HHS Public Access

Author manuscript

Exp Aging Res. Author manuscript; available in PMC 2016 March 01.

Published in final edited form as:

Exp Aging Res. 2015 ; 41(5): 475–495. doi:10.1080/0361073X.2015.1085748.

Data Harmonization in Aging Research: Not So Fast

Margaret Gatz,
Department of Psychology, University of Southern California, Los Angeles, California, USA; and Karolinska Institutet, Stockholm, Sweden

Chandra A. Reynolds,
University of California, Riverside, Riverside, California, USA

Deborah Finkel,
Indiana University Southeast, New Albany, Indiana, USA

Chris J. Hahn,
Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA

Yan Zhou,
Mary S. Easton Center for Alzheimer's Disease Research, Department of Neurology, David Geffen School of Medicine at UCLA, Los Angeles, California, USA

Catalina Zavala, and
University of California, Riverside, Riverside, California, USA

for the IGEMS Consortium

Abstract

Background/Study Context—Harmonizing measures in order to conduct pooled data analyses has become a scientific priority in aging research. Retrospective harmonization where different studies lack common measures of comparable constructs presents a major challenge. This study compared different approaches to harmonization with a crosswalk sample who completed multiple versions of the measures to be harmonized.

Methods—Through online recruitment, 1061 participants aged 30 to 98 answered two different depression scales, and 1065 participants answered multiple measures of subjective health. Rational and configural methods of harmonization were applied, using the crosswalk sample to determine their success; and empirical item response theory (IRT) methods were applied in order empirically to compare items from different measures as answered by the same person.

Results—For depression, IRT worked well to provide a conversion table between different measures. The rational method of extracting semantically matched items from each of the two scales proved an acceptable alternative to IRT. For subjective health, only configural harmonization was supported. The subjective health items used in most studies form a single robust factor.

Correspondence concerning this article should be addressed to Margaret Gatz, University of Southern California, Department of Psychology, Los Angeles, CA 90089-1061. gatz@usc.edu.

Conclusion—Caution is required in aging research when pooling data across studies using different measures of the same construct. Of special concern are response scales that vary widely in the number of response options, especially if the anchors are asymmetrical. A crosswalk sample that has completed items from each of the measures being harmonized allows the investigator to use empirical approaches to identify flawed assumptions in rational or configural approaches to harmonizing.

Keywords

Crosswalk table; data harmonization; data sharing; integrative data analysis; item response theory; pooled data analysis; Rasch analysis; depression; self-rated health

Researchers increasingly are coming to appreciate that testing nuanced models of aging processes will only be possible with contributions from the large datasets that result from combining and harmonizing data across several studies (Fortier, Doiron, Wolfson, & Raina, 2012). Consequently, the National Institute on Aging (NIA), along with other institutes, places an emphasis on data sharing and harmonization across studies. Harmonization can be prospective or retrospective. Prospective approaches include creating toolboxes of measures that all researchers are encouraged to use (e.g., PhenX Toolkit: Hamilton et al., 2011; NIH Toolbox: Choi et al., 2012). Where data have already been collected, harmonization must be done retrospectively, and methods developed to take into account differences in measures.

When it is possible to equate across studies with minimal inference, rational methods of harmonization are often used. For example, in the United States, it is common to ask for highest education completed (e.g., less than high school, high school, some college, etc.), whereas in Great Britain respondents might be asked what educational qualifications had been obtained (e.g., A-level, teaching qualification, etc.). For purposes of harmonization, years of education can be derived from each (Lee, Zamarro, Phillips, Angrisani, & Chien, 2011). In other rational approaches, when the same questions are used across studies but with differences in response options, answers are often recoded to the same number of categories, e.g., recoding a 4-category ordered response scale used in one study (such as “rarely or none of the time”, “some or a little of the time”, “occasionally or a moderate amount of time”, or “most or all of the time”) to a “yes”/”no” response scale used in another study(e.g., Bath, Deeg & Poppelaars, 2010). Recoding is not always accurate, however, and may lose important data (Sharp, Suthers, Crimmins, & Gatz, 2009); for example, does “no” correspond to only the “rarely” alternative, or does “no” correspond to the lower two options?

Another common practice is to standardize scores from different scales used by different studies, e.g., z-scores, percentiles, or proportion of items endorsed (e.g., Curran et al., 2008), creating a seemingly common metric for pooled analyses. The drawback is that item or demographic differences in samples are ignored. For example, the 50th percentile in an older adult sample may not correspond to the 50^{th} percentile in a young adult sample.

When there is evidence for configural invariance across studies (i.e., consistent pattern of factor loadings across studies), it becomes possible to conduct pooled analyses at the latent factor level (Davidov, Schmidt, & Schwartz, 2008). This approach is exemplified by

deriving a first principal component from various test batteries to construct a general cognitive factor “g” or “IQ” (see Johnson, Bouchard, Krueger, McGue, & Gottesman, 2004; Finkel, Pedersen & McGue, 1995). However, recent pooled analyses of multi-national cognitive data suggest complexities can arise, perhaps due to differences in measure construction (Molenaar et al., 2013).

Many limitations may be addressed by having a linked sample where there is at least a partial overlap in items completed by each sample (Bauer & Hussong, 2009; Curran & Hussong, 2009; van Buuren, Eyres, Tennant & Hopman-Rock, 2003). Then, methods based on item response theory (IRT) allow estimation of trait scores (Embretson & Reise, 2000). This approach has been used to construct depression scores at each wave of a longitudinal study when there were different numbers of response categories at different waves (Jones & Fonda, 2004),and to co-calibrate two commonly used measures of functional independence (Velozo, Byers, Wang, & Joseph, 2007). In the absence of existing bridge items in studies being brought together, it is possible to recruit a new linked sample, administer the measures to be harmonized, use IRT to build a measurement crosswalk, and produce a conversion table that can be applied to the original studies.

The recently developed consortium on Interplay of Genes and Environment across Multiple Studies (IGEMS: Pedersen et al., 2013) is currently facing many of the issues associated with retrospective harmonization. The IGEMS consortium is investigating social context and gene-environment interplay in late-life functioning using combined data from 9 twin studies in 3 countries with participants aged 25 to 102 at baseline. In general, while the studies had constructs in common, there were few exact overlaps in items and/or response scales. This situation, which illustrates what many investigators face in conducting integrated analyses, motivated the current study.

Collection of new data from a linked crosswalk sample allowed us to test various approaches to harmonization for depression and subjective health, two variables frequently used in studies of adulthood and aging (including the IGEMS samples). Each variable represented a distinct challenge in harmonization: depression was measured with different established scales in the different studies, while subjective health questions in the different studies had similar item stems but varying response scales. The aim of the current paper is to present results using the crosswalk sample to compare the success of various approaches to harmonization. Note that the current paper uses no data from the IGEMS samples, only from the crosswalk sample in which participants completed all of the different versions of the measures.

Methods

Participants and Procedures[1]

Respondents in the crosswalk sample came from three primary sources. First, we turned to Amazon’s Mechanical Turk (www.MTurk.com), a crowdsourcing website (see Paolacci, Chandler & Ipeirotis, 2010; Buhrmester, Kwang & Gosling, 2011). However, the MTurk

[1]More details about procedures, analyses and results are available in an on-line supplement.

sample failed to deliver a substantial number of participants in their 70s and older. Our second source was the Healthy Minds volunteer list comprised of participants recruited from the local community; we supplemented the list with additional older adults from the same community. Third, the Alzheimer's Association TrialMatch website connects people with Alzheimer's disease, caregivers, family members, interested community volunteers, and physicians to clinical trials or other studies based on location and eligibility criteria. We advertised for participants who did not have dementia, with no restrictions as to location.

We created an on-line survey using Qualtrics software (2013). For MTurk, the posted task included a brief informed consent after which the participant was linked to the Qualtrics survey. Healthy Minds participants and other local volunteers were contacted individually or through an eNewsletter, with the option of doing the survey online or by paper and pencil. TrialMatch posted a description of the research with a link to the Qualtrics survey.

By definition, a crosswalk is a within subject design. Using Qualtrics capabilities, blocks of questions were presented in counter-balanced order, separated by unrelated questions (Table S1). For separators, we used sets of three vocabulary items (Shipley, 1940), with these items also serving to screen for those who might be responding randomly or who were not sufficiently literate in English. In all, 8% of those who completed the survey on-line were excluded for missing one or more vocabulary words.

The sample for depression (N = 1061) is described in Table 1, and the sample for subjective health (N = 1065) is described in Table 2. Each time we re-posted the MTurk task, in order to move it to the top of the list of tasks, we varied what was included in the Qualtrics survey. Consequently, for the most part different MTurk and Healthy Minds participants are included in the depression and subjective health samples, although the TrialMatch participants, who became available during the latter part of the recruitment period, are the same in both samples. Approximately half of MTurk respondents were from outside the U.S., compared to less than 1% of the other participants.

Measures

Depression—Two measures of depression were used, each constituting a block of questions in the Qualtrics survey: the Center for Epidemiologic Studies-Depression (CES-D) scale (Radloff, 1977) and the Cambridge Mental Disorders of the Elderly Examination (CAMDEX; Roth et al., 1986). The CES-D score is the sum of 20 items, each answered on a 0 to 3 scale to indicate the frequency of experiencing the symptom during the past week, with four positive-worded items reversely scored in calculating the sum score, such that a higher score indicates more depressive symptoms. Previous research on the CES-D supports a four-factor structure representing depressed mood, psychomotor retardation and somatic symptoms, well-being (the four reverse-scored items), and interpersonal difficulties (Radloff, 1977; Gatz, Pedersen, Plomin, Nesselroade & McClearn, 1992).

The CAMDEX has seen more use in Europe than in the United States. As used in the Danish twin studies in IGEMS, there are 17 items assessing frequency of depressive symptoms. Most items are on a 3-point scale (from 1 = "no" to 3 = "most of the time"), with two "yes"/"no" items. Items are reversed such that a higher score indicates more depressive

symptomatology, with a possible sum from 17 to 49. The items comprise two factors, affect, including sad mood and lack of well-being, and somatic, including cognitive difficulties, slowing, and loss of energy (McGue& Christensen, 1997).

Subjective Health—In studies of older adults, generally the most common item used to assess subjective health [SH] is some version of "How would you rate your overall health? [SRH]." Researchers may add additional items that allow participants to indicate how their health affects their daily activities ("Is your health preventing you from doing things you like to do? [Activity]") or to rate their health compared to others ("Compared to others your age, how would you rate your overall health? [Comparative]"). Typical of subjective health questions in gerontological research, however, quite different response scales were used across the IGEMS studies, ranging from 3 to 7 response categories. Notably, even when the number of response categories was the same, the semantic labels might differ; e.g., with five response categories, sometimes 2="good" while other times 3="good". Therefore, to capture this variability in response scales, 10 subjective health items were included in the crosswalk study. Items were grouped into four blocks (Block A, B, C, and D, as seen in Table S1 and Table 3), each block including SRH plus Comparative and/or Activity items.

Analytic Methods

We initially carried out common harmonization procedures that did not require the crosswalk but used the crosswalk sample to judge their success. Two different rational approaches were applied to the two constructs: For depression, we extracted semantically matched items from each scale being harmonized. For example, the CES-D includes "I felt depressed;" the CAMDEX includes "Do you at the moment feel sad, depressed, or miserable?" The CES-D includes "I had trouble keeping my mind on what I was doing;" the CAMDEX includes "Do you find it more difficult to concentrate than usual?" The CES-D includes "I felt lonely;" the CAMDEX includes "Have you felt lonely lately?" Depression scores were created from these semantically comparable items by summing standardized item scores and transforming to T scores.

For subjective health, we recoded response options for each SRH, Comparative, or Activity item to a common response scale with the same number of categories, using the semantic labels to create a basis for matching one response scale to another. These rational approaches do not entail use of any data; the investigator looks at the scales to be harmonized and thinks logically how best to recode them so that the data can be combined.

Configural approaches were applied based on factor analysis: For depression, we used factors already reported in the literature; we did not conduct any new factor analyses. Thus, we scored the CAMDEX for the previously established affect (9 items) and somatic (7 items) factors and scored the CES-D for the previously established depressed mood factor (7 items) and the psychomotor retardation and somatic symptoms factor (7 items). Item scores were summed, and the subscale score standardized. For subjective health, factor analyses were conducted in the crosswalk sample in combinations of SH items as used by the various IGEMS studies, with factor invariance evaluated across gender, age, and participant source

(Davidov et al., 2008; Horn & McArdle, 1992; Reise & Widaman, 1999), and SH scores were constructed, weighting the items based on the factor analysis results.

The crosswalk sample was used to evaluate how well rational and configuration harmonization worked by comparing scores on the rationally recoded scales to one another and comparing scores on the configurally constructed scales to one another.

We then applied empirical harmonization methods that took advantage of participants' fully completing all measures of each construct. Rasch IRT modeling is commonly used for harmonizing alternative measures of the same construct or long and short forms of the same measure. A key feature is that IRT defines a scale for the underlying latent variable that is being measured by a set of items. Because items are calibrated with respect to this same scale, one can calculate co-calibrated total scores for participants on each of two measures that represent a common latent trait. We conducted the Rasch IRT using Winsteps (Version 3.72.3; Chicago, Illinois). We tested whether model fit in applying the IRT random equivalence equating method, or rating scale model (RSM), was comparable to the partial credit model (PCM); if the fit indices were comparable, we used RSM since it is relatively more parsimonious. RSM assumes that responses to the item categories reflect an underlying ordered continuum that is the same across all items whereas PCM does not make this strict assumption.

For the CES-D and CAMDEX, first a Rasch analysis was conducted separately on each test to obtain 'person ability estimates', or latent trait score estimates, indicating an individual's standing on an underlying depressive symptoms continuum based on their item response patterns. To transform CAMDEX to the scale of CES-D, we calculated rescaling parameters from the means and standard deviations of latent trait score estimates for each measure. We repeated the Rasch analysis for CAMDEX with the inclusion of the rescaling parameters to evaluate if the same 'frame of reference' was achieved as the CES-D. Test characteristic curves were then calculated to achieve a conversion table where raw scores from each test are matched to the same latent trait score value, using interpolation in order to present integer scores. (see Supplement Section I.a.)

Random equivalence equating methods were also applied for self-rated health. In this case, a Rasch analysis was conducted using the partial credit model for each question block (Block A, B, C and D) given the variety of rating scales used across the items (cf. Velozo et al., 2007). The rescaling transformations and crosswalks achieved were conducted in reference to Block A. (see Supplement Section I.b.)

Results

Depression

Sample characteristics, means and standard deviations for depression scores for each age by gender subgroup are shown in Table 1. Cronbach's alpha was .92 for CES-D and .91 for CAMDEX. There were no significant differences on either measure by order of presentation of the two measures or by gender. There were significant differences by age stratum on both measures, with those under age 60 scoring higher than those aged 60 and older, and

significant differences on both measures for data source, with Mechanical Turk respondents significantly higher on both measures compared to either TrialMatch or all of the other sources. However, there was no interaction between age stratum and source (Table S2). The CES-D and the CAMDEX were correlated .87 for the total sample.

Rational—Eight item pairs were identified based on the similarity of item wording in the CAMDEX and CES-D (Table S3). These items were used to create CAMDEX-8 and CESD-8. Means and standard deviation for CAMDEX-8 and CESD-8 T scores are shown in Table 1. Cronbach's alpha was .86 for the CAMDEX-8 and .86 for the CESD-8. The pattern of differences by age, gender, and source was the same as for the full scales. Because of the crosswalk sample, we can determine how well this rational approach works. Phi coefficients between item pairs ranged from .54 to .94. The CAMDEX-8 and CESD-8 correlated .86 with one another.

Configural—Means and standard deviations for affect and somatic subscales for CAMDEX and CES-D are shown in Table 1. The pattern of differences by age, gender, and source was the same as for the full scales. Because of the crosswalk sample, we can determine how well this configural approach works. The CAMDEX and CES-D affect subscales correlated .82 with one other, while the CAMDEX and CES-D somatic subscales correlated .71 with one other (Table S4).

Empirical—We applied IRT to the CES-D and CAMDEX items using both RSM and PCM. As the global root mean square error (RMSE) and relative amounts of explained and unexplained variances are essentially identical between the two models, we proceeded with the more parsimonious RSM model despite the significant chi-squared difference tests. Model comparisons are in Supplement (Table S5). Under the RSM model, the CES-D measure had a person reliability of 0.92 and the CAMDEX 0.89, acceptable infit and outfit mean-square statistics at the item level (Wright, Linacre, Gustafson, & Martin-Löf, 1994; Linacre, 2012). The average CES-D item infit was 1.02 and for the CAMDEX it was 1.00. CES-Dinfit and outfit mean-square statistics at the item level were below 1.7 for 19 of 20 items, as recommended for clinical scales (Wright et al., 1994), although values between 1.5–2.0 do not weaken measurement (Linacre, 2012). CES-D item 11 (*my sleep was restless*) had an outfit mean-square statistics of 1.76. All 17 CAMDEX items show infit and outfit mean-square statistics below 1.7..

Figure 1 maps 'item difficulties' (meaning how likely someone was to endorse depression) of CES-D and CAMDEX. The items least likely to be endorsed included sometimes feeling life was not worth living (CAMDEX) and having crying spells (CES-D). Items most likely to be endorsed included restless sleep (CES-D) and preferring to be on one's own (CAMDEX). Some items that would be expected to be similar were indeed similar in endorsement (i.e., items measuring being worried or bothered; feeling happy [reverse scored]; feeling hopeful or optimistic [reverse scored]). However, other rationally similar items did not share rates of endorsement (i.e., items measuring sense of failure or worthlessness, or feeling depressed). After bringing the items to a common scale, every point on the raw scale of CES-D was matched with corresponding raw scores on CAMDEX through the linked latent trait score values (Tables S6 and S7). The resulting conversion

table can be used to conduct pooled analyses across the different IGEMS studies (Table S8). For example, a CES-D raw score of 10 corresponds to a CAMDEX raw score of 23.3 because both of these raw scores were associated with the same latent trait score values. Table 1 shows scores on the harmonized depression score in CES-D units. We then examined conversion graphs by age group, gender, and source of the data (Table S9, Figure S1). The curves were highly overlapping, indicating that separate crosswalk tables are not required for different subgroups of respondents.

Subjective Health

Descriptive statistics for each item are presented in Table 3. No significant effects of order of presentation or gender were detected on the 10 subjective health items. Correlations with age were generally small (less than −.20) but significant. Responses to the SRH and Comparative items differed by source, with healthier average responses in the MTurk sample, but there were no significant interactions between source and age group. Harmonization methods are discussed in order from least to most effective.

Rational—We aligned the response scales for similar items based on the semantic labels associated with the numeric responses. If the transformation were successful, then the same individual should have approximately the same score on every version of that item. For example, for SRH, all responses were translated to a 7-point scale, with a common score of "2" always associated with the label "good", a score of "6" with the label "bad", etc. Descriptive statistics for original and translated items are reported in Table 3. Mean scores on the translated items differed significantly for SRH and Comparative ($p < .01$) and all but one comparison was significant for Activity, indicating that rational aligning of response scales was not a satisfactory approach to harmonization.

Empirical—We conducted initial comparisons of RSM to PCM Rasch model across the full set of 10 items. Model comparisons indicated that the RSM model resulted in severe loss of fit coupled with a higher root means squared residual (see Supplement Table S5). Hence, we fitted a Rasch partial credit model analysis to the SRH items by question block using random equivalence equating. Upon examining the separate Rasch analyses by question block, outfit for Blocks A and B (each containing 3 items) indicated the presence of possible outliers for two items (one in each set) where outfit mean-square statistics exceeded 1.7. Nineteen individuals were dropped from all further analyses based on Mahalanobis distance criteria for the full set of subjective health items. Outfit issues were resolved for Block A, but not Block B, in which one item retained an outfit of 2.8. All other subjective health items had infits ranging from 0.74 to 1.12 and outfits ranging from 0.64 to 1.17, suggesting good psychometric characteristics. Person reliability (Cronbach α) for each question block ranged from 0.73 to 0.85.

Items were brought to a common scale for the latent trait (Figure S2). The Comparative items all clustered very closely in terms of 'difficulties', i.e., likelihood to endorse poorer health. The SRH items and the Activity items had similar rates of endorsement, with the exception respondents were more likely to indicate poorer health on SRH-A and SRH-D, but less likely to indicate poorer health on Activity-A. After bringing the items to a common

scale, every possible point on the raw scale of Block A was matched with the corresponding raw scores on Blocks B, C, and D through the linked person measures, creating a conversion table in Block A units (Table S11). However, crosswalk results by age group or sex did not produce identical raw score conversions at the upper end of the person measure distributions (worse health), especially where data are sparse (Figure S3). Thus, a separate conversion table may be required for different subgroups of respondents depending on item composition.

Configural—In order to carry out a configural analysis, we re-created five different subjective health scales each comprised of a combination of SRH, Comparative, and Activity items, corresponding in content to the subjective health scales found in the different IGEMS studies (see Table 4). Four of these scales each included three items; e.g., scale #3 included SRH-A, Activity-B, and Comparative-A. Scale #1 had only two items. Correlations among SH items were remarkably similar across SH scales for both younger (age < 60) and older (age ≥ 60) adults (Table S10). We conducted separate factor analyses of the SH items within each scale, maintaining the original response scales. Results indicate an impressive uniformity in factor loadings and variance explained across the five different scales (see top of Table 4). Factor structure did not differ across gender or age group.

We then created five new subjective health scales corresponding to the same sets of items in each scale but using the factor loadings rather than original item scoring. With the crosswalk sample, we could examine correlations for the summed factor scores across the five different scales for both younger and older adults (see bottom of Table 4). Correlations across these subjective health scales created from summed factor loadings are all .82 or greater, and the average correlation is .89 for both younger and older adults. Findings were marginally less strong for the two-item scale than three-item scales. These correlations allow us to conclude that the latent subjective health factor constructed from the different items and response scales from each of the five scales tap the same latent construct. We further note that the difference between using unit weighting, a common practice, and factor loadings to create subjective health scales was non-significant.

Discussion

Combining data across studies to permit pooled analyses frequently requires investigators to find a way to co-calibrate or harmonize different measures of the same trait or construct. We undertook a comparison of different approaches to retrospective harmonization, for measures of depressive symptoms and subjective health. Using results from a crosswalk sample, for each measure we could observe how well rational and configural harmonization worked, and how much is gained by recruiting the crosswalk sample required to support empirical approaches.

For depression, rational harmonization entailed constructing short scales with semantically comparable items selected from the scales to be harmonized. When evaluated using the crosswalk sample, this approach worked well, with very similar correlations between the two short scales compared to inter-scale correlations from the IRT. However, correlations at the item level were not consistently strong, and, as the person-item map from the IRT

analyses showed, not all of the items identified as rationally comparable across scales had corresponding item difficulties, and thus were not as comparable to one another as the rational approach assumed.

For subjective health, there were essentially identical items, but the response scales were the focus of rational harmonization. When evaluated using the crosswalk sample, matching across the content of the anchors proved unsuccessful. The biggest problem for harmonization was asymmetric response options, e.g., “excellent”, “very good”, “good”, “poor”. Here, participants who were attending to the semantic labels would answer differently from participants who were focused on the numerical scale.

Configural harmonization derives a similar factor or factors from each respective measure; standardizing these factor scores makes it possible to pool studies. For depression, the configural approach was somewhat less effective than the other approaches, especially for the somatic factor, likely because the CAMDEX somatic subscale comprises items tapping anhedonia and psychomotor retardation while the CES-D somatic subscale includes both somatic symptoms of depression and psychomotor retardation items.

Configural harmonization worked well across the different subjective health items. The assumption of an underlying construct tapped by a few items supports the generally accepted practice of standardizing and summing items to create scales within studies and then pooling data across studies.

Empirical harmonization uses IRT to establish a common metric for translating between two measures of the same construct. We applied Rasch IRT models using the random equivalence equating method which permits equating when no common items exist. For depression, the empirical approach provided a crosswalk that was robust to gender, age group, and the source of the participants (Mechanical Turk or other registries).

On the other hand, for subjective health, empirical approaches did not offer a better alternative. In particular, it appeared that giving respondents a choice between “yes” and “no” made respondents less likely to endorse the statement that their health adversely affected their daily activities. Further, results were not robust to age and sex. Both rational and empirical approaches assume consistency in how participants use rating scales in their responses; absent that consistency, any item-level attempt at harmonization will be flawed.

Given the additional work required to obtain a crosswalk sample and conduct IRT analyses, when is this approach important to undertake? With a crosswalk, one can be more confident that two scales are measuring the same thing. Having empirically harmonized scales could be particularly important to finding the same associations among scales with a mutual set of covariates, among traits in a mediational pathway, or for outcomes such as frailty or how many more years to live. Moreover, empirical harmonization may be especially helpful in longitudinal studies where different items or even completely different instruments were used at different waves (see McArdle, Grimm, Hamagami, Bowles & Meredith, 2009). Failure to harmonize could lead to flawed conclusions about change over time. Finally, if access to resources, treatments, interventions, or opportunities hinge on particular cutpoint values, then empirical harmonization is essential in order to create a crosswalk that

translates across the respective measures. However, we also learned that IRT cannot overcome extreme differences in response scales.

There are a number of limitations of the study: First, due to the age range, we found it necessary to use various methods and sources to recruit crosswalk participants. However, the crosswalk findings held across source of participants. Second, the recruitment methods were not designed to result in a representative sample, since the key comparison is within subject. Moreover, calibrating items with IRT calibration requires a large and heterogeneous but not a representative sample (Hambleton & Jones, 1993). Third, the harmonization was conducted with an English-speaking sample and does not address cultural or translation differences that could affect harmonization. Fourth, the data collection was cross sectional. Therefore, we cannot report how well harmonization will hold up longitudinally. Fifth, although we focused on context created by response options, for example, showing that subjective health items demonstrated that “within-item” context affected self-reporting, we did not focus on the context in which questions were asked, i.e., inter-item context (Schwarz, 1999). Finally, we would emphasize that in this report, we are providing examples, but not a comprehensive test of all methods of harmonization.

In conclusion, we carried out various approaches to harmonization with two different constructs, using a crosswalk sample to perform empirical harmonization analyses and to test how well rational and configuration harmonization succeeded. The results demonstrated that not only item content but also response scales affect harmonization, and that semantically collapsing response options in order to achieve a harmonized scale may not always reflect how people actually use those respective scales, an outcome that can only be established with a crosswalk sample. We discovered that the optimal method of data harmonization depends on the measurement characteristics of the scales being harmonized, such that a single prescription will not be sufficient. Whatever approach is used, a theoretically strong and coherent latent construct is necessary to support harmonization.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

IGEMS is supported by the National Institutes of Health Grant No. R01 AG037985.

Appendix 1

Members of the consortium on Interplay of Genes and Environment across Multiple Studies (IGEMS) include: Nancy L. Pedersen(Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden, and Department of Psychology, University of Southern California, Los Angeles, CA), Kaare Christensen (Department of Epidemiology, University of Southern Denmark, Odense, Denmark), Anna Dahl(Institute of Gerontology, School of Health Sciences, Jönköping University, Jönköping, Sweden), Deborah Finkel (Department of Psychology, Indiana University Southeast, New Albany, IN), Carol E. Franz (Department of Psychiatry, University of California, San Diego, La

Jolla, CA), Margaret Gatz (Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden, and Department of Psychology, University of Southern California, Los Angeles, CA), Briana N. Horwitz (Department of Psychology, California State University, Fullerton, CA), Boo Johansson (Department of Psychology, University of Gothenburg, Gothenburg, Sweden), Wendy Johnson (Department of Psychology and Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK), William S. Kremen (Center of Excellence for Stress and Mental Health, VA San Diego Healthcare Center, La Jolla, CA, and Department of Psychiatry, University of California, San Diego, La Jolla, CA), Michael J. Lyons (Department of Psychology, Boston University, Boston, MA), Matt McGue (Department of Psychology, University of Minnesota, Minneapolis, MN), Jenae M. Neiderhiser (Department of Psychology, The Pennsylvania State University, University Park, PA), Inge Petersen (Department of Epidemiology, University of Southern Denmark, Odense, Denmark), and Chandra A. Reynolds (Department of Psychology, University of California-Riverside, Riverside, CA).

References

Bath PA, Deeg D, Poppelaars J. The harmonisation of longitudinal data: a case study using data from cohort studies in The Netherlands and the United Kingdom. Ageing & Society. 2010; 30:1419–1437.

Bauer DJ, Hussong AM. Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. Psychological Methods. 2009; 14:101–125. [PubMed: 19485624]

Buhrmester M, Kwang T, Gosling S. Amazon's Mechanical Turk: A new source of inexpensive, yet high quality, data? Perspectives on Psychological Science. 2011:3–5. [PubMed: 26162106]

Choi, SW.; Podrabsky, T.; McKinney, N.; Schalet, BD.; Cook, KF.; Cella, D. PROSetta Stone® Analysis Report: a Rosetta Stone for Patient Reported Outcomes. Volume 1. Chicago, IL: Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University; 2012. from http://www.prosettastone.org/AnalysisReport/Pages/default.aspx [Retrieved May 15, 2013]

Curran PJ, Hussong AM. Integrative data analysis: The simultaneous analysis of multiple data sets. Psychological Methods. 2009; 14:81–100. [PubMed: 19485623]

Curran PJ, Hussong AM, Cai L, Huang W, Chassin L, Sher KJ, Zucker RA. Pooling data from multiple longitudinal studies: The role of item response theory in integrative data analysis. Developmental Psychology. 2008; 44:365–380. [PubMed: 18331129]

Davidov E, Schmidt P, Schwartz SH. Bringing values back in: The adequacy of the European Social Survey to measure values in the 20countries. Public Opinion Quarterly. 2008; 72:420–445.

Embretson, SE.; Reise, SP. Item Response Theory for Psychologists. Mahwah, New Jersey: Lawrence Erlbaum Associates; 2000.

Finkel D, Pedersen N, McGue M. Genetic influences on memory performance in adulthood: Comparison of Minnesota and Swedish twin studies. Psychology and Aging. 1995; 10:437–446. [PubMed: 8527064]

Fortier I, Doiron D, Wolfson C, Raina P. Harmonizing data for collaborative research on aging: Why should we foster such an agenda? Canadian Journal on Aging. 2012; 31:95–99. [PubMed: 22373784]

Gatz M, Pedersen NL, Plomin R, Nesselroade JR, McClearn GE. The importance of shared genes and shared environments for symptoms of depression in older adults. Journal of Abnormal Psychology. 1992; 101:701–708. [PubMed: 1430610]

Hambleton RK, Jones RW. Comparison of classical test theory and item response theory and their applications to test development. Educational Measurement: Issues and Practice. 1993; 12(3):38–47.

Hamilton CM, Strader LC, Pratt JG, Maiese D, Hendershot T, Kwok RK, Hammond JA, Huggins W, Jackman D, Pan H, Nettles DS, Beaty TH, Farrer LA, Kraft P, Marazita ML, Ordovas JM, Pato CN, Spitz MR, Wagener D, Williams M, Junkins HA, Harlan WR, Ramos EM, Haines J. The PhenX Toolkit: Get the most from your measures. American Journal of Epidemiology. 2011; 174:253–260. [PubMed: 21749974]

Horn JL, McArdle JJ. A practical guide to measurement invariance in aging research. Experimental Aging Research. 1992; 18:117–144. [PubMed: 1459160]

Johnson W, Bouchard TJ Jr, Krueger RF, McGue M, Gottesman II. Just one g: Consistent results from three test batteries. Intelligence. 2004; 32:95–107.

Jones RN, Fonda SJ. Use of an IRT-based latent variable model to link different forms of the CES-D from the Health and Retirement Study. Social Psychiatry and Psychiatric Epidemiology. 2004; 39:828–835. [PubMed: 15669664]

Lee J, Zamarro G, Phillips D, Angrisani M, Chien S. RAND-Harmonized ELSA Data Documentation, Version: A. 2011 from https://mmicdata.rand.org/meta/codebooks/RH_ELSA_Codebook.pdf.

Linacre JM. A User's Guide to WINSTEPS® MINISTEP Rasch-Model Computer Programs. Program Manual 3.75.0. 2012 ISBN 0-941938-03-4.

McArdle JJ, Grimm KJ, Hamagami F, Bowles RP, Meredith W. Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. Psychological Methods. 2009; 14(2):126–149. [PubMed: 19485625]

McGue M, Christensen K. Genetic and environmental contributions to depression symptomatology: Evidence from Danish twins 75 years of age and older. Journal of Abnormal Psychology. 1997; 106:439–448. [PubMed: 9241945]

Molenaar D, van der Sluis S, Boomsma DI, Haworth CM, Hewitt JK, Martin NG, Plomin R, Wright MJ, Dolan CV. Genotype by environment interactions in cognitive ability: a survey of 14 studies from four countries covering four age groups. Behavior Genetics. 2013; 43(3):208–219. [PubMed: 23397253]

Qualtrics software, Version 45634 of the Qualtrics Research Suite. Copyright © 2013 Qualtrics, Provo, UT, USA. http://www.qualtrics.com.

Paolacci G, Chandler J, Ipeirotis PG. Running experiments on Amazon Mechanical Turk. Judgment and Decision Making. 2010; 5:411–419.

Pedersen NL, Christensen K, Dahl A, Finkel D, Franz CE, Gatz M, Horwitz BN, Johansson B, Johnson W, Kremen WS, Lyons MJ, Malmberg B, McGue M, Neiderhiser JM, Petersen I, Reynolds CA. IGEMS: The Consortium on Interplay of Genes and Environment across Multiple Studies. Twin Research and Human Genetics. 2013; 16:481–489. [PubMed: 23186995]

Radloff L. A self-report depression scale for research in the general population. Applied Psychological Measurement. 1977; 1:385–401.

Reise SP, Widaman KF. Assessing the fit of measurement models at the individual level: A comparison of item response theory and covariance structure approaches. Psychological Methods. 1999; 4:3–21.

Roth M, Tym E, Mountjoy CQ, Huppert FA, Hendrie H, Verma S, Goddard R. CAMDEX. A standardised instrument for the diagnosis of mental disorder in the elderly with special reference to the early detection of dementia. British Journal of Psychiatry. 1986; 149:698–709. [PubMed: 3790869]

Schwarz N. Self-reports: How the questions shape the answers. American Psychologist. 1999; 54:93–105.

Sharp ES, Suthers KM, Crimmins E, Gatz M. Does “No” mean “Sometimes”? How older adults respond to the same depression symptoms with different response formats. Clinical Gerontologist. 2009; 32:371–378. [PubMed: 20563229]

Shipley WC. A self-administering scale for measuring intellectual impairment and deterioration. Journal of Psychology: Interdisciplinary and Applied. 1940; 9:371–377.

van Buuren S, Eyres S, Tennant A, Hopman-Rock M. Assessing comparability of dressing disability in different countries by response conversion. European Journal of Public Health. 2003; 13(3 Supplement):15–19. [PubMed: 14533743]

Velozo CA, Byers KL, Wang YC, Joseph BR. Translating measures across the continuum of care: Using Rasch analysis to create a crosswalk between the Functional Independence Measure and the Minimum Data Set. Journal of Rehabilitation Research & Development. 2007; 44:467–478. [PubMed: 18247243]

Wright BD, Linacre JM, Gustafson J-E, Martin-Löf P. Reasonable mean-square fit values. Rasch Measurement Transactions. 1994; 8(3):370.

Figure 1.

Map representing item difficulties (i.e., rates of endorsement) of CES-D and CAMDEX items.

Exp Aging Res. Author manuscript; available in PMC 2016 March 01.

Table 1

Sample characteristics and descriptives for the depression sample for total CAMDEX and CES-D scores, semantically equivalent CAMDEX and CES-D scales, and factorally informed subscales.

	Men <60 N = 213	**Men 60+ N = 194**	**Women <60 N = 342**	**Women 60+ N = 312**	**Total N = 1061**
Age (years)	39.41 (8.88)	70.23 (8.00)	44.25 (9.25)	67.86 (6.61)	54.96 (15.68)
	Range 30–59	Range 60–98	Range 30–59	Range 60–94	
Bachelor's degree or higher (%)	62.92	73.71	65.50	67.31	67.01
Difficulties meeting expenses (%)	32.86	10.31	26.32	10.90	20.17
Source (N)					
MTurk	192	32	215	27	466
Healthy Minds and other volunteers	5	46	7	61	119
TrialMatch	16	116	120	224	476
CAMDEX (raw score)	25.68 (6.76)	23.4 (5.2)	25.46 (6.86)	23.16 (5.73)	24.47 (6.36)
CES-D (raw score)	14.54 (11.40)	10.30 (8.37)	14.36 (11.38)	10.53 (9.31)	12.53 (10.48)
CAMDEX-8 (T score)	51.22 (7.61)	48.87 (5.71)	51.08 (7.67)	48.69 (6.46)	50.00 (7.08)
CESD-8 (T score)	51.09 (7.89)	48.58 (5.94)	50.97 (7.55)	49.07 (6.65)	50.00 (7.17)
CAMDEX					
Affect (T score)	52.33 (10.80)	47.82 (7.67)	51.69 (10.96)	47.91 (8.89)	50.00 (10.00)
Somatic (T score)	50.97 (10.61)	49.46 (9.39)	51.03 (10.51)	48.54 (9.17)	50.00 (10.00)
CES-D					
Affect (T score)	50.91 (10.69)	47.62 (7.88)	52.12 (11.25)	48.53 (8.62)	50.00 (10.00)
Somatic (T score)	51.22 (10.48)	49.05 (9.13)	50.87 (10.09)	48.80 (9.94)	50.00 (10.00)
Harmonized depression score	14.68 (12.10)	10.68 (8.98)	14.26 (12.21)	10.45 (10.03)	12.57 (11.19)

Notes: CAMDEX-8 and CESD-8 = eight items from each scale identified as semantically comparable, with scores shown as the sum of the standardized items transformed into a T score. For CAMDEX, items 6, 8, 11, 13, 15, 16, and 17 are included. For CES-D, items 20, 5, 12, 1, 6, 9, and 8, respectively, are included. Factorally informed subscales are shown as the sum of items on the factor, standardized and transformed into a T score. The Affect subscale of the CAMDEX includes items 1, 11, 12, 13, 14, 15, 16, 17, and 18. The Somatic subscale of the CAMDEX includes items 3, 4, 5, 6, 8, 9, and 10. Item 7 is not included on either subscale. The depressed mood subscale of the CES-D includes item 3, 6, 9, 10, 14, 17, and 18. The psychomotor retardation and somatic symptoms subscale of the CES-D includes item 1, 2, 5, 7, 11, 13, and 20. The other CES-D items comprise the well-being and interpersonal difficulties subscales. The harmonized depression score is the conversion of the CAMDEX into CES-D units.

Table 2

Sample characteristics and descriptives for the subjective health sample.

	Men <60 N = 249	Men 60+ N = 200	Women <60 N = 317	Women 60+ N = 299	Total N = 1065
Age (years)	39.94 (8.82)	70.27 (7.92)	45.01 (9.45)	67.73 (6.51)	54.93 (15.50)
Bachelor's degree or higher (%)	61.04	74.50	57.72	70.91	65.35
Source (N)					
MTurk	227	22	191	13	453
Healthy Minds and other volunteers	6	61	6	64	137
TrialMatch	16	117	120	222	475

Table 3

Subjective health items and descriptive statistics for original items and for items after semantic transformation of response options.

Block	Response Options	Mean (SD) for original scoring	Mean (SD) for semantically transformed item
SRH: *How would you rate your overall health?*			
A	1=good, 2=reasonable, 3=bad	1.28 (0.51)	2.56 (1.02)
B	1=very good, 2=good, 3=acceptable, 4=bad, 5=very bad	1.95 (0.82)	2.20 (1.29)
C	1=excellent, 2=very good, 3=good, 4=fair, 5=poor	2.41 (0.97)	1.74 (1.07)
D	1=very good, 2=good, 3=rather good, 4=average, 5=rather bad, 6=bad, 7=very bad	2.28 (1.23)	2.28 (1.23)
Activity: *Is your health condition is preventing you from doing things you like to do?*			
A	1=no, 2=yes	1.31 (0.46)	2.25 (1.85)
B	1=not at all, 2=partly, 3=to a great extent	1.47 (0.61)	1.94 (1.22)
C	1=no, never, 2=no hardly ever, 3=yes now and then, 4=yes nearly always, 5=yes always	2.17 (0.96)	2.17 (0.96)
Comparative: *Compared to others your age, how would you rate your overall health* (A&B) *I am as healthy as anyone I know* (D)			
A	1=better, 2=about the same, 3=worse	1.57 (0.64)	1.57 (0.64)
B	1=excellent, 2=good, 3=fair, 4=poor	1.88 (0.73)	1.19 (0.45)
D	1=definitely true, 2=mostly true, 3=don’t know, 4=mostly false, 5=definitely false	2.13 (0.97)	1.35 (0.68)

Notes: A, B, C, D represent question blocks as presented to participants in the Qualtrics survey: all items labeled A appeared in question block A.

Table 4

Factor loadings of items in each subjective health scale and correlations between factors

	Scale[a]				
Variables	**#1**	**#2**	**#3**	**#4**	**#5**
Factor Loadings					
SRH	--	.91	.87	.90	.87
Comparative	.86	.83	.81	.83	.86
Activity	.86	.78	.79	.82	.80
Variance explained	73.5%	71.5%	67.3%	71.7%	71.6%
Correlations[b]					
#1		.85	.84	.85	.86
#2	.85		.96	.95	.89
#3	.82	.96		.92	.86
#4	.85	.95	.92		.89
#5	.86	.89	.86	.89	

Notes:

[a] Scale 1 = Activity-A and Comparative-B. Scale 2 = SRH-D, Activity-B, and Comparative-A. Scale 3 = SRH-A, Activity-B, and Comparative-A. Scale 4 = SRH-B, Activity-C, and Comparative-A. Scale 5 = SRH-C, Activity-B, and Comparative-D.

[b] Correlations for participants aged 60 and older are presented above the diagonal; correlations for participants aged less than 60 are below the diagonal.