), 99=Not applicable/not assessed for this item.* Since not having enough energy represents graver symptom and thus should be given a higher score, we recoded the response to *0=Yes, 1=No.*

Scoring types and scales

By reviewing each instrument, we found discrepancies in scoring procedures. For example, the Neuropsychiatric Inventory (NPI), Neuropsychiatric Inventory -Questionnaire (NPI-Q), Neuropsychiatric Inventory-Clinician rating scale (NPI-C), GDS (Geriatric Depression Screening Scale), CDDS (Cornell Depression in Dementia Scale), Care-recipient Behavioral Occurrence and Care-giver Upset (BOUP), Care-recipient Behavioral Occurrence and Care-giver Upset & Problem Behavioral Checklist (BOCGU) used binary coding; the Blessed Dementia Rating (Blessed) instrument adopted categories for scoring a behavioral symptom. Moreover, some instruments such as the Revised Memory and Behavior Checklist, Dementia Quality of Life Instrument (DEMQOL), Dementia Quality of Life Instrument-Proxy (DEMQOL-Proxy) use both binary and categorical scoring. Some items are counts because they were summary scores for behavioral symptoms. We identified four such summary score items and excluded them from our datasets.

One example of a discrepancy in scoring procedures among instruments is: one item from the Blessed Dementia Rating (Blessed) used score 0, 1, 2, 3 to code the presence of no serious behavioral language problem, shouting, cursing, or verbal aggression, with higher scores indicating greater severity; the Care-recipient Behavioral Occurrence and Care-giver Upset (BOUP) instrument used binary codes of 0 and 1 to indicate the presence of such behavior. By cycling through each item and comparing their minimum and maximum values across studies, we identified 20 items that had different scoring or coding procedures across studies. Therefore, we hard-coded some modifications to ensure cross-study consistency in scoring. To be more specific, we kept the 0 option and truncated scores equal or larger than 1, to 1. For studies that coded the lowest score value as 1 and not 0, we shifted all scores down by 1 point.

To reduce the risk of small cells and outliers in analysis, we also performed such truncation for items with small counts in certain cells, if the items were only present in one study. We collapsed one ordinal item into a binary response in REACH II; 3 items in ADAMS; 2 items in COPE; 2 items in TAP.

Model fitting

We examined model fit statistics, including empirical and model-estimated frequencies and standardized residuals. We leveraged this output to flag items that had a high impact on model fit and items responsible for the high standardized residuals (i.e. greater than 0.3 or smaller than −0.3). For bivariate residual correlations, we flagged items with residuals greater than 3 or smaller than −3. Specifically, we identified 4 sets of couplet items in ACT; 6 sets of couplet items in ADAMS; 36 sets of couplet items in ALZQOL; 2 sets of couplet items in COPE; 6 sets of couplet items in REACH II; 13 sets of couplet items in TAP; and 40 sets of couplet items in NACC. Inspection of these items revealed items that are conceptually similar, and are available in Supplemental materials. Such high residuals are potential artifacts

Table 2 Fit statistics of two-parameter IRT models within each study

Study	No. items	RMSEA	CFI	SRMR	Bifactor
ADSPlus	12	0.059	0.927	0.1	No
ACT	20	0.043	0.868	0.111	Yes
ADAMS	18	0.048	0.913	0.11	Yes
ALZQOL	25	0.033	0.952	0.123	Yes
COPE	19	0.037	0.925	0.103	Yes
NACC	23	0.069	0.753	0.123	Yes
REACH II	17	0.056	0.85	0.103	Yes
TAP	16	0	1	0.118	Yes

of violations of unidimensionality of the set of underlying items. Table 2 shows final model fit statistics for each study's IRT model. Judging by these criteria, our model fits range from acceptable to excellent in each study.

Discussion

Our study's goal was to describe and document the detailed procedures undertook in pre-statistical harmonization of behavioral instruments, to document our findings uncovered by the procedures in a reproducible manner [7]. To establish comparability of items across instruments and studies, we conducted extensive manual review of instruments and scrutinized raw data using automated procedures. We addressed several sources of heterogeneity across studies. Even when seemingly comparable items were asked across instruments and studies, differences in scoring schemes rendered them essentially non-compatible with each other. Table 3 contains a summary of procedures we undertook and may serve as a checklist for future studies requiring pre-statistical harmonization.

We summarize our recommendations for procedures to be taken and potential solutions for issues uncovered in Table 3. We recommend the following three general guidelines to avoid potential pitfalls and streamline the process for applied researchers who intend to develop a data harmonization project. First, establishing comparability of individual items is warranted even when standardized tests and instruments are used because administration differences can result in key cross-study differences. Researchers need to carefully document and scrutinize sources of variance, such as discrepancies in scoring and coding procedures that may lead to erroneous results. To facilitate this review process, obtaining abundance information from the source studies is especially important. Second, use a harmonized coding scheme that is both easy to use and retain all meaningful information [10]. In our study, we used the lowest common denominator (i.e. presence) to select variables that represent comparable behavioral symptoms. This means we discarded additional information regarding other aspects of the symptoms (i.e. frequency or severity). Finally, researchers should be cautious about sources of misfit in statistical models. Surely, misfit is usually attributable to a misspecified model, however in an integrative analysis across multiple data sources such misfit in parametric modeling can also be used to point towards non-equivalent items. Not detecting and accounting for issues around item non-comparability and conditional dependency may lead to failures in achieving acceptable model fit. Our study offers a template for conducting pre-statistical harmonization and fostering reproducibility.

Acknowledging and addressing limitations in raw data are critical steps before conducting data pooling. Our findings have three implications for harmonization of similar datasets involving survey data. First, pooling non-comparable items such as items

Table 3 Recommended procedures

Recommended procedures:
- Merge raw data from multiple sources with minimal pre-processing;
- Check whether item responses are comparable across sources;
- Clean data to establish item comparability:

○ Ensure constant directionality/polarity:
■ Review content and response options;
■ Run correlation matrices, flag items with sizable negative correlations;
■ Reverse code as necessary.
○ Ensure consistency in scoring type and scales:
■ Review response options;
■ Cross-tabulate items across datasets to evaluate whether items have different minimum and maximum values by dataset;
■ Exclude summary scores and counts in favor of more granular data;
■ Truncate, collapse response categories as necessary.
○ Eliminate conditional dependency:
■ Review content and logic flows;
■ Perform parametric modeling, scrutinize output for residuals;
■ Exclude conditional items.
○ Address missingness/skewness:
■Tabulate frequency of each item being endorsed;
■ Filter out items with coded missingness;
■ Filter out items with same min and max within a dataset;
■ Truncate, collapse response categories as necessary;
■Exclude items with no variability.

- Establish configural invariance:

○ Estimate parametric models within each dataset;
○ Scrutinize output for residuals;
○ Include residual covariances for items having high covariance residuals.

with reverse polarity and items with different coding schemes across studies may introduce bias by artificially creating variance between individuals from different studies. Second, conditional items or those with too much missingness or skewness may lead to spurious correlations and large residuals in statistical models. Particularly, items identified as being logically conditional on other items are especially problematic because they are highly correlated with each other. On top of that, conditional items usually were assessing the frequency or severity of a present behavior symptom. This is problematic because such count items often have high skewness, especially if a condition is rare. Finally, conditional items provide essentially duplicate information to other items.

This study highlights the importance of each aspect involved in the pre-statistical harmonization process of behavioral instruments. Specifically, conducting careful review of each instrument and item is critical in discovering potential sources of limitations in raw data. This step identified common items to be pooled both across and within studies, uncovered discrepancies in coding schemes, detected skip patterns. Performing preliminary IRT analysis within each study helped ensure reasonable model fit before pooling across studies. This step detected items that may violate the local independence or the unidimensionality assumption of IRT models [27]. Regardless of the statistical methods of harmonizing behavioral instruments of choice, detailed approaches described in this paper to uncover and tackle issues that are specific to harmonizing behavioral instruments are important to consider before carrying out the analysis.

After conducting pre-statistical harmonization, one conducts statistical harmonization to derive scores for a construct that are commonly scaled across multiple data sources. This score or set of scores can then be used for substantive research questions [17, 19–25, 36].

One limitation of the current study is that we did not combine items that indicate frequency or severity of a behavioral symptom with the screener item (i.e. item that indicate presence of the behavioral symptom). Instead, we excluded all conditionally dependent items. Additionally, we distilled all ordinal items into binary scale. Using indicator coding simplifies our analysis, but may lead to loss of resolutions and item quality. Another potential limitation is that identifying common items across datasets can be subjective. However, we leveraged expert reviews of items to assign items, which is considered state-of-the-art. In our next stage of analysis, after deriving a factor score for each participant, we were able to quantify the amount of error based on the quality and missingness of the respective item in a given study battery. We found a considerable number of participants have imprecisely estimated factor scores, especially in ADSPlus, ADAMS and REACH II. This observation could be a reflection of the inherent nature of psychometric measurements used to assess problematic dementia behaviors. However, this warrants careful interpretation of the harmonized factor scores, and may point to the need of sensitivity analysis in the future.

Conclusions

Data harmonization is an essential step towards effective use of existing data. In this study of pre-statistical harmonization, we pooled data on measures and items of dementia related behavioral symptoms captured in clinical assessments, a national survey, and randomized trials of non-drug dementia care interventions. An important next step is to reproduce the pre-statistical harmonization procedures described in this paper in other domains of interest, such as measures on functional and cognitive abilities of dementia patients across datasets.

Abbreviations

ACT	The Advancing Caregiver Training project
ADAMS	The Aging, Demographics and Memory Study
ADS Plus	The Adult Day Services Plus study
ALZQOL	The Alzheimer's Quality of Life study
Blessed	The Blessed Dementia Rating instrument
BOUP	The Care-recipient Behavioral Occurrence and Care-giver Upset instrument
BOCGU	The Care-recipient Behavioral Occurrence and Care-giver Upset & Problem Behavioral Checklist
CDDS	The Cornell Depression in Dementia Scale
COPE	The Care of Older Persons in Their Environment study
DEMQOL	The Dementia Quality of Life Instrument
DEMQOL-Proxy	The Dementia Quality of Life Instrument-Proxy
GDS	The Geriatric Depression Screening Scale
IRT	Item Response Theory
NACC	The National Alzheimer's Coordinating Center
NPI	The Neuropsychiatric Inventory
NPI-C	The Neuropsychiatric Inventory-Clinician rating scale
NPI-Q	The Neuropsychiatric Inventory-Questionnaire
REACH	The Resources for Enhancing Alzheimer's Caregiver Health project
TAP	The Tailored Activity Program
UDS	The Uniform Data Set

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-021-01431-6.

Additional file 1.

Acknowledgments

The NACC database is funded by NIA/NIH Grant U24 AG072122. NACC data are contributed by the NIA-funded ADRCs: P30 AG019610 (PI Eric Reiman, MD), P30 AG013846 (PI Neil Kowall, MD), P50 AG008702 (PI Scott Small, MD),

P50 AG025688 (PI Allan Levey, MD, PhD), P50 AG047266 (PI Todd Golde, MD, PhD), P30 AG010133 (PI Andrew Saykin, PsyD), P50 AG005146 (PI Marilyn Albert, PhD), P50 AG005134 (PI Bradley Hyman, MD, PhD), P50 AG016574 (PI Ronald Petersen, MD, PhD), P50 AG005138 (PI Mary Sano, PhD), P30 AG008051 (PI Thomas Wisniewski, MD), P30 AG013854 (PI Robert Vassar, PhD), P30 AG008017 (PI Jeffrey Kaye, MD), P30 AG010161 (PI David Bennett, MD), P50 AG047366 (PI Victor Henderson, MD, MS), P30 AG010129 (PI Charles DeCarli, MD), P50 AG016573 (PI Frank LaFerla, PhD), P50 AG005131 (PI James Brewer, MD, PhD), P50 AG023501 (PI Bruce Miller, MD), P30 AG035982 (PI Russell Swerdlow, MD), P30 AG028383 (PI Linda Van Eldik, PhD), P30 AG053760 (PI Henry Paulson, MD, PhD), P30 AG010124 (PI John Trojanowski, MD, PhD), P50 AG005133 (PI Oscar Lopez, MD), P50 AG005142 (PI Helena Chui, MD), P30 AG012300 (PI Roger Rosenberg, MD), P30 AG049638 (PI Suzanne Craft, PhD), P50 AG005136 (PI Thomas Grabowski, MD), P50 AG033514 (PI Sanjay Asthana, MD, FRCP), P50 AG005681 (PI John Morris, MD), P50 AG047270 (PI Stephen Strittmatter, MD, PhD).

Dr. Qian-Li Xue[1,2](qxue@jhmi.edu) contributed to reviewing and approving of the final manuscript as the second reader. Michelle Chung[2] (schung2@jhmi.edu) participated in the effort to standardize and harmonize items, and developed the data analytic file to merge datasets together.
[1] Johns Hopkins School of Medicine, Baltimore, MD.
[2] Johns Hopkins University Center on Aging and Health, Baltimore, MD.

Authors' contributions

Diefei Chen participated in the design of the analytic plan, conducting the analysis, interpreting the results and drafted the manuscript. Dr. Eric Jutkowitz obtained the funding, participated in the design of research plan, oversaw and led the qualitative review of instruments. Skylar Iosepovici participated in the qualitative review of instruments. John Lin participated in the qualitative review of instruments. Dr. Alden Gross obtained the funding, participated in the design of research plan, oversaw and led the design of the analytic plan, conducted the analysis, interpreted the results, and approved the final manuscript. The author(s) read and approved the final manuscript.

Funding

This work is part of a larger research project, Microsimulations to Compare Effectiveness and Cost-Effectiveness of Nondrug Interventions to Manage Clinical Symptoms in Racially/Ethnically Diverse Persons with Dementia, funded by National Institute on Aging. The research protocol was approved by JHSPH IRB((IRB No. IRB00013668) and Brown University (IRB No. 00000556).
COPE was Funded by the National Institutes of Health (R01 AG061945–01).
TAP was Funded by the National Institutes of Health (R21 MH069425).
ACT was funded by the National Institutes of Health (R01 AG13687).
REACH II was funded by the National Institutes of Health (AG13305, AG13289, AG13313, AG20277, AG13265, and NR004261.
ALZQOL was funded by the Alzheimer's Association (IIRG-07-28686) and the National Institute on Health (R01 AG22254).
ADS Plus was funded by the National Institute on Health (R01 AG049692–01).

Availability of data and materials

The datasets used during the current study are available in JHU OneDrive. The crosswalk we created, which can be used to reproduce the pre-statistical harmonization described in this manuscript, can be obtained from the corresponding author upon request.

Declarations

Ethics approval and consent to participate

This is a secondary analysis of several existing datasets for which the current research is consistent with the scope of the original consent processes. The written informed consent of original studies that have been included in the current research were obtained from all participants according to protocols approved by the Institutional Review Boards of JHSPH, the Institutional Review Boards of Brown. The methods used were carried out in accordance with the relevant guidelines and regulations specified in the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 13 June 2021 Accepted: 8 October 2021
Published online: 25 October 2021

References

1. About NACC data | National Alzheimer's Coordinating Center. [cited 2021 Feb 9]. Available from: https://naccdata.org/requesting-data/nacc-data
2. Bangdiwala SI, Bhargava A, O'Connor DP, Robinson TN, Michie S, Murray DM, et al. Statistical methodologies to pool across multiple intervention studies. Behav Med Pract Policy Res. 2016;6(2):228–35 [cited 2021 Mar 28]. Available from: https://academic.oup.com/tbm/article/6/2/228-235/4563145.
3. Belle SH, Burgio L, Burns R, Coon D, Czaja SJ, Gallagher-Thompson D, et al. Enhancing the quality of life of dementia caregivers from different ethnic or racial groups: a randomized, controlled trial. Ann Intern Med. 2006;145(10):727–38.
4. Besser L, Kukull W, Knopman DS, Chui H, Galasko D, Weintraub S, et al. Version 3 of the National Alzheimer's coordinating center's uniform data set. Alzheimer Dis Assoc Disord. 2018;32(4):351–8 [cited 2021 Jun 12]. Available from: https://journals.lww.com/00002093-201810000-00015.
5. Bialosiewicz S, Murphy K, Berry T. Do our measures measure up? The critical role of measurement invariance; 2013. p. 37.
6. Briceño EM, Gross AL, Giordani B, Manly JJ, Gottesman RF, Elkind MSV, et al. P4-368: pre-statistical harmonization of cognitive measures across six population-based cohorts: ARIC, CARDIA, CHS, FHS, MESA, AND NOMAS. Alzheimers Dement. 2018;14(7S_Part_30):P1611–2 [cited 2021 Aug 14]. Available from: http://doi.wiley.com/10.1016/j.jalz.2018.07.192.
7. Committee on Reproducibility and Replicability in Science, Board on Behavioral, Cognitive, and Sensory Sciences, Committee on National Statistics, Division of Behavioral and Social Sciences and Education, Nuclear and Radiation Studies Board, Division on Earth and Life Studies, et al. Reproducibility and replicability in science. Washington, D.C.: National Academies Press; 2019. [cited 2021 May 16]. Available from: https://www.nap.edu/catalog/25303
8. Crane PK, Trittschuh E, Mukherjee S, Saykin AJ, Sanders RE, Larson EB, et al. Incidence of cognitively defined late-onset Alzheimer's dementia subgroups from a prospective cohort study. Alzheimers Dement. 2017;13(12):1307–16 [cited 2021 Feb 21]. Available from: http://doi.wiley.com/10.1016/j.jalz.2017.04.011.
9. Doiron D, Burton P, Marcon Y, Gaye A, Wolffenbuttel BHR, Perola M, et al. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. Emerg Themes Epidemiol. 2013;10(1):12 [cited 2021 Jan 6]. Available from: https://ete-online.biomedcentral.com/articles/10.1186/1742-7622-10-12.
10. Esteve A, Sobek M. Challenges and methods of international census harmonization. Historical methods. J Quant Interdiscip Hist. 2003;36(2):66–79 [cited 2021 Mar 28]. Available from: http://www.tandfonline.com/doi/abs/10.1080/01615440309601216.
11. Fortier I, Raina P, Van den Heuvel ER, Griffith LE, Craig C, Saliba M, et al. Maelstrom research guidelines for rigorous retrospective data harmonization. Int J Epidemiol. 2016:dyw075 [cited 2021 Jan 7]. Available from: https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/dyw075.
12. Gitlin LN, Hodgson N, Piersol CV, Hess E, Hauck WW. Correlates of quality of life for individuals with dementia living at home: the role of home environment, caregiver, and patient-related characteristics. Am J Geriatr Psychiatry. 2014;22(6):587–97.
13. Gitlin LN, Reever K, Dennis MP, Mathieu E, Hauck WW. Enhancing quality of life of families who use adult day services: short- and long-term effects of the adult day services plus program. The Gerontologist. 2006a;46(5):630–9 [cited 2021 Feb 9]. Available from: https://academic.oup.com/gerontologist/article-lookup/doi/10.1093/geront/46.5.630.
14. Gitlin LN, Winter L, Dennis MP, Corcoran M, Schinfeld S, Hauck WW. A randomized trial of a multicomponent home intervention to reduce functional difficulties in older adults. J Am Geriatr Soc. 2006b;54(5):809–16.

15. Gitlin LN, Winter L, Dennis MP, Hodgson N, Hauck WW. A biobehavioral home-based intervention and the well-being of patients with dementia and their caregivers: the COPE randomized trial. JAMA. 2010;304(9):983–91.
16. Gitlin LN, Winter L, Vause Earland T, Adel Herge E, Chernett NL, Piersol CV, et al. The tailored activity program to reduce behavioral symptoms in individuals with dementia: feasibility, acceptability, and replication potential. Gerontologist. 2009;49(3):428–39.
17. Griffith, et al. Methods research report - harmonization of cognitive measures in individual participant data and aggregate data meta-analysis; 2013. p. 182.
18. Griffith LE, van den Heuvel E, Fortier I, Sohel N, Hofer SM, Payette H, et al. Statistical approaches to harmonize data on cognitive measures in systematic reviews are rarely reported. J Clin Epidemiol. 2015;68(2):154–62 [cited 2021 Jan 6]. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0895435614003497.
19. Griffith LE, van den Heuvel E, Raina P, Fortier I, Sohel N, Hofer SM, et al. Comparison of Standardization Methods for the Harmonization of Phenotype Data: An Application to Cognitive Measures. Am J Epidemiol [Internet]. 2016 Nov 15 [cited 2021 Oct 18];184(10):770–8. Available from: https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kww098.
20. Gross AL, Jones RN, Fong TG, Tommet D, Inouye SK. Calibration and validation of an innovative approach for estimating general cognitive performance. Neuroepidemiology. 2014a;42(3):144–53 [cited 2021 Aug 9]. Available from: https://www.karger.com/Article/FullText/357647.
21. Gross AL, Jones RN, Inouye SK. Development of an expanded measure of physical functioning for older persons in epidemiologic research. Res Aging. 2015a;37(7):–671, 94 [cited 2021 Aug 9]. Available from: http://journals.sagepub.com/doi/10.1177/0164027514550834.
22. Gross AL, Kueider-Paisley AM, Sullivan C, Schretlen D, International Neuropsychological Normative Database Initiative. Comparison of approaches for equating different versions of the mini-mental state examination administered in 22 studies. Am J Epidemiol. 2019;188(12):2202–12 [cited 2021 Aug 9]. Available from: https://academic.oup.com/aje/article/188/12/2202/5584412.
23. Gross AL, Mungas DM, Crane PK, Gibbons LE, MacKay-Brandt A, Manly JJ, et al. Effects of education and race on cognitive decline: an integrative study of generalizability versus study-specific results. Psychol Aging. 2015b;30(4):863–80 [cited 2021 Feb 21]. Available from: http://doi.apa.org/getdoi.cfm?doi=10.1037/pag0000032.
24. Gross AL, Power MC, Albert MS, Deal JA, Gottesman RF, Griswold M, et al. Application of latent variable methods to the study of cognitive decline when tests change over time. Epidemiology. 2015c;26(6):878–87 [cited 2021 Jan 7]. Available from: http://journals.lww.com/00001648-201511000-00015.
25. Gross AL, Sherva R, Mukherjee S, Newhouse S, Kauwe JSK, Munsie LM, et al. Calibrating longitudinal cognition in Alzheimer's disease across diverse test batteries and datasets. Neuroepidemiology. 2014b;43(3–4):194–205 [cited 2021 Jan 6]. Available from: https://www.karger.com/Article/FullText/367970.
26. Hamilton CM, Strader LC, Pratt JG, Maiese D, Hendershot T, Kwok RK, et al. The PhenX toolkit: get the most from your measures. Am J Epidemiol. 2011;174(3):253–60 [cited 2021 Jan 7]. Available from: https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwr193.
27. Hill CD, Edwards MC, Thissen D, Langer MM, Wirth RJ, Burwinkle TM, et al. Practical issues in the application of item response theory: a demonstration using items from the pediatric quality of life inventory (PedsQL) 4.0 generic core scales. Med Care. 2007;45(5):S39–47 [cited 2021 Aug 8]. Available from: https://www.jstor.org/stable/40221457.
28. Hooper D, Coughlan J, Mullen MR. Structural Equation Modelling: Guidelines for Determining Model Fit. Electron J Bus Res Methods 2008;6(1):53–60. available online at www.ejbrm.com.
29. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural equation modeling. Multidiscip J. 1999;6(1):1–55. [cited 2021 Aug 15]. Available from:. https://doi.org/10.1080/10705519909540118.
30. Kline RB. Principles and practice of structural equation modeling. 3rd ed. New York: Guilford Press; 2011. p. 427. (Methodology in the social sciences)
31. Kobayashi LC, Gross AL, Gibbons LE, Tommet D, Sanders RE, Choi S-E, et al. You say tomato, I say radish: can brief cognitive assessments in the U.S. health retirement study be harmonized with its international partner studies? Neupert S, editor. J Gerontol: Ser B. 2020;gbaa205. [cited 2021 Mar 19]. Available from: https://academic.oup.com/psychsocgerontology/advance-article/doi/10.1093/geronb/gbaa205/6009078.
32. Langa KM, Plassman BL, Wallace RB, Herzog AR, Heeringa SG, Ofstedal MB, et al. The aging, demographics, and memory study: study design and methods. Neuroepidemiology. 2005;25(4):181–91 [cited 2021 Feb 9]. Available from: https://www.karger.com/Article/FullText/87448.
33. Long JS. The workflow of data analysis using Stata. College Station: Stata Press; 2009. p. 379.
34. Lord FM. The relation of test score to the trait underlying the test. Educ Psychol Meas. 1953;13(4):517–49. [cited 2021 Aug 6]. Available from:. https://doi.org/10.1177/001316445301300401.
35. Maydeu-Olivares A. Evaluating fit in IRT models; 2015. p. 111–27.
36. Mukherjee S, Mez J, Trittschuh EH, Saykin AJ, Gibbons LE, Fardo DW, et al. Genetic data and cognitively defined late-onset Alzheimer's disease subgroups. Mol Psychiatry. 2020;25(11):2942–51 [cited 2021 Aug 15]. Available from: https://www.nature.com/articles/s41380-018-0298-8.
37. Muthén LKMB. Mplus User's Guide. 8th ed. Los Angeles: Muthén & Muthén; 2017.
38. Peng RD. Reproducible research in computational science. Science. 2011;334(6060):1226–7 [cited 2021 May 16]. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.1213847.
39. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten simple rules for reproducible computational research. Bourne PE, editor. PLoS Comput Biol. 2013;9(10):e1003285 [cited 2021 May 16]. Available from: https://dx.plos.org/10.1371/journal.pcbi.1003285.
40. Susukida R, Amin-Esmaeili M, Mayo-Wilson E, Mojtabai R. Data management in substance use disorder treatment research: implications from data harmonization of National Institute on Drug Abuse-funded randomized controlled trials. Clin Trials. 2021;18(2):215–25.
41. Takane Y, de Leeuw J. On the relationship between item response theory and factor analysis of discretized variables. Psychometrika. 1987;52(3):393–408. [cited 2021 Aug 6]. Available from:. https://doi.org/10.1007/BF02294363.
42. Zhu AH, Moyer DC, Nir TM, Thompson PM, Jahanshad N. Challenges and opportunities in dMRI data harmonization. In: Bonet-Carne E, Grussu F, Ning L, Sepehrband F, CMW T, editors. Computational diffusion MRI. Cham: Springer International Publishing; 2019. p. 157–72. (Mathematics and Visualization).

Publisher's Note