Methods Research Report

Harmonization of Cognitive Measures in Individual Participant Data and Aggregate Data Meta-Analysis

Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • **www.ahrq.gov**

Methods Research Report

Harmonization of Cognitive Measures in Individual Participant Data and Aggregate Data Meta-Analysis

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Contract No. 290-2007-10060-I

Prepared by:
McMaster University Evidence-based Practice Center
Hamilton, Ontario, Canada

Investigators:
Lauren Griffith, Ph.D.
Edwin van den Heuvel, Ph.D.
Isabel Fortier, Ph.D.
Scott Hofer, Ph.D.
Parminder Raina, Ph.D.
Nazmul Sohel, Ph.D.
Hélène Payette, Ph.D.
Christina Wolfson
Sylvie Belleville, Ph.D.

AHRQ Publication No. 13-EHC040-EF
March 2013

This report is based on research conducted by the McMaster University Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2007-10060-I). The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help health care decisionmakers—patients and clinicians, health system leaders, and policymakers, among others—make well-informed decisions and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

This report may be used, in whole or in part, as the basis for development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

Persons using assistive technology may not be able to fully access information in this report. For assistance contact EffectiveHealthCare@ahrq.hhs.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Suggested citation: Griffith L, van den Heuvel E, Fortier I, Hofer S, Raina P, Sohel N, Payette H, Wolfson C, Belleville S. Harmonization of Cognitive Measures in Individual Participant Data and Aggregate Data Meta-Analysis. Methods Research Report. (Prepared by the McMaster University Evidence-based Practice Center under Contract No. 290-2007-10060-I.) AHRQ Publication No.13-EHC040-EF. Rockville, MD: Agency for Healthcare Research and Quality; March 2013. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies and strategies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodological issues in systematic reviews. These methods research projects are intended to contribute to the research base in and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality. The reports undergo peer review prior to their release as a final report.

We welcome comments on this Methods Research Project. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by email to epc@ahrq.hhs.gov.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang M.D., M.P.H.
Director, EPC Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Parivash Nourjah, Ph.D.
Task Order Officer
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Acknowledgments

The researchers at the McMaster EPC would like to acknowledge the following people for their contributions. We are grateful to our Task Order Officer, Parivash Nourjah, for her support and guidance. Members of the Technical Expert Panel were instrumental in the formation of the parameters and goals of this review. We would like to thank the staff of the Maelstrom Research Program, specifically Dany Doiron, for their contributions to the review. We would also like to thank those who worked so conscientiously, retrieving and screening citations, abstracting data, preparing figures, and editing the report: Julianna Beckett, Bryan Cheeseman, Roxanne Cheeseman, Mary Gauld, Meghan Kenny, Leah Macdonald, and Maureen Rice.

Key Informants

Steph van Buuren, Ph.D.
Professor, Social and Behavioural Sciences
University of Utrecht
Utrecht, the Netherlands

Andrea Piccinin, Ph.D.
Associate Professor, Dept. of Psychology
University of Victoria
Victoria, British Columbia, Canada

Christopher Schmid, Ph.D.
Associate Professor of Medicine at Tufts University School of Medicine
Concentration Leader for the Clinical Research Program
Sackler School of Graduate Biomedical Sciences
Boston, MA

Ronald Stolk, Ph.D.
Professor and Department Head, Clinical Epidemiology
University of Groningen
Groningen, the Netherlands

Technical Expert Panel

Steph van Buuren, Ph.D.
Professor, Social and Behavioural Sciences
University of Utrecht
Utrecht, the Netherlands

Joseph Beyene, Ph.D.
Associate Professor, Dept. of Clinical Epidemiology & Biostatistics
McMaster University Chairholder: John D. Cameron Endowed Chair in Genetic Epidemiology
McMaster University
Hamilton, Ontario, Canada

Vincent Ferretti, Ph.D.
Principal Investigator and Senior Scientist, Informatics and Bio-computing Platform
Ontario Institute of Cancer Research
Toronto, Ontario, Canada

John Gallacher, Ph.D.
Reader in Environmental Epidemiology and Director of the Active Age Research Group
Cardiff University
Cardiff, United Kingdom

Peter Granda, Ph.D.
Assistant Director, General Archive Manager, International Archive of Education Data
University of Michigan (Inter-University Consortium for Political and Social Research)
Ann Arbor, MI

George Kelley, DA, FACSM
Professor and Director, Meta-Analytic Research Group
Department of Community Medicine
West Virginia University
Morgantown, WV

Julian Little, Ph.D.
Professor and Chair, Epidemiology & Community Medicine
Canada Research Chair in Human Genome Epidemiology
University of Ottawa
Ottawa, Ontario, Canada

Jack McArdle, Ph.D.
Professor of Psychology and Gerontology
Director of the National Growth and Change Study
University of Southern California
Los Angeles, CA

Harmonization of Cognitive Measures in Individual Participant Data and Aggregate Data Meta-Analysis

Structured Abstract

Objectives. The aim of this study was to identify approaches to statistical harmonization which could be used in the context of summary data and/or individual participant data meta-analysis of cognitive measures and to apply and evaluate these different approaches to cognitive measures from three studies.

Data Sources. MEDLINE®, Embase, Web of Science and MathSciNet with a supplemental search using the Google search engine. The references of relevant articles were also checked and a search for more recent articles that cited the articles already identified as being of interest was undertaken.

Review methods. A two-pronged approach was taken for this environmental scan. First, a search of studies that quantitatively combined data on cognition was conducted. The second component was to identify general literature on statistical methods for data harmonization. Standard environmental scan methods were used to conduct these reviews. The search results were rapidly screened to identify articles of relevance to this review. The references of relevant articles were checked and a search for more recent articles that cited the articles already identified as being of interest was undertaken.

Results. Three general classes of statistical harmonization models were identified: (1) standardization methods (e.g., simple linear-, Z-transformations, T-scores, and C-scores); (2) latent variable models; and (3) multiple imputation models. Cross-sectional data from three studies including 9,269 participants were included in the applied analyses to examine the relationship between physical activity and cognition. A harmonization process was undertaken to determine the combinability of data across studies. The latent variable analysis underscored the difficulty harmonizing these cognition data. In general consistency was found among the statistical harmonization methods; however, there was some evidence that heterogeneity can be masked when specific standardization methods were used.

Conclusions. This study provides empirical evidence to inform methods of combining complex constructs using aggregate data (AD) or individual participant data meta-analysis. The results underscore that very careful consideration of inferential equivalence needs to be undertaken prior to combining cognition data across studies. Of the three methods of statistical harmonization for cognition data, T-score standardization is the least desirable compared with the centered score method or latent variable methods. Finally, assessment of the assumptions underlying statistical harmonization is not possible without some individual-level data which are required to assess the potential for bias in combining complex outcomes using AD meta-analysis.

Contents

Tables

Figures

Appendixes

Report Outline

There are occasions when researchers would like to combine the results of constructs that are measured on different scales, such as cognitive measures, depression, and quality of life. Combining data measuring these complex constructs can be particularly challenging, and a rigorous approach as well as specialized methods of harmonization, including statistical harmonization, are required. The use of these methods of harmonization in the conduct of systematic reviews is virtually nonexistent. In this report, we were particularly interested in combining measures of memory across studies. As there are currently no guidelines available to determine the best way to combine these types of measures, we undertook a series of projects with the underlying objective to better understand the issues around the statistical harmonization of cognitive measures.

Although there were many commonalities in cognitive measures in the data sets, they were measured using different instruments. The first study objective was to identify approaches to statistical harmonization which could be used in the context of aggregate data and/or individual participant data meta-analysis of cognitive measures. To meet this objective, two environmental scans were conducted. The first, an environmental scan of meta-analyses including cognitive measures, was conducted to identify the types of methods that have been used to combine cognition data in meta-analyses. In practice, we found that most studies either restricted their analyses to cognitive measures with a common scale or combined effect sizes across studies. None of the studies formally probed into whether the cognitive measures should be harmonized. To identify more sophisticated methods that would allow the exploration of whether the cognitive data should be combined, a second environmental scan was undertaken to assess what types of statistical harmonization methods were used in the general literature. These methods were assessed with a technical expert panel to determine their applicability harmonizing cognitive measures. The environmental scans are included in "Methods and Results: Environmental Scans to Identify Methods of Combining Cognition Data and Statistical Harmonization Methods (Objective 1)."

Of the methods identified for statistical harmonization available, it was determined that the latent variable method was promising for our problem. Multiple imputation methods also had potential, but because the focus was on summary constructs rather than individual items, it was not applicable to our situation. Before this method of harmonization could be compared with the ones traditionally used in meta-analysis, data sets that were inferentially equivalent had to be created. Like many studies of harms, the designs of the included studies were not randomized controlled trials, thus it is important to consider covariates in our analyses. The process of harmonizing the covariate information is described in "Methods and Results: Process of Preparing Data for Statistical Harmonization (Objective 2)."

Prior to conducting meta-analyses to compare the methods of statistical harmonization it was also necessary to determine whether the cognition measures were inferentially equivalent. This step was not generally taken in meta-analyses of cognition data identified in the environmental scan. Latent variable regression analysis was used to determine if the cognitive measures were measuring one consistent construct across the data sets and thus should be combinable.

The different methods for statistical harmonization in the context of a traditional meta-analysis were then compared. To create combinable measures, we used methods traditionally used to combine cognitive measures, two standardization methods, and the latent variable method identified by the environmental scan and endorsed by the Technical Expert Panel. The objective of these analyses was to examine how well the results of these statistical harmonization methods agree with each other (correlation analysis) and whether they would provide similar results if a traditional meta-analysis were undertaken. In the latent variable analysis to assess inferential equivalence, it was identified that the cognitive measures may not reflect one underlying construct, thus it was also important to see how well these different methods of statistical harmonization identified heterogeneity in the meta-analysis as well. These analyses are described in “Methods and Results: Implementing and Evaluating Three Methods of Statistical Harmonization Applied to Cognitive Measures (Objective 3).”

Finally in the discussion, we present the overall conclusions and the strengths and limitations of the study and the guidance suggested by the study results.

Introduction

Individual Participant Data (IPD) meta-analysis has become an increasingly popular method to combine unique participant data from randomized controlled trials and observational studies.[1,2] IPD meta-analyses increase the power to detect differential treatment effects across individuals in a randomized controlled trial (RCT) and allow for adjustment of confounding factors in the meta-analysis of observational studies, but they are time consuming and costly to conduct. The main advantage of IPD meta-analysis is that researchers can assess the influence of participant-level covariates on all collected outcomes and measured time points of interest, not all of which are reported in the literature.[1] IPD meta-analysis is particularly relevant to comparative effectiveness reviews (CERs) when conducting sub-group analyses and when combining evidence from RCTs and observational studies examining benefits, harms, adherence, or persistence.[3] However, combining individual participant data is scientifically and technically very challenging. Integration or comparison of individual participant data requires the generation of compatible (or harmonized) datasets across studies. The value of these harmonized datasets is necessarily dependent on: (1) the quality of the data collected by individual studies; (2) the potential for studies to create the variables needed to achieve statistical analysis foreseen; and (3) the acceptable level of heterogeneity across study designs and data collection methods. Such heterogeneity can result from a vast range of study-specific characteristics including, for example, the targeted population, sampling frame, tools and standard operating procedures used by the study investigators to collect data, data collection timeline, etc. To combine or compare individual participant data, it is thus important to limit integration of study-specific data to studies constructed upon clinically and methodologically compatible designs and methods. But there is no standard definition of whether studies are similar enough. The decision to combine study data to produce an overall estimate of effect depends on whether a meaningful answer to the scientific question addressed can be obtained. Harmonization is thus to be considered as a process composed of a series of complementary steps which must be applied with rigorous procedures and decisionmaking in order to ensure validity and reproducibility of the harmonization outputs.

However, in many situations a systematic review is carried out where there are multiple methods and tools used to measure the same underlying construct such as cognition or physical function. This situation is very common when one is examining the comparative effectiveness of interventions in diseases such as dementia, depression, or attention deficit hyperactivity disorder. An example that highlights these issues is in the area of vitamin D, cognition, and dementia. Balion, et al. recently conducted a systematic review that examined the association between vitamin D, cognition, and dementia.[4] There have been two systematic reviews[5,6] which suggest that cognitive function was not associated with 25-hyroxyvitamin D (25(OH)D). However, the authors of both reviews decided not to quantitatively pool data to come up with a summary effect because of the potential heterogeneity across studies. Their conclusions were based on whether there were statistically significant associations between 25(OH)D and one specific measure of cognition, the Mini-Mental State Examination (MMSE), although many measures of cognition were reported. Even though the authors did not pool data their qualitative interpretation of data assumed inferential equivalence of exposure and outcome measures.

Part of the difficulty in summarizing this literature is that there are potentially important methodological differences among the studies, such as the type of assay used to measure 25(OH)D, the threshold values used to define vitamin D insufficiency, and the different measures of cognition utilized. To harmonize the threshold values of vitamin D, additional summary data were requested from the authors by Balion to assess the inferential equivalence of the assay to determine whether results of different assay types could be pooled across studies.

As far as outcome measures of cognition were concerned, the challenges were much larger as the primary studies included a multitude of outcome measures that ranged from a general mental status score, such as the MMSE, to a complete neuropsychological battery. Since many studies used the MMSE, the primary meta-analysis was restricted to that subgroup of studies. A sensitivity analysis was used to examine if the relationship differed when other general cognitive measures were included. In these analyses, a difference in the relationship between 25(OH)D and MMSE based on the type of assay used was found; a consistency was found when the analysis focused exclusively on MMSE and when all general measures of cognition were included. In one analysis, potential methodological differences among the studies were examined, and in the second analysis, Balion, et al. assessed if the same construct was measured. To do this, it was assumed that MMSE measured the same general cognition construct among the studies and that other similar tools measured the same construct. When one is restricted to traditional aggregate data (AD) meta-analysis one is limited to subgroup and sensitivity analyses to examine whether the variables are inferentially equivalent across studies. When one has access to IPD, however, a more complex analysis can be undertaken to assess the inferential equivalence before data are pooled or to have the ability to create new inferentially equivalent variables that allow unbiased pooling of the data.

In the current report, we consider, in the context of the harmonization process, two different data processing methods: (1) qualitative harmonization and (2) statistical harmonization. The choice of which method to employ depends on the nature of the measure to be harmonized. On the one hand, the generation of compatible or inferentially equivalent information across studies can involve creating simple study-specific cut-points for a variable like age or troponin I level or combining different response categories across studies to make them compatible. This is what we refer to as qualitative harmonization and is done using processing algorithms that derive data collected by different studies into a common format. Combining data on more complex constructs such as cognition, physical function, depression, or anxiety can be particularly challenging. To harmonize such constructs, it is essential to use specialized methods such as statistical harmonization or processing. With this second method, statistical models are applied to derive common format data. With both methods, a rigorous step by step approach to harmonization will ensure reproducibility of the harmonized databases and greatly increase the quality and precision. The application of such methods of harmonization in the conduct of systematic reviews is virtually nonexistent.

Although statistical methods for harmonization have been proposed, there has not been a comparative assessment of the strengths and weaknesses of these specialized methods as it relates to the conduct of systematic reviews. Such an assessment could help users identify the most appropriate method given a specific context. For this report, an environmental scan was performed on the methods used to combine different complex measures across studies, and then a few of these methods were applied to create combinable derived variables using three large population-based cohort studies. While a rigorous generic harmonization and decisionmaking process was used to select studies and combine data, the present report focuses on the usage of specialized statistical harmonization.

IPD analysis is an area of interest for the Agency for Healthcare Research and Quality (AHRQ), but there are currently no guidelines for its use.[7] A conceptual background on the need for harmonization and methods used to harmonize information across studies, as well as a background on the different types of cognitive measures that are used in CERs is provided.

Data Harmonization

Ensuring data compatibility and inferential equivalence through harmonization allows integrating information from different studies/databases and can thereby permit pooling of data from a large number of studies to obtain statistically valid results. It also allows one to properly explore the similarities and discrepancies across studies, jurisdictions, or countries, and improve the validity and reliability of comparative effectiveness research.

IPD harmonization essentially aims to attain, or improve, compatibility of information collected from similar but independent sources.[8] This can be achieved by making use of different approaches.[9] If we consider when the harmonization process takes place in the lifecycle of studies, we can distinguish between prospective and retrospective harmonization approaches. Under prospective harmonization, study investigators will agree upon common measures and protocols before beginning data collection. Agreement on a core set of common measures and collection procedures prior to data collection will facilitate future integration or comparison of data across standalone studies. However, IPD meta-analysis in comparative effectiveness literature is generally achieved by making use of retrospective harmonization. Retrospective harmonization takes place after collection of study-specific data has been initiated, generally without any attempt to prospectively ensure a certain level of compatibility across studies. This retrospective approach thus necessitates a rigorous documentation of studies participating in the exercise and a meticulous process to harmonize and integrate study-specific data under a common format. Under both prospective and retrospective harmonization, the ultimate potential to integrate information is directly related to the level of heterogeneity amongst study-specific populations, designs, and standard operating procedures used by study investigators to collect data. An explicit harmonization process is generally not undertaken by the systematic reviewers. When combining IPD to achieve meta-analysis, it is thus essential to interpret whether individual variables and measures are similar, both qualitatively and quantitatively. Thus, prior to analysis a process of data harmonization is required.

To achieve IPD harmonization, investigators of any retrospective harmonization initiative will need to follow a series of practical steps. Firstly, once a research question guiding the harmonization initiative has been identified, investigators identify and document the characteristics of participating studies, such as study designs and data access and usage policies, and all relevant information describing samples, data items, and collection methods, such as data dictionaries or codebooks, questionnaires, and standard operating procedures. This documentation allows the identification of sources of study heterogeneity and provides the elements required to achieve proper evaluation of the harmonization potential across studies. Secondly, based on documentation obtained and the scientific aims of the harmonization initiative, variables targeted to serve as reference for data harmonization across studies are selected. A priori selection of variables targeted for harmonization is generally guided by a balance between enabling integration of a significant number of studies to provide the benefits of large sample sizes, while restricting integration to studies providing the lowest level of heterogeneity possible. Finally, following the identification of reference variables and the selection of studies collecting the valid information required to construct these variables, various methodologies can be applied to transform study-specific data items under the target variable format. As we have seen, these include both qualitative harmonization and statistical harmonization methodologies. Qualitative harmonization involves processing study-specific data items using logical algorithms, and is often applied to create dichotomous or categorical variables (e.g., ethnic background, type of cancer, or smoking status). In addition to qualitative harmonization methods, statistical harmonization methods may be used to harmonize complex constructs, such as cognition, and also requires processing study-specific data items to create an inferentially equivalent construct before data pooling can begin. The choice of which harmonization method to use therefore depends on the nature of the measures to be harmonized.

In addition to providing key scientific benefits and increasing potential for cross-national and international collaborations, IPD harmonization enables and encourages secondary use of existing and emerging research infrastructures. There is without any doubt a growing interest from researchers and funding agencies for the development of more efficient IPD harmonization methodologies and resources. These methodologies will be essential to improve the quality and research potential of current IPD harmonization capacities, and therefore support the achievement of the next generation of specialized international research initiatives.

Harmonization of Data on Cognitive Functioning

In the context of comparative effectiveness reviews and technology assessment reviews, one is almost by definition undertaking a retrospective harmonization. When examining pharmacological treatment of dementia, for example, Raina et al. examined three broad categories of pharmacologic treatments: (1) cholinergic neurotransmitter modifying agents, (2) noncholinergic neurotransmitter/neuropeptide modifying agents, and (3) other pharmacological agents.[10] As many studies used a wide variety of cognitive function measures as outcomes, the authors had to determine which measures could be statistically combined without any methodological guidance. In this report 20 different "general" scales were identified and only the most commonly reported MMSE[11] and the Alzheimer's Disease Assessment Scale,[12] were included in quantitative meta-analyses, therefore resulting in a loss of information.

A major step in combining or comparing results across studies involves identifying comparable variables and assessing the potential for harmonizing the data. The similarity of a measure can vary at a number of levels, and even when using the same measure, large operational differences can be found.[13] When considering combining data sets from across the world, these differences can be magnified. Regardless of whether the same measure has been used, differences are inevitably introduced due to language, administration, and item relevance. Furthermore, sampling characteristics can be strikingly different such that results from different studies may reflect different sections of the population. A balance must be found between optimal similarity of administration, similarity of meaning, and significance of meaning thereby avoiding unreasonable loss of information or lack of depth. The process of retrospective harmonization is essential. Systematic reviewers must take these issues into account prior to deriving common variables that can be combined to create an overall estimate of effect of a given intervention or exposure.

Types of Cognitive Measures

Cognitive ability measures can be classified into different structures. The most psychometrically validated structure is the Cattell-Horn-Carroll theory which identifies 10 broad stratum abilities comprising over 70 narrow abilities.[14] A variety of tasks are also used to measure the construct of executive functioning, considered to be a complex function of multiple cognitive processes involving planning, attention, working memory, verbal fluency, inhibition, flexibility, initiation, and monitoring of actions. Many of the measures of the cognition reported in primary studies assess somewhat different underlying constructs, and therefore create a substantial challenge for systematic reviewers.

Some of the most common measures are the following:

Memory

Short-term memory is the ability to apprehend and retrieve elements within a few seconds. Long-term storage and retrieval memory is the ability to store and retrieve information over longer term periods. Although many of the studies use standard measures such as the Wechsler Adult Intelligence Scale (WAIS) Digit Span subtests,[15] the Wechsler Memory Scale Logical Memory Test,[15] or the memory items from the MMSE,[11] a large majority of the immediate and delayed word list recall tests are based on different words and different numbers of stimuli and exposures.

General Mental Status Exams

The primary purpose of mental status exams is as a screening measure for identifying individuals with cognitive impairment. Such exams, therefore, contain items focused at the lower end of cognitive function with maximum scores typically obtained by nonimpaired adults (producing ceiling effects). The MMSE[11] is a commonly used scale in longitudinal studies and clinical trials. Because it is a screening device, this measure will not provide good specificity compared with other cognitive assessments.[16] Mental status exams are commonly used in medical and epidemiological studies that might not have obtained other cognitive measures. MMSE-based tests have been developed with a higher range of measurement (e.g., the Cambridge Cognition Examination[17] and the Modified Mini-Mental State Exam [3MS][18]).

Verbal/Crystallized Knowledge.

Many longitudinal studies use at least one of the WAIS Revised,[19] including Vocabulary, Synonyms, and Similarities. However, a number of studies have made use of idiosyncratic or short forms (e.g., three item subsets) of these scales. The National Adult Reading Test[20] and the Mill Hill Vocabulary Scale[21] are also markers of this cognitive ability.

Fluid Reasoning

This ability includes measures of the ability to reason, form concepts, and solve novel problems. Measures of fluid ability include the WAIS Block Design subtest, Cattell Culture Fair Test,[22] the Raven Progressive Matrices,[21,23] and WAIS Figure Logic.[15] Other matrices and rotation tests such as Thurstone Primary Mental Abilities Spatial Orientation and Card Rotations[24] are also used in a variety of longitudinal studies.

Speed

A variety of rapid scanning or cognitively simple tasks are used to measure the reaction time or speed of response. Substitution coding modalities (e.g., WAIS Digit Symbol, Symbol–Digit, Symbol–Letter),[15] alphabet coding, and simple/complex reaction time measures dominate the speed domain.

Working Memory/Attention

Working memory involves the manipulation of information that is held in temporary storage and is a fundamental process involved in many other cognitive tasks. WAIS Digit Span Backward and Serial 7’s from the MMSE are the most common measures that can be considered to measure working memory.

Verbal Fluency

Verbal fluency tasks tap a complex set of cognitive functions and do not fit clearly into either verbal ability or memory constructs. Verbal fluency is measured mainly by category word fluency (e.g., animals) and first letter (e.g., FAS) word fluency tasks that require individuals to generate as many words as they can within a specified time limit. Word fluency has also been used as a measure of executive function.

Potential Methods for Harmonization of Cognitive Variables

The type of response of a participant to an item depends on how the item is formulated. In the simplest form the response of a participant to an item is considered either correct or wrong (i.e., dichotomous), while the most informative setting would provide a continuous response to an item (i.e., numerical value). For this latter setting factor models or analysis would be used to analyze this type of data while for the former setting item response theory (including the well-known Rasch model) is typically applied. Note that item response theory has been extended to polytomous items (a categorical outcome with more than two outcomes) as well.

Factor Models

Factorial invariance and the larger enterprise of measurement are foundational aspects of empirical research.[25,26] This is especially true where important objects of study are latent constructs whose nature and validity are simultaneously derived from consideration of a set of interrelationships with other latent and observed constructs. The establishment of measurement equivalence is essential for studies that implicitly require the quantitative comparability of constructs across samples differing in birth cohort, country, culture, and over time.

Factorial Invariance

Any study with the aim of making comparisons across time or groups assumes measurement equivalence, but it is possible to test this assertion using factorial invariance. A logical hierarchy of constraints in factor models begins with configural invariance,[27] requiring that the same number of factors and pattern of salient factor loadings be equivalent across groups. It is identified by fixing the factor estimate (i.e., mean, variance) in one group and propagating the identification by constraining the corresponding parameter (i.e., intercept, loading) of a reference indicator across groups. Using this as a baseline, Meredith's hierarchy of constraints[25,26,28] is fit to the data: (1) weak factorial invariance, or "metric" invariance, involves equivalence of factor-variable regressions (i.e., factor loadings); (2) strong factorial invariance adds constraints on manifest intercept (mean) terms and requires that one factor mean be fixed to specify the metric of the latent variables; and (3) strict factorial invariance further constrains unique variances to be equivalent. Apart from the identifying constraints, factor variances, covariances, and factor means must be freely estimated, as factorial invariance concerns only the measurement model. A sequence of nested models from least to most constrained is recommended. For multi-occasion models in the context of considerable age-based change, bias may result unless age heterogeneity is small, compared with the amount of change expected.[26,29] In such cases, chronological age differences will be controlled statistically for evaluating factorial invariance.

Factor models with polytomous (ordered categorical) indicators and robust procedures for item-level invariance tests are important extensions of factorial invariance testing procedures. Recent software enhancements relax the assumption of continuous and normally distributed indicators and permit direct analysis of binary and ordinal (polytomous) indicators. Although we will make use of factor-based models for test of factorial invariance, we note its equivalence to the graded response model within item response theory.[30,31] Recent developments also permit direct estimation of categorical factor models with incomplete data.

Item Response Models

Item response theory provides a formal model of the individual's response to items comprising a scale. It offers many advantages in the development and refinement of psychometric instruments. One major benefit of an item response theory framework is that the placement of items onto a continuum of difficulty (severity) facilitates comparably scaled construct scores from two measures with overlapping but disjointed item sets. While evidence of cross-cultural or cross-age factorial invariance validates the comparability of construct scores across studies, factorial invariance procedures are silent regarding options when invariance is not demonstrated for some items. Factor models (i.e., item loadings) are estimated in the context of the other items and factors in the measurement model. It is problematic that the same item may have different loadings as the context changes with the addition or removal of specific items or other factors from the measurement model. Although an item's item response theory difficulty and threshold parameters are estimated in the context of a sufficiently large item pool, and this is analogous to the context of factor-analytic models, obtaining comparable scores from disjoint item sets is much more straightforward in an item response theory framework, and will be an important tool for data harmonization.

There is a close relationship between item response theory and certain classes of factor models.[31,32] Further synthesis of item response theory and factor-analytic measurement models is an important aspect of this data harmonization methodology. Factor-analytic models are preferable when constructs are multidimensional (e.g., somatic, cognitive, emotion dimensions of depression) and items are factorially complex, which is often the case. Factor models also readily extend to first and second order latent constructs, whereas item response theory models focus on first order unidimensional latent constructs, estimating the measurement parameters of an item in the context of one such construct at a time.

Latent Variable Models

Factor analysis and item response theory are considered latent variable models within the field of statistics and the latent variable models themselves form a special class within the larger set of linear and generalized linear mixed models.[33-35] Some measures, such as memory, working memory, and word fluency are not represented by a continuous response nor by a dichotomous or polytomous response. They represent some kind of summary score or overall test score that is based on a count of the number of items tested with specific properties (e.g. the correct recall of specific words tested). These overall test scores should therefore not be analyzed with factor analysis or item response theory.

The individual items for the memory tests could in principle be analyzed with item response theory, if the responses of these items would be available for participants, but such models would not provide direct relationships between the overall test score and the memory construct. The overall test score can also be analyzed with a latent variable model using, for instance, a binomial distribution to be able to represent counts. In these latent variable models it can also be assumed that the ability of the participant would determine the probability of an outcome of a response in combination with the specific parameters for the memory tests, similar to factor analysis and item response theory.

The ability of a participant would be represented by a single or unideminsional latent variable and it would naturally vary with participants, forming some kind of population distribution for the construct of interest of the participants. This distribution could be represented by a normal distribution (which is also typical in factor analysis and item response theory) or by any other alternative distribution such as a log normal distribution. The probability of an outcome of the memory test could be taken equal to a binomial distribution (as just mentioned) with a logit link function to connect the specific parameters of the memory test together with the latent variable to the probability parameter of the binomial distribution. The number of specific parameters for the memory test that can be used in the data analysis may depend on how many memory tests (or overall test scores) would actually be used on the participants to measure the construct memory. Two types of parameters, that may have similar interpretations as the parameters in the two-parameter item response theory, could be essential. One type of parameter would indicate how difficult it would be to correctly respond to the memory tests. This parameter would relate to the origin or location of the latent variable. A second parameter would indicate the discrimination of the memory test, because some memory tests may discriminate between participants better than other memory tests. This parameter would then relate to the variability of the latent variable.

Primary Objectives

The primary objectives of this project are: (1) to conduct environmental scans to identify approaches to statistical harmonization which could be used in the context of aggregate data and/or individual participant data meta-analysis of cognitive measures; (2) to conduct retrospective qualitative harmonization of available data sets to prepare them for statistics harmonization; and (3) to evaluate different approaches to statistical harmonization when applied to cognitive measures.

Methods and Results: Environmental Scans To Identify Methods of Combining Cognition Data and Statistical Harmonization Methods (Objective 1)

Introduction

To address our first study objective, to identify approaches to statistical harmonization which could be used in the context of aggregate data and/or individual participant data meta-analysis of cognitive measures, we undertook two environmental scans in consultation with key informants and a Technical Expert Panel (TEP). By environmental scan we mean that the research question is not narrow, the search terms are quite broad, and single reviewers are involved in both consideration of eligibility of articles and in data extraction. In addition, the articles have not been reviewed for methodological quality in the usual sense of a systematic review, but methodological properties of the methods used are heavily scrutinized.

The purpose of our first scan was to identify what methods were currently being used to quantitatively combine cognition data in systematic reviews. Of particular interest was when the cognitive measures combined in the meta-analysis differed among studies to assess the current methods being employed by researchers to aggregate these nonuniform measures. In practice, we found that most studies either restricted their analyses to cognitive measures with a common scale or combined effect sizes across studies. None of the studies formally probed into whether the cognitive measures should be harmonized. To identify more sophisticated methods that would allow us to explore whether the cognitive data should be combined we undertook a second environmental scan to assess what types of statistical harmonization methods were used in the general literature. This scan was not restricted to the harmonization of cognitive measures; it was a general search of harmonization methods used in any context. Each environmental scan is described in detail below. The environmental scans culminated in the identification of statistical methods for harmonization that would be used to address study objectives two and three.

Literature Scan—Studies Quantitatively Combining Data on Cognition

The first search was undertaken in a number of databases: Medline®, EMBASE®, Web of Science, and MathSciNet® (January 2001 to September 2011). The search terms used for the first part of the search were “cognition” and “meta-analysis”. A similar search was undertaken using Google search engine. The search results were rapidly screened to identify articles of relevance to this review. The references of relevant articles were also checked and a search for more recent articles that cited the articles already identified as being of interest was undertaken.

Inclusion and Exclusion Criteria

Any study that quantitatively combined individual-level or aggregate-level data on cognitive measures and was published in English was eligible. Cognitive measures were defined as one or more standardized neuropsychometric tests (i.e., measuring global function, executive function, psychomotor speed, attention, memory, or intelligence).

Review Process

The scan of cognition meta-analyses resulted in 120 citations. A single rater reviewed the titles and abstracts of all articles to identify which articles contained a quantitative summary of cognitive data. The full-text was retrieved and reviewed for each article passing the title and abstract screening. Study level characteristics were extracted by one reviewer. These included a description of the populations, study design and number of studies included in the meta-analysis, the intervention of interest (if appropriate), the inclusion criteria, the types of cognitive measures and domains measured, and how the data were statistically combined.

Results

The first level of the scan involved searching the subject heading "cognition" and then limiting the search to those studies that were meta-analyses. There were 121 potential meta-analyses of cognition measures identified. The abstract screen focused on identifying studies that reported a quantitative summary of cognitive data. There were 47 abstracts that passed this level of screening and the full text articles were retrieved. The full text screening involved reducing the number of studies to include only those that combined different cognitive measures, resulting in a total of 33 articles, which are summarized in Table 1.[36-68] All meta-analyses used aggregate data. Most of the meta-analyses included observational studies (19, 57.6 percent); 14 (42.4 percent) were restricted to randomized controlled trials (RCTs). The populations ranged from school-aged children to adults aged 55 and older. The primary focus of the studies varied greatly, but most used the cognitive tests as an outcome associated with a putative harmful agent (e.g., mobile phone electromagnetic fields) or positive factor (e.g., being an expert athlete), or after an intervention (e.g., comparing off-pump vs. on-pump coronary artery revascularization). The cognitive measures differed over the meta-analyses. Most meta-analyses included multiple instruments that measured different aspects of cognition, (e.g., executive function, psychomotor speed).

All meta-analyses including cognition outcomes either restricted their analyses to a subset of studies for which cognitive measures with a common scale were used or effect sizes were combined across studies. In all cases, the cognitive measures were continuous variables. The most common method of analysis was to combine standardized mean differences or effect sizes across studies. When the measures of cognition were consistent across studies or were comparable tests with a normalized scale, a weighted mean difference was used. Four studies[38,40,41,62] used meta-regression; three of the four used a standardized effect size, Cohen's d, as the dependent variable;[38,41,62] and one used a weighted mean difference of normalized comparable tests[40]

Literature Scan—Statistical Methods for Data Harmonization

Identifying literature on statistical methods of data harmonization is challenging as there are no standard keywords or mesh terms used to identify this literature in the bibliographic databases. The initial set of keywords and searching terms were reviewed by the key informants and the TEP to identify additional potential search terms. A focused search was undertaken in a number of databases: Medline®, EMBASE®, Web of Science, and MathSciNet®(January 2001 to September 2011). The final set of search terms used were "individual patient data," OR IPD, OR pooling, OR "multiple imputation," OR "data harmonization," OR "meta-analysis methods." A similar search was undertaken using Google search engine. The search results were rapidly screened to identify articles of relevance to this review. The references of relevant articles were checked and a search for more recent articles that cited the articles already identified as being of interest was undertaken. These references were further supplemented by articles identified by key informants and the TEP increase the comprehensiveness of the search.

Inclusion and Exclusion Criteria

Any study that reported statistical methods for retrospective harmonization of survey data was included. For the purpose of this review, harmonization was defined as "procedures aimed at achieving and improving the comparability of different surveys".[69] We adapted this definition to include study designs other than surveys. For completeness, these studies were supplemented with articles on the conduct and methodology of individual participant data (IPD) meta-analysis, methods for evaluating equivalence (i.e., whether instruments measure the same construct or latent variable, latent trait, or factor across groups or over time), imputation methods, and examples of data harmonization.

Review Process

Because the numbers were relatively small (63 articles), all identified statistical harmonization methodology articles underwent full text screening for relevance by at least two raters. Data extracted from the methodology articles included a description of the statistical method used, the context in which it was used, and the pros and cons of the method.

Results

The scan of statistical methods used for harmonization resulted in 63 unique articles. Of the 63 articles, 53[2,8,69-119] (84.1 percent) met the inclusion criteria. The 10 excluded articles that were not directly relevant to this review are listed in Appendix A. Seven of the 53 articles (13.2 percent) described methods for statistical harmonization (Table 2). Ten articles (18.9 percent) focused on the conduct of IPD meta-analysis and an additional 6 articles (11.3 percent) focused on IPD meta-analysis methodology. Six articles (11.3 percent) reviewed imputation methods and the appropriateness of their use and 2 articles (3.8 percent) described methods for evaluating equivalence of item functioning across study subgroups. A summary of these supplemental studies are in Table 3. Finally, 22 articles (40.7 percent) reported the results of 16 unique statistical harmonization analyses undertaken in different contexts (Table 4).

There were three general classes of statistical methods identified in this scan. One class used a simple linear- or z-transformation to create a common metric for combining constructs measured using different scales across datasets. A summary of the assumptions and the application of this type of model is in Table 5. An example of this class is in the Comparison of Longitudinal European Studies on Aging. When harmonization was deemed appropriate, some constructs were converted to a 0 to 1 scale by dividing a continuous score by its maximum score.

A second class of methods posits that there is a latent factor(s) that underlies a set of measured items that can be modeled using linear factor analysis (if the items are continuous), two parameter logistic item response theory (if the items are binary), or a polytomous Rasch model (if the items are ordinal), or moderated nonlinear factor analysis (MNFA) if there is a mix of binary, ordinal, and/or continuous items. These methods are described in the articles by van van Buuren, et al.,[76] Bauer, et al.,[70] and Gorsuch.[72] In each case, the first step is to construct a "conversion key" using one of the statistical models described above. This step models the relationship between the latent construct and the measured items. The second step uses the conversion key to convert the information onto a common scale. Measurement equivalence must then be assessed across samples.[97]

The MNFA method proposed by Bauer is the most generalizable as it can accommodate different types of item data—binary, ordinal, or continuous—within a single model.[70] All of these approaches require that items can be "chained" together among studies, such that each study must have at least some items that overlap with another study. Another potential limitation is that the methods require independent data within studies and may not be appropriate for repeated measures in a longitudinal study. The authors using these methods tended to randomly choose one observation per person if more than one was available in the dataset. The methods proposed by van Buuren[76] and Bauer[70] are described in detail in Table 6.

The approach described by Gorsuch[72] was not originally meant for the statistical harmonization of data, but could have some application in this context (Table 2). Gorsuch proposed extension analysis to compute the relationship among common factors to variables that were not included in the factor analysis. For example, this would be used in a situation where a factor analysis would include proven items but no new experimental items. This method, however, would still require that items could be chained together.

Most of the examples from this class of models were restricted to a single observation per individual. McArdle, et al.[73] combined an item response theory approach with a latent growth/decline curve modeling to allow for repeated measures. Their method allowed for a varying number of data points per individual and for the instruments to change over time (Table 2).

The final class of methods, multiple imputation, is described by Burns, et al.[71] (Table 5). In this situation, the authors were interested in combining Mini-Mental State Examination (MMSE) scores with missing data across nine Australian longitudinal studies of aging. The MMSE score comprises 11 items and the proportion of missing at least one MMSE item varied greatly by contributing study and wave of data collection. Furthermore, the missingness of information was related to demographic characteristics, especially age and education. Burns, et al. used an imputer model that utilized multiple imputation with chained equations to impute appropriate missing MMSE item scores. A detailed summary of the methods proposed by Burns, et al. is in Table 6. General issues around methods of imputation are reviewed by Peyre, et al.[92] and Spratt, et al.[94] (see Table 3). This method required that the same measures were included across studies, but the approach may be able to be extended to more general situations.

Examples of Analysis of Harmonized Data

Table 4 presents a summary of 22 publications arising from 16 data harmonization projects. In many of these projects the harmonization was done in terms of standardizing response options and determining whether questions were comparable across cohorts. For example, Minicuci, et al.[112] compared disability-free life expectancy using survey data collected in three populations. Data on five activities of daily living (ADL) questions that were common to all surveys were used and the response options for these questions were dichotomized to create a common scale. Pluijm, et al.[113] similarly combined ADL data across six countries. There was overlap in the ADL items among the four items comprising the Katz ADL index; all four items were present in four of the six country surveys. In countries where the two items were not measured, the data for these were extrapolated from other "comparable" ADL items. Finally, subjects were excluded if two or more items were missing. Hot deck methods were used to impute values when one of the items was missing due to nonresponse.

Bath, et al.[99] harmonized cognitive data from the Longitudinal Aging Study Amsterdam (LASA) and the Nottingham Longitudinal Study on Activity and Ageing (NLSAA). LASA used the Mini Mental State Exam (MMSE: 30 point scale) and the NLSAA used the Clifton Assessment Procedures for the Elderly (CAPE: 12 point scale). In the analysis, the authors simply created derived variables MMSE/30 and CAPE/12, and combined across studies.

Many of the studies used item response theory-based methods for analysis. van Buuren, et al.[118] used response conversion to harmonize international disability information, while Crane, et al.[102] used item response theory to co-calibrate cognitive scales. Both Curran, et al.[103] and Grimm, et al.[106] combined item response theory and growth curve models. Curran fit these models to data of developmental internalizing symptomology and Grimm examined the association between early behavioral and cognitive skills with later achievement. McArdle, et al.[109] used linear structural equations modeling with incomplete data to analyze repeated measures twin-data to evaluate biometric genetic hypotheses in the context of intellectual growth and change. The authors incorporated a twin analysis including means and age effects, longitudinal analyses based on latent growth components, and biometric-genetic analyses for components of growth using linear structural equations models.

Schenker, et al.[116] combined clinical examination data with self-reported survey data. The National Health and Nutrition Examination Survey asked self-report questions on health conditions and obtained clinical measures based on physical examinations. The National Health Interview Survey was larger and obtained a rich set of variables for use in multivariate analyses, but the study relied on self-report questions for the information on health conditions. Multiple imputation was used to properly reflect the sources of variability in subsequent analyses.

The Fibrinogen Studies Collaboration[105] combined data from 31 cohort studies using a two-stage approach. In the first stage partially and, where possible, fully adjusted estimates were obtained from each study, together with their standard errors. This method addresses the issue of when studies included in an IPD meta-analysis include some, but not all, important confounding variables. This may be relevant if we have some studies fully measuring the construct and other studies only partially measuring the construct; this could be analyzed with this bivariate approach. This approach, however, does require the exposure and outcome of interest to be available across all studies. In the second stage, they combine the study-specific estimates.

Other Related Areas of Application

In educational and achievement testing, the problems and approaches related to dealing with data collected with different instruments are discussed in terms of score linking or score equating.[120] Different editions of tests are designed to measure the same constructs, but almost by definition will differ in their psychometric properties. For example, if one edition is more difficult than another, examinees would be expected to receive lower scores on the harder form. Score equating seeks to eliminate the effects on scores of these unintended differences in test form difficulty. There is a long history of research in the area of score equating which is summarized in Dorans, et al.[120] Whereas the body of literature developed in the area of data harmonization in health research may be more applicable to outcomes and exposures in the majority of research studies, there are potential links in measuring constructs such as cognitive performance.

Statistical Harmonization Methods for Application

In consultation with our TEP, three methods identified in the environmental scans were selected to create combinable cognition constructs across datasets. The methods were chosen based on their frequency of use in the literature, and their appropriateness for use with summary cognitive constructs. Methods were also selected based on the ease of application. We reasoned that if standardization methods were sufficient, one would not need to use the specialized methods that require more sophisticated statistical analyses and nonstandard software. The first two methods standardize the cognition variables to a common metric. The specific methods chosen have been used in the context of comparing or combining cognition measures in the literature. The final method chosen was a latent variable approach. This approach uses a generalized linear mixed model to identify a univariate underlying latent variable for subjects common to all datasets. The missing data-type methods were also considered but are more amenable to constructs that are collected using a large number of items across studies, most of which are overlapping, for example, measuring disability using basic and instrumental activities of daily living. As our construct of interest was an overall score we did not pursue this type of statistical harmonization. These methods are described in detail in the "Methods and Results: Implementing and Evaluating Three Methods of Statistical Harmonization Applied to Cognitive Measures (Objective 3)" section.

Methods and Results: Process of Preparing Data for Statistical Harmonization (Objective 2)

Introduction

The statistical pooling of cognitive data is a complex process and requires many steps before it can be undertaken. Before we could compare our chosen methods of statistical harmonization we had to create data sets that were inferentially equivalent. To meet our second objective of implementing the steps of a retrospective harmonization in a specific example of combining cognitive data from three studies, we undertook the process of qualitative harmonization.

In this section, the process of pre-statistical harmonization of the cognition data is described. This includes the selection of data sets, identification of relevant cognitive domains and instruments, the identification of other variables of interest (e.g., variables associated with cognition) and potential confounding variables (e.g., sex, age, and education), the qualitative harmonization of the noncognitive variables, and preparation of data for statistical harmonization before pooling of cognitive variables can be undertaken.

Qualitative harmonization is used in this section whenever we deemed that inferentially equivalent variables could be created using processing algorithms. For example, age can be grouped into standard categories across studies. In qualitative harmonization, a process is first undertaken to determine what data can be validly combined across studies. If it is determined that data can be combined across studies then processing algorithms are created and implemented on data collected by each individual study to produce a standard (i.e., common format) set of variables. Most of these transformations involve grouping continuous values or grouping different response categories across studies into standard categories. It is important to mention that qualitative harmonization can only be applied to simple constructs (e.g., quantity of cigarettes smoked, marital status). It is not applicable to harmonize more complex measures such as different rating scales across studies.

Prestatistical Harmonization Methods

Identification of Potential Datasets

Acquiring individual participant data (IPD) from research studies is a time consuming activity.[121] Collecting and assembling IPD from studies of work-related mechanical exposures and low back pain, for example, took nearly 3 years.[121] To identify candidate data sets, we balanced feasibility to acquire data in a timely fashion with the availability of sufficient cognitive constructs to conduct a statistical harmonization. We proposed including data sets that were accessible to study co-investigators, included complex cognitive measures of which at least some were overlapping among the data sets, and contained data on additional variables of interest and confounding variables. After discussion with the Technical Expert Panel (TEP), individual level data from three Canadian studies were obtained: the Canadian Study on Health and Aging (CSHA), the Canadian Community Health Survey on Healthy Aging (CCHS), and the Quebec Longitudinal Study on Nutrition and Aging (NuAge). A brief description of each of these data sets follows. Although we have included data sets collected within a single country, the studies are all population-based. Canada has two official languages and is culturally diverse such that one may find the type of heterogeneity usually expected between countries.

Canadian Study on Health and Aging

The CSHA is a national, population-based study of dementia in Canadian adults aged 65 or older (Table 7).[122] In the first wave of the CSHA in 1991 (CSHA-1), face-to-face interviews were conducted with 10,263 older adults across Canada: 9,008 were living in the community and 1,255 were living in institutions. The 10,263 comprised representative random samples of people aged 65 or over, drawn in 39 urban centers and nearby rural areas in the 10 Canadian provinces.

In CSHA-1, a cognitive screening test, the Modified Mini-Mental State Exam (3MS), was administered to all community residents. Community-dwelling residents who scored less than 78 on the 3MS, plus a sample of people who scored 78 or above, were invited to undergo the clinical assessment. The clinical assessment was also conducted for participants who could not complete the 3MS, and all those living in institutions. At the clinical assessment, a nurse re-administered the 3MS. Participants scoring 50 or more on the 3MS were administered a standardized neuropsychological assessment by a trained psychometrician.[123] The Neuropsychological Assessment included tools to measure the following domains of cognition: memory, abstract thinking, executive function, judgment, aphasia, amnesia, and construction (Table 8).[19,124-129] These tests were shown to accurately predict incident Alzheimer's disease after 5 and 10 years.[130] Our study includes the 1,730 CSHA participants who had complete data for the neuropsychological battery at CSHA-1.

Quebec Longitudinal Study on Nutrition and Aging

NuAge is a 5-year observational study of 1,793 men and women aged 68–82 years in good general health at recruitment (Table 7). Community-dwelling men and women, living in the regions of Montreal, Laval, and Sherbrooke in Quebec, Canada, were included if they spoke French or English, were free of disabilities in activities of daily living, had no cognitive impairment (Modified Mini-Mental State Examination [3MS] score, >79), were able to walk one block or climb one flight of stairs without rest, and were willing to commit to a 5-year study period. Those who had heart failure greater than or equal to Class 2, chronic obstructive pulmonary disease requiring oxygen therapy or oral steroids, inflammatory digestive diseases, or cancer treated by radiation therapy, chemotherapy, or surgery in the past 5 years were excluded. Face-to-face recruitment interviews were through 2003 to 2005.

NuAge includes many sub-studies on some relevant, complex research problems, one of which was to examine the effects of nutrition quality on cognitive decline. In 2006 to 2007 and 2008 to 2009, a neuropsychological battery was administered to a subset of 464 NuAge participants. Only francophone participants with a 3MS above the age-adjusted cut-off for risk of dementia at the third NuAge visit were recruited. Additional exclusion criteria for this sub-study included head trauma resulting in unconsciousness, stroke leading to hospitalization, cardiac arrest with resuscitation, epilepsy, subdural hematoma, sub-arachnoid hemorrhage, brain tumor or metastasis, central nervous system infection (e.g., meningitis), Guillain-Barré syndrome, multiple sclerosis, and toxicity (e.g., CO2 or methanol intoxication). The battery included instruments to measure memory, psychomotor speed, and executive function (Table 8).[19,131-135] This study includes data from the 432 respondents of the first administration of the neuropsychological battery in 2006 to 2007 with complete cognition data.

Canadian Community Health Survey on Healthy Aging

The CCHS-Healthy Aging includes community-dwelling people aged 45 years and over living in the 10 Canadian provinces (Table 7).[136] Excluded from the sample were residents of the three territories, persons living on Indian reserves, Crown lands, in institutions, full-time members of the Canadian Forces, and residents of some remote regions. Data collection took place in participants' homes from December 1, 2008 through November 30, 2009 using computer-assisted personal interviewing. The content of the CCHS-Healthy Aging was developed collaboratively by Statistics Canada and researchers from the Canadian Longitudinal Study on Aging (CLSA). As part of the Statistics Canada-CLSA collaboration, CCHS participants were asked whether their survey data could be shared with the CLSA. This article includes data from CCHS participants who were between the ages of 45 and 85 years and who agreed to share their data with the CLSA, henceforth referred to as the CCHS-CLSA sample.

The cognition module was administered in English and French to consenting, nonproxy respondents. CCHS questionnaires included four cognitive instruments to measure memory and executive function (Table 8).[127,137,138] The categorization of levels of cognitive functioning in the Canadian household population aged 45 or older based on these data has been validated.[139] Standardized scores for these instruments were found to be related to self-reported general and mental health status, memory, and problem-solving ability, activities of daily living, life satisfaction, loneliness, depression, and chronic conditions in cross-sectional data. To make the study populations more comparable, the analyses included 7,107 CCHS-CLSA participants who were 65 or more years old and had complete data for the cognition module.

Identification of Relevant Cognitive Domains and Instruments

As described previously, there are several domains of cognition and instruments that have very different applications and properties. Constructs most often combined in meta-analyses of cognitive measures include general cognitive tests, such as the 3MS. Although the general tests can often be used for screening, tools that measure specific domains of cognition are required to measure specific aspects of cognitive function such as memory and executive function.

Table 8 displays the cognitive domains and specific instruments used in the CSHA, CCHS and NuAge. The underlying assumption of meta-analysis is that one is combining comparable information across studies. As such, an a priori decision was made to restrict the analyses to instruments that measured the same cognitive domain. As many of the statistical harmonization methods require overlapping items, instruments that were used in at least two of the three studies were chosen. The idea was that even if no items are common to all studies, the studies can be "chained" together. For example, if item sets A and B are available in study 1, item sets B and C are available in study 2 and item sets C and D are available in study 3. These studies are chained by A-B-B-C-C-D. To meet these requirements, we chose to focus our statistical harmonization analyses on the memory domain and the specific instruments including the Rey Auditory Verbal Learning Test (RAVLT) and the Buschke Cued Recall Procedure (BCRP).

Commonly, a meta-analyst has access to only the summary measures for most studies. For example, the RAVLT is a summation of correctly recalled words from a list of 15. If one had access to the complete data set, they might also be able to analyze each of the 15 nouns separately. Summary measures were used for this project as this best represents the type of data one would be including in a Comparative Effectiveness Review. One additional requirement for latent variable analysis, however, is that more than one variable per study is required for convergence, similar to item response theory models. In both the CSHA and NuAge, multiple cognitive measures met this criterion. The only measure that met all of the criteria in the CCHS was the RAVLT. To help identify and attain convergence of our model, we also included the Health Utilities Index Memory/Thinking attribute which is a measure of memory, but is not included in another study. The Health Utilities Index (HUI®) is an indirect measure of memory, but will still contain information on the memory construct and thus is informative for our latent variable model.

The Rey Auditory Verbal Learning Test

The RAVLT[124]was used to measure short-term memory in the CSHA and the CCHS. The RAVLT is a 15-item word learning test that assesses both learning and retention. A participant is read a list of words and asked to listen carefully. The participant is then asked to recall as many words as they can in any order. There are generally a number of immediate trials and then a delayed trial. In the CSHA there were five trials and a delay; in the CCHS, there was one trial and a delay.

The RAVLT is one of the most widely used neuropsychological test instruments[140] and extensive normative data is available for both the English and French versions.[127,141] The RAVLT has good test-retest reliability ($0.51 \leq r \leq 0.86$),[142] though reliability as low as 0.12 has been reported,[143] and has been shown to be extremely sensitive in detecting early cognitive decline.[144,145] Overall, patients diagnosed with probable Alzheimer's disease recalled fewer words than normal controls. Similarly, those with mild cognitive impairments also recalled significantly fewer words than normal controls.[146] Schoenberg and colleagues[147] reported respectable classification accuracy on the delayed recall trial using a cut-off of z-scores <1.5 SD and clinician diagnosis for Alzheimer's disease as the criteria (sensitivity = 82.9 percent, specificity = 82.8 percent, positive predictive value [PPV] = 86.2 percent, and negative predictive value = 78.9 percent).

The Health Utilities Index Cognition Attributes

HUI is a generic, preference-scored, comprehensive system for measuring health status and health-related quality of life, which produces utility scores.[148] The HUI Mark III assesses functional health in eight domains: vision, hearing, speech, ambulation, dexterity, emotion, pain and cognition.[149,150] The HUI has been used in many settings and has been shown to have strong validity and reliability,[151] even in patients with dementia.[152] The cognition subscale is derived from two questions regarding usual ability to remember things and to think and solve day-to-day problems. The HUI cognition subscale has been shown to be correlated with other measures of Intelligence Quotient and achievement in children,[153] as well as with the RAVLT and other neuropsychological memory tests in adults.[139]

The Buschke Cued Recall Procedure

The BCRP tests memory under conditions of free recall and cued recall. The particularity of the test is that it includes a preliminary stage that ensures appropriate encoding of the material. Subjects are shown a picture sheet with four images of objects belonging to four different categories. They are asked to point out and name the object in each category. For example, "Which one is a piece of clothing?" Following this, the sheet is taken away and the examiner ensures that the items have been properly encoded by asking immediate cued recall of each item. This is continued for all items on the list to ensure that the information has been learned and effectively encoded. After a distracter task, the subject is asked to recall the items that were viewed (free recall). The interviewer then provides cues for items missed during free recall. The free and cued recall procedures are repeated over a number of trials. A delayed recall trial is administered after a break. The CSHA used English and French versions of the 12 item Buschke memory test (BMT)[125] and NuAge used a French version of the Free and Cued Selective Reminding Test (FCSR), 16 item test adapted from Grober and Buschke.[154] In both cases, the scores are simply the number of items recalled spontaneously or in response to cues. Both CSHA and NuAge included three immediate trials and one delayed trial.

The BMT and FCSR were designed to minimize apparent memory deficits that would be confounded with problems in attention or encoding[154] by ensuring that all items have been appropriately encoded and by providing cues during learning and retrieval to lessen the cognitive effort required by older adults.[155] Among measures of memory functioning, the free recall measure was the best predictor of functional impairment in older persons with suspected memory disorders.[156] All measures from the Buschke memory test have been shown to differentiate among individuals with various levels of cognitive impairment.[157] The 16-item version has also been shown to be able to distinguish patients with mild cognitive impairment (MCI) who converted to Alzheimer's disease from MCI nonconverters.[158] In seven studies comparing individuals with Alzheimer's disease and healthy controls, both free recall and total recall have shown to have high sensitivity (62 to 100 percent) and specificity (94 to 100 percent).[159]

Identification of Other Variables of Interest and Potential Confounding Variables

We wanted to examine how our methods of statistical harmonization impact the relationship between the cognition construct with other variables. To do this, a set of variables was identified for which there is a strong understanding of the underlying relationship with cognition. These variables included sociodemographic and lifestyle factors (age, sex, education, income, country of birth, physical activity, smoking status, alcohol consumption) and anthropometric and health conditions (height, weight, body mass index, hip circumferences, heart rate, diastolic and systolic blood pressure, and self-reported diagnosis of high blood pressure, stroke, diabetes, myocardial infarction, and family history of high blood pressure, stroke, diabetes, myocardial infarction). These variables were selected based on the demonstrated relationship with cognition in the literature[160] and endorsed by the TEP.

The DataSHaPER (DataSchema and Harmonization Platform for Epidemiological Research; www.datashaper.org) approach to qualitatively harmonize variables of interest was employed.[161,162] Using this approach, a set of variables targeted for a harmonization project is first defined. Such a set of variables is called a DataSchema. A priori rules are then defined for each variable of a DataSchema and used to formally determine if the information collected in a given study can be used to generate a given DataSchema variable. This step of the process therefore establishes what data can be validly combined across studies. The process of applying the rules and assessing the compatibility of variables with the DataSchema is called pairing. Selection and definition of DataSchema variables, rule creation, and pairing are based on protocols involving iteration between domain experts, research assistants, and a validation panel. The compatibility of each study's data and each variable in a DataSchema is assessed on a three-level scale of matching quality: "complete," "partial," or 'impossible" match. Table 9 describes the elements required for each level of compatibility. All variables of interest that are a "complete" or "partial" match to the DataSchema were included.

Results—Prestatistical Harmonization

Table 10 presents a description of the 34 variables of interests and the pairing results for the three studies. Family history of chronic conditions was not collected in any of the studies. Of the remaining variables, there were complete matches in all studies for 10 targeted variables: age, sex, household income, country of birth, height, weight, body mass index, occurrence of diabetes, and level of physical activity (3 categories and 4 categories). There were complete or partial matches across the 3 participating studies for 3 variables: alcohol consumption, occurrence of high blood pressure, and myocardial infarction. Each of the education variables had at least one impossible match; however a simplified three-level categorical variable indicating the number of years of education (low (0 to 8 years), medium (9 to 13 years), and high (>13 years) could be derived in all data sets. Because income was only available for a subset of the CSHA participants, it was not included in the analyses. Stroke was measured in the three datasets, but the definition was sufficiently different (stroke vs. all cerebrovascular event) that it could not be used in the analyses.

Sociodemographic and health-related characteristics of participants of the CSHA, CCHS-CLSA, and NuAge included in the analyses are presented in Table 11. The average age of the CCHS-CLSA participants (73.2 years) and NuAge participants (73.7 years) was less than that of CSHA participants (79.7 years). The CSHA participants tended to have a lower level of education and income (adjusted to 1992) compared with the CCHS-CLSA and NuAge participants. Income, as previously mentioned, was only available for a subset of CSHA participants and was not included in any analyses. As well, fewer participants reported being born in Canada in CSHA. The CCHS-CLSA and NuAge more often reported being a current alcohol drinker than the CSHA participants; however, there were some differences in the way the question was asked across studies (Appendix B). More CSHA participants reported a low level of physical activity compared with CCHS-CLSA and NuAge. The CSHA participants also reported a lower level of physical activity and alcohol consumption. The CCHS-CLSA participants reported high blood pressure and diabetes more often than CSHA or NuAge participants.

The pre-statistical harmonization culminated in three data sets with combinable data on 13 variables of interest. Appendix B includes a complete summary of how the variables were operationalized in their respective studies and the final derived variable used in subsequent analyses.

Methods and Results: Implementing and Evaluating Three Methods of Statistical Harmonization Applied to Cognitive Measures (Objective 3)

Introduction

In the previous section we were able to conduct qualitative harmonization and develop algorithms to derive inferentially equivalent information across studies by combining categories of categorical variables or instituting common cut-points for continuous variables. Combining data on complex constructs such as cognition, however, can be particularly challenging and require other methods of harmonization, including statistical harmonization. In this case the continuous variables are all measuring an underlying construct which we assume is same across data sets, but this assumption must be tested.

To address our final objective, to evaluate different approaches to statistical harmonization when applied to cognitive measures, we undertook a number of statistical analyses. To create combinable measures we used methods identified in the environmental scans and endorsed by the Technical Expert Panel (TEP). Two standardization methods that have been commonly used to combine and compare cognitive variables across data sets and the latent variable method were used to create combinable cognition data across studies.

The pre-statistical harmonization culminated in datasets that are amenable to the creation of harmonized cognitive constructs that could be combined across studies. Prior to comparing the methods of statistical harmonization, however, we needed to determine whether our cognition measures were inferentially equivalent. Latent variable regression analysis was used to determine if the cognitive measures were measuring one construct that is consistent across the data sets. In such a setting, the cognitive measures would be combinable, but this is less clear when either the measures do not represent one construct, or whether the construct is not consistent.

A second set of analyses was used to compare the different methods for statistical harmonization in the context of a traditional meta-analysis. The objective of the second set of analyses was to examine how well the results of these statistical harmonization methods agree with each other (correlation analysis) and whether they would provide similar results if a traditional meta-analysis was undertaken.

Methods of Statistical Harmonization

Calculation of T-Scores

Our first standardization method was utilized in both the Canadian Study on Health and Aging (CSHA)[130] and the Canadian Community Health Survey on Healthy Aging (CCHS)[139] to convert raw cognitive measures to demographically-corrected standardized T-scores. T-scores are dependent on the underlying distribution of cognitive measures in each study and have been used to create norms and compare different cognitive measures on a common scale. The method for creating a T-score is a two-step process described in Tuokko, et al.[130] Briefly, the T-score is created by first normalizing each cognitive test score distribution to have a mean of 10 and a standard deviation of three. Each of these test-scaled scores is then regressed separately on age, sex, and education within each study. These variables were chosen because performance on many neuropsychological measures is related to age, sex, and education[142] and these measures are often normalized with the performance of a control group, characterized by age, sex, and education. Age was included as a continuous variable, but because education was measured differently across the three cohorts, we used a three-level categorical variable (low: 0 to 7 years, medium: 8 to 13 years, and high: >13 years) to represent education in our regression models. The predicted values from the regression model were then used to create a "residual" scaled score (i.e., actual scaled score—predicted scaled score) for each participant. These residual scores represent how much better or worse the individual did on the test compared with what would be predicted based on their demographic characteristics (age, education, and sex). In a second step, the residual scores were converted to T-scores for each participant using the following formula:

$$T - score = \left(\frac{residual\ score}{standard\ deviation(residual\ score)}\right) * 10 + 50$$

The resulting T-scores are normally distributed and have a mean of 50 and a standard deviation of 10. T-scores can be interpreted as how an individual's score on each cognitive measure compares to the average score of participants of the same sex and age, and with the same educational background. Because they have been adjusted for the demographic factors in their creation, the T-scores should not be related to age, education, or sex.

Calculation of Standardized Demographic Category-Centered Scores

A second method was used to calculate study-specific scores standardized relative to a consistent group across datasets. In this case, the mean and standard deviation for a common demographically determined group that is presumed to be homogeneous with respect to the cognitive measures are used to standardize, or "center" the individual cognitive measures. Again, the demographic factors known to be associated with cognitive measures—age, education, and sex—were used to identify a homogeneous group with a sufficient sample size to provide stable estimates of the means and standard deviations of the cognitive measures. Appendix C, Table 1 displays the number and percent of participants by 5 year age categories, sex, and education level for each study. Based on these factors, there were two possible subgroups; females 70 to 74 with 8 to 13 years of education and females 75 to 79 with 8 to 13 years of education. These subgroups had similar numbers in each of the datasets. The 75 to 79 age group had a lower minimum number (n=41 in the Quebec Longitudinal Study on Nutrition and Aging [NuAge] compared with n=48 for the 70 to 74 year old subgroup), but had a larger number over all of the datasets (n=533 compared with n=488 for 70 to 74 year old subgroup). Since our goal was to identify a homogenous subgroup, we then compared the standard deviations (SD) of the cognitive measures for each of the subgroups (Appendix C Table 2). The 70 to 74 year old subgroup had a smaller SD for all measures except the Rey in the CCHS to the Canadian Longitudinal Study on Aging (CLSA) and the Buschke Total score in CSHA. Based on these data we used females 70 to 74 with 8 to 13 years of education as our reference, or "centering" group.

The mean and SD for each cognitive measure in the reference was estimated and the scaled score (C-score) was calculated for each participant as follows:

$$C - score = \left(\frac{raw\ score - mean\ of\ females\ 70 - 74\ with\ 8 - 13\ years\ of\ education}{standard\ deviation\ of\ females\ 70 - 74\ with\ 8 - 13\ years\ of\ education}\right)$$

This method retains the associations with age, sex, and education, while standardizing the scores to a common subgroup within each study. Like the T-score standardization, this method does not take into account the differences between the measurement properties of the scales.

Calculation of Latent Variables

In this method, it is assumed that the overall test scores of a participant are influenced by a univariate continuous latent variable unique to that participant. The overall test scores are viewed as counts representing a correct number of scored test items. Conditionally on the latent variable, the overall test scores follow a binomial distribution. The performance on correctly scoring the individual test items may change even within subjects, which justifies the binomial distribution. The mathematical model is represented by:

$$(1)\qquad \begin{array}{l} Y_{ij} = y \mid Z_i = z \sim B\left(N_j, p_{ij}(z)\right) \\ Z_i \sim N\left(0, \tau_i^2\right) \end{array}$$

with Y_{ij} the number of correctly scoring test items for memory test *j* of subject *i* and Z_i the latent variable for subject *i*. The latent variable Z_i is assumed to be normally distributed. The probability $p_{ij}(z)$ may depend on subject *i* through different covariates, such as age, sex, and education. It is assumed that this probability of correctly scoring an item is of the logistic type:

$$(2)\qquad \operatorname{logit}\left(p_{ij}(z)\right) = \beta_{0j} + \sum\nolimits_{k=1}^{K} \beta_k \cdot x_{ik} + z\ ,$$

with β_{0j} the intercept for test *j*, β_k's the parameters for the covariates, and x_{ik}'s the *K* covariates for subject *i*. The standard deviation τ_i of the latent variable may also depend on covariates, such as age, sex, and education. It is assumed that the standard deviation has the following form:

$$(3)\qquad \log(\tau_i) = \eta_0 + \sum\nolimits_{k=1}^{K} \eta_k \cdot x_{ik}\ .$$

The intercept parameters in (2) are related to the difficulty of the memory test *j*. A larger value would indicate that the memory test would be easier to conduct, while smaller values would tell us that the corresponding memory test is harder. The ability to investigate whether memory tests would be able to discriminate between participants better than other memory tests would require a separate intercept for memory tests in formula (3), that is, in the variability of the latent variable. Estimation of discrimination parameters requires at least the use of four different memory tests. In the current model it was assumed that all memory tests would discriminate in the same way between participants.

The parameters in (2) would indicate difference in performances on the memory tests for particular subgroups of participants. For instance, it is expected that age would reduce the performance on the memory tests. The parameters in (3) on the other hand would indicate if the memory tests discriminate better for subgroups of participants, for example, whether the performance differences in memory tests are greater for males than for females (or the other way around) or whether older people would demonstrate larger differences in performance than younger people.

Model (1) with relationships (2) and (3) are related to the work of Van Buuren, et al.[76] and Bauer, et al.,[70] although they described procedures for individual items instead of overall test scores. Table 12 depicts the different cognitive measures calculated for each dataset. Essentially, for each of the memory measures we had the raw data, the T-score, the centered score (C-score), and the latent variable score.

Assessing the Cognitive Measures as a Single Construct

To assess if the cognitive measures included in our analyses measured a single construct, we did a number of analyses. In the cases where cognitive measures were collected in multiple studies, like the Buschke Free and Buschke Total recall in CSHA and NuAge, it can be investigated if the tests behave similarly across studies, when corrected for the covariates. This implies that the difference in two difficulty parameters β_{0j_1} and β_{0j_2} is the same across studies. To be able to compare different studies, homogeneity of the tests is required, because otherwise the different studies may not have measured the same latent variable (a form of measurement invariance). Only then is it possible to compare the parameters β_k's across studies to investigate the heterogeneity between studies. This lack of homogeneity in memory tests can be investigated with a likelihood ratio test. Furthermore, a goodness-of-fit of the latent variable model is obtained by calculating Pearson's chi-square statistic between the observed and predicted outcomes.

The value for the latent variable predicted for each subject was then compared with the T-scores. The T-scores are considered the response of the subject and the latent variable is viewed as a subject characteristic. A linear regression analysis is performed to determine the relationship between the T-scores and the latent variables. If the latent variable is measuring a single construct, one would expect a similar monotonic relationship between the latent variable and each of the T-scores. This analysis is conducted per T-score and study. The estimated regression models can be compared with respect to the estimated parameters for the latent variable.

Results of Assessing the Cognitive Measures as a Single Construct

A latent variable model was first fitted to CCHS-CLSA, CSHA, and NuAge datasets separately. Let Y_{ij} be the number of correct words for memory test j of subject i and denote the latent variable for memory by Z_i for subject i. The conditional distribution of Y_{ij} given the latent variable $Z_i = z$ is given by a binomial distribution with the parameters given by the maximal number of words N_j and the probability $p_{ij}(z)$. The latent variable Z_i is assumed to be normally distributed. In mathematical terms, the model is described by:

$$(1) \qquad \begin{aligned} &Y_{ij} = y \mid Z_i = z \sim B\left(N_j, p_{ij}(z)\right) \\ &Z_i \sim N\left(0, \tau_i^2\right) \end{aligned}$$

The probability $p_{ij}(z)$ may depend on subject i through the variables age, sex, and education level. It is assumed to be of the logistic type:

$$(2) \qquad \operatorname{logit}\left(p_{ij}(z)\right) = \beta_{0j} + \beta_A \text{age} + \beta_S \text{sex} + \beta_{EM} \text{edu}_{\text{medium}} + \beta_{EH} \text{edu}_{\text{high}} + z$$

The standard deviation τ_i of the latent variable may also depend on the three variables age, sex, and education level and is of the form:

(3) $$\log(\tau_i) = \eta_0 + \eta_A \text{age} + \eta_S \text{sex} + \eta_{EM} \text{edu}_{\text{medium}} + \eta_{EH} \text{edu}_{\text{high}}$$

The intercepts β_{0j} for test j are related to the probability of answering the correct number of words for males with a low education and thus relates to the difficulty of the memory test. For a 70 year old woman, the values of β_S and $70 \cdot \beta_A$ need to be added to this intercept, since male is the reference group for sex and age was not centralized around the study mean. The intercept η_0 for the standard deviation of the latent variable belongs to the same group with the same restriction to age. Note that $\text{edu}_{\text{medium}}$ and edu_{high} were dummy variables for estimation of the effects of medium and high levels of education compared with a low level of education.

Model (1) with relationships (2) and (3) is investigated per study for its goodness-of-fit. The test scores for each subject can be predicted on the basis of the model. Pearson’s chi-square statistic is calculated between the observed and predicted outcomes. A possible lack-of-fit can be caused by many things, but one explanation is that memory is not represented by just one latent variable. This implies that the tests do not measure one univariate construct. Higher dimensional latent variable models would then be useful, but are more complicated to fit numerically, and they require three or more memory tests in each study.

The parameters for model (1) with relationships (2) and (3), together with an approximate 95% confidence interval are provided in Table 13 for all three studies.

From Table 13 (in the CSHA column), it follows that the Rey memory test is significantly ($p<0.001$) more difficult than the Buschke memory tests (p-values not shown in Table 13). The parameter for the Rey memory test is estimated negatively (-1.444), which implies that the Rey memory test is more difficult or has a lower probability of correctly recalling the words than the Buschke free memory test (which operates as the reference test). The Buschke total recall is easier than the Buschke free recall since the parameter estimate of 2.693 is positive. Apparently, subjects with a medium level of education do not significantly do better than subjects with a low education level for CSHA ($p=0.271$) and NuAge ($p=0.091$), but they do better in the CCHS study ($p<0.001$). High education affects the memory performance positively in all three studies (CCHS: $p<0.001$, CSHA: $p<0.001$, NuAge: $p<0.001$). An interesting observation is that the females in NuAge and CCHS have a better memory than the males (CCHS: $p<0.001$, NuAge: $p<0.001$), but this could not be demonstrated in the CSHA study ($p=0.928$). Age was negatively associated with memory in all three studies (CCHS: $p<0.001$, CSHA: $p<0.001$, NuAge: $p=0.014$).

The discrimination or variance of the latent variable is affected by medium education in the CSHA study ($p=0.038$), but this is not demonstrated in the other two studies (CCHS: $p=0.604$, NuAge: $p=0.257$). Also, sex seems to influence the variability in CSHA ($p<0.001$), but not in the other studies (CCHS: $p=1=0.130$, NuAge: $p=0.879$). The variance of the latent variable is affected by age in CCHS ($p=0.003$) and CSHA ($p<0.001$), but not in NuAge ($p=0.139$).

The latent variable model does demonstrate a strong lack-of-fit using Pearson’s goodness-of-fit statistic. This statistic was calculated over all four tests simultaneously and was equal to 22170 on 19910 degrees of freedom ($p<0.001$). However, the ratio of the statistic with respect to the degrees of freedom is only 1.114, which is close to one. In generalized linear models, this type of ratio would indicate an acceptable goodness-of-fit.

It is mentioned by Bauer, et al.[70] that harmonizing latent variable models requires measurement invariance. Strong factorial invariance means that the mean and variance of the latent variable for each item would be the same across studies. This invariance is not met since a comparison of the parameter estimates included in the standard deviation and the mean of the latent variable across studies are not statistically similar (see the estimates and their confidence intervals in Table 13). On the other hand, the standardized latent variable can still be compared, since the differences with respect to age, sex, and education have been eliminated. It can also be viewed as heterogeneity across studies. Another aspect or form of measurement invariance is consistency or homogeneity of the memory tests. Even though the memory tests may differ across studies, it is not unrealistic that the relative complexity between the memory tests remains consistent across studies. Indeed, for comparing the relative complexity of the memory tests, the heterogeneity between studies is eliminated since the relative complexity is a comparison of the tests within subjects and not across subjects. In this report, this type of invariance can only be investigated by comparing the difference in the two Buschke tests between CSHA and NuAge. If this consistency would hold true, then both studies have implemented these tests in the same way, or they may operate in both studies in the same way, even though there may exist study heterogeneity. In this report, it means that the estimates 2.693 and 2.291 for CSHA and NuAge should be statistically the same. Homogeneity in the relative complexity of the Buschke tests was conducted with a likelihood ratio test. This test investigates if the latent variable model with the two estimates 2.693 and 2.291 is similar to the latent variable model where the two estimates can be reduced to one and the same estimate. The likelihood ratio test was determined at 49 with one degree of freedom and indicated that homogeneity in the relative complexity of the Buschke tests is rejected ($p<0.001$).

The univariate latent variable model demonstrated that there exists heterogeneity in the memory tests across studies in two ways. Firstly, differences in test scores between subgroups of subjects on each memory test separately (measurement invariance) are not consistent across studies, and secondly, differences between memory tests for any subgroup (relative complexity of memory tests) are also not consistent between studies.

Results of Latent Variable Versus T-Scores and Other Variables

A comparison between the latent variable model and the other methods is needed to investigate if the methods do in principle harmonize the studies in the same way, albeit their different statistical formulations. Indeed, if the T-scores can be perfectly predicted by the latent variable, the T-scores are just another view towards the same latent variable model.

Latent variables were created using model (1) with relationships (2) and (3), with the restriction that the effect for Buschke total recall is the same for both studies (assuming homogeneity of the Buschke tests across studies). Under this model the latent variable Z_i for each subject can be predicted using the best linear unbiased prediction or the empirical Bayes estimator. The latent variable will be standardized with the standard deviation τ_I to make them comparable across studies. These will be referred to as the standardized latent variables.

The standardized latent variables were compared with the T-scores per study. This was done through a graphical investigation (Figures 1 to 7) by taking the T-scores as a response variable and the latent variable as a covariate since this is considered the true value for memory behavior for individuals. The investigation shows some kind of quadratic relationship between the T-scores and the latent variable. These quadratic relationships were estimated per T-score and study. This means that the latent variable model is strongly related to the method of T-scores. However, the predictions are not perfect, which could indicate that the T-scores contain information that is not captured by the latent variable model. Furthermore, it is clear from Figures 1 to 7, that the T-score for the Health Utilities Index (HUI®) has a completely different relationship with the latent variable (the T-score for the HUI is found in Figure 1. The T-scores for the relationship between other measures in the latent variable are presented in Figures 2 to 7). This implies that the HUI index may not represent the same information on memory tests, which may not be surprising considering the nature of the HUI compared with the other memory tests. However we do believe that the HUI provides valuable information on the construct of memory in the latent variable model. Indeed, projecting the dots in Figure 1 onto the horizontal axis would still provide one unimodal distribution (e.g., normal distribution) for the latent variable instead of several distributions that are distinct in clear subgroups or intervals. This would not be the case when the dots would be projected onto the vertical axis. This implies that the HUI is less suitable for the use of T-scores, since it represents a more qualitative form of information, but it is suitable for the latent variable model since it extracts the relevant information from HUI for the construct of memory.

Methods of Comparison of Statistical Harmonization

The objective of the second set of analyses was to examine how well the results of these statistical harmonization methods agree with each other (correlation analysis) and whether they would provide similar results if a traditional meta-analysis was undertaken. When one uses meta-analysis to estimate a combined effect from a group of similar studies, there needs to be a check that the effects found in the individual studies are similar enough that one can be confident that a combined estimate will be a meaningful description of the set of studies. In the latent variable analysis to assess inferential equivalence, however, we identified that the cognitive measures may not reflect one underlying construct, thus it was also important to see how well these different methods of statistical harmonization identified heterogeneity in the meta-analyses. One would expect the individual effect estimates among the studies to vary by chance; some variation is expected. The question is whether there is more variation than would be expected by chance alone. When this excessive variation occurs, it is called statistical heterogeneity. Based on our latent variable analysis we would expect that statistical heterogeneity would be present. In particular, we hypothesized that the Rey and the Buschke Free would be most similar as they are both measures of free recall. The Buschke Total and the HUI were slightly different measures of the memory construct. In this section we also present descriptive information on the exposure, outcome, and covariates of interest.

Individual Dataset Analyses

We first correlated the resultant values of the constructs harmonized using different methods with each other and with the raw cognition scores using Pearson linear correlation coefficients. This analysis examined the extent to which information was lost in transforming the cognition variables. A correlation of one would imply we have not transformed our data in any way as it would be a perfect linear translation. A correlation of zero would indicate that our transformed variable had no relationship with the original scores.

We then used linear regression analysis to examine bivariate relationships between the 13 variables identified (with acceptable pairing status) with the raw and derived cognition variables. These analyses were done to determine which variables were related to memory scores, and thus could potentially explain heterogeneity if it were to be present. For each regression we report the coefficients for intercept and slope, the p-value for the slope, and the proportion of variance explained, R^2. A multivariable model was then created for each cognitive variable using all 13 variables. Any covariate that had a statistically significant relationship with any of the cognitive measures at the $p<0.05$ level in at least one of the studies was included in a common set of covariates that was used to create adjusted estimates for meta-analyses (see next section).

Combined Dataset Analysis—Aggregate Data Meta-Analysis

For each dataset, a number of summary effect estimates were calculated that were used in traditional aggregate data (AD) meta-analyses (Table 12).We chose to compare the statistically harmonized memory measures in participants reporting no or low physical activity to those reporting moderate or high levels of physical activity. Physical activity was chosen because of its known association with cognition[160] and based on the results of the regression analyses from the individual dataset analyses.

A meta-analysis was conducted for each possible combination of derived cognitive measures across the three studies. Because the CCHS-CLSA included two measures (Rey [R] and HUI [H]), the CSHA included three measures (R, Buschke-Free [BF] and Buschke-Total [BT]), and NuAge included two (BF and BT), there were 12 possible combinations of cognitive measures: R-R-BF, R-R-BT, R-BF-BF, R-BF-BT, R-BT-BF, R-BT-BT, H-H-BF, H-H-BT, H-BF-BF, H-BF-BT, H-BT-BF, and H-BT-BT. Furthermore, these combinations could be analyzed using raw data, T-scores, or C-scores. The effects of using unadjusted means and means adjusted for a common set of covariates identified as potential confounders in the regression analysis were explored. For example, we started with an unadjusted effect estimate based on the raw data in each study. A mean and standard deviations were calculated for the “low” and “high” physical activity groups. A linear regression model was then used to estimate least square means and standard deviation for the “low” and “high” physical activity groups after adjusting for the common set of potential confounders. One additional analysis was done for the raw data to simulate what one would typically find if a systematic review and AD meta-analysis were conducted. Often one is limited to summary data that are not commonly adjusted across datasets. We created one additional set of AD adjusted estimates for the raw data which was the least square means adjusted for all variables that were statistically significant with that outcome within a study. For example, in CSHA, the BF AD least square means were adjusted for age, height, weight, and country of birth, while the NuAge AD BF least square means were adjusted for sex, age, height, and weight.

As the purpose of this report is to explore methods of statistical harmonization, it was assumed that these particular cognitive tests indicate the same construct and could be quantitatively combined across studies. Meta-analyses using random effects models proposed by DerSimonian and Laird[163,164] for the weighted mean difference and Hedges’ g were conducted using MetaAnalyst 3.0.[165] We used a test based on the deviations of the individual study estimates from the summary estimate of effect (the Q statistic) as our primary method to test for heterogeneity.[166] To supplement this test, the I^2, a statistic to quantify heterogeneity, was calculated to describe the proportion of the variance in the point estimate due to heterogeneity rather than sampling error.[167] Although there are no strict rules for interpreting I^2, a rough guide is that an $I^2>50$ percent may represent substantial heterogeneity.[168] Ten separate meta-analyses were conducted for each of the 12 combinations of derived cognitive measures, unadjusted, unadjusted using participants with complete data for all potential confounders, and adjusted means were combined using the raw data, and derived T-scores and C-scores (9 analyses). The 10th analysis included the raw data adjusted for study and outcome specific covariates. For ease of comparison, the Hedges' g analysis is presented for all cognitive measures. Meta-analysis results for the cognitive measures in their original units are included in Appendix D.

Finally, a single latent variable reflecting the memory construct was calculated for each participant of the three studies. Because there was a single latent memory construct hypothesized to underlie the cognition measures, a single meta-analysis was conducted for the unadjusted latent variable estimates. An unadjusted analysis including participants with complete data for all potential confounders and an adjusted latent variable meta-analysis were also conducted.

Results of Comparison of Statistical Harmonization

Individual Dataset and Descriptive Analyses

Table 14 presents the frequencies of the correct number of words identified for the four memory tests for CSHA and NuAge. The test scores are used individually to develop T-scores and C-scores, and they are used simultaneously to determine a latent variable for memory.

Table 15 displays Pearson correlation coefficients among the cognitive measures and between each cognitive measure and the latent variable. If the measures were inferentially equivalent, one would expect a high correlation among the individual measures, and a high and consistent correlation between each of the cognitive measures and the latent variable. In the CCHS, the HUI and Rey were not strongly correlated, (r=0.12), although both measures were correlated with the resulting latent variable (r=0.79 Rey and r=0.56 HUI). In both CSHA and NuAge, the Buschke Free and Buschke Total were moderately correlated (r≥0.58). The Rey and the Buschke Scores in CSHA were weakly correlated (r=0.33 (free) and r=0.27(total)). In the CSHA and NuAge, the latent variable was highly correlated with the Buschke scores (r≥0.80); in the CSHA, the correlation between the Rey and the latent variable was 0.58.

Table 16 provides information on the covariates that are used in the statistical analyses and how each was operationalized in the adjusted analysis. It should be noted that the T-scores and the latent variable approach only correct for age, sex, and education. Table 17 summarizes the bivariate relationships between the raw cognitive measures and the variables of interest by study. None of the variables, except for age, were statistically significant for all measures in all studies. The gender effect differed by measure. For example, in the CCHS it was significant for the Rey but not the HUI. Level of education was related to Rey and HUI, but was less strongly related to the Buschke. The relationship between the cognitive measures and anthropometric measures, weight, height, and body mass index differed within and between studies. Having a higher level of physical activity was strongly related to better cognitive measures in all studies except NuAge; however the sample size of the NuAge study was smaller than the other studies. The relationship between the chronic conditions and cognitive measures also differed among the studies. Self-reported diagnosis of diabetes and myocardial infarction were all related to the Rey and the HUI in CCHS, but high blood pressure was only significantly associated with the Rey. Appendix E includes the bivariate regression results for the T-scores and C-scores.

When one conducts meta-analyses using observational data, it is important to consider the impact of potential confounders on the overall analysis. If there are strong confounders and the study-specific effect estimates are not uniformly adjusted, one would expect statistical heterogeneity to be introduced. The data presented in Table 17 indicates that there were no consistently strong potential confounders among the studies.

Combined Dataset Analysis—Aggregate Data Meta-Analysis

In evaluating the different approaches to statistical harmonization, we applied each method to our cognitive measures. We then examined the relationship between physical activity and cognition across the three datasets. In each dataset we compared the mean cognition score between two groups, low physical activity and high physical activity. In this case, the effect size is the difference in mean score between the groups.

Tables 18a to18c provide a summary of the data comparing a high level of physical activity and a low level of physical activity included in the AD meta-analyses. For example, the average Rey score for those with a high level of physical activity in the CCHS was 4.8 (1.95) compared with 4.4 (1.91) in the low physical activity group. Table 18a displays the raw data, Table 18b includes the average values for the physical activity groups after adjustment for a common set of covariates, and Table 18c includes the average values for the activity groups after adjustment for only those statistically significant variables for that study and outcome combination. The adjusted mean differences were slightly attenuated compared with the unadjusted means, but there remained a statistically significant difference for all cognitive measures in CCHS-CLSA and CSHA. In the NuAge study, the direction of the relationship was reversed (i.e., those with lower levels of physical activity had higher cognitive scores), but the difference was not statistically significant. There was very little qualitative difference between the mean scores adjusted for the common set of covariates, compared with the difference when only statistically significant covariates were used for adjustment.

Tables 19a to 19l present the AD meta-analysis results for the 12 possible combinations of cognitive measures. Table 19m provides an overall summary of these results presented in Tables 19a to 19l. The overall estimated effect size was small, ranging from 0.08 to 0.18. Only two of the 60 analyses including the HUI were statistically significant. The cognitive measure combinations most likely to result in a statistically significant overall estimate were the R-R-BF and R-BT-BF (6 out of 10 comparisons). In most analyses significant heterogeneity was found. Only 23 of the 120 analyses had a p-value greater than 0.05, and five had a p-value greater than 0.10. The analyses with the least heterogeneity were associated with the R-R-BF and R-BT-BF combinations (5 of 10 analyses with $p<0.05$), and the H-BT-BF combination (4 of 10 analyses with $p<0.05$). Of the 23 analyses that did not indicate statistically significant heterogeneity at the $p<0.05$ level, 14 included T-scores, and all five analyses indicating a lack of heterogeneity at the $p<0.10$ level included T-scores. Consequently, the only four analyses with an I^2 value less than 50 percent (indicating an "acceptable" level of homogeneity) also included T-scores. Figure 8 is an example of the forest plot where heterogeneity was not observed. The results were similar when T-scores and C-scores were combined in their natural units using weighted mean differences. The results of these analyses are summarized in Appendix D.

Table 20 includes the AD meta-analysis results in which a single latent variable representing the common cognitive construct is measured for individuals within each study. These are equivalent to the AD analyses presented in tables 19a to 19l, but because all of the cognitive measures were combined in a single latent variable, only one set of analyses is presented. Like the T-score, the latent variable is minimally adjusted for age, sex, and education level. Because the cognitive measures were not inferentially equivalent, one would expect significant statistical heterogeneity when we forced the creation of a single latent variable. The average effect size of 0.13 was not qualitatively different based on adjusted for additional variables of interest. In each case, the I^2 and the p-value of the Q statistic indicated significant heterogeneity.

Results of Sensitivity Analyses

This section describes some additional statistical analyses to investigate the robustness of some of our results described in previous sections. This section is organized by different subsections, each subsection relates to a specific topic of the results.

Language Differences and Living Conditions in Latent Variable Model

Buschke total recall is a memory test that is easier than the Buschke free recall memory test. The differences in difficulty were however not consistent for the two Canadian studies CSHA and NuAge. The measurement invariance principle was questioned. Reasons for this inconsistency in the memory tests could have different origins, but two explanations could come from the sample differences between CSHA and NuAge. The latter study contained only French speaking community-dwelling participants while CSHA had English and French speaking participants in combination with community-dwelling and institutionalized people. Therefore, the inconsistency in Buschke memory tests was investigated for French speaking community-dwelling participants only.

In the full analysis of the latent variable model that is reported in Table 13, the difference in difficulty for Buschke total recall between CSHA and NuAge was estimated at 0.40 (2.69-2.29). In our sensitivity analysis on French speaking community dwelling participants this estimated difference changed to 0.41, which is not smaller than in the full analysis. The likelihood ratio test again gives a p-value smaller than 0.001. This implies that the inconsistency in the Buschke memory tests between CSHA and NuAge could not be explained by language differences nor by differences in the living conditions of participants (institutionalized or not).

The latent variable model in the full analysis could have incorporated an effect for language differences and differences in living conditions (institutionalized or not) on the memory performance and the discrimination of the memory tests. For CSHA, this analysis demonstrates that language did not significantly affect the average performance on the memory tests (p=0.055), but institutionalized participants performed worse (p<0.001) than community-dwelling participants. Furthermore, the memory tests did discriminate somewhat better between French speaking participants and English speaking participants (p=0.005), which means that French speaking participants demonstrated a larger variation in memory test scores than English speaking participants. The institutionalized participants showed substantially more variation in the performance of the memory tests than community-dwelling participants (p<0.001).

A comparison of the estimated latent variables for participants that were based on the inclusion and exclusion of language and living circumstances did not demonstrate a large difference. Indeed, Pearson’s correlation coefficient was estimated at 0.979 (95% CI, 0.976 to 0.980), which indicates that the latent variable is minimally affected by language and living conditions. This implies that the results from our selected latent variable model are considered sufficient. It should be mentioned that a correction for living circumstances would be somewhat arbitrary since several participants may not be institutionalized at the start of CSHA, but they could have been institutionalized only a few months later. Furthermore, institutionalization is also related to age, for which we already corrected.

IPD Meta-Analysis

The meta-analysis on the raw scores, T-scores, C-scores, and latent variables in Tables 18 through 20 were concerned with a traditional meta-analysis on the aggregate data. Since the individual participant data were available, we could have done an IPD meta-analysis as well. Therefore, we provided this analysis as a sensitivity analysis to see how much it would provide similar results to the AD meta-analysis.

For the IPD meta-analysis, we unified raw scores into 0 to 100 scales so that scores can be comparable between tests and between studies. We conducted two-stage (Appendix F) and one-stage (Appendix G) IPD meta-analyses. The two-stage approach is similar to our original approach as summary effect estimates are calculated for each study. An overall effect estimate is calculated as a weighted average of the study-specific estimates. We also used a one-stage meta-analysis on all data simultaneously using a linear mixed model. This model incorporates the effect of physical activity with the outcome variable, unadjusted or adjusted for covariates, and a random intercept for studies to model possible heterogeneity across studies.

The results of the two-stage and one-stage IPD meta-analyses are summarized in Appendix F and Appendix G. A similar summary table was provided for the AD meta-analysis discussed in previous section to be able to compare the two approaches. The results using these models were very similar to each other. Comparing the coefficients for physical activity between the IPD and AD meta-analyses demonstrates that the AD meta-analyses give wider confidence intervals than the IPD meta-analyses.

Discussion

To address our overall objectives, this report comprises two environmental scans of the literature on statistical harmonization methods and an application of three methods of statistical harmonization to cognition data from three population-based cohort studies. This application was designed to examine the properties of these different methods of statistical harmonization and how the choice of method impacts meta-analysis results.

Environmental Scans

Summary and detailed information extracted from the articles that were selected from our environmental scan are provided in Tables 1 through 6. They contain articles on meta-analysis of cognitive measures, on statistical harmonization, and on studies of harmonized data. The environmental scan of meta-analyses including cognitive measures revealed that all summary data meta-analyses including cognition outcomes either restricted their analyses to a subset of studies for which cognitive measures with a common scale were used or combined effect sizes across studies. None of the studies formally explored if the cognitive measures should be harmonized.

In our environmental scan of the methods of statistical harmonization we found three general classes of methods. The first class uses a simple linear- or z-transformation to standardize the scale of constructs across datasets. The second class of methods posits that there is a latent factor(s) that underlies a set of measured items that can be modeled, while the third class of methods was an "incomplete data" approach in which multiple imputation procedures or maximum likelihood estimation could be used to impute values for missing items. These items are then used to calculate a common scale that could be combined across studies. Each method has strengths and weaknesses and all require at least some overlap in items or scales among studies to be able to harmonize the data. In the literature scan no examples were found where methods other than creating an effect size or a standardized score were used to combine cognitive scores in an aggregate data meta-analysis.

In general, there was little focus in the literature on methods used to determine the inferential equivalence of variables prior to statistical harmonization. This "pre-statistical" harmonization step may have, in fact, been conducted, but was not often reported. Granda, et al.[69] describe general approaches to harmonization. The authors describe issues around determining cultural equivalence as a component of inferential equivalence. For example, Pluijm, et al.[113] describe harmonizing measures of activities of daily living in older people across six countries. For some specific activities, questions used to collect data were similar, but there were cultural differences in meaning attached to the performance of the activities. For example, in Southern European countries, older people receive help for cutting their toenails even if they do not have any difficulty in completing the task. The implication is that even when variables are standardized by such efforts as the Core Outcome Measures in Effectiveness Trials (COMET) Initiative,[169] careful pre-statistical harmonization is required.[9]

Data from three Canadian population-based studies were compared for this report. There were methodological and design differences among the studies. For example, Canadian Community Health Survey on Healthy Aging (CCHS) recruited community-dwelling participants, while Canadian Study on Health and Aging (CSHA) included both community-dwelling participants and institutionalized participants, and the Quebec Longitudinal Study on Nutrition and Aging (NuAge) included participants who, at baseline, did not have serious chronic conditions. As well, although the data did not arise from different countries, there was the potential that language differences across the study populations could have an impact on the cognitive tests. The NuAge cognitive battery was administered only in French, while the CSHA and CCHS participants could choose to have the cognitive modules in French or English. In the CSHA, Steenhuis and Østbye found that, particularly on language-based cognitive tests, persons tested in French tended to have lower scores than persons tested in English.[170] If this was an indication that the underlying construct being measured by the French and English versions of the test differed, this could impact the decision to create a single latent variable or even to combine these studies. Since part of our interest was to see whether the different methods of statistical harmonization were able to identify "important" heterogeneity, we combined these data. It was of interest, however, that although adjustment for covariates slightly attenuated the mean cognitive scores within the studies, it had little impact on the qualitative conclusions of the meta-analyses.

The item response theory methods from the environmental scan also focused on harmonizing data when a number of individual items are available for each study, with some overlap of items between studies. We chose to use summary measures instead of individual items of the tests in this report as these best represent the type of data one would be including in a comparative effectiveness review (CER). This choice required us to have more than one summary variable per study to attain model convergence. Note that in item response theory models three items are required to be able to estimate the latent trait properly. Therefore, in CCHS-Canadian Longitudinal Study on Aging (CLSA) we included an additional memory subscale, the Health Utilities Index (HUI®). In our latent variable analysis it was assumed that the overall test scores of a subject were influenced by a univariate continuous latent variable and the overall test scores are viewed as counts representing a correct number of scored test items. The HUI, however, is slightly different in that it is not a count but an ordinal scale. It was interesting that although the results indicated that the T-score for the HUI had a completely different relationship with the latent variable, the relationship between the Rey from CCHS-CLSA (after including the HUI in the latent variable model) and CSHA with their comparable T-scores, was similar. It should be noted, however, that the quadratic relationship between the T-score and the latent variable model was still more similar for the Buschke tests between the CSHA and NuAge than the relationship between the Rey tests between the CSHA and CCHS-CLSA.

Comparison of Statistical Harmonization Methods

Analyses included three methods of statistical harmonization identified from the environmental scan. Two methods standardized the cognition variables to a common metric. The final method used a generalized linear mixed model to identify a univariate underlying latent variable for subjects that would measure the construct common to all studies. Information on the latent variable model was presented in Table 13 and comparisons between the different choices of outcome measures (T-scores, C-scores, latent variable) are provided in Tables 15 and 17.

We did not explore the missing data-type methods as they are more amenable to constructs that are collected using a large number of items across studies, most of which are overlapping. As our construct of interest was an overall score we did not pursue this type of statistical harmonization. It should be noted, though, that the latent variable model is estimable with incomplete data, when item or scale pairing has been observed in at least one study, and can be used to predict missing items.

In our aggregate data meta-analyses that were reported in tables 18 through 20, we found some difference based on the method of harmonization. In terms of the magnitude of effect in the unadjusted analyses, the Hedges' g based on the raw data and the C-scores were most similar. The T-score effect sizes tended to be somewhat attenuated, but this is due to the inherent adjustment for age, sex, and education level of the T-scores. Adjustment for a common set of covariates attenuated the effect sizes for all methods, but there were no systematic differences in the magnitude of the effect sizes across the methods of harmonization.

Because of the methodological differences among the studies it was anticipated that residual heterogeneity would exist that would not be accounted for by the variables included in the analyses. It was also hypothesized that the Rey and the Buschke Free would be most similar as they both are measures of free recall. If heterogeneity did exist, it would be most evident when the Buschke Total and the HUI were being combined with the Rey and Buschke Free. In most cases there was a clear indication of heterogeneity. Using a criterion of P_Q<0.05, 99/120 (83 percent) of the analyses indicated there was statistically significant heterogeneity. Although, 5 of the 21 cases in which heterogeneity was not indicated combined the Rey and Buschke Free only, this was not consistent with a prior hypothesis that these cognitive measures would be most similar.

When heterogeneity was not found at the P_Q<0.05 level, 14/23 (61 percent) of the analyses involved the T-Score. When a more conservative level of P_Q<0.10 was used, all four analyses in which heterogeneity was not indicated using the P_Q<0.10 criterion involved a T-score. As the T-score is adjusted for age, sex, and education, it may tend to make the study-specific measures more similar. The other analyses indicating a lack of heterogeneity included adjusted estimates; however the only analysis with an I^2 <0.5 included T-scores. The relative inability to detect heterogeneity when using the T-score standardization makes it a less desirable method of statistical harmonization.

The latent variable analyses were able to further explore these relationships. The effect of sex, age, and education on the latent variable was different across studies. Both the mean and the standard deviation of the latent variable were different, which indicates clear heterogeneity across studies. This heterogeneity was demonstrated in most, but not all, meta-analyses. Furthermore, evidence was found that the Rey was significantly more difficult than the Buschke Free recall in CSHA but, as expected, it was also demonstrated that the Buschke Total recall is substantially easier than the Buschke Free recall in the CSHA and NuAge. This may suggest consistency in the Buschke test across the studies, which would be one form of measurement invariance.[70] Unfortunately, there was a significant lack of homogeneity for Buschke tests across the CSHA and NuAge, which means that there was no measurement invariance for the Buschke tests. This type of analysis which informs the potential combinability of constructs, however, is not possible without at least some individual-level data.

The ability to detect heterogeneity was used as a criterion in assessing the desirability of statistical methods of harmonization in summary data meta-analysis. When undertaking such an analysis, the underlying assumption is that the construct being measured is the same across measures and datasets. Usually in a meta-analysis, one does not have access to the individual participant data that allows the exploration of this assumption. If the assumption was valid, and there was no bias introduced by using more straightforward methods of statistical harmonization, such as standardization, would be the most desirable method as it is easily implemented and can be computed using standard software. The criterion to measure the appropriateness of a statistical harmonization method, therefore, may differ by context.

Issues in Applying Statistical Harmonization Methods

Data from observational studies are presented in this report; methods to retrospectively harmonize outcome, exposure, and covariate data were used. If one were applying harmonization methods to a meta-analysis of RCTs, using unadjusted measures of effect is generally more appropriate, and thus the inclusion of covariate data would not be warranted. There are situations, however, when one is interested in effect modification in which combinable covariate data are required. As well, in the context of evaluating harms, one is often limited to nonexperimental data. In such a case, both qualitative harmonization (the processing of study-specific data items using logical algorithms, often applied to create dichotomous or categorical variables) and statistical harmonization may be required.

Any study making comparisons across time or groups assumes measurement invariance. It is possible to test this assertion using factorial invariance when some items are measured across studies. A logical hierarchy of constraints in factor models begins with configural invariance,[27] requiring that the same number of factors and pattern of salient factor loadings be equivalent across groups. It is identified by fixing the factor estimate (i.e., mean, variance) in one group and propagating the identification by constraining the corresponding parameter (i.e., intercept, loading) of a reference indicator across groups. Using this as a baseline, Meredith's[25] hierarchy of constraints[26,28] is fit to the data: (1) weak factorial invariance (also known as "metric" invariance) involves equivalence of factor-variable regressions (i.e., factor loadings); (2) strong factorial invariance adds constraints on manifest intercept (mean) terms and requires that one factor mean be fixed to specify the metric of the latent variables; and (3) strict factorial invariance further constrains unique variances to be equivalent. Factor variances, covariances, and nonreference factor means must be freely estimated, as factorial invariance concerns only the measurement model and not the structural model. A sequence of nested models from least to most constrained is recommended. Factorial invariance was not evaluated here, given the limited number of tests available across these three studies, but heterogeneity of the Buschke test across studies was observed.

Issues in Applying Statistical Harmonization to Other Types of Data

In this project we included only cross-sectional data. This issue of invariance is further complicated when one considers repeated measures of constructs over time. In our environmental scan, we found that some methods were limited to one observation per participant. For example, the method of Bauer, et al.[70] required independent observations. Other authors used growth curve models to accommodate repeated observations.[73,103,106] Regardless of the method, careful consideration needs to be taken to assure the invariance in a construct over time. When studies use different measures and different followup schedules, the potential for assessing longitudinal invariance across studies may be confounded by differences in measurement and potential differences in the detected pace of natural development. An example of a related issue is practice effects for repeated cognitive measures over time[171] Essentially, on repeat testing, improvement can occur because of an intervention or natural recovery but also because of gains related to prior exposure to the testing materials. The practice effect, a complex process related to decreased anxiety, familiarity with the testing situation, and gains due to strategy and content familiarization, is usually not uniform across measures[172] and can be associated with other factors such as age or initial ability level.[173] As well, some researchers try to diminish practice effects by using alternate forms in serial neuropsychological testing,[174] but this often leads to additional difficulties due to nonequivalent forms. In addition, for multi-occasion models in the context of considerable age-based change, bias may result unless age heterogeneity is small compared with the amount of change expected.[26,29] In such cases, chronological age differences must be adjusted statistically when evaluating factorial invariance or analyses can be performed in age-homogenous subsamples. If heterogeneity across studies completely explains the differences in the latent variable only, differences in tests would be consistent across studies. All of these factors contribute to the complexity of the harmonization exercise.

Although we focused on cognitive measures in this project, our results are relevant to combining any complex construct. The same harmonization issues are relevant when considering outcome measures such as physical functioning,[13,113] quality of life,[171,175] or exposures such as nutritional intake.[101,176] Whereas repeated measures of these constructs may not be associated with "practice" effects, similar issues will arise when instruments are culturally interpreted among clinical sites or adapted over time.

Alternative Statistical Methods for Harmonization

Latent variable models as a class of statistical models, which include factor analysis, item response theory, and the current latent variable model for count data, is a very general and appropriate class of models for harmonization of psychological measurements. They transform the test results into a continuous measurement of the latent trait that is believed to be underneath the test measures. Our example demonstrated this using the different memory tests. The latent variable models could handle continuous, counts, ordinal, and binary outcomes, although combinations of these types of outcomes would most likely require substantial programming in a sophisticated software package. Truncated continuous variables can also be handled with latent variable models, since this would lead to some form of the Tobit model. The latent variable models would typically implement a single one-dimensional trait, but there is essentially no restriction to use unidimensional traits. This means that multidimensional traits can be implemented as well and these types of models would become similar to structural equation models. Although latent variable models are very general, they do not contain Bayesian methods or multiple imputation methods. These latter two have not been explored in our study but may provide alternative approaches for harmonization.

For instance, Bayesian methods for item response theory have been developed by Fox and Glas.[177] It is argued that Bayesian methods for item response theory are more flexible than the maximum likelihood approach, which is typically the method of estimation for latent variable models, since Bayesian approaches could handle more complicated data such as complex hierarchical structures. Bayesian methods would use Markov Chain Monte Carlo simulations to be able to estimate the model parameters. Thus Bayesian methods may be very useful, in particular if more complex data would be present. This could be especially true when longitudinal data must be harmonized.

The advantage of multiple imputations is on the other hand less clear a priori. Missing values are mainly "missing by design" because test results are just not implemented in certain studies. For instance, CCHS did not use the Buschke tests on memory, but only the Rey test was part of the design. This means that the missingness process may possibly be viewed as missing at random. In this setting likelihood approaches are appropriate and correct, which means that latent variable models would be appropriate when implemented with likelihood estimation. Furthermore, latent variable models can be viewed themselves as single imputation methods.

They determine a latent variable and a relationship to the tests results which gives the opportunity to determine the test scores for the missing tests. The benefit of multiple imputation methods may possibly come from improved estimates of the standard errors. It would be of interest to investigate if multiple imputation approaches may provide other conclusions towards issues as heterogeneity.

Overall Conclusions

This report provides empirical evidence to inform methods of combining complex constructs using aggregate data (AD) or individual participant data (IPD) meta-analysis of cognitive measures. Combining data on complex constructs such as cognition can be particularly challenging, and specialized methods of harmonization, including statistical harmonization, are required. The use of these methods of harmonization in the conduct of systematic reviews, however, is virtually nonexistent.

Overall there was a general consistency of our AD meta-analysis and IPD latent variable analysis results. Both methods underscored the difficulty of harmonizing these cognition data. There were multiple examples when using the AD methods, however, in which important heterogeneity was not identified. This masking of heterogeneity happened most often when using a T-score to standardize the cognition scores compared with a C-score or summary latent variable.

When one uses the T-score to create combinable data, there is an assumption that the sample characteristics are identical across studies. For the C-score this assumption is made only for a subgroup of participants. These assumptions may have no effect on an analysis of the data within a study but become an issue when pooling data across studies. The T-scores further reduce the variability within studies by standardizing the cognition constructs to age, sex, and education, but the C-scores are standardized to a common homogeneous age, sex, and education subgroup and the within- population variability related to age, sex and education remain. Within-study standardization to a common metric is generally not recommended given the potential differences in scale distributions and sample composition. The C-Score method aligns the within-study standardization procedure to a common subgroup but differences in measurement scales remain untested, with differences due to unadjusted variables. The latent variable approach is the only approach that allows for the examination of measurement invariance. We recommend item response theory (IRT) or factor-based type of models be used for harmonization procedures but also realize that this may not always be possible for particular research questions. New data collection efforts may be required to obtain the necessary individual-level data to confidently harmonize different indicators of common constructs.

The following guidance is suggested from our results:

- Very careful consideration of inferential equivalence needs to be undertaken prior to combining cognition data across studies. Qualitative harmonization is an essential step prior to statistical harmonization, when undertaking an AD or IPD meta-analysis.
- These results are likely applicable to meta-analyses of other complex constructs such as quality of life, depression, physical functioning, and nutritional status.
- Of the three methods of statistical harmonization for cognition data, T-score standardization is the least desirable compared with the C-score method or latent variable methods. The latent variable method is most desirable as it is the only method that allows the assessment of measurement invariance.

- Assessment of the assumptions underlying statistical harmonization is not possible without some individual-level data. This analysis is required to assess the potential for bias in combining complex outcomes using AD meta-analysis.
- Rigorous measurement approaches for evaluating measurement equivalence on item-level data using items response theory or latent variable approaches are recommended.

Table 1. Summary articles describing meta-analyses of cognitive measures

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Angevaren, M. 2008[37] Individuals 55 years and older participating in RCTs assessing the effectiveness of physical activity on cognitive function RCTs 11 studies included	Physical activity programs	Studies were included if: • they were published RCTs comparing aerobic physical activity programs with another intervention or no intervention • participants were 55 years and older [p.3]	• Simple RT • Choice RT • TMT part A and B • Digit symbol substitution test • Rand memory test story recall • Ross Information Processing Assessment • WAIS • BVRT • Digit span backward and forward • 16 words delayed recall • RAVLT delayed recall trial • WMS • Word comparison • Task switching paradigm • Verbal fluency • Face recognition • Stroop color word test • Stopping task • Digit vigilance • Tracking • Letter search • Finger tapping • Visual search • Pursuit rotor task [p. 23-24]	• Cognitive speed • Verbal memory functions • Visual memory functions • Working memory • Memory function • Executive functions • Perception • Cognitive inhibition • Visual attention • Auditory attention • Motor function [p. 23-24]	• Neuropsychological tests included in the RCTs were organized into a number of categories measuring the same construct • The weighted mean difference was used if studies applied the same cognitive tests and if the outcome measurements were on the same scale. In all other cases, the SMD was calculated [p. 4]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Eilander, A. 2010[44] Children ages 0-18 years participating in studies evaluating the impact of micronutrient supplementation on cognitive performance RCTs 19 studies included	Micronutrient supplementation	Studies were included if: • they were randomized placebo-controlled trials evaluating the effect of micronutrient supplementation on cognitive performance • they were trials focused on healthy children • they reported cognitive results as primary or secondary outcome measures of the intervention • they were trials where children were supplemented with ≥ 3 micronutrients with a placebo [p. 116]	• BAS • DG • CPAS-R • CTBS • GMT • MISIC • NAR • OOHMT • PGI • PMAT-FC • NEPSY • RAVLT • SDMT • WAIS • WIAT • WISC-III/R [p. 123]	• Fluid intelligence • Crystallized intelligence • Short term memory • Visual perception • Retrieval ability • Cognitive processing speed • Sustained attention • Motor skills • Academic performance [p. 116]	• If trials were using more than one cognitive measure to assess one cognitive domain, a standardized mean difference was calculated • Effect sizes were calculated for individual trials by dividing the difference between the mean change in intervention and control group by the pooled SD • Overall mean effect size was calculated by applying the random-effects model [p.116-117]
Falkingham, M. 2010[45] Anemic and nonanemic children and adults RCTs 14 studies included	Intervention groups received an oral iron supplement	Studies were included if: • participants were human and at least 6 years old • participants were randomized to an iron supplementation vs. a control • the length of intervention was at least 4 weeks • the additive effect of iron was clear • and some objective measure of cognitive performance had to be evaluated [p. 2]	• Raven's Color Progressive Matrices • Peabody Picture Vocabulary • Rey Auditory Verbal Learning • Wechsler's Digit Span • Mazes test • Hopkins Verbal Learning Test • Bourden-wisconsin concentration [p.4]	• Attention and concentration • IQ • Memory • Psychomotor • Scholastic achievement [p. 5]	• The inverse variance method was used for the meta-analysis • Standardized mean differences were used in random effects meta-analysis • Heterogeneity was measured using the I^2 statistic [p. 3]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Guilera, G. 2009[48] Patients diagnosed with schizophrenia RCTs 18 studies included	Antipsychotic medications	Studies were included if: • they were RCTs comparing atypical with typical antipsychotics in adults • they focused on participants with a diagnosis of schizophrenia, schizo-affective or schizophreniform disorder • they used a standardized neurological test listed in the Lezak manual [p. 3-4]	• Stroop, CPT, TMT-A and B, CANTAB Rapid Visual Information Processing Test • WCST, WISC-R, WAIS, HVOT • WMS-R, CVLT, RCAVLT, HVLT • NART, Peabody Picture Vocabulary Test • Finger tapping, Grooved Pegboard test • Benton Judgment of Lines [p. 3]	• Attention and vigilance • Automaticity and procedural learning • General intellectual functioning • Language and verbal comprehension • Perceptual processing • Psychomotricity • Reasoning and problem solving • Speed of processing • Verbal learning and memory • Visual learning and memory • Working memory [p. 3]	• Effect sizes were combined using a random effects model to create a weighted mean estimate for each of the cognitive domains and the global cognitive index • Pooled effect sizes were weighted by applying the inverse variance method [p. 5]
Hogervorst, E. 2010[49] Postmenopausal women RCTs 38 studies included	Hormone replacement therapy	Studies were included if: • they were RCTs assessing hormone replacement therapy and its effect on cognitive function in postmenopausal women • they used a placebo treated control group • they included cognitive measures [p. 65]	• paragraph recall, story recall • COWAT, FAS • Face recognition, BVRT, RAVLT • SRT, Digit Span • Stroop, TMT-B • MMSE [p. 66]	• Verbal memory • Verbal fluency • Visual memory • Concentration • Executive function • Visuospatial [p. 66]	• A table driven meta-analytical approach was used to address the difficulty in comparing individual tests • Analyses were conducted using general linear models with post hoc Tukey tests or M-W *U*-tests and Chi-Square for categorical data • Associations between continuous data were assessed by conducting Spearman's rank correlations [p. 65-66]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Hogervorst, E. 2009[50] Postmenopausal women that have been diagnosed with Alzheimer's disease or other dementia syndromes RCTs 7 studies included	Interventions included administering estrogens alone or combined with a progestogen	Studies were included if: • they were double-blind RCTs focusing on the effect of ERT and HRT on cognitive function • the treatment period was at least 2 weeks • participants were postmenopausal women with Alzheimer's disease or other types of dementia [p. 1]	• MMSE, BIMC, ADAS-Cog, HSD • WMS, BSRT, CERAD, Digit span • VRT, visual span, face recognition, ROVMT • Boston naming test, Token test • TMT-A and B, DSST, Stroop, Digit Span backward [p. 2]	• General cognitive function • Verbal memory • Visual memory • Language • Speed and efficiency of information processing [p. 4]	• Weighted mean difference was used in the meta-analysis when studies used the same treatment and test outcome measure. A fixed effect model was used if there was no significant heterogeneity • Standardized mean difference was used with either fixed effect or random effects models in studies where they used different types of treatment or different tests to measure the same construct [p. 6]
Li, H. 2011[51] Individuals with mild cognitive impairment RCTs 17 studies included in meta-analysis [p.287]	Cognitive training for people with mild cognitive impairment	• Study should focus intervention on MCI group and include pre- and post-test data of the intervention group • Must include cognitive stimulation/training or cognitive rehabilitation method • Study must provide means, standard deviations, t test or F test and sample size of intervention group • Study should have at least 1 dependent variable of cognitive test or functional ability assessment [p.286]	• Episodic and semantic memory tests • Trail Making Test (Part A and Part B) • WCST • Figure rey-copy, pattern and picture reproduction, facial recognition test • MMSE [p. 286-287]	• Memory • Executive functioning • Attention/processing speed • Visuospatial ability • Changes of the Mini Mental State Examination (MMSE) during the intervention • Emotional state—focusing on depression and anxiety [p. 286-287]	• Effect sizes (Cohen's *d*) for the differences in post- and pre-test, followup test and pre-test performances of the intervention were calculated (differences between the means of pre-test and post-test, divided by the pooled standard deviation) • When the studies included several tests of a certain domain, each domain of tests was averaged to one pooled effect size. [p.287]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Karsdorp, PA. 2007[53] Children and adolescents with congenital heart disease (CHD) RCTs 11 studies included	No intervention	Studies were included if: • they were published in a peer-reviewed English or German journal • they only used patients with CHD • participants were between 2 and 19 years old (mean age ≥4) • participants had all had surgery or interventional catheterization • they used the CBCL (parent form) and/or measures of cognitive function • they provided data required to calculate effect sizes • they used a control group [p. 529]	• BAS • BSID • DAS • HAWIE • HAWIK • HAWIVA • KABC • LIS • MSCA • SB • WISC • WPPSI [p.533]	• Specific domains not reported	• Meta-analytic procedures focused on effect sizes modeled after the techniques of Hunter and Schmidt • Standardized mean difference was used as the estimate of effect size • Weight mean effect size was calculated to account for differences in sample size [p. 529]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Lethaby, A. 2008[55] Healthy women who had undergone natural or surgical menopause RCTs 16 studies included	Hormone replacement therapy	Studies were included if: • they were double-blind RCTs assessing the effect of estrogen replacement therapy or hormone replacement therapy on cognitive function • the treatment period was at least 2 weeks • participants were postmenopausal women [p. 1]	• CAMCOG, Folstein Mini-Mental State Examination, MMSE • WMS, CVLT, Boston naming test • VRT, BVRT, VMT • TMT-A, DSST, Finger tapping, Grooved Pegboard test, Stroop • Letter cancellation tests • WCST, WAIS, COWAT, Digits backward [p. 5]	• Global cognitive function • Verbal memory and language • Visuospatial • Speed tests, Attention and Manual Dexterity Semantic Memory • Mental rotation tests and accuracy • Executive function [p. 5]	• Using the fixed effects model, the weighted mean difference was calculated for continuous data • Odds ratios were calculated for dichotomous data through the use of a fixed effects model • WMD was used as the outcome measure in trials that were measuring the same outcomes • When trials were not comparable (i.e., different tests within the same cognitive domain, different kinds of participants or different interventions) the standardized mean was calculated using a random effects model [p. 7]
Marasco, SF. 2008[56] Patients with coronary artery bypass grafting RCTs 8 studies included	Off-pump (beating heart) coronary artery bypass grafting vs. on-pump	Studies were included if: • they were prospective randomized control trials comparing off-pump vs. on-pump coronary artery revascularization • neurocognitive testing was conducted [p. 962]	• Rey Auditory Verbal Learning • Grooved Pegboard • TMT—Part A and B • WAIS III • Digit Symbol substitution test [p. 963]	• Verbal memory • Motor capacity • Divided attention and executive function • Information processing [p. 963]	• Outcomes were analyzed as continuous variables • Weighted mean difference was calculated for each outcome • A fixed effect or random effects model was selected based on the degree of heterogeneity [p. 962]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Martin, M. 2011[36] Healthy older people (age 60 yrs and older) and people with mild cognitive impairment RCTs 36 studies included	Mental training, problem solving training, speed training, cognitive restructuration technique	Studies were included if: • they described cognitive training specific domains of cognitive function such as memory, attention, or speed • they were published, written in English or German, and presented in a journal article • they were an RCT • they had a minimum of 2 measurements and assessment [p. 4]	• problem solving • verbal episodic memory • Luria • TMT • visuomanual coordination • short term memory • immediate recall • recent logic execution memory • abstraction proverbs • phonematic fluency • IADL • Guild Memory Test • supermarket test • subjective memory tests • Geriatric Depression Scale • alpha span (working memory) total score • Brown-Peterson: secondary memory, primary memory • free recall of digit spans • UFOV • Road Sign Test • letter comparison • global auditory memory score • letter series test • word series test • letter sets test • Hopkins Verbal Learning Test • Rivermead Behavioural Memory Test	• Memory • Attention • Speed [p. 4]	• When rating scales used in the trials had a number of categories > 10 the data were treated as continuous outcomes arising from a normal distribution • For binary outcomes the odds ratio was used to measure treatment effect • A weighted estimate of the typical treatment effect across trials was calculated [p. 7]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Martin, M. 2011[36] (continued)			• Hopkins Prospective Memory Task questionnaires: Memory Controllability Inventory, Memory Functioning Questionnaire • short story recall • letter and semantic verbal fluency • Raven matrices • Rey figure • immediate and delayed recall of words • shopping list • name-faces • Benton visual retention test • total word recall • long-term retrieval • vocabulary subtest of Wechsler Adult Intelligence Scale-Revised • face-name recall • NEO-PI • Physical and cognitive variables Wechsler Memory Scale		

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Metternich, B. 2010[57] Patients reporting subjective memory complaints (SMC) RCTs 14 studies included	Nonpharmacological interventions (i.e., Mental Training, Psycho-educational programs, physical training)	Studies were included if: • they focused on nondrug interventions • they were published in peer-reviewed journals in English, Dutch, German or French • they were RCTs • if they reported sufficient data to conduct the meta-analyses [p. 8]	• MIA, MCI, MFQ • Face-Name Task • CVLT, HVLT, Visual Verbal Learning Test, WMS, Buschke Selective Reminding Test, Guild Memory Test [p. 8]	• Subjective memory • Objective memory [p. 8]	• Standardized mean differences of change scores were calculated for all comparisons • Due to the small sample size of some of the trials, Hedges' adjusted g was used • To account for clinical and methodological diversity, a random effects model was used to summarize individual effect sizes [p. 9]
Repantis, D. 2010[62] Participants enrolled in studies looking at using modafinil and MPH for neuro-enhancement RCTs 91 studies included	Modafinil and MPH use	Studies were included if: • they were published single- or double-blind RCTs comparing MPH or modafinil with a placebo • participants showed no signs of psychiatric disorder, cognitive decline or other diseases [p. 188]	Not reported	• Attention and vigilance • Memory and learning • Executive functions and information processing [p. 201-202]	• A standardized effect difference was calculated for the appropriate test parameters of each study • A linear mixed model was applied in the analysis, to account for heterogeneity and correlation within studies • A meta-analysis and meta-regression were conducted based on the linear mixed model [p. 189]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Woodward, ND. 2007[67] Prospective, double-blind randomized trials 16 studies included	Atypical antipsychotic drugs	• Studies were included if: • they involved participants diagnosed with schizophrenia or schizoaffective disorder • they used a prospective study design with a baseline assessment and a minimum of one followup • the trial lasted at least 1 week • no antipsychotic meds were administered with the exception of the study meds • there was a minimum baseline sample size of 10 • the study was published in a peer-reviewed journal as of April 2004 • findings of neuropsychological change to treatment were reported for at least one of the identified tests [p. 213]	• TMT-A and B • Continuous performance test • Digit symbol substitution/modalities test • WCST • CVLT, RVLT, BVLT • Controlled Oral Word Association Test, Category Instance Generation Test • Finger Tapping/Oscillation Test • Grooved Pegboard Test [p. 214]	• Attention • Processing speed • Executive function • Verbal learning • Delayed verbal recall • Verbal fluency • Motor skill [p. 214]	• Effect sizes were calculated for overall cognitive function by calculating a Global Cognitive Index • In cases where studies failed to report a standardized cognitive summary score, the Global Cognitive Index was determined by averaging effect sizes across all neuropsychological tests used in the study • A weighted average effect size was calculated for the Global Cognitive Index and specific neuropsychological tests by combining effect sizes across studies following the fixed effects model [p. 214-215]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Campbell, LK. 2007[43] Children that have received treatment for Acute Lymphocytic Leukemia (ALL) Observational 28 studies included	Treatment vs. healthy comparison group	Studies were included if: • they were published in English • they included original data on post-treatment neurocognitive functioning of childhood ALL patients • they included valid and reliable neurocognitive measures • they included published normative data [p. 65]	• WPPSI, WISC, WAIS, Standford-Binet, MSCA, K-BIT, KABC • WRAT, Woodcock-Johnson Tests of Achievement • Digit span, TMT, Stroop • WISC • Finger tapping, Grooved Pegboard, Purdue Pegboard • WRAML, CVLT, RAVLT, Buschke Selective Reminding Task • VMI, Rey, Weschler Block design, Judgment of Line Orientation • Benton Visual Retention [p. 66]	• Overall cognitive functioning • Academic achievement • Attention • Executive functioning • Information processing speed • Psychomotor skill • Verbal memory • Visuospatial skill • Visuospatial memory [p. 66]	• A random effects model was applied to calculate Hedges' g for each study outcome • Effect sizes were averaged to calculate a single effect size for the analysis if multiple measures were applied to evaluate the same neurocognitive domain in a study [p.66]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Goodman, M. 2002[46] Workers that have been exposed to lead and who have not been exposed Observational 22 studies included	Exposed to lead in the workplace	Studies were included if: • there was a central tendency for blood concentration <70 µg/dl • totals of exposed and unexposed participants were reported • test score means and measures of dispersion were provided for both exposed and unexposed participants [p. 218]	• block design test • logical memory test • digit symbol substitution test • visual interference • similarities • Bento n visual retention • Paired associates • Visual reproduction • Flicker fusion • Arithmetic • Symbol digit learning • Digit span (forward and backward) • TMT—A and B • Simple reaction time • Picture completion • Grooved Pegboard • Tapping rate [p. 221]	• Specific domains not reported	• Fixed and random effects models were both used for each analyses due to the variation in testing procedures, scoring practices, variations in study samples, and assumptions of homogeneity [p. 219]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Grant, I. 2003[47] Cannabis-using and nondrug using participants in studies looking at neurocognitive effects of cannabis use Observational 11 studies included	Cannabis use vs. no cannabis use	Studies were included if: • they included a group of "cannabis only" users • they included a control group • they included appropriate information for calculating effect size • neuropsychological tests were included in outcome measures • cannabis group was drug-free on day of testing • they included information regarding other substance use in cannabis group • they included history of neurological or psychiatric problems • they included data on length of abstinence from cannabis before testing [p. 681]	• WAIS-R Digit Span and Digit Vigilance • WAIS-R Vocabulary, Verbal Fluency • WCST, RAVENS • WAIS-R Block Design, Object Assembly • Grooved Pegboard, Finger Tapping • CVLT, RAVLT [p. 683]	• Simple reaction time • Attention • Verbal/language • Abstraction/executive functioning • Perceptual motor • Simple motor • Learning • Forgetting/retrieval [p. 683]	• The method prescribed by Hedges and Olkin was applied based on the assumption that different tests measuring the same cognitive domain would likely be correlated • A fixed effects model was used due to the likelihood of heterogeneity in data across studies • An overall neurocognitive effect size was calculated by pooling effect sizes across domains [p. 683]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Valentini, E. 2010[64] Participants of studies assessing the psychomotor effects of mobile phone electromagnetic fields (EMF) Observational 24 studies included	Exposure to EMF	Studies were included if: • they were in English and reported human provocation/laboratory studies on the effect of modulated RF-EMF on behavioral outcomes • they had an experimental design using real and sham exposure to EMF • the data concentrated on acute effects of GSM/UMTS-like exposures • they reported speed measures they used comparable tasks or used them more than once	• Trail Making Test—version B • Two-choice reaction time, ten-choice reaction time • Simple reaction time task • Subtraction • Sentence verification • Vigilance task [p. 711]	• Attention and speed of processing • Divided and sustained attention • Working memory • Semantic memory [p. 709]	• Statistical values were extracted to calculate the standardized mean difference • Pooled effect size was calculated by weighing each effect size by its corresponding sample size (Hedges and Olkin) [p. 711]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Wheaton, P. 2009[66] Adults with Traumatic Brain Injury (TBI) Independent groups repeated measures design and independent groups design 22 studies included	Drug treatment for TBI	Studies were included if they: • were published in a journal in English • included a TBI control group matched to the treatment group based on age and severity of injury • were not a case study; • had TBI treatment and control groups that had suffered nonpenetrating TBIs • both groups were given measures of cognition and/or behaviour to assess outcome • treatment group was given drug treatment in the early stages following the injury • no subject had sustained a TBI prior to the current injury, no pre-existing impairments, no history of mental health problems or substance abuse • participants not recently treated with pharmaceuticals to improve behaviour or enhance cognition • results reported in a manner that allowed for calculation of effect size • subjects were 16 years or older [p. 469]	• PASAT • Story memory [p. 473]	• Attention • Memory • General cognition [p. 474]	• Cohen's *d* was used to calculate effect size • Effect size was calculated for each score for all outcome measures of cognition and behaviour. If there were multiple scores for a particular outcome measure then they were combined and a mean effect size was calculated for that measure • Effect sizes from different studies using the same measure were averaged to permit evaluation of the combined findings [p. 470]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Barth, A. 2008[39] Subjects enrolled in studies investigating the effects of electromagnetic fields emitted by GSM mobile phones Single or double-blind experimental study design [p.343] 10 studies were included in the meta-analysis [p. 342]	No intervention	• Treatment group and control group with baseline measures or repeated measurements of subjects with mobile phone switched on and switched off • Mean and SD of the dependent variables documented for both groups or both times or test statistics • Single-blind or double-blind experimental study design • Exposure by a GSM mobile phone specified at a range of 900 MHz to 1800 MHz under two conditions: on vs. off • Enrolled participants were considered healthy • Study must include at least one neuropsychological test that is used in another study [p. 343]	• SRT • CRT • VIG • SUB • VER • N-Back test • TMT • Digit span forward and backward • Spatial span forward and backward [p.343]	• Information processing • Reaction time • Attention • Memory • Executive functions [p. 343]	• Results of all studies were first transformed into their respective effect sizes, indicating whether there was an effect of mobile phones—these effect sizes were transformed into the meta-analytic delta-measure • An assumption was made that there was a common population effect in the different studies and that a single effect size could be calculated for each individual study • When homogeneity for single effect sizes was not provided, a random effects model was applied [p. 344]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Brands, A. 2005[42] Patients with Type 1 diabetes Cross-sectional design 33 studies	No intervention	Studies were included if: • they were published in English after 1980 and before 2004, • they focused solely on adults 18 years and older with type 1 diabetes, • had a defined control group, • measured cognitive performance using standard neuropsychological or reliable experimental testing methods at normal glucose values, • original studies and test scores were provided for the experimental and control groups or if statistics such as the exact *t* or *F* values were provided [p. 726]	• Measures not reported	• Overall intelligence • Working memory • Learning immediate memory • Delayed memory • Psychomotor efficiency • Speed of information process • Motor speed • Attention • Cognitive flexibility • Visual perception • Language [p. 731]	• Effect sizes were calculated for every result • A combined *d* value was included in the meta-analysis. • The statistic Q was calculated to determine the heterogeneity of the sample • Tests measuring the same cognitive domain were taken together in the analysis. All cognitive domains were pooled into an overall *d* value and then separate meta-analyses were conducted for the different cognitive domains [p. 727]
Sibley, B. 2003[63] Elementary school-aged children Quasi-experimental and cross-sectional 44 studies included in meta-analysis [p.245]	No intervention	• Studies exploring the relationship between physical activity and cognition or academic performance were included • English studies conducted before January 2002 were included [p. 245]	• Measures not reported	• Perceptual skills • IQ • Achievement • Verbal tests • Math tests • Memory • Developmental level/academic readiness [p. 247]	• Effect sizes were calculated for each study and then an overall ES was calculated • Each level of the moderator variables had an effect size calculated • Homogeneity tests were done by partitioning the total variance into between groups variance and within groups variance [p. 245-247]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Balint, S. 2009[38] Adult ADHD patients and control group participants in studies looking at differences in neuropsychological performance in ADHD vs. normal control subjects Observational 25 studies included	No intervention	Studies were included if: • they included at least one measure of attention • they compared performance of ADHD subjects with a normal control group • participants were adults (>18 years) • raw data for effect size calculation included in the paper • ADHD diagnosis obtained using DSM-III-R or DSM-IV criteria • articles were published in English [p. 1338]	• Stroop • TMT • WAIS-R Digit Span and Digit Symbol subtests • CPT [p. 1338-1340]	• Attention (simple, focused and sustained) [p. 1338-1340]	• Pooled effect size was calculated across the studies to determine the differences between the ADHD group and control group • A random effects meta-regression was used to calculate the pooled effect size of the difference between the ADHD vs. the control group [p. 1340]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Bhutta, A. 2002[40] School-aged children who were born preterm Case-control studies 15 studies were included in the meta-analysis [p.729]	No intervention	• Studies had to have a case-control design • Had to report cognitive data, behavioral data or both • Had to perform evaluations after the fifth birthday of participants • Had to have an attrition rate of less than 30% • Had to be published in 1980 or later [p.729]	• BAS • MIQS • WPPSI • KABC • WISC [p. 731]	• IQ • Other domains not reported	• For each study, the nonstandardized difference between mean cognitive test score of cases and controls was weighted by the inverse of the variance for this difference • The weighted mean differences were then pooled across studies to compute an overall mean cognitive difference between cases and controls • Random-effects and fixed-effects least-square regression models were used for combining the results in the meta-analysis [p.729]
Bora, E. 2009[41] Participants in studies assessing neuropsychological deficits in euthymic patients Observational 62 studies included	No intervention	• Studies were included if: • they included neuropsychological data on remitted adults with bipolar disorder or first-degree relatives of patients with bipolar disorder • they used a healthy control group • they included mean test scores and standard deviations • they used at least one cognitive measure that was used in at least three studies in both bipolar patients and healthy relatives of bipolar patients [p. 3]	• RAVLT, CVLT, VLT • WMS-R • CPT • TMT-A and B • FAS • WCST, CANTAB • WAIS-R Digit Span • Stroop • ROCF • NART [p. 8]	• Verbal learning and memory • Visual memory • Sustained attention • Processing speed • Verbal fluency • Set shifting • Working memory • Response inhibition • Visuospatial abilities • General intelligence [p. 8]	• The standardized mean difference method was used • A random effects model was used as there was heterogeneity for many of the analyses • Meta-regression analyses were conducted with the random effects model using restricted-information maximum likelihood method with a significance level of p<0.05 [p. 12]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Jansen, C. 2005[52] Cancer patients who had or were currently receiving chemotherapy Observational 16 studies included	No intervention	Studies were included if: • they used original study data • they used an adult sample • they conducted neuropsychological testing of cancer patients who had or were presently receiving chemotherapy • they used reliable, valid, and standardized neuropsychologic tests • they reported sufficient information on at least one domain of cognitive function to estimate effect size [p.2223]	• CPT • DRS • HRNB • HSCS • RBANS • RCFT • TMT • WAIS • WMS [p. 2227]	• Attention or concentration • Executive function • Speed of information processing • Language • Motor function • Visuospatial skill • Verbal memory • Visual memory [p. 2224]	• Potential for bias was corrected by weighting the standardized mean difference effect size for each test by the sample size and pooled variance • For studies using more than one test to measure a specific cognitive domain, an average effect size was calculated for that domain • Test producing numerous scores had an average effect size calculated for that test [p. 2224]
Krabbendam, L. 2005[54] Patients with bipolar disorder or schizophrenia Observational 31 studies were included	No intervention	Studies were included if they: • evaluated cognitive performance using standardized neuropsychological testing procedures • compared adult participants with schizophrenia and with bipolar disorder • provided test results of both participant groups, or exact *p*-values, *t*-values, or *F*-values • were published in a peer-reviewed English language journal [p.138]	• Digit Span, Letter-Number Span • DSST, Trailmaking Test Part A • Trailmaking Test Part B, Stroop Color-Word interference • Wisconsin Card Sorting Test [p.139]	• Verbal working memory • Verbal fluency • Mental speed • Executive control • Concept formation and shifting [p. 139]	• An effect size was determined for each test parameter • In situations where means and SD were not provided, *d*-values were calculated from exact *p*-values, *t*-values, or *F*-values • In instances where multiple tests were reported for one cognitive domain, they were combined into one *d*-value • Homogeneity was also tested using the *Q*-statistic [p. 139]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
McDermott, LM. 2009[58] Patients diagnosed with depression Observational 14 studies included	No intervention	Studies were included if: • they included a sample of participants diagnosed with major or minor depression • they reported means and standard deviations of neuropsychological test scores • they assessed the impact of severity of depression on performance on neuropsychological tests [p. 2]	• WCST, word fluency, TMT-B, COWAT, Hayling test B, Stroop, MCST, digit span (forwards and backward), CANTAB • Semantic fluency test, Verbal fluency test, TMT-A, MFFT-20, ZVT, TAP, digit symbol substitution test, processing speed, Grooved Pegboard • RAVLT, AVLT, signal recognition, recognition test [p. 3]	• Episodic memory • Executive function • Processing speed • Semantic memory • Visuospatial memory [p. 3]	• The population effect size was calculated using the study effect sizes and sample sizes • A separate meta-analysis was conducted to replicate results for each cognitive domain and two sets of timed and untimed tests [p. 3]
Naguib, JM. 2009[59] Children with type 1 diabetes Case-control design 24 studies included	No intervention	Studies were included if: • the sample was people ≤19 years old with type 1 diabetes • they used a case-control design • used standardized neuropsychological tests of seven cognitive domains [p. 271]	Not reported	• Intelligence • Visuospatial • Language and education • Memory and learning • Psychomotor activity • Attention • Executive function [p. 274]	• Cohen's d statistic was calculated for every test of a particular study • In each cognitive subdomain a meta-analysis was conducted when a minimum of three studies had evaluated the same subdomain • A fixed effects model of meta-analysis was applied [p. 275]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Nieto, RG. 2011[60] Patients with early onset schizophrenia or pediatric bipolar disorder Observational 12 studies included	No intervention	Studies were included if: • they published in English • they included a healthy comparison group • data were available to calculate effect sizes [p. 267]	Not reported	• Attention • Working memory • Executive control • Visual memory • Verbal learning and memory • Visuospatial skills • Verbal fluency • Processing speed • Motor skills [p. 268]	• Standardized mean differences were calculated using Hedges and Olkin's method (difference between the means of each of the diagnosed groups and control group divided by the pooled standard deviation) • A random effects model was used for the meta-analysis [p. 268]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Quinn, TJ. 2011[61] Adult patients participating in studies studying the association between circulating hemostatic measures and cognitive impairment Cross-sectional, cohort and case control, and longitudinal studies were included 21 studies included	No intervention	Studies were included if: • the sample included only adults (18 years or older) • they were original research • the analysis included measures of at least one circulating blood biomarker • the biomarker was pertinent to hemostasis • the analysis included at least one measure of cognitive function, cognitive impairment or dementia • the data included was case-control, cross-sectional, longitudinal or a combination [p. 1476]	• DST • LM • RAVENS • VFT [p. 1478]	• Speed of processing • Verbal declarative memory • Nonverbal reasoning • Executive function • General composite cognitive function [p. 1478]	• Fixed effect and random effects models were both used. Random effects models were favored when there was substantial heterogeneity and three or more studies [p. 1478]

Table 1. Summary articles describing meta-analyses of cognitive measures (continued)

Source Population Study Design # of Studies Included in Analysis	Intervention	Inclusion Criteria	Types of Cognitive Measures	Different Domains Looked at (i.e., Memory)	How Data Were Combined
Voss, MW. 2010[65] Participants in studies exploring the relationship between sport expertise and cognition Observational 20 studies included	No intervention	Studies were included if: • they were published in English • used a controlled laboratory examination of cognitive skills • they compared expert athletes with a matched control group of nonexpert athletes [p. 815]	• Not reported	• Attentional cuing • Processing speed [p. 814]	• Effect sizes for each study were calculated using Hedges' formula • A random effects model was used to determine the effect of sport expertise on cognition. An average of the effect sizes for multiple cognitive measures was calculated to determine the mean effect size estimate per study [p. 818]
Zhang, JP. 2010[68] Healthy participants in studies focusing on genetic variation in DTNBP1 and general cognitive ability Observational 8 studies included	No intervention	Studies were included if: • they reported the relationship between DTNBP1 polymorphisms and cognition in humans • they involved healthy control participants • they included the full-scale IQ score [p. 1127]	• CANTAB • COWAT • CPT-I/P • CVLT • MWT-B • WAIS-III • WAIS-R • WMS-III • WRAT-3 [p. 1128]	• General cognitive ability	• Hedges' g was applied as the effect size measure • Raw IQ scores were used to calculate effect sizes in each study as different studies used different covariates • Pooled effect sizes across studies were calculated using a random effects model • *Q* and I^2 statistics were used to assess heterogeneity [p. 1129]

ADAS-Cog = Alzheimer's Disease Assessment Scale-Cognitive; ADHD = attention deficit/hyperactivity disorder; ALL = acute lymphocytic leukemia; AVLT = Auditory Verbal Learning Test; BAS = British Ability Scale; BIMC = Blessed Information-Memory-Concentration; BSID = Bayley Scales of Infant Development; BSRT = Buschke Selective Reminding Test; BVLT = Buschke Verbal Learning Test; BVRT = Benton Visual Retention Test; CAMCOG = Cambridge Cognition Examination; CANTAB = Cambridge Neuropsychological Test Automated Battery; CBCL = Child Behaviour Checklist; CERAD = Consortium to Establish a Registry for Alzheimer's Disease; CHD = congenital heart disease; Choice RT = Choice Reaction Time; COWAT = Controlled Oral Word Association Test; CPAS-R = Cognitive Psychomotor Assessment System-Revised; CPT = Continuous Performance Test; CPT-I/P = Continuous Performance Test-Identical Pairs Version; CRT = Choice Reaction Task; CTBS = Comprehensive Test of Basic Skills; CVLT = California Verbal Learning Test; DAS = Differential Ability Scale; DG = Differentiele Geschiktheidsbatterij; DRS = Dementia Rating Scale; DSM (-III-R, -IV) = Diagnostic and Statistical Manual of Mental Disorders (3rd Edition-Revised, 4th Edition); DSST = Digit symbol substitution test; DTNBP1 =; EMF = electromagnetic fields; ERT = Estrogen replacement therapy; ES = effect size; FAS = FAS, letter fluency test; GCI = Global Cognitive Index; GSM/UMTS = Global System for Mobile Communications (originally Groupe Spécial Mobile)/Universal Mobile Telecommunications System; GMT = Group Mathematics Test; GSM = Global System for Mobile Communications (originally Groupe Spécial Mobile); HAWIE = Hamburger Wechsler Intelligence Test for Adults; HAWIK = Hamburger Wechsler Intelligence Test for Children; HAWIVA = Hamburger Wechsler for Children in Pre-school Age; HRNB = Halstein-Reitan Neuropsychologic Battery; HRT = hormone replacement therapy; HSCS = High Sensitivity Cognitive Screen; HSD = Hasegawa dementia scale; HVLT = Hopkins Verbal Learning Test; HVOT = Hooper Visual Organization Test; IQ = intelligence quotient; KABC = Kaufman Assessment Battery of Childhood; K-BIT = Kaufman Brief Intelligence Test; LIS = Leiter International Scale; LM = logical memory; MCI = Memory Controllability Inventory; MCST = Modified Card Sorting Test; MFFT-20 = Matching Familiar Figures Test-20; MFQ = Memory Functioning Questionnaire; MHz = megahertz; MIA = Metamemory in Adulthood Questionnaire; MIQS = McCarthy IQ Scale; MISIC = Malin’s Intelligence Scale for Children; MMSE = Mini Mental State Examination; MPH = methylphenidate; MSCA = McCarthy Scales of Children’s Abilities; MWT-B = Mehrfachwahl-Wortschatz-Intelligenztest-Version B; NAR = Neale Analysis of Reading; NART = National Adult Reading Test; NEPSY = Developmental Neuropsychological Assessment; OOHMT = Otis Ottawa d’Habitele Mentale Test; PASAT = Paced Auditory Serial Addition Test; PGI = Post Graduate Institute (India); PMAT-FC = Primary Mental Abilities Test for Filipino Children; RAVENS = Raven’s progressive matrices; RAVLT = Rey Auditory Verbal Learning Test; RBANS = Repeatable Battery for the Assessment of Neuropsychologic Status; RCAVLT = Rey and Crawford Auditory Verbal Learning Test; RCFT = Rey-Osterrieth Complex Figure Test; RCT = randomized controlled trial; RIPA = Ross Information Processing Assessment; ROCF = Rey Osterreich Complex Figure; ROVMT = Rey-Osterrieth Visual Memory Test; RVLT = Rey Visual Learning Test; SB = Stanford Binet Scale; SD = standard deviation; SDMT = Symbol Digit Modalities Test; Simple RT = Simple reaction time; SMC = subjective memory complaints; SMD = standardized mean difference; SRT = Simple Reaction Task; SUB = Subtraction; TAP = Test of Attentional Performance; TBI = traumatic brain injury; TMT = Trail Making Test; VER = Sentence Verification; VFT = verbal fluency test; VIG = Vigilance; VLT = verbal learning test; VMI = visual-motor integration; VMT = Visuospatial Memory Test; VRT = Visual Retention Test; vs. = versus; WAIS = Wechsler Adult Intelligence Scale; WAIS-R = Wechsler Adult Intelligence Scale-Revised; WCST = Wisconsin Card Sorting Test; WIAT = Wechsler Individual Achievement Test Screener; WISC (-III, -R) = Wechsler Intelligence Scale for Children (3rd Edition, Revised); WMD = weighted mean difference; WMS = Wechsler Memory Scale; WMS-III = Wechsler Memory Scale-Third Edition; WMS-III = Wechsler Memory Scale-Third Edition; WMS-R = Wechsler Memory Scale-Third Edition-Revised; WPPSI = Wechsler Preschool and Primary Scale of Intelligence; WRAML = Wide Range Assessment of Memory and Learning; WRAT = Wide Range Achievement Test; WRAT-3 = Wide Range Achievement Test-Third Edition; ZVT = Zahlen-Verbindungs-Test

Table 2. Summary articles describing statistical methods for data harmonization

Study	Method	Context	Description	Pro	Con
Bauer, DJ. 2009[70]	Linear factor analysis (LFA)	This article compares different psychometric methods for developing commensurate measures in the context of integrative data analysis (the simultaneous analysis of data obtained from two or more independent studies)	**LFA** • A latent factor(s) is/are posited to underlie a set of observed, continuous variables • Must test the invariance of the latent factor(s) when comparing across groups	• Methodology well known • As long as there is a subset of invariant items, the factor can be combined across studies • Can include noncommon items in analyses • Even if no items are common to all studies, can still be conducted if studies can be "chained" together. For example, if item sets A and B are available in study 1, item sets B and C are available in study 2 and item sets C and D are available in study 3	• The validity of the results is dependent on the method used to identify the model. For example, the use of the reference item method (i.e., constraining the intercept and loading of one item to 0 and 1 in both groups) implicitly assumes the reference item is invariant across groups. Another option is to set the factor mean and variance to 0 and1 in one group then estimate factor mean and variance in the other group while placing equality constraints on one or more item intercepts and loadings. This second procedure only works well when the number of noninvariant items is small relative to the number of invariant items • Requires continuous indicators • Requires more than one common item to measure equivalence • Sample units must be independent of one another (i.e., no repeated measures)

Table 2. Summary articles describing statistical methods for data harmonization (continued)

Study	Method	Context	Description	Pro	Con
Bauer, DJ. 2009[70] (cont'd)	Two-parameter logistic (2-PL) using Item response theory (IRT)		**2-PL IRT** • Assumes a single latent trait underlies a set of binary responses. Some items may be more difficult than others and some items may be more strongly related to the latent trait than others. Each item is assumed to have a conditional Bernoulli distribution including discrimination and difficulty parameters • It is typically assumed the latent trait has a standard normal distribution	• Widely used methodology (especially in testing) • As long as there is a subset of invariant items, the factor can be combined across studies • Can include noncommon items in analyses • Even if no items are common to all studies, can still be conducted if studies can be "chained" together. For example, if item sets A and B are available in study 1, item sets B and C are available in study 2 and item sets C and D are available in study 3	• Requires binary indicators. For example, it could be used to measures dimensions of psychopathology by using a set of symptoms indicators • Requires more than one common item to measure equivalence • Sample units must be independent of one another (i.e., no repeated measures)
	Moderated nonlinear factor analysis (MNLFA)		**MNLFA** Uses generalized factor analysis models that can incorporate binary, ordinal, continuous and count items. The key parameters of the factor model (i.e., the indicator intercepts, factor loadings, and factor mean and variance) are permitted to vary as a function of one or more exogenous variables	• Can accommodate different scale types among items	• Requires more than one common item to measure equivalence Sample units must be independent of one another (i.e., no repeated measures)

Table 2. Summary articles describing statistical methods for data harmonization (continued)

Study	Method	Context	Description	Pro	Con
Burns, RA. 2011[71]	Multiple Imputation	Combining MMSE scores with missing data across 9 Australian longitudinal studies of aging (Dynamic Analyses to Optimize Aging [DYNOPTA] project) Participants missing at least one MMSE item varied greatly by contributing study and wave of data collection [range 0.5%-95%] Missing information related to demographic characteristics, especially age and education	MI used imputer model in multiple imputation with chained equations (MICE). The model included: gender, years of education, study and study interactions. Created 5 imputation datasets and took the average of the 5 imputed plausible values Software: MICE add-on to STATA version 10	• Does not require special software to run MICE subsequent analyses (i.e., one final dataset with average values) • Compared with single imputation methods, MI accounts for uncertainty in the missing value by using a set of plausible values	• Requires the same measures across studies • Requires item level data
Gorsuch, R. 1997[72]	Extension analysis in exploratory factor analysis	In exploratory factor analysis extension analysis refers to computing the relationship among common factors to variables that were not included in the factor analysis. For example, this would be used in a situation where a factor analysis would include proven items but now new experimental items.	A factor analysis is conducted on the core variables. The correlations between the core variables and the extension variables to estimate the factor pattern of the extension variables with the factors derived from the core variables.	• Widely used methodology	• Requires the core set of variables to be common to all studies • Requires the extension set of variable to be common to all studies • Only appropriate for continuous variables

Table 2. Summary articles describing statistical methods for data harmonization (continued)

Study	Method	Context	Description	Pro	Con
McArdle, JJ. 2009[73]	IRT combined with latent growth/ decline curve modeling	This method was applied to longitudinal data from different cognitive test batteries to examine how to best model changes in cognitive constructs over a life span. The data come from 3 classic studies on intellectual abilities (Berkeley Growth Study (BGS), Guidance–Control Study (GCS), and Bradway–McArdle Longitudinal (BML) Study). In total, 441 persons were repeatedly measured as many as 16 times with age at measurement ranging from 2 to 72 years. Vocabulary data were analyzed for 419 participants and memory data were analyzed for 416 participants. The cognitive constructs measured were vocabulary and memory using 8 different intelligence test batteries (1916 Stanford–Binet (SB), SB Form L, SB form LM, Wechsler–Bellevue (WB) Intelligence Scale Form I, Wechsler Adult Intelligence Scale (WAIS), WAIS-Revised, and the Woodcock–Johnson (WJ) Psycho-Educational Battery–Revised). Although the tests were common items among the tests, different tests were between studies and over time within studies. All three studies include participant from the Bay area of California.	The authors consider several techniques for linkage across measurement scales and across multiple groups and fit a unidimensional Rasch model to item responses and a latent curve model together with changing latent scores over age and groups. The latent growth/decline curve model had a separate within-time measurement equation and over-time functional change equation. Because some items used in these analyses have graded outcome scores (i.e., 0, 1, or 2), a partial credit model was used for the IRT model. The parameters of both IRT and latent curve models were simultaneously estimated based on a joint model likelihood approach	• Can be used without complete overlap of items • Items can be linked by data (i.e., bridge studies and bridge items), by assuming equivalence, or by a combination of the two • Allows for a varying number of data points per person • Allows instruments to change over time within an individual • Method allows one to separate out differences in scales over time from changes in constructs over time	• Requires overlapping information across studies

Table 2. Summary articles describing statistical methods for data harmonization (continued)

Study	Method	Context	Description	Pro	Con
Minicuci, N. 2003[74]	Multiple methods including recategorization and z-score transformations	Constructed a harmonized measures using data from six countries [Finland, Italy, the Netherlands, Spain, Sweden and Israel] contributing data to the Comparison of Longitudinal European Studies on Aging (CLESA) Study. The first goal of the study was to create a common data base (CDB) with a framework to include behavioral, social, psychological, and health status measures. A common measure was created if at least 3 countries had measured the construct of interest (this led to the creation of different harmonized options for the same construct). The CDB is a harmonized file containing 11,557 records and 111 variables on socio-demographic characteristics, health habits, health status, physical functioning, social networks and support, and health and social service utilization	For each type of variable, harmonization guidelines were developed including: the definition and variable names of the standardized measures and their modalities and relative coding; a list was made of the name of the original variable(s) for each country, with their original modalities; and the algorithm used to obtain the harmonized variables from the original variables When harmonization was deemed appropriate, the most common methods for harmonization were to recategorize variables into a common set of response option and to create a common scale, e.g., 0-1, by dividing a continuous score by its maximum score Another related method of conversion is to create z-scores for each construct by subtracting the overall mean and dividing the raw score by the standard deviation.	• Relatively simple and does not require specialized statistical software • Does not require any common items across studies	• Does not take into account the difference in distributions/variability across populations • Assumes the underlying constructs are the same and measured equally well across populations

Table 2. Summary articles describing statistical methods for data harmonization (continued)

Study	Method	Context	Description	Pro	Con
Gross, A. 2012[75]	Mean, linear, and percentile transformations	Constructed mean, linear and percentile equating using data from two large-scale, multi-site cohorts: the Advanced Cognitive Training for Independent and Vital Elderly (ACTIVE) and the Alzheimer's Disease Neuroimaging Initiative (ADNI). ACTIVE is a longitudinal randomized trial of cognitive training in cognitively intact, community dwelling adults age 65 and older ADNI was a five-year observational cohort study of Alzheimer's disease (AD), with the primary goal of assessing the extent to which serial magnetic resonance imaging, positron emission tomography, other biological markers, and cognitive tests can be used to predict progression to mild cognitive impairment (MCI) and AD	Used a two stage approach. In the first stage, an equating sample was selected from which to collect necessary characteristics of test distributions and derive the equating algorithm. In the second stage, equating algorithms were applied to the full study sample in a way that preserved attrition, aging, cohort, and group differences but eliminated form differences. Equated scores were then compared visually using plots of mean recall over time and cumulative probability plots and statistically using tests of equivalence of means in reference groups as well as estimates of within-person change using latent growth models.	• Can be used with longitudinal data • Can be used to adjust	• Assumes that the population producing responses on different scaled tests at each time point have the same underlying ability • Linear equating assumes normally distributed variables (not a limitation for equipercentile equating) • Outcome must be continuous
van Buuren, S. 2005[76]	Response conversion (RC)	This method was applied to binary and original data measuring walking disability measured across 10 European countries.	RC is a two-step method. The first step is to construct a conversion key using a statistical model (e.g., polytomous Rasch model). This step models the relationship between the common scale and the measured items. The second step uses a conversion key to convert information onto a common scale	• Can be used without complete overlap of items • Items can be linked by data (i.e., bridge studies and bridge items), by assuming equivalence, or by a combination of the two	• Requires overlapping information across studies

2PL-IRT = two-parameter logistic using item response theory; BGS = Berkeley Growth Study; BML = Bradley-McArdle Intelligence Scale; CDB = common database; CLESA = Comparison of Longitudinal European Studies on Aging; DYNOPTA = Dynamic Analyses to Optimizing Ageing; GCS = Guidance-Control Study; IRT = item response theory; LFA = linear factor analysis; MI = multiple imputation; MICE = multiple imputation with chained equations; MMSE = Mini Mental State Examination; MNLFA = moderated nonlinear factor analysis; RC = response conversion; SB = Stanford-Binet; WAIS (-R) = Wechsler Adult Intelligence Scale (Revised); WB = Wechsler-Bellevue; WJ = Woodcock Johnson Psycho Educational Battery-Revised

Table 3. Summary of supplemental articles on individual participant data meta-analysis and methods to support statistical harmonization (continued)

Topic	Citation	Summary
General articles on conducting IPD	Blettner, M. 1999[77]	• Described the strengths and limitations of four methods of summarizing data: qualitative summary, meta-analysis of published data, re-analysis of IPD, and prospectively planned pooled analyses • Harmonization of data not mentioned
	Cooper, H. 2009[78]	• Discussed the relative merits of conducting an IPD vs. aggregated data (AD) analysis • IPD permits subgroup analysis and quality assurance of the original analysis reported in the literature, but is more costly
	Curran, PJ. 2009[79]	• Discussed issues around conducting integrative data analysis (IDA) as defined as the statistical analysis of a single data set that consists of two or more separate samples that have been pooled into one • Identified two possible methods to deal with heterogeneity due to measurement: nonlinear factor analysis (NFA) and item response theory (IRT) • They recognize the issue of combining complex constructs, such as depression, but they do not provide detailed information on how to handle this issue in pooling data from different studies
	Friedenreich, CM. 1993[86]	• Presented a methodology for the pooling and analysis of epidemiologic studies using individual subject level data • Discussed random and fixed effect models, examining homogeneity of effects, explaining any heterogeneity, sensitivity analyses and quality assessment
	Ioannidis, JPA. 2002[80]	• Discussed advantages and disadvantages of IPD meta-analysis of time to event data in genetic epidemiology • Standardization of information across studies using a priori definitions was listed as an advantage as a standardized set of variables was available for all studies • Other issues around harmonization were not mentioned
	Riley, RD. 2010[2]	• Discussed the rationale, conduct and reporting of IPD meta-analyses • Did not discuss the issue of different variables/measures being available among datasets
	Schmid, CH. 2003[81]	• Discussed issues around conducting an IPD analysis using data from multiple international RCTs evaluating the effect of ACE inhibitors for treatment of nondiabetic renal disease
	Simmonds, MC. 2005[82]	• Reviewed methods used to conduct IPD meta-analyses conducted during 1999-2001. • Harmonization of data not mentioned
	Van der Steen, JT. 2008[83]	• Discuss benefits and pitfalls of pooling databases from comparable observational studies of lower respiratory infection in nursing home residents in the U.S. (Missouri) and the Netherlands (Amsterdam) • Identified issues in comparability in measurements in terms of: 1) question wording and response options, 2) clinical meaning, 3) response distributions • If response distributions to the same question differed by population, they tried to do qualitative interviews with physicians to determine whether the variable had a different meaning between the countries • Did not discuss specific methodology for constructing new variables when differences there were differences between the variables in the two databases
	van Walraven, C. 2010[84]	• Discussed reward and challenges of IPD meta-analysis. • Reward: outcome and analytical harmonization • Challenge: getting and harmonizing data

Table 3. Summary of supplemental articles on individual participant data meta-analysis and methods to support statistical harmonization (continued)

Topic	Citation	Summary
IPD analysis Methods	Bennett, DA. 2003[85]	• Reviewed analytic methods for prospective cohort studies using time to event data for single studies and IPD meta-analyses • Discussed issues around missing data (event times and covariates) for individual studies as well as for IPD meta-analyses • Suggested running a simulation sensitivity analysis to determine the extent of biasing and underestimation of standard errors using different methods for imputation of event times • In their example, the authors did not employ imputation methods for covariates due to the size of the data set • The authors reported the number and nature of the missing covariate values according to key variables such as cohort, censoring status, age at recruitment.
	Granda, P., Blasczyk, E. 2010[8]	• Defined general approaches to harmonization • Input harmonization aims to achieve standardizes measurement processes and methods in all national or regional populations • Output harmonization uses different national or regional measurements possibly derived from nonstandard measurement tools • An ex-post strategy to output harmonization (i.e., surveys made comparable after the fact, retrospective harmonization) requires a conversion process • This conversion process should be transparent , well documented ,and reversible • Focus should be given to both variable level and survey level harmonization • Need to develop criteria to assess the quality of harmonization
	Granda, P., Wolf, C., Hadorn, R. 2010[69]	• Discussed strategies and issues around harmonization of survey data • Provided methods for assessing the quality of harmonization or the degree to which the original information is preserved in the harmonized data • This is most applicable to direct harmonization (i.e., when a single harmonized variable is created directly a single questionnaire item)
	Hofer, SM. 2009[87]	• Discussed the challenges of meta-analytic and pooled data approaches using cognitive aging literature as an example • Discussed concurrent calibration (cocalibration) of data using IRT models or with latent variable approaches based on item- or scale-level data across studies • Feasibility of pooling variable is limited when variables are not operationally defined in the same way • Using standardized variables (T scores) or proportion correct requires assuming the measurement properties of the variables are relatively comparable and linear • Also need to consider population characteristics such as age, birth cohort, education ranges • Proposed a coordinated analysis approach to enhance communication and collaboration among researchers, facilitate reproducible research, archive analysis and measurement alignment process, to provide a stronger basis for cumulative science, and to permit quick entry into completed analyses
	Jones, AP. 2009[88]	• Discussed methods used to combine longitudinal clinical trial data across studies using IPD and aggregate data methods • Did not discuss the issue of different variables/measures being available among datasets
	Mathew, T. 2010[89]	• Compared One-step (linear function of the mean obtained from a linear model of IPD) vs. Two-step (linear function of the mean obtained from linear model of summary data) meta-analysis models using IPD • It provides a nice overview of IPD meta-analysis • Did not discuss the issue of different variables/measures being available among datasets

Table 3. Summary of supplemental articles on individual participant data meta-analysis and methods to support statistical harmonization (continued)

Topic	Citation	Summary
Comparison of imputation methods	Burgess, S. 2011[90]	• Described four Bayesian methods for imputing missing data based on a missing at random (MAR) assumption in the context of genetic epidemiology: multiple imputations, single nucleotide polymorphism (SNP) imputation, latent variables, and haplotype imputation • Results of a simulation study and application to the British Women's Heart and Health Study were presented • Method analogous to the 2-stage least-squares method except it accounts for the observational correlation between phenotype and outcome. This analysis was done using WinBUGS • Precision was improved using four imputation methods—equivalent to 25% increase in sample size • All imputation methods give similar results
	Donegan, S. 2010[91]	• Reviewed the reporting and methodological quality of indirect comparisons (which could be considered an extreme missing data situation) • Authors conducted a systemic review including 43 reviews in which clinical effectiveness of two interventions were indirectly compared • In general, the underlying assumptions of conducting an indirect comparison analysis were not routinely described or tested
	Peyre, H. 2011[92]	• Compared imputation method for data Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) within one study • Methods compared: personal mean score (PMS), multiple imputation (MI), hot deck (HD) imputation and full information maximum likelihood (FIML) • MI and FIML superior to PMS and HD in terms of accuracy and precision • HD tended to underestimate and PMD associated with insignificant bias
	Siddique, J. 2011[93]	• The authors used an imputation approach to calibrate rater bias in the diagnostic assessment of posttraumatic stress disorder (PTSD) • Nurse practitioners were twice as likely to diagnose PTSD than a clinical psychologist—and each patient was randomly assigned to be rated by only one rater • A Bayesian random effects censored ordinal probit model was used to identify a latent moderate class of patients • A Markov chain Monte Carlo algorithm was used to estimate the posterior distribution of the model parameters and generate multiple imputations for the recalibrated diagnosis variable
	Spratt, M. 2010[94]	• Examined how the choice of imputation model and the number of imputations affected estimates of prevalence and associations in a study of wheezing among 81-month-old children in Avon Longitudinal Study of Parents and Children • Preliminary analyses of the association of measured variables with missingness and outcome variables are required to determine the plausibility of the assumptions underlying both complete-case and multiple-imputation-based analyses. • They applied a covariate (socioeconomic status) which was available on all subjects as an intermediate variable to generate multiple imputations procedures • Analyses of MI should often be based on 25 or more imputed values in order to reduce the impact of random sampling inherent in the MI process
	Sterne, JAC. 2009[95]	• Reviewed the reasons why missing data may lead to bias and loss of information • Discussed situations in which multiple imputation may help reduce bias and increase precision as well as the potential pitfalls • Proposed guidelines for reporting analyses using multiple imputation

Table 3. Summary of supplemental articles on individual participant data meta-analysis and methods to support statistical harmonization (continued)

Topic	Citation	Summary
Methods for evaluating equivalence	Crane, PK. 2008[96]	- Compared item- and scale-level strategies for handling demographic heterogeneity when measuring executive function - Examined the extent to which item-level and scale-level adjustment for demographic variables influenced the relationships with various composite executive function scores with an external criterion (MRI) - The authors created composite scores for executive function using classical test theory and item response theory in which demographic differences were ignored or taken into account - Candidate scores were compared using 3 linear regression models; model A included demographic terms as independent variables, model B include MRI variables, and model C included both - R^2 was used to estimate effect sizes
	Teresi, JA. 2007[97]	- Discussed methods based on IRT that can be used to examine differential item functioning (DIF) within study subgroups - The method used was the item response theory log-likelihood ratio (IRTLR) approach - This method could also be extended to testing DIF among study populations

ACE = angiotensin-converting-enzyme; DIF = differential item functioning; FIML = fill information and maximum likelihood; HD = hot deck; IDA = integrative data analysis; IPD = individual patient data; IRT = item response theory; IRTLR = item response theory log-likelihood ratio; MAR = missing at random; MCAR = missing completely at random; MI = multiple imputation; MNAR = missing not at random; MRI = magnetic resonance imaging; NFA = nonlinear factor analysis; PMS = personal mean score; PTSD = post traumatic stress disorder; RCT = randomized controlled trial; SNP = single nucleopeptide polymorphisms; U.S. = United States of America

Table 4. Examples of studies presenting harmonized data

Topic	Citation	Summary
Examples of analyses of harmonized data	Anstey, KJ. 2010[98]	• Harmonized data from the Dynamic Analyses to Optimizing Ageing (DYNOPTA) project • Harmonized data [including cognitive measures] from 9 Australian cohorts using response conversion (see van Buuren above) • Did not give details on the analysis
	Bath, PA. 2010[99]	• Harmonized data from the Longitudinal Aging Study Amsterdam (LASA) and the Nottingham Longitudinal Study on Activity and Ageing (NLSAA) [including cognitive measures] • LASA used the Mini Mental State Exam (MMSE: 30 point scale) and the NLSAA used the Clifton Assessment Procedures for the Elderly (CAPE: 12 point scale) • The derived variables were simply MMSE/30 and CAPE/12
	Beer-Boorst S. 2000a[100] Beer-Boorst, S. 2000b[101]	• Developed a common surveillance system to allow for the comparison of lifestyle and biological risk factors from different populations across Europe including seven collaborating centers [European Alimentation (EURALIM)] • Common variables included: diet, health, lifestyle and demographic variables • Did not discuss method of harmonization
	Crane, PK. 2008[102]	• Used IRT to cocalibrate cognitive scales from three large community-based studies (the Cardiovascular Health Study [CHS], the Adult Changes in Thought Study [ACT] and the Indianapolis site from the Indianapolis-Ibadan Dementia Project • The primary objective was to cocalibrate the Mini Mental State Examination (MMSE), Modified Mini Mental State Exam (3MS), Cognitive Abilities Screening Instrument (CASI), and The Community Screening Instrument for Dementia (CSI 'D') • Used McDonald's bifactor model to evaluate whether the scales were unidimensional • Identified anchor items that were comparable across tests—only included identical items (e.g., interlocking pentagons) • Used Samejima's graded response model to estimate the probability of each response category for each item for any level of cognitive functioning. This formula was used to determine the most likely response for every cognitive functioning level
	Curran PJ. 2008[103]	• Used IRT to fit a series of growth curve models to a single pooled sample that consists of data drawn from three separate studies of developmental internalizing symptomology • The studies examined children with and without alcoholic parents (The Michigan Longitudinal Study [MLS], the Adolescent/Adult Family Development Project [AFDP], and the Alcohol and Health Behavior Project [AHBP]) • There were 21 unique dichotomous self-report items to define internalizing symptomology; four items were present in all studies • Dimensionality Step: Factor analysis was used to examine the dimensionality of the 21 items by conducting an exploratory factor analysis in each study to assess unidimensionality based on traditional measures including eigenvalues, scree plots and estimates of incremental variance • Calibration Step: Fitted a standard 2PL IRT model to the 21 dichotomous items from a single randomly selected assessment for each participant in the pooled sample • DIF Step: Estimated a series of multiple group IRT models as a function of developmental status, gender and study group membership • Scoring Step: calculated individual time-specific scale scores for every participant at every time point at which they were assessed using a modal a posteriori method.

Table 4. Examples of studies presenting harmonized data (continued)

Topic	Citation	Summary
Examples of analyses of harmonized data (continued)	Darby S. 2006[104]	• Authors used data from 13 studies of residential radon and lung cancer carried out in Europe (Austria, the Czech Republic, Finland [2], France, Germany [2], Italy, Spain, Sweden [3], and the UK) • Data were assembled according to a common format and uniform definitions were used except for study-specific definitions for social status [did not detail how this was put into a common metric] • Data were analyzed using a linear odds model; models were fit using conditional maximum likelihood (similar to conditional logistic regression)
	The Fibrinogen Studies Collaboration. 2009[105]	• The authors combine data on the association between plasma fibrinogen and coronary heart disease in 31 cohort studies using proportional hazards (Cox) model, stratified by cohort, sex and (for the two RCTs) trial arm • All studies provided data on fibrinogen level, age, smoking status, total cholesterol, SBP and BMI • Some studies also provided data on HDL and LDL cholesterol, alcohol consumption, triglycerides and history of diabetes • The authors use a two-stage process. At the first stage partially and (where possible) fully adjusted estimates are obtained from each study, together with their standard errors (a key issue is estimating the within study correlation of the two estimates) • At the second stage, the results are combined in a bivariate meta-analysis • This study addresses the issue of when studies included in an IPD meta-analysis include some, but not all, important confounding variables • The proposed bivariate model, with estimates of the parameter of interest either fully or partially adjusted for confounding factors may be useful also for more difficult constructs. Some studies measuring the construct fully and other studies measuring the construct only partially could possibly analyzed with this bivariate approach
	Grimm KJ. 2010[106] (see also Duncan, GJ. 2007[107]	• The authors examined the associations between early behavioral and cognitive skills with later achievement using data from 3 longitudinal studies (the NICHD Study of Early Child Care and Youth Development [SECCYD], the National Longitudinal Survey of Youth–Children and Young Adults [NLSY-CYA], and the Early Childhood Longitudinal Study Kindergarten Cohort [ECLS-K]) • Behavior scales differed among the studies and categorization of children into "normative", "problematic" and "clinical" groups was done using set cut-offs or based on the observed distribution of the data (e.g., T-scores) • The authors used a combined item-response and growth curve model to account for differential reliability
	Khachaturian, AS. 2010[108]	• The authors describe the challenges and opportunities for developing a national database for successful aging • One of the main challenges is defining a "case" for population-based prevention studies • Clinical assessment conducted by experts produce accurate diagnoses, but are very costly, labor intensive and require highly trained personnel • Population studies, because of a lower yield, necessitates greater efficiency and lower-cost less-highly trained personnel • Need for multi-stage assessment among subject, other informants (e.g., family members), as well as clinical assessments by clinicians and nonclinicians (including surveys by mail and telephone cognitive assessments) • The critical questions identified were: 1) can these assessment approaches be refined in order to detect or predict individuals who may develop future impairments, or declination, in cognition or behavior, or even scaled down for high volume throughput; 2) can technologies be developed to allow the most passive, nonintrusive assessment of the individual's cognitive and behavioral function; and 3) will the collected longitudinal data afford the possibility to measure intra-person change, vis-a-vis Bayesian-modeling approaches • The issue of measuring within-person change over time was also highlighted as the ultimate aim is to predict the trajectory of an individual's cognitive-behavioral-functional health, the rate of decline, and the point at which one crosses the threshold from an asymptomatic stage to a phenotype resembling pre-MCI, then to MCI, and then to AD

Table 4. Examples of studies presenting harmonized data (continued)

Topic	Citation	Summary
Examples of analyses of harmonized data (continued)	McArdle, JJ. 1998[109] [see also McArdle, JJ. 1994;[110] McArdle, JJ. 1997[111]]	• The authors use methods based on linear structural equations models with incomplete or missing data to analyze longitudinal twin data for two cognitive variables to evaluate a biometric genetic hypothesis in the context of a developmental model of intellectual growth and change (biometric genetic analysis of intellectual abilities [BGIA]) • In this study, the same measurement scales (block design and vocabulary measures) were used over time, however the number of observations, the age at first administration, and the interval between administrations differed within the twin pairs and among the sets of twins. The raw scores were transformed into percentage-correct scales (0-100). • The authors used all available data, including participants with incomplete and possibly nonrandomly missing data • The authors incorporated a twin analysis including means and age effects; a longitudinal analyses based on latent growth components; and a biometric-genetic analyses for components of growth using linear structural equations models
	Minicuci, N. 2011[112]	• Compared measures of Disability Free Life Expectancy (DFLE) across different surveys conducted in Bulgaria (National Health Interview Survey [NHIS]), Italy (Multidisciplinary Survey among Italian Families [IMF-S]) and Latin America (the Salud, Bienestar y Envejecimiento [SABE]) • Harmonized 5 ADL questions common to all surveys • Dichotomized responses to create a common scale
	Pluijm, SMF. 2005[113]	• Constructed a harmonized measure of ADL using data from six countries contributing data to the Comparison of Longitudinal European Studies on Aging (CLESA) Study • There was overlap in the ADL items among countries, but only 2 of the 11 possible items were asked in all surveys • Items that were incompatible across countries because of cultural differences were excluded from the harmonization process • Harmonization focused on the four items comprising the Katz ADL index; all four items were present in four of the six country surveys; five- and six-item scales were constructed in the countries that had the additional items in common • In countries where the two items were not measured, the data for these was extrapolated from other “comparable” ADL items • Because they used different response options among the surveys all items were dichotomized to put them on a consistent scale • Subjects were excluded if 2 or more items were missing; hot deck methods were used to impute values when one of the items was missing due to nonresponse • Reliability and validity of the four item scale was assessed
	Ruggles, S. 2003[114] Esteve, A. 2003[115]	• The Integrated Public Use Microdata Series (IPUMS)-International involved working with census data from different time periods and institutional origins • The first stage of harmonization involved standardizing the data formats and correcting errors • The second stage of harmonization involved harmonizing the codes for all variables across datasets • Variable-level harmonization involved recoding variables to maximize comparability across datasets. • In this example, the content included among the datasets was greatly overlapping, but different numeric classification systems were used

Table 4. Examples of studies presenting harmonized data (continued)

Topic	Citation	Summary
Examples of analyses of harmonized data (continued)	Schenker, N. 2007[116]	• The authors describe several situations in which data from multiple surveys were used to enhance estimation of measures of health • The four projects involved: (1) combining estimates from a survey of households and a survey of nursing homes to extend coverage; (2) using information from an interview survey to bridge the transition in race reporting in the United States census; (3) combining information from an examination survey and an interview survey to improve on analyses of self-reported data; and (4) combining information from two interview surveys to enhance small-area estimation • In project 3, the authors discussed methods for combining information from two surveys conducted by the National Center for Health Statistics to improve on analyses of self-reported data on health conditions • One of the surveys, the National Health and Nutrition Examination Survey, was unusual in that it not only asked self-report questions on health conditions during face-to-face interviews, but it also obtained clinical measures based on physical examinations • The other survey, the National Health Interview Survey, was larger, and it obtained a rich set of variables for use in multivariate analyses, but it relied on self-report questions for its information on health conditions • ‘Measurement error’ models that predict clinical outcomes from self-reported answers and covariates were fitted to data from the National Health and Nutrition Examination Survey, and the fitted models were then applied to data from the National Health Interview Survey to adjust for possible inaccuracies due to self-reporting • Multiple imputation was used to properly reflect the sources of variability in subsequent analyses

Table 4. Examples of studies presenting harmonized data (continued)

Topic	Citation	Summary
Examples of analyses of harmonized data (continued)	Slimani, N. 2002[117]	• Harmonized data from 10 Western European countries (Denmark, France, Germany, Greece, Italy, Norway, Spain, Sweden, The Netherlands, UK) which from the European Prospective Investigation into Cancer and Nutrition (EPIC) project • Information on usual individual dietary intakes was obtained using different dietary assessment methods developed and validated in each participating country • A calibration approach was adopted to adjust for possible systematic over- or underestimation in dietary intake measurements and correct for attenuation bias in relative risk estimates • A single 24-hour dietary recall was collected from a random sample of 5-12% (1.5% in the UK) of the EPIC cohorts, weighted according to the cumulative number of cancer cases expected per fixed age and sex stratum • Standardized software (EPIC-SOFT) was developed to assess dietary intake reported across the EPIC centers
	van Buuren, S. 2003[118] Hopman-Rock, M. 2000[119]	• Used response conversion to harmonize international disability information from ERGOPLUS (Rotterdam) and EURIDISS (3 countries in Europe) • The first step was to create a conversion key; used Rasch modeling (a partial credit model) to estimate the parameters for the conversion key • The second step involved the conversion of the observed data onto the common scale

2PL IRT = two-parameter logistic using item response theory; 3MS = Modified Mini Mental State Exam; ACT = Adult Changes in Thought Study; AD = Alzheimer's disease; ADL = activities of daily living; AFDP = Adolescent/Adult Family Development Project; AHBP = Alcohol and Health Behavior Project; BGIA = biometric genetic analysis of intellectual abilities; BMI = body mass index; CAPE = Clifton Assessment Procedures for the Elderly; CASI = Cognitive Abilities Screening Instrument; CHS = Cardiovascular Health Study; CLESA = Comparison of Longitudinal European Studies on Aging; CSI 'D' = Community Screening Instrument for Dementia; DFLE = Disability Free Life Expectancy; DIF = differential item functioning; DYNOPTA = Dynamic Analyses to Optimizing Ageing; ECLS-K = Early Childhood Longitudinal Study Kindergarten Cohort; EPIC = European Prospective Investigation into Cancer and Nutrition; EURALIM = European Alimentation; HDL = high-density lipoprotein; IMF-S = Multidisciplinary Survey among Italian Families; IPD = independent patient data; IPUMS = Integrated Public Use Microdata Series; IRT = item response theory; LASA = Longitudinal Aging Study Amsterdam; LDL = low-density lipoprotein; MCI = mild cognitive impairment; MLS = Michigan Longitudinal Study; MMSE = Mini Mental State Examination; NHIS = National Health Interview Survey; NLSAA = Nottingham Longitudinal Study on Activity and Ageing; NLSY-CYA = National Longitudinal Survey of Youth-Children and Young Adults; RCT = randomized controlled trial; SABE = Salud, Bienestar y Envejecimiento; SBP = systolic blood pressure; SECCYD = NICHD Study of Early Child Care and Youth Development; UK = United Kingdom

Table 5. Assumptions for the different classes of statistical harmonization methods

Method	Assumptions	How Can It Be Applied
Standardization Methods 6 studies used this class of methods, e.g., Minicuci, N. 2003[74]	- Scales have an underlying normal distribution - The scales have a similar distribution (i.e., being in the 5th percentile of one scale is equivalent to being in the 5th percentile of another)	Can be applied in most situations with continuous variables and does not require specialized software Does not require common items across studies Need to transform back to a chosen scale(s) for interpretation
Item Response Theory Latent Variable Model 15 studies used this class of methods, e.g., Van Buuren, S. 2005;[76] Bauer, DJ. 2009;[70] McArdle, J. 2009[73]	- Underlying constructs are unidimensional - Some items must be common across datasets or at least can be "chained" together - The items are equally discriminating (only for IP and Rasch models) - Factorial invariance If repeated measures: - Item difficulty is invariant with respect to time or age - Item discrimination does not change across time or age	Can be applied to continuous, binary and ordinal data but requires some specialized software Can accommodate different scale types among items However can be extended to include longitudinal data as per McArdle, et al. by integrating IRT and latent curve modeling using a joint model likelihood approach
Missing data by design with multiple imputation 3 studies used this class of methods, e.g., Burns, RA. 2011[71]	- Missingness is assumed to be at random (i.e., MAR) - Some items must be common across datasets or at least can be "chained" together	Can be applied to continuous, binary and ordinal data but requires some specialized software and multiple datasets Can accommodate different scale types among items Can be used if scales are not unidimensional

IRT = item response theory; MAR = missing at random

Table 6. Overview of some approaches for harmonization of constructs

Van Buuren S, Eyres S, Tennant A, Hopman-Rock M, Improving comparability of existing data by response conversion, Journal of Official Statistics, 2005, 21(1), 53-72.[76]

- Focus is on health surveys using questions with ordinal categories
- Construct of interest is “walking disability” but the approach is applicable also to other constructs
- Incomparability of data occurs when questions on the same construct are not identically formulated over different studies. Example of incomparability

  UK Health survey:

  How far can you walk without stopping/experiencing severe discomfort, on your own, with aid if normally used?

  1) can’t walk
  2) a few steps only
  3) more than a few steps
  4) less than 200 yards
  5) 200 yards or more

  Dutch Health survey

  Can you walk 400 meters without resting (with walking stick if necessary)?

  1) yes, no difficulty
  2) yes, minor difficulty
  3) yes, major difficulty
  4) no
- The approach of harmonization requires at least one “bridge variable”, which is a question that is comparable over two studies. Linking *m* studies requires at least one bridge variable, when all studies include this question, and at most *m*-1 bridge variables, when only one unique bridge variable is available between two studies.
- Two studies with 306 and 292 participants, respectively, were used to illustrate their harmonization method of “response conversion”. Three questions on walking disability were used for illustration: SI01, HAQ8, and GAR9. The HAQ8 is the bridge variable, while SI01 is only observed in one study and GAR9 is only observed in a second study.

  SI01: *I walk shorter distances or often stop for a rest*
  0 = no
  1 = yes

  HAQ8: *Able to walk outdoors on flat grounds*
  0 = without any difficulty
  1 = with some difficulty
  2 = with much difficulty
  3 = unable to do so

  GAR9: *Can you, fully independently, walk outdoors (if necessary with a cane)?*
  0 = yes, no difficulty
  1 = yes, with some difficulty
  2 = yes, with much difficulty
  3 = no, only with help from others

Table 6. Overview of some approaches for harmonization of constructs (continued)

Van Buuren S, Eyres S, Tennant A, Hopman-Rock M, Improving comparability of existing data by response conversion, Journal of Official Statistics, 2005, 21(1), 53-72.[76] (continued)

- A latent variable (Rasch) model was selected to describe the outcome probabilities for these three questions. Let Y_{hij} be the ordinal response for question or item j (=SI01, HAQ8, GAR9), observed in study h (=1,2) for participant i (=1,2,…, n_h). Note that Y_{1i3} and Y_{2i1} are essentially missing for all participants in the corresponding study. The proposed model is referred to as the “conditional logit” model (see Agresti, 2002):

  $$P\left(Y_{hij}=c \mid Z_{hi}=\theta\right)=\exp\left(\sum_{k=0}^{c}\left(\theta-\delta_{jk}\right)\right)\Big/\sum_{r=0}^{K_j}\exp\left(\sum_{k=0}^{r}\left(\theta-\delta_{jk}\right)\right), \qquad (1)$$

  with $c=0,1,.....,K_j$ the range of outcomes for item j , $\delta_{j0},\delta_{j1},.....,\delta_{jK_j}$ the fixed item specific parameters, and Z_{hi} a random latent variable for participant i in study h . Conditionally on the latent variable Z_{hi} the items are considered independent.

  The fixed item specific parameters require a constraint, since δ_{j0} is unidentifiable in this formulation. Indeed, both the numerator and denominator in (1) can be divided by the expression $\exp(\theta-\delta_{j0})$ which would exclude the parameter δ_{j0} from the formula. Thus one particular constraint is for instance $\delta_{j0}=0$. The article describes other constraints on the item specific parameters, but they seem awkward, since they involve a constraint on the latent variable.
- The article discusses several priors or distributions for the latent variable Z_{hi} , but the preferred prior is the shifted lognormal distribution with mean 1, variance 0.5, and shift -5. The comparison of their choices of priors was somewhat strange, because these priors varied in mean and standard deviations. It seems more relevant to compare standardized priors to see the effect of the prior instead of the effect of location shift or scale differences. Furthermore, they did not discuss separate parameters in the priors for their different studies, which could have been useful for their “response conversion” and for their investigation of equivalence of items over studies (“tension coefficient”).
- Using model (1) and the selected prior, “response conversion” is now defined through the conditional expectation of the latent variable given the observed outcome of the item response. Essentially, for any item j (including the bridge variable), they calculate the conditional expectation $E\left(Z_{hi} \mid Y_{hij}=c\right)$ for each of the categories $c=0,1,.....,K_j$. Since the prior on the latent variable is independent of studies, the conditional expectation for the bridge variable is unique and independent of studies. The expected value of the latent variable is now

  $$\mu_h=E\left(Z_{hi}\right)=E\left(E\left(Z_{hi} \mid Y_{hij}\right)\right)=\sum_{c=0}^{K_j}E\left(Z_{hi} \mid Y_{hij}=c\right)P\left(Y_{hij}=c\right)\Big/\sum_{c=0}^{K_j}P\left(Y_{hij}=c\right). \qquad (2)$$

  When the probabilities $P\left(Y_{hij}=c\right)$ are estimated from the observed frequencies of an item in the study, the estimated mean value $\hat{\mu}_h$ for the items in the studies may all be different. Since all of these estimates represent the mean walking disability (in any of the items) they can be compared across studies without using the bridge variable. In a sense, the difference in mean walking disability demonstrates the heterogeneity in walking disability between studies.

  This approach would provide different mean walking disability for different items within the same study. This means that different estimates in mean disability arise when different items from same study are determined. The authors explain that this is reasonable and that it can be expected from the technical details, but this could have been avoided by including a shift parameter in the prior (that would indicate the difference between studies). Both the fixed item specific parameters and this location shift in the prior could have been estimated simultaneously using maximum likelihood estimation. This would provide a systematic difference in the latent walking disability which is constant across studies whatever choice of items is used.
- Their case study is substantially more complicated, since they combine multiple studies with multiple bridge variables. The method section discusses no specific approaches to detect possible lack of fit of the model (differential item functioning), but this was investigated in the example. They came up with a “tension coefficient”, which investigates the equivalence of items across studies. More research is required to investigate the distributional aspects of this tension coefficient, or to compare it to other approaches.

Table 6. Overview of some approaches for harmonization of constructs (continued)

Bauer DJ, Hussong AM, Psychometric approaches for developing commensurate measures across independent studies: traditional and new models, Psychological Methods, 2009, 14(2), 101-125.[70]

- The proposed approach is applicable to all types of scales (continuous, count, ordinal, binary) including combinations of scales for different items in one study.
- Construct of interest is “alcohol involvement in adolescences”, but it is applicable to any other construct. They mention for instance dimensions of personality and psychology.
- Harmonization is defined as recoding variables such that they are scored with identical values in each study.

  Example: One study may collect categorical data on family income (<10,000; 10,000—19,999; etc.) while the other record the numerical value. A harmonized variable would be a value collected in categories. This is a minimal step in the approach but it is not considered sufficient since the numerical value could be underreported, while this may be less of an issue when family income is asked in categories. This type of difference in harmonized variables may also occur when interpretation of items is affected by region or placement of items among other items.
- They provide three approaches, two old methods and one new method. The old methods are linear factor analysis (LFA) for items with a continuous scale and 2-parameter item response theory (2P-IRT) for items in a binary scale. The new method, moderated nonlinear factor analysis (MNLFA) makes it possible to combine different scales and moderate the model parameters with subject and study covariates. These approaches require “common items”, which are identical or harmonized items for a construct that is available across studies.

  The mathematical detail of the LFA and 2P-IRT were presented by the authors only for one study, but we will formulate them for multiple studies. It is assumed that studies share common items such that all studies are linked. When all items have the same type, these psychometric models can be used to make and test commensurate measures across studies.
- Linear Factor Analysis:

  Let Y_{hij} be the continuous outcome for item j (=1,2,…., J_h) of subject i (=1,2,…., n_h) in study h (=1,2,…., H). The statistical model is

  $$Y_{hij} = \nu_{hj} + \lambda_{hj} Z_{hi} + \varepsilon_{hij}\,, \qquad (3)$$

  with ν_{hj} the intercept and λ_{hj} the factor loading for item j in study h, $Z_{hi} \sim N\left(\mu_h, \tau_h^2\right)$ the latent variable for subject i in study h, and $\varepsilon_{hij} \sim N\left(0, \sigma_{hj}^2\right)$ the residual. The latent variable and the residuals are assumed independent and the outcomes for items at one subject are independent conditionally on the latent variable.
- Two types of constraints on the model parameters were presented, but it seemed that the authors preferred to fix the distributional parameters of the latent variable for one study, e.g., $\mu_1 = 0$ and $\tau_1 = 1$. They also defined strong factorial invariance and partial factorial invariance. The studies are strong factorial invariant when the item specific parameters for all the common items are independent of studies. When this would hold only for a subset of common items, it is referred to as partial factorial invariant. They suggest testing this with the likelihood ratio test. Their strategy is to start with assuming strong factorial invariance and then it relaxes the constraints on the item specific parameters one by one item, until no real model improvement is obtained. Several references were provided to support this procedure. *The article gives the impression that “commensurate measures” across studies are only attainable when there is at least partial factorial invariance. Without partial factorial invariance (or strong factorial invariance) there is no way that the studies can be combined because the studies essentially measure something else. Difference in distributional parameters of the latent variable across studies indicates heterogeneity between studies.*
- 2 parameter item response theory

  The outcome Y_{hij} for item j (=1,2,…., J_h) of subject i (=1,2,…., n_h) in study h (=1,2,…., H) is now binary. The statistical model is

  $$P\left(Y_{hij} = 1 \mid Z_{hi} = \theta\right) = \exp\left(\nu_{hj} + \lambda_{hj}\theta\right) / \left(1 + \exp\left(\nu_{hj} + \lambda_{hj}\theta\right)\right), \qquad (4)$$

  with all item specific parameters and latent variable the same as for the latent factor analysis. The authors presented an alternative parameterization, but it is similar to (4). Factorial invariance and approaches to testing item equivalence for common items are also approached in the same way as for LFA.

Table 6. Overview of some approaches for harmonization of constructs (continued)

Bauer DJ, Hussong AM, Psychometric approaches for developing commensurate measures across independent studies: traditional and new models, Psychological Methods, 2009, 14(2), 101-125.[70] (continued)

- Moderated nonlinear factor analysis (MNLFA)
  The LFA and 2P-IRT are examples of generalized linear mixed models (McCullagh and Nelder, 1989), which means that link functions are used to connect the distribution of the outcome variable for an item to a linear predictor of explanatory variable. Let Y_{ij} be the outcome for item j (=1,2,…., J) of subject i (=1,2,….,n). In this formulation the authors only assume common items across all studies. The data has been pooled and an index for study does not occur anymore since studies will be incorporated through dummy variables. The author's mention that this restriction to common items is no limitation since uncommon items can be seen as missing data for some studies. This means that they can combine through indirect comparisons and/or use imputation. The model is now

$$\begin{aligned} E\left(Y_{ij} \mid Z_i = \theta\right) &= g_j^{-1}\left(\nu_{ij} + \lambda_{ij}\theta\right) \\ Z_i &\sim N\left(\mu_i, \tau_i^2\right) \end{aligned} \qquad (5)$$

  with the link function g_j depending on the item j and the item specific parameters ν_{ij} and λ_{ij} and distributional parameters μ_i and τ_i depending on the subject i through subject and study specific variables x_{1i} , x_{2i} , …., x_{Qi} :

$$\nu_{ij} = \nu_{0j} + \sum_{q=1}^{Q} \nu_{qj} x_{qi}$$

$$\lambda_{ij} = \lambda_{0j} + \sum_{q=1}^{Q} \lambda_{qj} x_{qi}$$

$$\mu_i = \alpha_0 + \sum_{q=1}^{Q} \alpha_q x_{qi}$$

$$\tau_i = \tau_0 \exp\left(\sum_{q=1}^{Q} \omega_q x_{qi}\right)$$

  For identifiability of the model it is necessary to require $\alpha_0 = 0$ and $\tau_0 = 0$. In case the distribution of the outcome Y_{ij} for item j is normal, the residuals variance σ_{ij}^2 is modeled by $\sigma_{ij} = \sigma_{0j} \exp\left(\sum_{q=1}^{Q} \delta_{qj} x_{qi}\right)$. For ordinal outcome variables, threshold parameters are introduced. The link functions for different scales are taken as the canonical link functions.

Table 6. Overview of some approaches for harmonization of constructs (continued)

Burns RA, Butterworth P, Kiely KM, Bielak AAM, Luszcz MA, Mitchell P, Christensen H, Von Sanden C, Anstey KJ, Multiple Imputation was an efficient method for harmonizing the Mini-Mental State Examination with missing item-level data, Journal of Clinical Epidemiology, 2011, 64, 787-793.[71]

- The proposed approach is applicable to all types of scales (continuous, count, ordinal, binary) including combinations of scales for different items in one study. However, the method does imply the assumption of missing at random (MAR), but this was only briefly mentioned in the discussion.
- The authors studied the harmonization of nine Australian longitudinal studies of aging, which all used the Mini-Mental State Examination (MMSE). MMSE is used to screen for dementia and to estimate the cognitive status of individuals. It consists of 30 items with binary outcomes (correct/incorrect) and involves items on “orientation”, “registration”, “attention”, and “calculation”.
- Each study has missing item-level data that may be associated with cognitive decline or impairment and impending mortality. Participants may fail to complete items due to other reasons (e.g., stroke, deafness, physical disability, etc.). The authors claim that mean item substitution and casewise deletion are inappropriate for handling this type of missing item-level data and they propose multiple imputation to maintain the largest possible sample of participants and the largest available amount of data.
- Percentages of missing item-level data varied strongly between studies from 0.5% to 95% participants with missing items. However, 33.8% of the participants missed just one item and only 6.8% of the participants had more than 10 items missing. For most items not more than 1% was missing but for some items as high as 15% was missing.
- Multiple imputation with chained equations (MICE), using the add-on in STATA version 10, was applied to all variables: MMSE items, age, sex, years of education, study, and study interactions. Essentially, MICE uses the conditional distribution of a variable, given all the other variables to generate new values, see Van Buuren, Brand, Groothuis-Oudshoorn, and Rubin (2006). Five imputation data sets were generated, but they only used the average value of the imputed data sets as the imputed value. Thus for the MMSE items, they imputed a probability (estimated from five imputations). Their approach seems similar to a mean substitution and therefore loses the strength of multiple imputation. They do recognize this (in their discussion), but they argue that this is more practical (and therefore appropriate). They tested the efficacy of their averaged MI approach on the complete cases, by randomly assigning MMSE items missing. Thus they only investigated their procedure by assuming that items were missing completely at random (MCAR).
- The analysis of the data demonstrated an effect of age and years of education on the missing data (by comparing the variable for participants with and without missing item-level data). Older participants and less years of education were more likely to have missing data. No items (univariate approach) seem to be related to clinical diagnosis for dementia (after correction of multiplicity), nor did clinical diagnosis for dementia affect the number of missing items for participants. The MI approach resulted in an upward shift in the mean MMSE, but a lower mean MMSE score remain true for participants with missing item-level data compared with participants with no missing item-level data after imputation. This is what the authors expected, but without arguing why this is intuitive. A priori it is not obvious that the differences should become closer.
- The efficacy study generated missing data at 5%, 10%, 15%, 25%, 50%, 75%, 90%, and 95% for a random sample of participants of size 10% and 20%. The size of participants seem without effect, but the level of missing items showed a dramatic reduction in accuracy beyond 50% of missing item-level data. The correlation between imputed and original scores drops rapidly if more missing items are missing. A t-test on the imputed values with the original values (for the 20% participants and 50% missing item-level data) was not significant. Furthermore, the use of MI elevated MMSE scores for those imputed, similar to the earlier analyses.
- The authors suggest that their MI model is correctly modifying nonresponders’ total scores, except maybe for the very old (95+). For this group less consistent results between imputed and original values were observed. They also believe that MI reflects a significant improvement to estimating total MMSE scores over previous estimation techniques. They claim that it is suitable for up to 50% missing item-level data.

2P-IRT = two-parameter logistic using item response theory; LFA = linear factor analysis; MAR = missing at random; MCAR = missing completely at random; MI = multiple imputation; MICE = multiple imputation with chained equations; MMSE = Mini Mental State Examination; MNLFA = moderated nonlinear factor analysis; UK = United Kingdom

Table 7. Characteristics of CSHA, CCHS-CLSA, and NuAge cohorts

Characteristic	CSHA	CCHS-CLSA	NuAge
Study Design	10 year Longitudinal Cohort with follow-up at 5 and 10 years	Cross-Sectional	5-year Longitudinal Cohort with yearly follow-up
Period of Recruitment	Feb 1991- May 1992	Dec 2008-Nov 2009	Jan 2004-April 2005
Sample Size at Baseline	10,263 (9,008 community-dwelling; 1,255 institutionalized)	20,087 (full CCHS sample 32,005)	1,793
Inclusion Criteria	Community-dwelling and residents of institutions aged 65 and older residing in Canada. All participants scoring below 78/100 on the Modified Mini-Mental State Examination and a subsample of persons scoring 78 or greater were invited to attend the clinical component of the CSHA (total n=2,339). Those who could not complete the 3MS (n=59) were sent for clinical evaluation. All participants were required to be fluent in either English or French.	Community-dwelling people aged 45 years and over living in one of the ten Canadian provinces. The content of the CCHS-Healthy Aging was developed collaboratively by Statistics Canada and researchers from the CLSA. As part of the Statistics Canada-CLSA collaboration, CCHS participants were asked whether their survey data could be shared with the CLSA. The CCHS-CLSA sample includes data from CCHS participants between the ages of 45-85 who agreed to share their data with the CLSA	Community-dwelling men and women 67-84 years old living in the regions of Montreal, Laval and Sherbrooke in Quebec, Canada who spoke French or English, were free of disabilities in activities of daily living, without cognitive impairment, able to walk one block or to climb one flight of stairs without rest and willing to commit to a 5-year study period
Exclusion Criteria		Individuals living on Indian Reserves and on Crown Lands, institutional residents, full-time members of the Canadian Forces, residents of certain remote regions	Those who had heart failure ≥ class II, chronic obstructive pulmonary disease requiring oxygen therapy or oral steroids, inflammatory digestive diseases or cancer treated either by radiation therapy, chemotherapy or surgery in the past 5 years
Sampling Frame	The community sampling frame used the computerized records of the provincial universal health insurance plans, except in the province of Ontario where an aggregated list based on election and other municipal records was used. Age-stratified random samples were drawn in each sampling area, following an optimum allocation procedure. The institutional sampling frame included nursing homes and chronic care facilities. Institutions were stratified by size and random samples of people aged 65 or over were drawn from them.	The CCHS—Healthy Aging used the 2006 Census as its sampling frame. All dwellings within the 10 Canadian provinces containing at least one household member aged 45 and over were included in the sampling population. A two-step strategy was used to allocate the sample to the provinces. First, 125 sample units based on groups of census area dissemination blocks were allocated to each domain of interest (10 age/sex groups) in each province. Second, the remaining units were allocated to the provinces using a power-allocation method	
Response Rate	85.7%	74.4%	58.6%
Subset Included in Current Study	n=1,730 participants with full neuropsychological battery	n=7,107 participant 65 years and older who completed the CCHS cognition module	n=432 participating in the nutrition quality and cognitive decline sub-study

3MS = Modified Mini Mental State Exam; CCHS = Canadian Community Health Survey; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; NuAge = Quebec Longitudinal Study on Nutrition and Aging

Table 8. Neuropsychological measures/tests in CSHA, CCHS-CLSA and NuAge

Domain	Test/Task	CSHA	CCHS-CLSA	NuAge
Memory	Buschke Cued Recall	•		•
	Wechsler Memory Scale: Information Subtest	•		
	Rey Auditory-Verbal Learning	•	•	
	Benton Visual Retention Test -Revised	•		
	Working Memory	•		
	Brown-Peterson Task			•
	Rey & Taylor recall			•
	Health Utilities Index Memory/Thinking Attribute		•	
Abstract thinking	WAIS—R Similarities Test (short form)	•		
Executive functioning	WAIS-R Digit Symbol Sub-test	•		•
	Animal Naming (semantic fluency)	•	•	
	Mental Alternation Test		•	
	Stroop Task			•
Judgment	WAIS-R Comprehension (short form)	•		
Aphasia	Tokens Test	•		
	Word Fluency	•		
Agnosia	Buschke Visual Identification	•		
Construction	WAIS-R Block Design (short form)	•		
	Rey & Taylor copy test			•

CCHS = Canadian Community Health Survey; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; NuAge = Quebec Longitudinal Study on Nutrition and Aging; WAIS-R = Wechsler Adult Intelligence Scale-Revised

Table 9. Classification of the level of compatibility between assessment items and DataSchema variables

Class*	Description
Complete	According to the pairing rules, the meaning, format or standard operating procedures used for collection of the assessment items allow construction of the variable as defined.
Partial	According to the pairing rules, the meaning, format or standard operating procedures used for collection of the assessment items allow the construction of the variable as defined, but with an unavoidable loss of information. This class includes two subcategories: ***Proximate*****:** if the only reason for the classification as partial is because categories are used to collect information for a DataSchema variable that is defined as continuous. ***Tentative*****:** whenever a variable is classified as partial for any other reason.
Impossible	If no relevant information is collected **(Impossible Not Covered)** or, based on the pairing rules, insufficient information exists to construct the variable as defined **(Impossible Covered).**

*In certain instances, a DataSchema variable is not pertinent in the context of a particular study (*e.g.*, the ‘occurrence of prostate cancer’ variable in the context of a study recruiting only women). In such cases, the variable is classified as “not applicable” for that study.

Table 10. Variable description and results of pairing for CSHA, CCHS-CLSA, and NuAge

Candidate Variable	Variable Description	CSHA	CCHS-CLSA	NuAge
Age	Participant's age at recruitment self reported by the participant.	CM	CM	CM
Sex	Gender of the participant.	CM	CM	CM
Highest Level of Education	Highest level of education completed by the participant.	CM	CM	IM
Number of Years Education	Number of years of education	CM	IM	CM
Household Income	Average current annual income, before taxes, of the participant's entire household.	PM	CM	CM
Household Income Categorical	Average current annual income, before taxes, of the participant's entire household based on CSHA categories.	CM	CM	CM
Country of Birth	Country where the participant was born.	CM	CM	CM
Ever Smoked Cigarettes	Indicator of whether the participant has ever smoked cigarettes.	PM	CM	IM
Current Cigarette Smoker	Indicator of whether the participant currently smokes cigarettes.	IM	CM	PM
Current Quantity of Cigarettes Smoked	Current average number of cigarettes smoked per week.	IM	CM	PM
Ever Alcohol Consumption	Indicator of whether the participant has ever consumed alcohol.	PM	CM	IM
Current Use Alcohol	Indicator of whether the participant currently consumes alcohol.	IM	CM	CM
Current of Ever Use of Alcohol	Indicator of whether the participant currently or ever consumes alcohol.	PM	CM	PM
Standing Height	The vertical measurement or distance from the foot to the head of the participant when he/she is standing.	CM	CM	CM
Weight	Weight of the participant.	CM	CM	CM
Body Mass Index	Weight (in kg) divided by height (in m) squared. Body mass index = (Weight) / (Standing height * 0.01)^2	CM	CM	CM
Hip Circumference	Measured distance around hips.	IM	IM	CM
Waist Circumference	Measured distance around waist.	IM	IM	CM
Heart Rate at Rest	Number of heart beats per minute measured at rest.	IM	IM	CM
Diastolic Blood Pressure at Rest	Diastolic blood pressure measured at rest.	CM	IM	CM
Systolic Blood Pressure at Rest	Systolic blood pressure measured at rest.	CM	IM	CM
Occurrence of High Blood Pressure	Occurrence of high blood pressure at any point during the life of the participant.	CM	CM	PM
Current Treatment for High Blood Pressure	Indicator of whether the participant is currently treated for high blood pressure.	CM	CM	IM
Occurrence of Stroke	Occurrence of stroke at any point during the life of the participant.	CM	CM	IM
Occurrence of Diabetes	Occurrence of diabetes at any point during the life of the participant.	CM	CM	CM
Occurrence of Myocardial Infarction	Occurrence of myocardial infarction at any point during the life of the participant	CM	CM	PM
Family History of High Blood Pressure	Occurrence of high blood pressure amongst members of the biological family of the participant (mother, father, siblings and children).	IM	IM	IM
Family History of Stroke	Occurrence of stroke amongst members of the biological family of the participant (mother, father, siblings and children).	IM	IM	IM
Family History of Diabetes	Occurrence of diabetes amongst members of the biological family of the participant (mother, father, siblings and children).	IM	IM	IM

Table 10. Variable description and results of pairing for CSHA, CCHS-CLSA, and NuAge (continued)

Candidate Variable	Variable Description	CSHA	CCHS-CLSA	NuAge
Family History of Myocardial Infarction	Occurrence of myocardial infarction amongst members of the biological family of the participant (mother, father, siblings and children).	IM	IM	IM
Level of Physical Activity	Categorical indicator of the participant's level of physical activity. Based on IPAQ scoring protocol (https://sites.google.com/site/theipaq/scoring-protocol).	IM	CM	CM
Total physical activity	Quantitative indicator of global physical activity in metabolic equivalent (MET)-minutes per week. Based on IPAQ scoring protocol (https://sites.google.com/site/theipaq/scoring-protocol)	IM	CM	CM
Level of Physical Activity (CSHA-based—ordinal)	Categorical indicator of the participant's level of physical activity using CSHA categories (ordinal; 3 categories)	CM	CM	CM
Level of Physical Activity (CSHA-based -binary)	Categorical indicator of the participant's level of physical activity using CSHA categories (binary)	CM	CM	CM

CCHS = Canadian Community Health Survey; CLSA = Canadian Longitudinal Study on Aging; CM = complete match; CSHA = Canadian Study of Health and Aging; IM = impossible match; IPAQ = International Physical Activity Questionnaire; kg = kilograms; m = meters; NuAge = Quebec Longitudinal Study on Nutrition and Aging; PM = partial match

Table 11. Demographic and health-related characteristics of CSHA, CCHS-CLSA, and NuAge at baseline in participants with cognition data

	CCHS-CLSA (n=7,107)	**CSHA (n=1,730)**	**NuAge (n=432)**
Age [mean (sd)]	73.2 (5.9)	79.7 (7.0)	73.7 (4.0)
Age Group (n, %)			
65-74	4,162 (58.6)	367 (21.2)	265 (61.3)
75-85	2,945 (41.4)	976 (56.4)	167 (38.7)
>85		387 (22.4)	-
Sex (n, % Female)	4,103 (57.7)	1,084 (62.7)	232 (53.7)
Highest Level of Education			
Low (0-8)	1,342 (19.0)	841 (48.9)	66 (15.3)
Medium (9-13)	2,664 (37.7)	619 (35.8)	171 (39.6)
High (14+)	3,055 (43.3)	270 (15.6)	195 (45.1)
Household Income (n, %)			
<$10k	360 (2.22)	39 (6.9)	3 (0.7)
$10k to $14,999	712 (4.40)	108 (19.1)	20 (4.6)
$15k to $19,999	1,000 (6.18)	35 (6.2)	19 (4.4)
$20k to $29,999	1926 (11.90)	66 (11.6)	66 (15.2)
$30k to $39,999	1,789 (11.06)	31 (5.5)	90 (20.8)
$40k to $49,999	1,484 (9.17)	21 (3.7)	57 (13.2)
$50k to $59,999	1,404 (8.68)	15 (2.7)	50 (11.6)
$60k to $69,999	1,166 (7.21)	9 (1.6)	19 (4.4)
$70k and more	4,582 (28.33)	8 (1.4)	54 (12.5)
Missing	1,756 (10.85)		54
Country of Birth (n, % Canadian)	5781 (81.4)	1166 (67.4)	387 (89.6)
Ever/current Alcohol Use (n, %)	6550 (92.2)	344 (22.8)	411 (95.1)
Level of Physical Activity (n, %)			
Low	1,342 (19.0%)	841 (48.6%)	66 (15.3%)
Medium	2,664 (37.7%)	619 (35.8%)	171 (39.6%)
High	3,055 (43.3%)	270 (15.6%)	195 (45.1%)
Height [mean,(sd)]			
Male	174.6 (7.1, n=2998)	170.5 (7.7, n=607)	168.5 (7.4)
Female	160.4 (6.4, n=4073)	157.3 (7.4, n=993)	155.4 (5.7)
All	166.5 (9.7, n=7,001)	162.3 (9.9, n=1,600)	161.5 (9.2)
Weight [mean,(sd)]			
Male	82.8 (14.3, n=2990)	72.6 (12.7, n=622)	80.0 (12.9)
Female	68.6 (13.8, n=4011)	60.3 (12.5, n=1032)	66.4 (12.8)
All	74.6 (15.7, n=7,001)	64.9 (13.9, n=1,554)	72.7 (14.5)
Chronic conditions (n, %)			
High Blood Pressure	3,993 (56.2)	614 (35.5)	206 (47.7) ^{p}
Stroke	283 (4.0)	211 (12.2)	0 ^{p}
Diabetes	1,258 (17.7)	228 (13.2)	40 (9.3) ^{p}
Myocardial Infarction	876 (12.4)	263 (15.2)	57 (13.4) ^{p}

CCHS = Canadian Community Health Survey; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; IM = impossible match; IPAQ = International Physical Activity Questionnaire; k = thousand;n = sample size; NuAge = Quebec Longitudinal Study on Nutrition and Aging; sd = standard deviation

^{P}Partial match.

Table 12. Cognition data available for analysis in CCHS-CLSA, CSHA and NuAge

Analysis Type	CCHS-CLSA	CSHA	NuAge
Within Dataset	raw-Rey	raw-Rey	
	t-Rey	t-Rey	
	c-Rey	c-Rey	
	raw-HUI		
	t-HUI		
	c-HUI		
		raw-Buschke (free)	raw-Buschke (free)
		t-Buschke (free)	t-Buschke (free)
		c-Buschke (free)	c-Buschke (free)
		raw-Buschke (total)	raw-Buschke (total)
		t-Buschke (total)	t-Buschke (total)
		c-Buschke (total)	c-Buschke (total)
	Latent-memory	Latent-Memory	Latent-Memory
Between Datasets	Harmonized latent variable [Rey]	Harmonized latent variable [Rey, Buschke (free, total)]	Harmonized latent variable [Buschke (free, total)]

CCHS = Canadian Community Health Survey; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; HUI = Health Utilities Index; n = sample size; NuAge = Quebec Longitudinal Study on Nutrition and Aging; sd = standard deviation

Table 13. Estimates of the model parameters of the latent variable model

Mean Parameters	CCHS-CLSA	CSHA	NuAge
Intercept	0.955 [0.774; 1.136]	2.653 [2.187; 3.119]	1.009 [-0.313; 2.332]
Effect Health Utility Index	2.813 [2.777; 2.848]	NA	NA
Effect REY free recall	NA	-1.444 [-1.489; -1.399]	NA
Effect Buschke total recall	NA	2.693 [2.629; 2.758]	2.291 [2.199; 2.383]
Effect medium education	0.201 [0.161; 0.240]	0.049 [-0.038; 0.136]	0.160 [-0.026; 0.346]
Effect high education	0.356 [0.317; 0.395]	0.176 [0.036; 0.317]	0.357 [0.163; 0.552]
Effect sex	0.210 [0.183;0.395]	-0.004 [-0.086; 0.078]	0.449 [0.311; 0.588]
Effect age	-0.029 [-0.031; -0.027]	-0.033 [-0.039; -0.027]	-0.022 [-0.040; -0.005]
Variance parameters	**CCHS-CLSA**	**CSHA**	**NuAge**
Intercept	-3.124 [-4.295; -1.953]	-2.161 [-2.714; -1.608]	-2.263 [-4.289; -0.237]
Effect medium education	-0.063 [-0.301; 0.175]	0.106 [0.006; 0.206]	0.207 [-0.151; 0.565]
Effect high education	-0.010 [-0.241; 0.221]	0.144 [-0.006; 0.293]	0.184 [-0.186;0.555]
Effect sex	0.148 [-0.044; 0.339]	0.286 [0.183; 0.388]	0.018 [-0.209; 0.244]
Effect age	0.023 [0.008; 0.038]	0.021 [0.014; 0.027]	0.020 [-0.007; 0.047]

CCHS = Canadian Community Health Survey; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; NuAge = Quebec Longitudinal Study on Nutrition and Aging

Table 14. Frequencies of the cognition tests in CCHS-CLSA, CSHA and NuAge

Correct Number	CCHS-CLSA (n=7,107)		CSHA (n=1,730)			NuAge (n=432)	
	Health Utility	Rey Free Recall	Rey Free Recall	Buschke Free Recall	Buschke Total Recall	Buschke Free Recall	Buschke Total Recall
0	5,055	83	79	74	4	1	0
1	145	200	126	40	5	0	0
2	1,483	617	226	64	10	3	0
3	351	1,108	373	88	7	14	0
4	70	1,441	307	149	10	28	1
5	1	1,438	227	188	14	58	0
6	NA	1,085	92	242	22	61	3
7	NA	592	45	301	36	69	4
8	NA	324	18	248	46	64	4
9	NA	134	1	185	73	53	10
10	NA	56	3	111	123	38	5
11	NA	19	0	35	286	20	24
12	NA	8	0	5	1,094	14	33
13	NA	2	0	NA	NA	8	46
14	NA	0	0	NA	NA	1	81
15	NA	0	0	NA	NA	0	84
16	NA	NA	NA	NA	NA	0	137
Total	7,105	7,107	1,497	1,730	1,730	432	432

CCHS = Canadian Community Health Survey; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; n = sample size; NA = not applicable; NuAge = Quebec Longitudinal Study on Nutrition and Aging

Table 15. Correlations among cognitive measures and between cognitive measures and the latent variable.

Study	Variables	HUI	BUSC1	BUSC_T	LZ
CCHS-CLSA	QREY HUI	0.12482	-	-	0.78996 0.56384
CSHA	QREY BUSC1 BUSC_T		0.33719	0.26896 0.67315	0.58046 0.86794 0.81743
NuAge	BUSC1 BUSC_T			0.58398	0.84326 0.83290

BUSC1 = BUSC 1st recall; BUSC_T = BUSC total recall; CCHS = Canadian Community Health Survey; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; HUI = Health Utilities Index; LZ = predictive value; NuAge = Quebec Longitudinal Study on Nutrition and Aging; QREY = REY 1st recall

Table 16. Information about variables used in the meta-analyses

Variable	Levels	CCHS-CLSA (n=7107)	CSHA (n=1730)	NuAge (n=432)	p-value
Age [mean, (sd)]	Numerical	73.2 (5.85)	79.7 (6.96)	73.7 (3.95)	< 0.001[a]
Height [mean, (sd)]	Male	174.6 (7.1, n=2998)	170.5 (7.7, n=607)	168.5 (7.4)	
	Female	160.4 (6.4, n=4073)	157.3 (7.4, n=993)	155.4 (5.7)	
	All	166.5 (9.7, n=7,001)	162.3 (9.9, n=1,600)	161.5 (9.2)	
Weight [mean, (sd)]	Male	82.8 (14.3, n=2990)	72.6 (12.7, n=622)	80.0 (12.9)	
	Female	68.6 (13.8, n=4011)	60.3 (12.5, n=1032)	66.4 (12.8)	
	All	74.6 (15.7, n=7,001)	64.9 (13.9, n=1,554)	72.7 (14.5)	
BMI	Numerical	26.9 (4.85)	24.6 (4.58)	27.8 (4.58)	<0.001[a]
Sex	Male	3004 (42.3%)	646 (37.3%)	200 (46.3%)	0.001[b]
	Female	4103 (57.7%)	1084 (62.7%)	232 (53.7%)	
Education Level	Low	1342 (19.0%)	841 (48.6%)	66 (15.3%)	<0.001[b]
	Medium	2664 (37.7%)	619 (35.8%)	171 (39.6%)	
	High	3055 (43.3%)	270 (15.6%)	195 (45.1%)	
Country of Birth	Canada	5781 (81.4%)	1166 (67.4%)	387 (89.6%)	<0.001[b]
	Other	1324 (18.6%)	564 (32.6%)	45 (10.4%)	
Physical Activity	Never/Low	1938 (27.3%)	991 (66.8%)	493 (33.2%)	<0.001[b]
	Mod/High	5169 (72.7%)	203 (13.7%)	341 (78.9%)	
Alcohol consumption	Never	557 (7.8%)	1163 (77.2%)	21 (4.86%)	<0.001[b]
	Other	6550 (92.2%)	344 (22.8%)	411 (95.1%)	
Occurrence Diabetes	Never	5847 (82.3%)	1463 (86.5%)	392 (90.7%)	<0.001[b]
	Other	1258 (17.7%)	228 (13.5%)	40 (9.26%)	
Occurrence of High Blood Pressure	Never	3112 (43.8%)	947 (60.7%)	226 (52.3%)	<0.001[b]
	Other	3993 (56.2%)	614 (39.3%)	206 (47.7%)	
Occurrence of Myocardial Infarction	Never	6217 (87.7%)	1402 (84.2%)	367 (86.6%)	0.008[b]
	Other	876 (12.4%)	263 (15.8%)	57 (13.4%)	

CCHS = Canadian Community Health Survey; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; NuAge = Quebec Longitudinal Study on Nutrition and Aging

[a] Analysis of variance (between CCHS-CLSA, CSHA, and NuAge).

[b] Pearson’s chi-square test (between CCHS-CLSA, CSHA, and NuAge).

Table 17. Summary of relationship of raw scores for cognitive measures with variables of interest

Variable	CCHS-CLSA n=7,107		CSHA n=1,730			NuAge n=432	
	Rey	HUI	Rey	BUSC1	BUSC_T	BUSC1	BUSC_T
Age in years (n)	7,106	7,104	1,497	1,730	1,730	432	432
Intercept	11.8354	6.4795	7.0990	13.6813	15.3429	13.3827	18.3016
Beta	-0.0981	-0.0151	-0.0463	-0.0938	-0.0542	-0.0819	-0.0578
R-square	0.0869	0.0073	0.0332	0.0615	0.0391	0.0184	0.0113
p-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.0048	0.0269
Gender (n)	7,106	7,104	1,497	1,730	1,730	432	432
Intercept	4.2756	5.3790	3.2968	6.3050	11.2523	6.6700	13.6350
Beta (Ref: male)	0.6503	-0.0089	0.1876	-0.1629	-0.3667	1.2610	0.7572
R-square	0.0272	0.00002	0.0027	0.0009	0.0087	0.0692	0.0310
p-value	<0.0001	0.7203	0.0442	0.2133	0.0001	<0.0001	0.0002
Highest level of education (n)	7,060	7,058	1,488	1,719	1,719	432	432
Intercept	3.7675	5.2601	3.1001	6.2735	11.0190	6.9091	13.6515
Beta (Ref: Low (0-8)							
Moderate (9-13)	0.7843	0.0886	0.4733	-0.1492	-0.0015	0.4006	0.3079
High (14+)	1.3601	0.1828	0.8941	-0.1575	0.0738	0.6530	0.6384
R-square	0.0660	0.004	0.0313	0.0007	0.0002	0.0084	0.0110
p-value	<0.0001	<0.0001	<0.0001	0.5271	0.8773	0.1632	0.0939
Country of birth (n)	7,104	7,102	1,497	1,730	1,730	432	432
Intercept	4.4698	5.3505	3.2076	6.0496	10.8972	7.40000	14.2000
Beta (Ref: other country)	0.2235	0.0289	0.3095	0.2274	0.1860	-0.0589	-0.1767
R-square	0.0020	0.0001	0.0070	0.0016	0.0021	0.0001	0.0006
p-value	0.0002	0.3598	0.0012	0.0922	0.0571	0.8759	0.6018
Ever/current alcohol consumption (n)	7,106	7,104	1,297	1,507	1,507	432	432
Intercept	4.2496	5.3680	3.4443	6.2562	11.0120	6.0000	13.5714
Beta (Ref: never)	0.4356	0.0063	0.0643	0.0054	0.1217	1.4161	0.4943
R-square	0.0036	0.000003	0.0002	0.0000	0.0007	0.0162	0.0025
p-value	<0.0001	0.8907	0.5853	0.9736	0.2982	0.0080	0.3041
Standing height in centimeter (n)	7,070	7,068	1,386	1,600	1,600	432	432
Intercept	6.9217	5.2908	2.9275	3.9735	8.9286	12.1409	14.8455
Beta	-0.0136	0.0010	0.0032	0.0141	0.0132	-0.0297	-0.0050
R-square	0.0046	0.0001	0.0003	0.0028	0.0048	0.0131	0.0005
p-value	<0.0001	0.4292	0.4994	0.0341	0.0055	0.0172	0.6574
Weight in Kg	7,000	6,998	1,428	1,654	1,654	432	432
Intercept	4.7616	5.3239	3.1631	4.7059	10.0507	9.1386	14.2068
Beta	-0.0014	0.0007	0.0040	0.0235	0.0152	-0.0246	-0.0023
R-square	0.0001	0.0001	0.0010	0.0154	0.0121	0.0224	0.0002
p-value	0.3389	0.3780	0.2332	<0.0001	<0.0001	0.0018	0.7502

Table 17. Summary of relationship of raw scores for cognitive measures with variables of interest (continued)

Variable	CCHS-CLSA n=7,107		CSHA n=1,730			NuAge n=432	
	Rey	HUI	Rey	BUSC1	BUSC_T	BUSC1	BUSC_T
Body mass index (n)	6,968	6,966	1,365	1,576	1,576	432	432
Intercept	4.3508	5.3571	3.2984	4.8682	10.3956	8.6985	14.0952
Beta	0.0012	0.0007	0.0063	0.0575	0.0273	-0.0486	-0.0019
R-square	0.0008	0.00001	0.0003	0.0101	0.0044	0.0087	0.0000
p-value	0.0165	0.7784	0.5477	<0.0001	0.0081	0.0529	0.9320
Occurrence of high blood pressure(n)	7,104	7,102	1,358	1,561	1,561	432	432
Intercept	4.7317	5.3851	3.4469	6.1542	10.9926	7.4912	14.0354
Beta (Ref: never)	-0.1427	-0.0191	-0.0563	0.1699	0.1198	-0.3018	0.0131
R-square	0.0013	0.00008	0.0002	0.0010	0.0010	0.0040	0.0000
p-value	0.0022	0.4397	0.5650	0.2144	0.2168	0.1905	0.9494
Occurrence of diabetes (n)	7,104	7,102	1,464	1,691	1,691	432	432
Intercept	4.7253	5.3873	3.4269	6.1955	10.9863	7.3367	14.0077
Beta (Ref: never)	-0.4169	-0.0746	-0.1144	0.0326	0.1979	0.1133	0.3673
R-square	0.0067	0.0008	0.0005	0.0000	0.0012	0.0002	0.0025
p-value	<0.0001	0.0202	0.3994	0.8624	0.1484	0.7758	0.3032
Occurrence of myocardial infarction (n)	7,092	7,090	1,440	1,665	1,665	424	424
Intercept	4.7196	5.3889	3.4459	6.2126	11.0057	7.2888	14.0327
Beta (Ref: never)	-0.5404	-0.1138	-0.1490	0.1258	0.2528	0.5006	0.1077
R-square	0.0083	0.0013	0.0010	0.0003	0.0025	0.0051	0.0003
p-value	<0.0001	0.0023	0.2391	0.4740	0.0434	0.1436	0.7234
Categorical indicator of the participants level of physical activity using CSHA categories	7,106	7,104	12,79	1,484	1,484	432	432
Intercept	4.2963	5.298	3.2058	5.7855	10.8261	7.5870	14.1739
Beta (Ref: none)							
Low level							
Moderate or high level	0.1992	-0.0651	0.4366	0.6480	0.3512	-0.0092	0.3816
	0.4578	0.1140	0.5714	1.1009	0.4862	-0.3025	-0.2179
R-square							
p-value	0.0082	0.0039	0.0231	0.0357	0.0141	0.0026	0.0076
	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.5737	0.1928

Table 17. Summary of relationship of raw scores for cognitive measures with variables of interest (continued)

Variable	CCHS-CLSA n=7,107		CSHA n=1,730			NuAge n=432	
	Rey	HUI	Rey	BUSC1	BUSC_T	BUSC1	BUSC_T
Categorical indicator of the participants level of physical activity using CSHA categories (4 categories) (n)	7,106	7,104	1,279	1,484	1,484	432	432
Intercept	4.2963	5.2980	3.2058	5.7855	10.8261	7.5870	14.1739
Beta (Ref: none)							
Low level	0.1992	-0.0651	0.4366	0.6480	0.3512	-0.0092	0.3816
Moderate level	0.3834	0.0975	0.5080	0.9744	0.3902	-0.1623	-0.1830
High level	0.6909	0.1656	0.7845	1.5215	0.8054	-0.5542	-0.2805
R-square	0.0115	0.0045	0.0246	0.0382	0.0169	0.0075	0.0080
p-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.3602	0.3274

BUSC1 = BUSC 1st recall; BUSC_T = BUSC total recall; CCHS = Canadian Community Health Survey; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; HUI = Health Utilities Index; kg = kilograms; n= sample size; NuAge = Quebec Longitudinal Study on Nutrition and Aging

Table 18a. Summary of aggregate data from CCHS-CLSA, CSHA, and NuAge studies used in traditional meta-analyses of unadjusted effect estimates for raw data

Study	Instrument	High Level of Physical Activity			Low Level of Physical Activity			Difference (95% CI)
		n	Mean	SD	n	Mean	SD	
CCHS-CLSA	Rey	5,169	4.8	1.95	1,938	4.4	1.91	0.38 (0.28, 0.48)
	HUI	5,169	5.4	1.00	1,938	5.3	1.11	0.14 (0.09, 0.19)
CSHA	Rey	449	3.8	1.85	830	3.3	1.72	0.47 (0.28, 0.68)
	Buschke Free	493	6.9	2.45	991	5.9	2.71	0.97 (0.68, 1.25)
	Buschke Total	493	11.3	1.52	991	10.9	2.07	0.41 (0.21, 0.62)
NuAge	Buschke Free	341	7.3	2.38	91	7.6	2.45	-0.30 (-0.85, 0.26)
	Buschke Total	341	14.0	2.20	91	14.4	1.94	-0.40 (-0,90, 0.10)

CCHS = Canadian Community Health Survey; CI = confidence interval; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; HUI = Health Utilities Index; n= sample size; NuAge = Quebec Longitudinal Study on Nutrition and Aging; SD = standard deviation

Table 18b. Summary of aggregate data from CCHS-CLSA, CSHA, and NuAge studies used in traditional meta-analyses of effect estimates adjusted for a common set of covariates for raw data

Study	Instrument	High Level of Physical Activity			Low Level of Physical Activity			Difference (95% CI)
		n	Mean	SD	n	Mean	SD	
CCHS-CLSA	Rey	5,045	4.2	1.81	1,860	4.0	1.81	0.24 (0.15, 0.34)
	HUI	5,043	5.4	1.05	1,860	5.2	1.05	0.11 (0.05, 0.17)
CSHA	Rey	363	3.8	1.77	659	3.4	1.77	0.35 (0.12, 0.58)
	Buschke Free	400	6.9	2.60	778	6.1	2.60	0.82 (0.50, 1.13)
	Buschke Total	400	11.5	1.86	778	11.2	1.86	0.31 (0.09, 0.54)
NuAge	Buschke Free	334	6.8	2.30	90	7.1	2.30	-0.30 (-0.83, 0.23)
	Buschke Total	334	13.9	2.13	90	14.2	2.13	-0.35 (-0.84, 0.15)

CCHS = Canadian Community Health Survey; CI = confidence interval; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; HUI = Health Utilities Index; n= sample size; NuAge = Quebec Longitudinal Study on Nutrition and Aging; SD = standard deviation, adjusted for age, sex, education, BMI, country of birth, alcohol consumption, diabetes, high blood pressure and myocardial infarction

Table 18c. Summary of aggregate data from CCHS-CLSA, CSHA, and NuAge studies used in traditional meta-analyses of effect estimates adjusted for a covariates that were statistically significant at p=0.05 in the original study

Study	Instrument	High Level of Physical Activity			Low Level of Physical Activity			Difference (95% CI)
		n	Mean	SD	n	Mean	SD	
CCHS-CLSA	Rey	5,123	4.2	1.84	1,918	4.0	1.84	0.24 (0.15, 0.34)
	HUI	5,125	5.3	1.04	1,918	5.2	1.04	0.10 (0.05, 0.16)
CSHA	Rey	447	3.8	1.74	824	3.4	1.74	0.36 (0.16, 0.56)
	Buschke Free	491	6.8	2.61	984	6.1	2.61	0.76 (0.47, 1.04)
	Buschke Total	491	11.4	1.94	984	11.1	1.94	0.28 (0.07, 0.49)
NuAge	Buschke Free	341	6.7	2.30	91	7.0	2.30	-0.34 (-0.87, 0.20)
	Buschke Total	341	13.8	2.09	91	14.2	2.09	-0.35 (-0.84, 0.13)

CCHS = Canadian Community Health Survey; CI = confidence interval; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; HUI = Health Utilities Index; n= sample size; NuAge = Quebec Longitudinal Study on Nutrition and Aging; SD = standard deviation, adjusted for Age, Gender, weight, height and alcohol

Table 19a. Summary of traditional meta-analysis results for Rey from CCHS-CLSA and CSHA, and Buschke Free from NuAge using Hedges' g

Method of Statistical Harmonization	Adjustment	Summary Effect Estimate (95% CI)	I^2	Q^*	p-value
Raw Data	Unadjusted	0.16 (0.01, 0.30)	0.78	8.96	0.01
	Unadjusted (Including Participants with Complete Data for Covariates)	0.16 (0.01, 0.30)	0.75	7.84	0.02
	Study-Specific Statistically Significant Covariates	0.11 (-0.02, 0.24)	0.72	7.18	0.03
	Common Set of Covariates	0.11 (-0.02, 0.23)	0.66	5.81	0.06
T-Score	Unadjusted†	0.12 (0.01, 0.23)	0.64	5.5	0.06
	Unadjusted† (Including Participants with Complete Data for Covariates)	0.12 (0.01, 0.22)	0.55	4.47	0.11
	Common Set of Covariates	0.11 (-0.02, 0.23)	0.66	5.81	0.06
C-Score	Unadjusted	0.16 (0.01, 0.30)	0.78	8.96	0.01
	Unadjusted (Including Participants with Complete Data for Covariates)	0.16 (0.01, 0.30)	0.75	7.84	0.02
	Common Set of Covariates	0.11 (-0.02, 0.23)	0.66	5.81	0.06

CCHS = Canadian Community Health Survey; CI = confidence interval; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; NuAge = Quebec Longitudinal Study on Nutrition and Aging

*All Q Statistics have 2 degrees of freedom.

†T-Scores minimally adjusted for age, sex and education level.

Table 19b. Summary of traditional meta-analysis results for Rey from CCHS-CLSA and CSHA, and Buschke Total from NuAge using Hedges'g

Method of Statistical Harmonization	Adjustment	Summary Effect Estimate (95% CI)	I^2	Q[*]	p-value
Raw Data	Unadjusted	0.13 (-0.04, 0.30)	0.84	12.20	0.002
	Unadjusted (Including Participants with Complete Data for Covariates)	0.13 (-0.04, 0.30)	0.82	10.99	0.004
	Study-Specific Statistically Significant Covariates	0.10 (-0.04, 0.24)	0.75	8.14	0.02
	Common Set of Covariates	0.10 (-0.04, 0.23)	0.72	7.06	0.03
T-Score	Unadjusted[†]	0.10 (-0.04, 0.23)	0.76	8.34	0.02
	Unadjusted[†] (Including Participants with Complete Data for Covariates)	0.10 (-0.04, 0.23)	0.72	7.14	0.03
	Common Set of Covariates	0.10 (-0.04, 0.23)	0.72	7.06	0.03
C-Score	Unadjusted	0.13 (-0.04, 0.30)	0.84	12.20	0.002
	Unadjusted (Including Participants with Complete Data for Covariates)	0.13 (-0.04, 0.30)	0.82	10.99	0.004
	Common Set of Covariates	0.10 (-0.04, 0.23)	0.72	7.06	0.03

CCHS = Canadian Community Health Survey; CI = confidence interval; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; NuAge = Quebec Longitudinal Study on Nutrition and Aging

[*] All Q Statistics have 2 degrees of freedom.

[†] T-Scores minimally adjusted for age, sex and education level.

Table 19c. Summary of traditional meta-analysis results for Rey from CCHS-CLSA and Buschke Free from CSHA, and NuAge using Hedges' g

Method of Statistical Harmonization	Adjustment	Summary Effect Estimate (95% CI)	I^2	Q[*]	p-value
Raw Data	Unadjusted	0.18 (-0.01, 0.36)	0.88	16.46	<0.001
	Unadjusted (Including Participants with Complete Data for Covariates)	0.18 (0.0001, 0.36)	0.85	13.37	0.001
	Study-Specific Statistically Significant Covariates	0.12 (-0.04, 0.29)	0.85	13.0	0.002
	Common Set of Covariates	0.13 (-0.05, 0.31)	0.85	12.9	0.002
T-Score	Unadjusted[†]	0.14 (-0.03, 0.31)	0.85	12.92	0.002
	Unadjusted[†] (Including Participants with Complete Data for Covariates)	0.14 (-0.03, 0.32)	0.82	11.35	0.003
	Common Set of Covariates	0.14 (-0.05, 0.32)	0.85	13.38	0.001
C-Score	Unadjusted	0.18 (-0.01, 0.36)	0.88	16.46	<0.001
	Unadjusted (Including Participants with Complete Data for Covariates)	0.18 (0.0001, 0.36)	0.85	13.37	0.001
	Common Set of Covariates	0.13 (-0.05, 0.31)	0.85	12.95	0.002

CCHS = Canadian Community Health Survey; CI = confidence interval; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; NuAge = Quebec Longitudinal Study on Nutrition and Aging

[*] All Q Statistics have 2 degrees of freedom.

[†] T-Scores minimally adjusted for age, sex and education level.

Table 19d. Summary of traditional meta-analysis results for Rey from CCHS-CLSA and Buschke Free from CSHA, and Buschke Total from NuAge using Hedges' g

Method of Statistical Harmonization	Adjustment	Summary Effect Estimate (95% CI)	I^2	Q^*	p-value
Raw Data	Unadjusted	0.16 (-0.05, 0.36)	0.90	20.32	<0.001
	Unadjusted (Including Participants with Complete Data for Covariates)	0.16 (-0.04, 0.36)	0.88	16.67	<0.001
	Study-Specific Statistically Significant Covariates	0.12 (-0.06, 0.29)	0.86	13.98	0.001
	Common Set of Covariates	0.12 (-0.06, 0.31)	0.86	14.29	0.001
T-Score	Unadjusted†	0.12 (-0.07, 0.31)	0.87	15.93	<0.001
	Unadjusted† (Including Participants with Complete Data for Covariates)	0.12 (-0.07, 0.32)	0.86	14.20	0.001
	Common Set of Covariates	0.13 (-0.07, 0.32)	0.86	14.72	0.001
C-Score	Unadjusted	0.16 (-0.05, 0.36)	0.90	19.88	<0.001
	Unadjusted (Including Participants with Complete Data for Covariates)	0.16 (-0.04, 0.36)	0.88	16.67	<0.001
	Common Set of Covariates	0.12 (-0.06, 0.31)	0.86	14.29	0.001

CCHS = Canadian Community Health Survey; CI = confidence interval; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; NuAge = Quebec Longitudinal Study on Nutrition and Aging

[*] All Q Statistics have 2 degrees of freedom.

[†] T-Scores minimally adjusted for age, sex and education level.

Table 19e. Summary of traditional meta-analysis results for Rey from CCHS-CLSA, Buschke Total from CSHA, and Buschke Free from NuAge using Hedges' g

Method of Statistical Harmonization	Adjustment	Summary Effect Estimate (95% CI)	I^2	Q[*]	p-value
Raw Data	Unadjusted	0.14 (0.02, 0.27)	0.73	7.35	0.03
	Unadjusted (Including Participants with Complete Data for Covariates)	0.15 (0.02, 0.28)	0.71	6.80	0.03
	Study-Specific Statistically Significant Covariates	0.09 (-0.01, 0.20)	0.63	5.47	0.07
	Common Set of Covariates	0.10 (-0.008, 0.21)	0.61	5.18	0.08
T-Score	Unadjusted[†]	0.11 (0.02, 0.19)	0.42	3.44	0.18
	Unadjusted[†] (Including Participants with Complete Data for Covariates)	0.11 (0.02, 0.20)	0.42	3.44	0.18
	Common Set of Covariates	0.10 (-0.01, 0.22)	0.63	5.40	0.07
C-Score	Unadjusted	0.14 (0.02, 0.27)	0.73	7.35	0.03
	Unadjusted (Including Participants with Complete Data for Covariates)	0.15 (0.02, 0.28)	0.71	6.80	0.03
	Common Set of Covariates	0.10 (-0.008, 0.21)	0.61	5.18	0.08

[*] All Q Statistics have 2 degrees of freedom

[†] T-Scores minimally adjusted for age, sex and education level

CCHS = Canadian Community Health Survey; CI = confidence interval; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; NuAge = Quebec Longitudinal Study on Nutrition and Aging

Table 19f. Summary of traditional meta-analysis results for Rey from CCHS-CLSA and Buschke Total from CSHA, and NuAge using Hedges' g

Method of Statistical Harmonization	Adjustment	Summary Effect Estimate (95% CI)	I^2	Q^*	p-value
Raw Data	Unadjusted	0.12 (-0.03, 0.27)	0.81	10.51	0.005
	Unadjusted (Including Participants with Complete Data for Covariates)	0.12 (-0.03, 0.28)	0.80	9.89	0.007
	Study-Specific Statistically Significant Covariates	0.09 (-0.03, 0.20)	0.69	6.39	0.04
	Common Set of Covariates	0.09 (-0.03, 0.22)	0.69	6.42	0.04
T-Score	Unadjusted†	0.08 (-0.04, 0.20)	0.67	6.14	0.05
	Unadjusted† (Including Participants with Complete Data for Covariates)	0.08 (-0.04, 0.21)	0.70	6.04	0.05
	Common Set of Covariates	0.09 (-0.04, 0.22)	0.70	6.64	0.04
C-Score	Unadjusted	0.12 (-0.03, 0.27)	0.81	10.52	0.005
	Unadjusted (Including Participants with Complete Data for Covariates)	0.12 (-0.03, 0.28)	0.80	9.89	0.007
	Common Set of Covariates	0.09 (-0.03, 0.22)	0.69	6.42	0.04

CCHS = Canadian Community Health Survey; CI = confidence interval; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; NuAge = Quebec Longitudinal Study on Nutrition and Aging

*All Q Statistics have 2 degrees of freedom.

†T-Scores minimally adjusted for age, sex and education level.

Table 19g. Summary of traditional meta-analysis results for HUI cognition from CCHS-CLSA, Rey from CSHA, and Buschke Free from NuAge using Hedges' g

Method of Statistical Harmonization	Adjustment	Summary Effect Estimate (95% CI)	I^2	Q*	p-value
Raw Data	Unadjusted	0.13 (-0.02, 0.28)	0.80	9.86	0.007
	Unadjusted (Including Participants with Complete Data for Covariates)	0.13 (-0.02, 0.27)	0.75	8.10	0.02
	Study-Specific Statistically Significant Covariates	0.09 (-0.04, 0.22)	0.74	7.58	0.02
	Common Set of Covariates	0.09 (-0.03, 0.22)	0.66	5.83	0.054
T-Score	Unadjusted†	0.11 (-0.01, 0.22)	0.66	5.9	0.051
	Unadjusted† (Including Participants with Complete Data for Covariates)	0.11 (-0.002, 0.21)	0.57	4.63	0.099
	Common Set of Covariates	0.09 (-0.03, 0.22)	0.66	5.83	0.054
C-Score	Unadjusted	0.13 (-0.02, 0.28)	0.80	9.86	0.01
	Unadjusted (Including Participants with Complete Data for Covariates)	0.13 (-0.02, 0.27)	0.75	8.10	0.02
	Common Set of Covariates	0.10 (-0.03, 0.22)	0.66	5.83	0.054

CCHS = Canadian Community Health Survey; CI = confidence interval; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; NuAge = Quebec Longitudinal Study on Nutrition and Aging

*All Q Statistics have 2 degrees of freedom.

†T-Scores minimally adjusted for age, sex and education level.

Table 19h. Summary of traditional meta-analysis results for HUI cognition from CCHS-CLSA, Rey from CSHA, and Buschke Total from NuAge using Hedges' g

Method of Statistical Harmonization	Adjustment	Summary Effect Estimate (95% CI)	I^2	Q^*	p-value
Raw Data	Unadjusted	0.11 (-0.06, 0.28)	0.84	12.67	0.002
	Unadjusted (Including Participants with Complete Data for Covariates)	0.11 (-0.06, 0.28)	0.81	10.77	0.005
	Study-Specific Statistically Significant Covariates	0.08 (-0.06, 0.22)	0.77	8.61	0.01
	Common Set of Covariates	0.08 (-0.05, 0.22)	0.71	6.98	0.03
T-Score	Unadjusted†	0.09 (-0.06, 0.23)	0.77	8.58	0.01
	Unadjusted† (Including Participants with Complete Data for Covariates)	0.08 (-0.05, 0.22)	0.72	7.07	0.03
	Common Set of Covariates	0.08 (-0.05, 0.22)	0.71	6.98	0.03
C-Score	Unadjusted	0.11 (-0.06, 0.28)	0.84	12.67	0.002
	Unadjusted (Including Participants with Complete Data for Covariates)	0.11 (-0.06, 0.28)	0.81	10.77	0.005
	Common Set of Covariates	0.08 (-0.05, 0.22)	0.71	6.98	0.03

CCHS = Canadian Community Health Survey; CI = confidence interval; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; NuAge = Quebec Longitudinal Study on Nutrition and Aging

*All Q Statistics have 2 degrees of freedom.

†T-Scores minimally adjusted for age, sex and education level.

Table 19i. Summary of traditional meta-analysis results for HUI cognition from CCHS-CLSA and Buschke Free from CSHA, and NuAge Using Hedges' g

Method of Statistical Harmonization	Adjustment	Summary Effect Estimate (95% CI)	I^2	Q*	p-value
Raw Data	Unadjusted	0.15 (-0.06, 0.36)	0.90	20.62	<0.001
	Unadjusted (Including Participants with Complete Data for Covariates)	0.15 (-0.05, 0.35)	0.88	16.54	<0.001
	Study-Specific Statistically Significant Covariates	0.11 (-0.08, 0.30)	0.88	16.56	<0.001
	Common Set of Covariates	0.12 (-0.07, 0.31)	0.86	14.53	0.001
T-Score	Unadjusted†	0.13 (-0.05, 0.31)	0.86	14.26	0.001
	Unadjusted† (Including Participants with Complete Data for Covariates)	0.13 (-0.05, 0.32)	0.84	12.56	0.002
	Common Set of Covariates	0.12 (-0.08, 0.32)	0.87	14.88	0.001
C-Score	Unadjusted	0.15 (-0.06, 0.36)	0.90	20.61	<0.001
	Unadjusted (Including Participants with Complete Data for Covariates)	0.15 (-0.05, 0.35)	0.88	16.54	<0.001
	Common Set of Covariates	0.12 (-0.07, 0.31)	0.86	14.53	0.001

CCHS = Canadian Community Health Survey; CI = confidence interval; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; NuAge = Quebec Longitudinal Study on Nutrition and Aging

*All Q Statistics have 2 degrees of freedom.

†T-Scores minimally adjusted for age, sex and education level.

Table 19j. Summary of traditional meta-analysis results for HUI cognition from CCHS-CLSA, Buschke Free from CSHA, and Buschke Total from NuAge using Hedges' g

Method of Statistical Harmonization	Adjustment	Summary Effect Estimate (95% CI)	I^2	Q^*	p-value
Raw Data	Unadjusted	0.13 (-0.09, 0.35)	0.92	23.61	<0.001
	Unadjusted (Including Participants with Complete Data for Covariates)	0.13 (-0.09, 0.35)	0.90	19.36	<0.001
	Study-Specific Statistically Significant Covariates	0.11 (-0.09, 0.30)	0.89	17.66	<0.001
	Common Set of Covariates	0.11 (-0.09, 0.31)	0.87	15.77	<0.001
T-Score	Unadjusted†	0.11 (-0.09, 0.31)	0.88	17.08	<0.001
	Unadjusted† (Including Participants with Complete Data for Covariates)	0.11 (-0.09, 0.31)	0.87	15.21	<0.001
	Common Set of Covariates	0.11 (-0.09, 0.32)	0.88	16.11	<0.001
C-Score	Unadjusted	0.13 (-0.09, 0.35)	0.92	23.61	<0.001
	Unadjusted (Including Participants with Complete Data for Covariates)	0.13 (-0.09, 0.35)	0.90	19.36	<0.001
	Common Set of Covariates	0.11 (-0.09, 0.31)	0.87	15.77	<0.001

CCHS = Canadian Community Health Survey; CI = confidence interval; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; NuAge = Quebec Longitudinal Study on Nutrition and Aging

*All Q Statistics have 2 degrees of freedom.

†T-Scores minimally adjusted for age, sex and education level.

Table 19k. Summary of traditional meta-analysis results for HUI cognition from CCHS-CLSA, Buschke Total from CSHA, and Buschke Free from NuAge using Hedges' g

Method of Statistical Harmonization	Adjustment	Summary Effect Estimate (95% CI)	I^2	Q^*	p-value
Raw Data	Unadjusted	0.12 (-0.004, 0.24)	0.72	7.01	0.03
	Unadjusted (Including Participants with Complete Data for Covariates)	0.12 (-0.001, 0.24)	0.70	6.04	0.049
	Study-Specific Statistically Significant Covariates	0.09 (-0.03, 0.21)	0.68	6.30	0.04
	Common Set of Covariates	0.09 (-0.02, 0.20)	0.60	4.97	0.08
T-Score	Unadjusted†	0.10 (0.02, 0.17)	0.38	3,21	0.20
	Unadjusted† (Including Participants with Complete Data for Covariates)	0.10 (0.01, 0.18)	0.36	3.14	0.21
	Common Set of Covariates	0.09 (-0.03, 0.21)	0.62	5.25	0.07
C-Score	Unadjusted	0.12 (-0.004, 0.24)	0.72	7.01	0.03
	Unadjusted (Including Participants with Complete Data for Covariates)	0.12 (-0.001, 0.24)	0.67	6.04	0.049
	Common Set of Covariates	0.09 (-0.02, 0.20)	0.60	4.97	0.08

CCHS = Canadian Community Health Survey; CI = confidence interval; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; NuAge = Quebec Longitudinal Study on Nutrition and Aging

[*] All Q Statistics have 2 degrees of freedom.

[†] T-Scores minimally adjusted for age, sex and education level.

Table 19l. Summary of traditional meta-analysis results for HUI cognition from CCHS-CLSA and Buschke Total from CSHA, and NuAge using Hedges' g

Method of Statistical Harmonization	Adjustment	Summary Effect Estimate (95% CI)	I^2	Q^*	p-value
Raw Data	Unadjusted	0.10 (-0.05, 0.24)	0.80	9.76	0.008
	Unadjusted (Including Participants with Complete Data for Covariates)	0.10 (-0.05, 0.24)	0.77	8.65	0.01
	Study-Specific Statistically Significant Covariates	0.08 (-0.05, 0.21)	0.73	7.32	0.03
	Common Set of Covariates	0.08 (-0.04, 0.20)	0.67	6.10	0.047
T-Score	Unadjusted†	0.07 (-0.04, 0.19)	0.65	5.72	0.06
	Unadjusted† (Including Participants with Complete Data for Covariates)	0.07 (-0.05, 0.19)	0.64	5.51	0.06
	Common Set of Covariates	0.08 (-0.05, 0.21)	0.69	6.39	0.04
C-Score	Unadjusted	0.10 (-0.05, 0.24)	0.80	9.76	0.008
	Unadjusted (Including Participants with Complete Data for Covariates)	0.10 (-0.05, 0.24)	0.77	8.65	0.01
	Common Set of Covariates	0.08 (-0.04, 0.20)	0.67	6.10	0.047

CCHS = Canadian Community Health Survey; CI = confidence interval; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study of Health and Aging; NuAge = Quebec Longitudinal Study on Nutrition and Aging

*All Q Statistics have 2 degrees of freedom.

†T-Scores minimally adjusted for age, sex and education level.

Table 19m. Summary of Tables 19a through 19l on traditional meta-analysis results

Selection of Memory Tests	Type of Outcome	Ranges of Values From the Analysis With Different Selections of Covariates			
		Effect Size	I^2	Q	p-value
CCHS:REY CSHA:REY NuAge: Buschke Free	Raw	0.11—0.16	0.66—0.78	5.81—8.96	0.01—0.06
	T-score	0.11—0.12	0.55—0.66	4.47—5.81	0.06—0.11
	C-score	0.11—0.16	0.66—0.78	5.81—8.96	0.01—0.06
CCHS:REY CSHA:REY NuAge: Buschke Total	Raw	0.10—0.13	0.72—0.84	7.06—12.20	0.002—0.03
	T-score	0.10—0.10	0.72—0.76	7.06—8.34	0.02—0.03
	C-score	0.10—0.13	0.72—0.84	7.06—12.20	0.002—0.03
CCHS:REY CSHA:Buschke Free NuAge: Buschke Free	Raw	0.12—0.18	0.85 -0.88	12.9—16.46	<0.001—0.002
	T-score	0.14—0.14	0.82—0.85	11.35—13.38	0.001—0.002
	C-score	0.13—0.18	0.85—0.88	12.95—16.46	<0.001—0.002
CCHS:REY CSHA:Buschke Free NuAge: Buschke Total	Raw	0.12—0.16	0.86—0.90	13.98—20.32	<0.001—0.001
	T-score	0.12—0.13	0.86—0.87	14.20—15.93	<0.001—0.001
	C-score	0.12—0.16	0.86—0.90	14.29—19.88	<0.001—0.001
CCHS:REY CSHA:Buschke Total NuAge: Buschke Free	Raw	0.09—0.15	0.61—0.73	5.18—7.35	0.03—0.08
	T-score	0.10—0.11	0.42—0.63	3.44—5.40	0.07—0.18
	C-score	0.10—0.15	0.61—0.73	5.18—7.35	0.03—0.08
CCHS:REY CSHA:Buschke Total NuAge: Buschke Total	Raw	0.09—0.12	0.69—0.81	6.39—10.51	0.005—0.04
	T-score	0.08—0.09	0.67—0.70	6.04—6.64	0.04—0.05
	C-score	0.09—0.12	0.69—0.81	6.42—10.52	0.005—0.04
CCHS:HUI CSHA:REY NuAge: Buschke Free	Raw	0.09—0.13	0.66—0.80	5.83—9.86	0.007—0.054
	T-score	0.09—0.11	0.57—0.66	4.63—5.9	0.051—0.10
	C-score	0.10—0.13	0.66—0.80	5.83—9.86	0.01—0.054
CCHS:HUI CSHA:REY NuAge: Buschke Total	Raw	0.08—0.11	0.71—0.84	6.98—12.67	0.002—0.03
	T-score	0.08—0.09	0.71—0.77	6.98—8.58	0.01—0.03
	C-score	0.08—0.11	0.71—0.84	6.98—12.67	0.002—0.03
CCHS:HUI CSHA:Buschke Free NuAge: Buschke Free	Raw	0.11—0.15	0.86—0.90	14.53—20.62	<0.001—0.001
	T-score	0.12—0.13	0.84—0.87	12.56—14.88	0.001—0.002
	C-score	0.12—0.15	0.86—0.90	14.53—20.61	<0.001—0.001
CCHS:HUI CSHA:Buschke Free NuAge: Buschke Total	Raw	0.11—0.13	0.87—0.92	15.77—23.61	<0.001
	T-score	0.11—0.11	0.87—0.88	15.21—17.08	<0.001
	C-score	0.11—0.13	0.87—0.92	15.77—23.61	<0.001
CCHS:HUI CSHA:Buschke Total NuAge: Buschke Free	Raw	0.09—0.12	0.60—0.72	4.97—7.01	0.03—0.08
	T-score	0.09—0.10	0.36—0.62	3.14—5.25	0.07—0.21
	C-score	0.09—0.12	0.60—0.72	4.97—7.01	0.03—0.08
CCHS:HIU CSHA:Buschke Total NuAge: Buschke Total	Raw	0.08—0.10	0.67—0.80	6.10—9.76	0.008—0.047
	T-score	0.07—0.08	0.64—0.69	5.51—6.39	0.04—0.06
	C-score	0.08—0.10	0.67—0.80	6.10—9.76	0.008—0.047

Table 20. Summary of traditional meta-analysis results for a common variable constructed from the latent variable analysis using Hedges' g

Method of Statistical Harmonization	Adjustment	Summary Effect Estimate (95% CI)	I^2	Q[*]	p-value
Latent Variable—Weighted Mean Difference	Unadjusted[†]	0.13 (-0.03, 0.29)	0.83	12.01	0.002
	Unadjusted[†] (Including Participants with Complete Data for Covariates)	0.13 (-0.03, 0.29)	0.81	10.37	0.006
	Common Set of Covariates	0.13 (-0.04, 0.30)	0.83	11.98	0.002

CI = confidence interval

[*] All Q Statistics have 2 degrees of freedom.

[†] Latent variable minimally adjusted for age, sex and education level.

Figure 1. Scatterplot of the T-scores versus latent variable: study = CCHS-CLSA, instrument = HUI

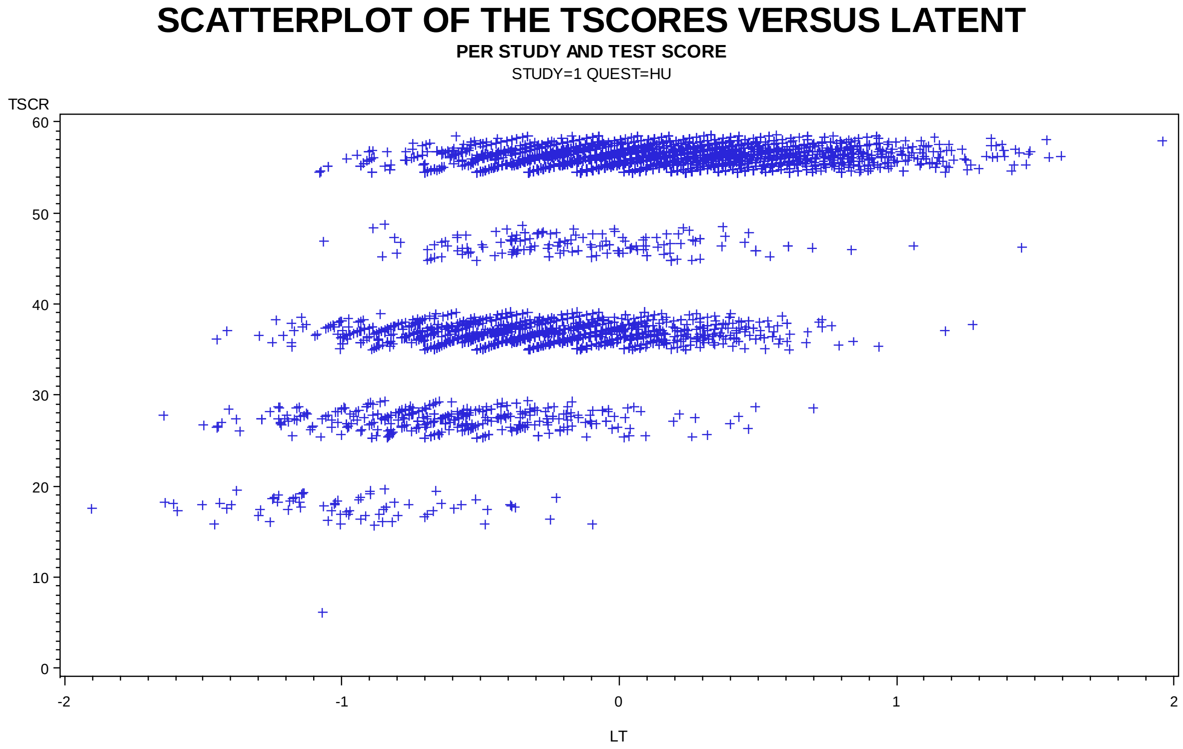


Figure 2. Scatterplot of the T-scores versus latent variable: study = CCHS-CLSA, instrument = Rey

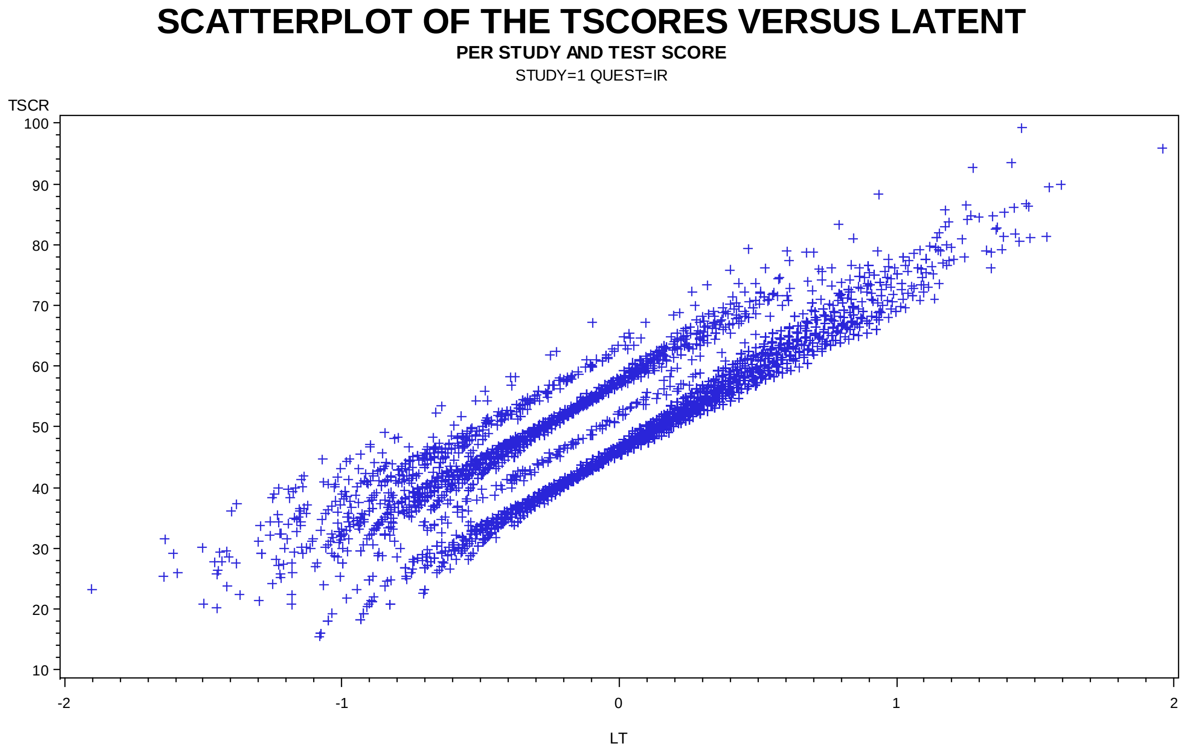


Figure 3. Scatterplot of the T-scores versus latent variable: study = CSHA, instrument = Buschke Free

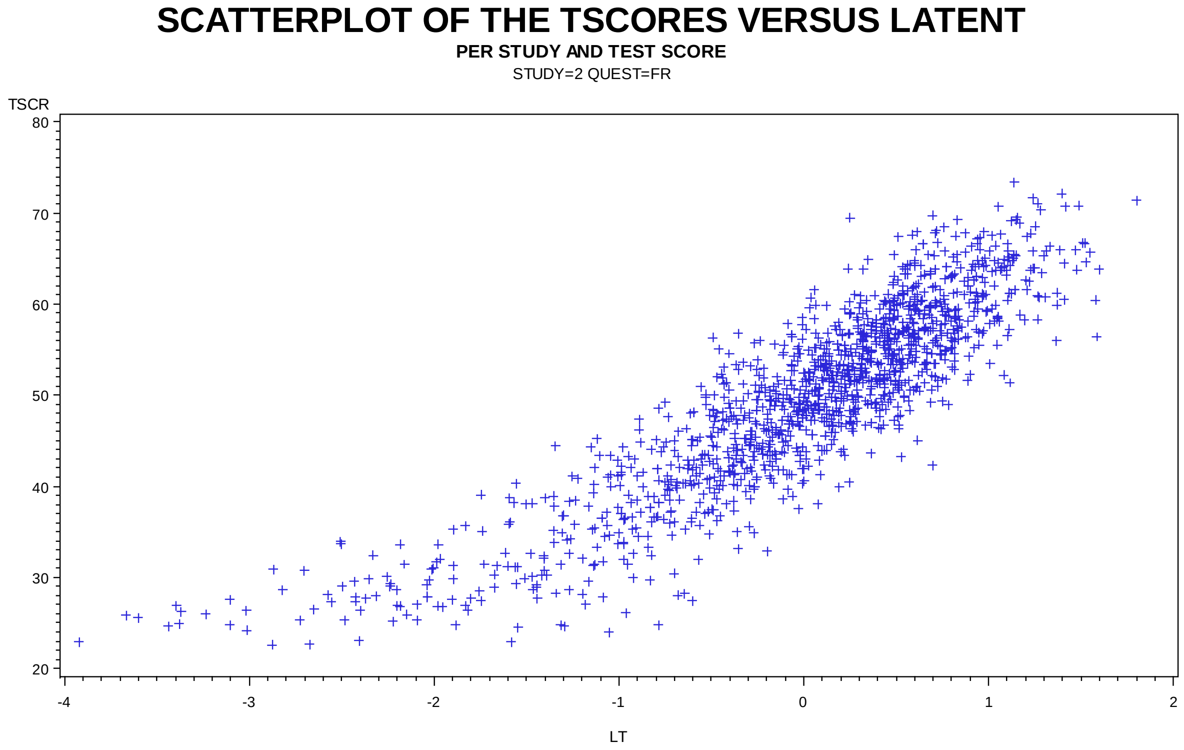


Figure 4. Scatterplot of the T-scores versus latent variable: study = CSHA, instrument = Rey

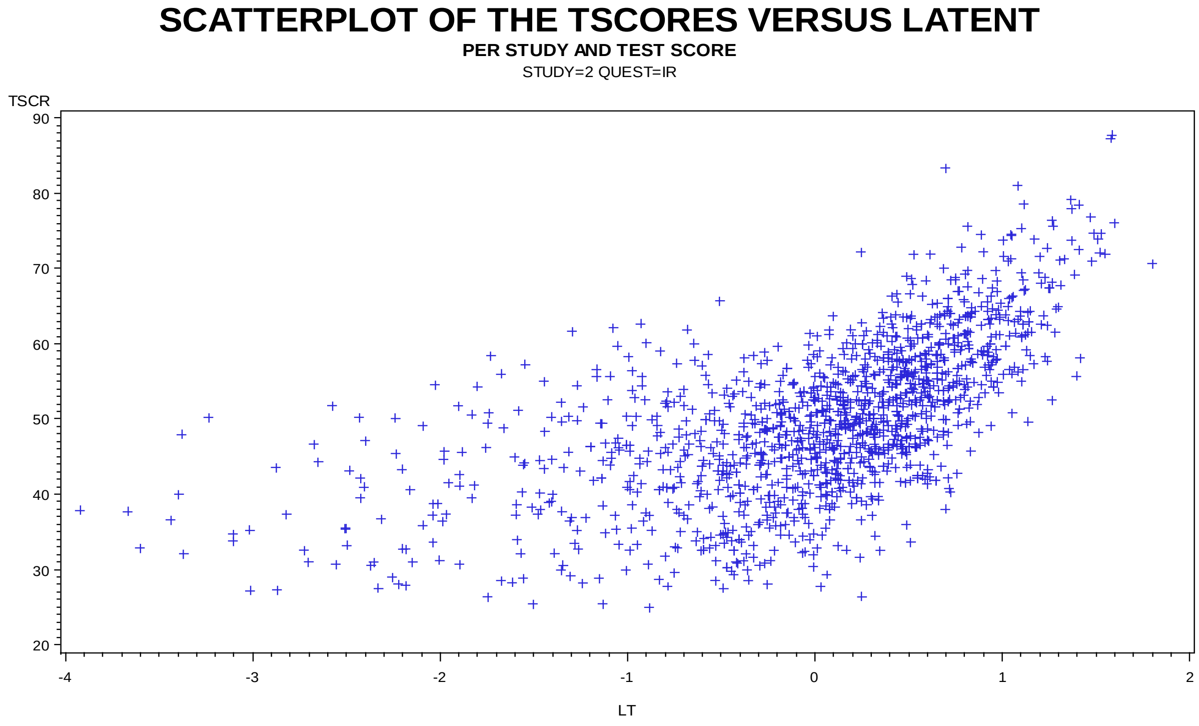


Figure 5. Scatterplot of the T-scores versus latent variable: study = CSHA, instrument = Buschke Total

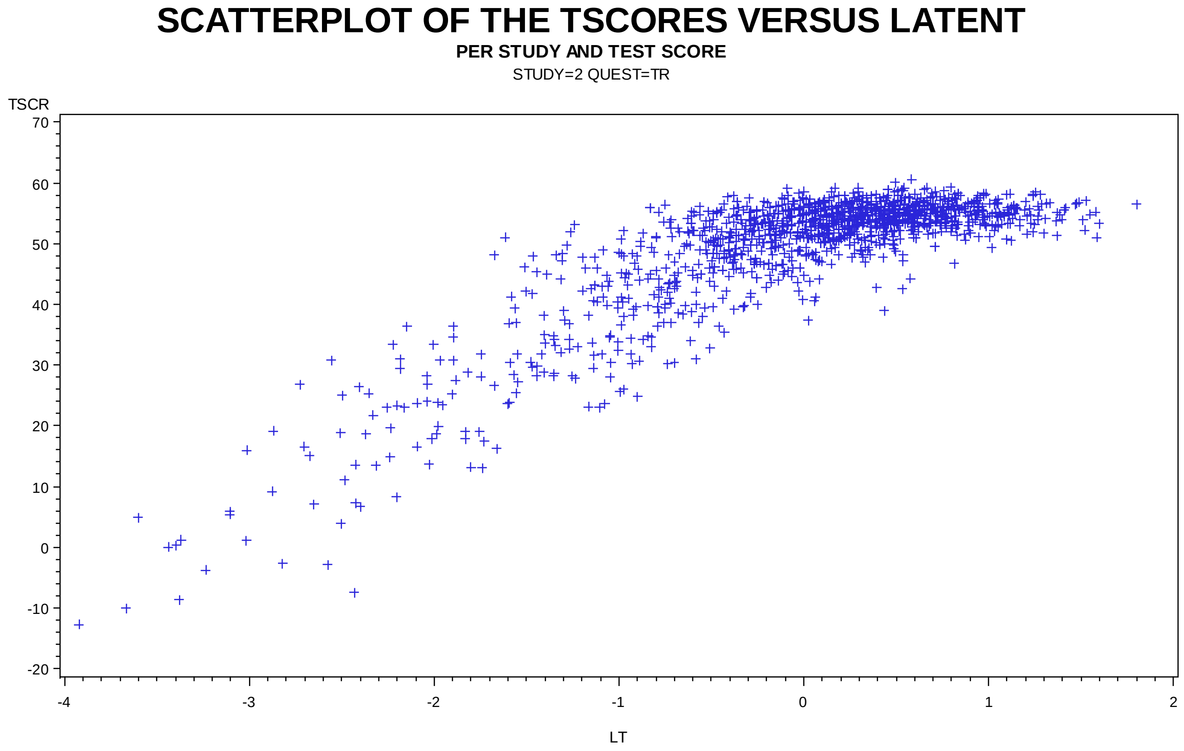


Figure 6. Scatterplot of the T-scores versus latent variable: study = NuAge, instrument = Buschke Free

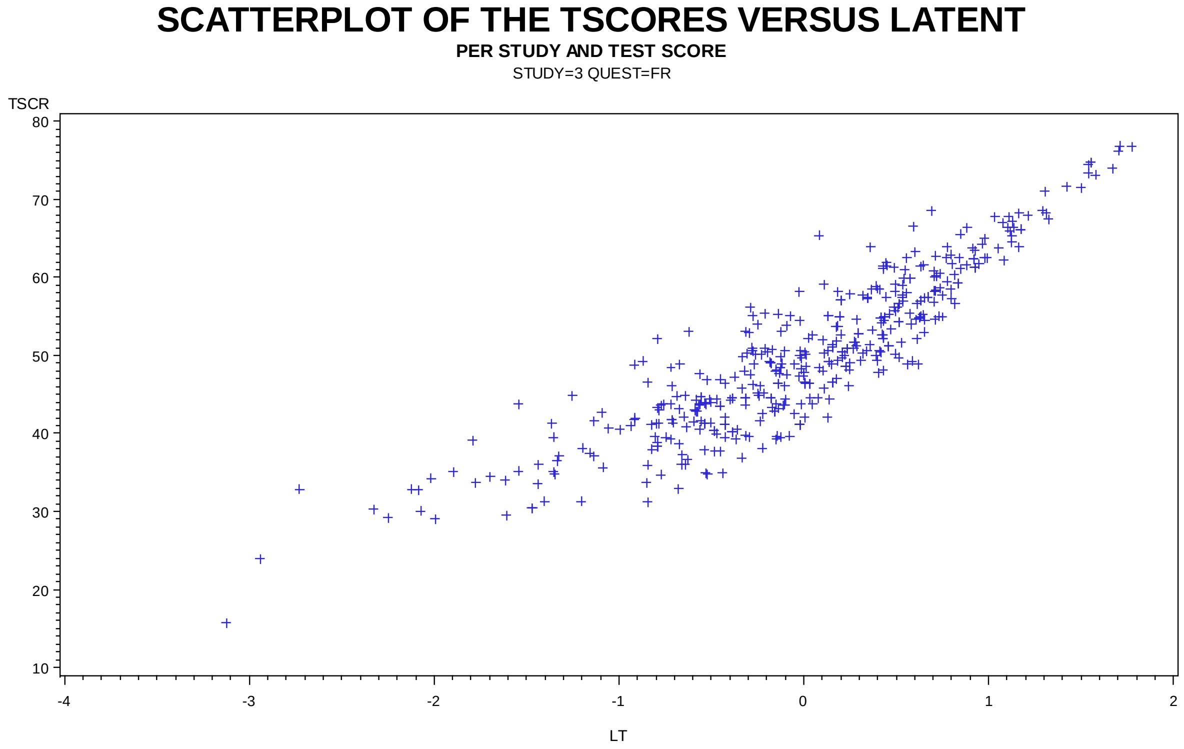


Figure 7. Scatterplot of the T-scores versus latent variable: study = NuAge, instrument = Buschke Total

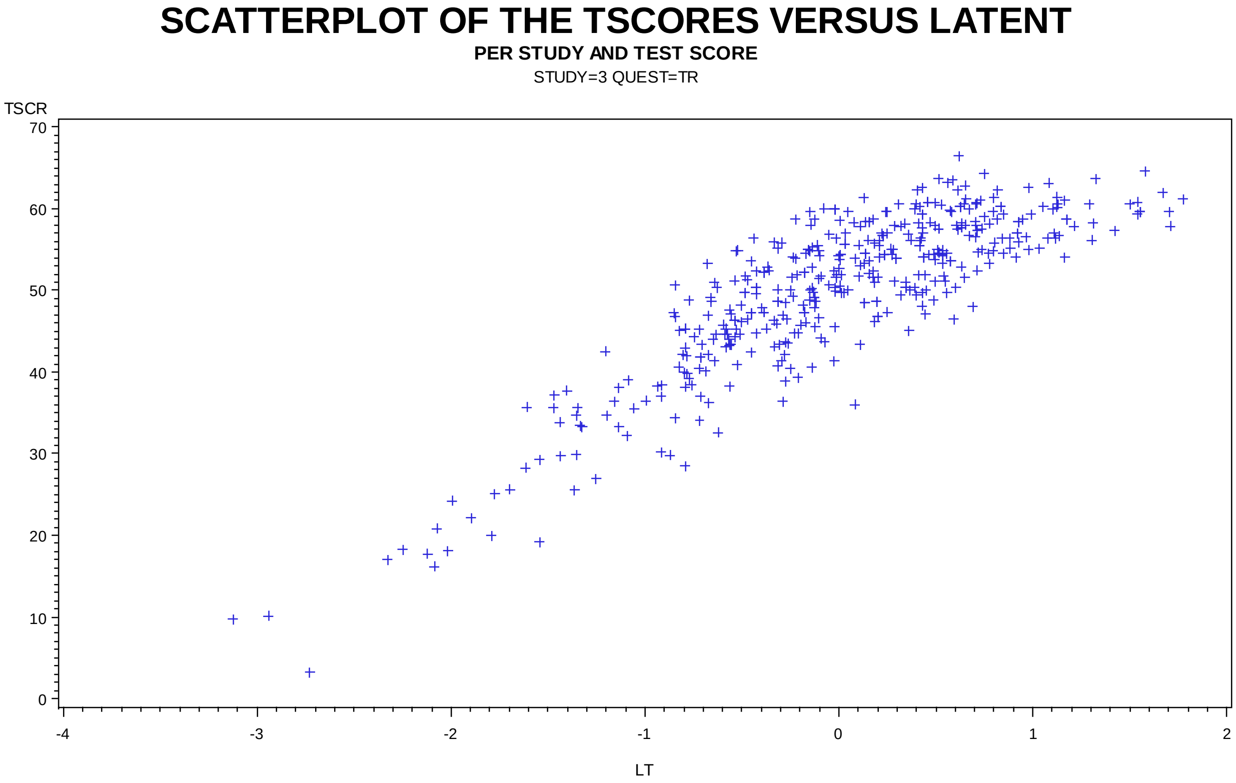
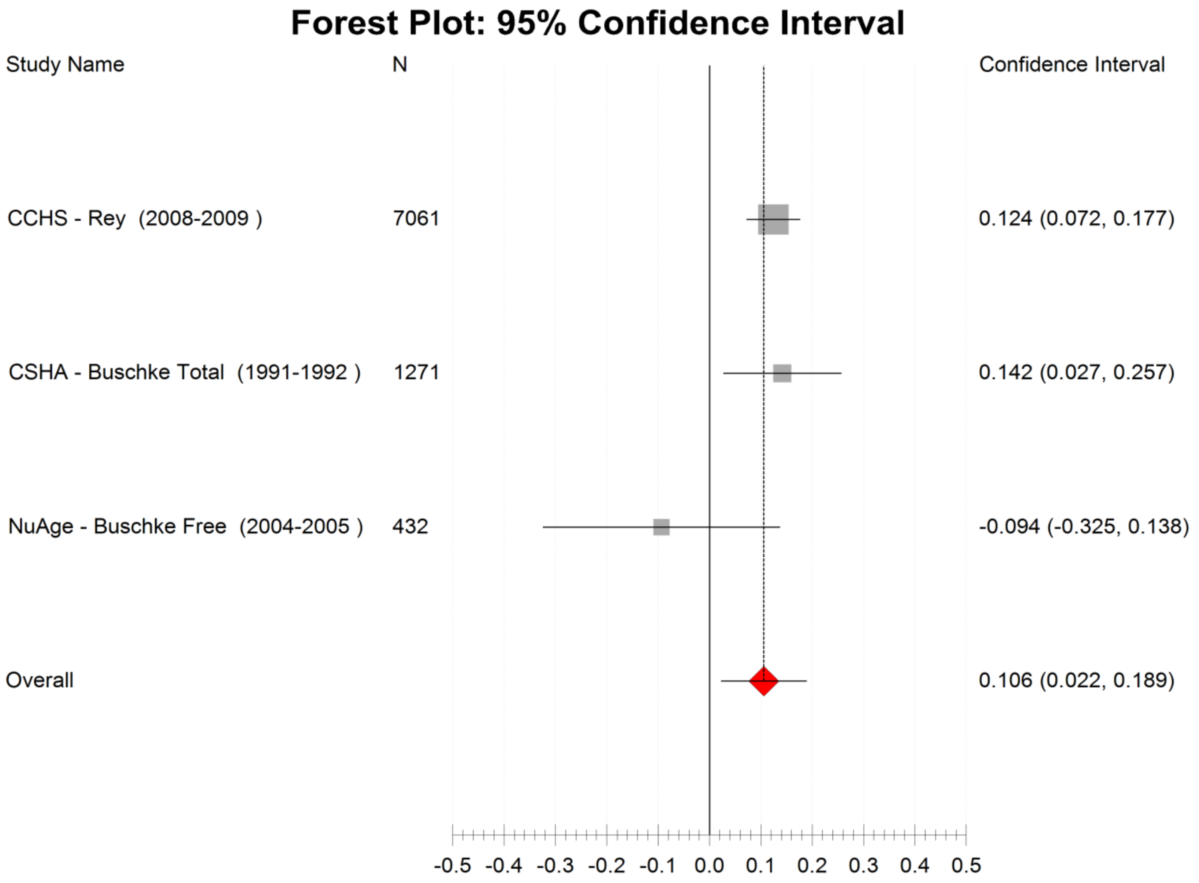


Figure 8. Forest plot representing the difference in the mean T-scores for cognition between high and low physical activity groups using the Rey in CCHS, Buschke Total in CSHA, and the Buschke Free in NuAge. The p-value$_{\text{Heterogeneity}}$=0.18 and the I^2 =42%



References

1. Oxman AD, Clarke MJ, Stewart LA. From science to practice—Metaanalyses using individual patient data are needed. JAMA. 1995;274(10):845-6. PMID:7650811.

2. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: Rationale, conduct, and reporting. BMJ. 2010;340:c221. PMID:20139215.

3. Slutsky J, Atkins D, Chang S, et al. AHRQ Series Paper 1: Comparing medical interventions: AHRQ and the Effective Health-Care Program. J Clin Epidemiol. 2010;63(5):481-3.

4. Balion C, Griffith L, Strifler L, et al. Vitamin D, cognition and dementia: A meta-analysis. Clin Biochem. 2011;1166:502 .

5. Annweiler C, Schott AM, Berrut G, et al. Vitamin D and ageing: neurological issues. Neuropsychobiol. 2010;62(3):139-50. PMID:20628264.

6. Barnard K, Colon-Emeric C. Extraskeletal effects of vitamin D in older adults: Cardiovascular disease, mortality, mood, and cognition. Am J Geriatr Pharmacother. 2010;8(1):4-33. PMID:20226390.

7. Fu R, Gartlehner G, Grant M, et al. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. J Clin Epidemiol. 2011;64(11):1187-97. PMID:21477993.

8. Granda P, Blasczyk E. Data harmonization. In: Guidelines for Best Practice in Cross-sectional Surveys, 2nd ed. 2010. XIII .

9. Fortier I, Doiron D, Burton P, et al. Invited commentary: Consolidating data harmonization--how to obtain quality and applicability? Am J Epidemiol. 2011;174(3):261-4. PMID:21749975.

10. Raina P, Santaguida P, Ismaila A, et al. Effectiveness of cholinesterase inhibitors and memantine for treating dementia: evidence review for a clinical practice guideline. Ann Intern Med. 2008;148(5):379-97. PMID:18316756.

11. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res. 1975;12(3):189-98. PMID:1202204.

12. Mohs RC, Rosen WG, Davis KL. The Alzheimer's disease assessment scale: An instrument for assessing treatment efficacy. Psychopharmacol Bull. 1983;19(3):448-50. PMID:6635122.

13. Wiener JM, Hanley RJ, Clark R, et al. Measuring the activities of daily living: Comparisons across national surveys. J Gerontol. 1990;45(6):S229-37. PMID:2146312.

14. Flanagan DP, Ortiz SO, Alfonso VC. Essentials of cross-battery assessment. 2nd ed. New York: Wiley; 2007.

15. Wechsler D. Administration and Scoring Manual for the Wechsler Adult Intelligence Scale. 3rd. San Antonio: The Psychological Corporation; 1997.

16. Royall DR, Chiodo LK, Polk MJ. Misclassification is likely in the assessment of mild cognitive impairment. Neuroepidemiology. 2004;23(4):185-91. PMID:15272221.

17. Roth M, Tym E, Mountjoy CQ, et al. CAMDEX. A standardised instrument for the diagnosis of mental disorder in the elderly with special reference to the early detection of dementia. Br J Psychiatry. 1986;149:698-709. PMID:3790869.

18. Teng EL, Chui HC. The Modified Mini-Mental State (3MS) examination. J Clin Psychiatry. 1987;48(8):314-8. PMID:3611032.

19. Wechsler D. Manual for the Wechsler Adult Intelligence Scale test—Revised. New York: Psychological Corporation; 1981.

20. Nelson HE, Wilson JR. The Revised National Adult Reading Test—Test manual. Windsor U.K: NFER-Nelson; 2012.

21. Raven JC. Mill Hill Vocabulary Scale. London; 1958.

22. Cattell RB. Theory of fluid and crystallized intelligence: A critical experiment. J Educ Psychol. 1963;54:1-22. .

23. Raven JC. Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 1: General Overview. Harcourt Assessment; 1998.

24. Thurstone LL. Psychological implications of factor analysis. Am Psychol. 1948;3(9):402-8. PMID:18880205.

25. Meredith W. Measurement invariance, factor analysis and factorial invariance. Psychometrika. 1993;(4):525-43.

26. Meredith W, Horn JL. The role of factorial invariance in modelling growth and change. In: Collins LM, Sayer AG, editors. New methods for the analysis of change, Washington, DC: American Psychological Association; 2001. p. 203-40.

27. Horn JL, McArdle JJ, Mason R. When is invariance not invarient: A practical scientist's look at the ethereal concept of factor invariance. South Psychol. 1983;(4):179-88.

28. Hofer SM, Horn JL, Eber HW. A robust five-factor structure of the 16PF: Strong evidence from independent rotation and confirmatory factorial invariance procedures. Pers Indiv Differ. 1997;(2):247-69.

29. Meredith W. Notes on factorial invariance. Psychometrika. 1964;(2):177-85. .

30. Muthen B. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika. 1984;(1):115-32.

31. Takane Y, de Leeuw J. On the relationship between item response theory and factor analysis of discretized variables. Psychometrika. 1987;(3):393-408.

32. McDonald RP. Test Theory: A Unified approach. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.; 1999.

33. Molenberghs G, Verbeke G. Models for discrete longitudinal data. New York: Springer; 2005.

34. Agresti A. Categorical data analysis. 2nd. Hoboken, New Jersey: John Wiley & Sons Inc.; 2002.

35. Agresti A. Analysis of ordinal categorical data. 2nd. Hoboken, New Jersey: John Wiley & Sons Inc.; 2010.

36. Martin M, Clare L, Altgassen AM, et al. Cognition-based interventions for healthy older people and people with mild cognitive impairment. Cochrane Database Syst Rev. 2011;1:CD006220. PMID:21249675.

37. Angevaren M, Aufdemkampe G, Verhaar HJ, et al. Physical activity and enhanced fitness to improve cognitive function in older people without known cognitive impairment. Cochrane Database Syst Rev. 2008;(2):CD005381. PMID:18425918.

38. Balint S, Czobor P, Komlosi S, et al. Attention deficit hyperactivity disorder (ADHD): Gender- and age-related differences in neurocognition. Psychol Med. 2009;39(8):1337-45. PMID:18713489.

39. Barth A, Winker R, Ponocny-Seliger E, et al. A meta-analysis for neurobehavioural effects due to electromagnetic field exposure emitted by GSM mobile phones. J Occup Environ Med. 2008;65(5):342-6. PMID:17928386.

40. Bhutta AT, Cleves MA, Casey PH, et al. Cognitive and behavioral outcomes of school-aged children who were born preterm: A meta-analysis. JAMA. 2002;288(6):728-37. PMID:12169077.

41. Bora E, Yucel M, Pantelis C. Cognitive endophenotypes of bipolar disorder: A meta-analysis of neuropsychological deficits in euthymic patients and their first-degree relatives. J Affect Disord. 2009;113(1-2):1-20. PMID:18684514.

42. Brands AMA, Biessels GJ, De Haan EHF, et al. The effects of type 1 diabetes on cognitive performance: A meta-analysis. Diabetes Care. 2005;28(3):726-35. PMID:15735218.

43. Campbell LK, Scaduto M, Sharp W, et al. A meta-analysis of the neurocognitive sequelae of treatment for childhood acute lymphocytic leukemia. Pediatr Blood Canc. 2007;49(1):65-73. PMID:16628558.

44. Eilander A, Gera T, Sachdev HS, et al. Multiple micronutrient supplementation for improving cognitive performance in children: Systematic review of randomized controlled trials. Am J Clin Nutr. 2010;91(1):115-30. PMID:19889823.

45. Falkingham M, Abdelhamid A, Curtis P, et al. The effects of oral iron supplementation on cognition in older children and adults: A systematic review and meta-analysis. Nutr J. 2010 Jan 25;9:4. doi: 10.1186/1475-2891-9-4.

46. Goodman M, LaVerda N, Clarke C, et al. Neurobehavioural testing in workers occupationally exposed to lead: Systematic review and meta-analysis of publications. J Occup Environ Med. 2002;59(4):217-23. PMID:11934948.

47. Grant I, Gonzalez R, Carey CL, et al. Non-acute (residual) neurocognitive effects of cannabis use: A meta-analytic study. J Int Neuropsychol Soc. 2003;9(5):679-89. PMID:12901774.

48. Guilera G, Pino O, Gomez-Benito J, et al. Antipsychotic effects on cognition in schizophrenia: A meta-analysis of randomised controlled trials. Eur J Psychiat. 2009;23(2):77-89.

49. Hogervorst E, Bandelow S. Sex steroids to maintain cognitive function in women after the menopause: A meta-analyses of treatment trials. Maturitas. 2010;66(1):56-71. PMID:20202765.

50. Hogervorst E, Yaffe K, Richards M, et al. Hormone replacement therapy to maintain cognitive function in women with dementia. Cochrane Database Syst Rev. 2009;(1):CD003799. PMID:19160224.

51. Li H, Li N, Li B, et al. Cognitive intervention for persons with mild cognitive impairment: A meta-analysis. Ageing Res Rev. 2011;10:285-96. PMID:21130185.

52. Jansen CE, Miaskowski C, Dodd M, et al. A metaanalysis of studies of the effects of cancer chemotherapy on various domains of cognitive function. Cancer. 2005;104(10):2222-33. PMID:16206292.

53. Karsdorp PA, Everaerd W, Kindt M, et al. Psychological and cognitive functioning in children and adolescents with congenital heart disease: A meta-analysis. J Pediatr Psychol. 2007;32(5):527-41. PMID:17182669.

54. Krabbendam L, Arts B, Van OJ, et al. Cognitive functioning in patients with schizophrenia and bipolar disorder: A quantitative review. Schizophr Res. 2005;80(2-3):137-49. PMID:16183257.

55. Lethaby A, Hogervorst E, Richards M, et al. Hormone replacement therapy for cognitive function in postmenopausal women. Cochrane Database Syst Rev. 2008;(1):CD003122. PMID:18254016.

56. Marasco SF, Sharwood LN, Abramson MJ. No improvement in neurocognitive outcomes after off-pump versus on-pump coronary revascularisation: A meta-analysis. Eur J Cardiothorac Surg. 2008;33(6):961-70. PMID:18424064.

57. Metternich B, Kosch D, Kriston L, et al. The effects of nonpharmacological interventions on subjective memory complaints: A systematic review and meta-analysis. Psychother Psychosom. 2010;79(1):6-19. PMID:19887887.

58. McDermott LM, Ebmeier KP. A meta-analysis of depression severity and cognitive function. J Affect Disord. 2009;119(1-3):1-8. PMID:19428120.

59. Naguib JM, Kulinskaya E, Lomax CL, et al. Neuro-cognitive performance in children with type 1 diabetes—a meta-analysis. J Pediatr Psychol. 2009;34(3):271-82. PMID:18635605.

60. Nieto RG, Xavier CF. A meta-analysis of neuropsychological functioning in patients with early onset schizophrenia and pediatric bipolar disorder. J Clin Child Adolesc Psychol. 2011;40(2):266-80. PMID:21391023.

61. Quinn TJ, Gallacher J, Deary IJ, et al. Association between circulating hemostatic measures and dementia or cognitive impairment: systematic review and meta-analyzes. J Thromb Haemost. 2011;9(8):1475-82. PMID:21676170.

62. Repantis D, Schlattmann P, Laisney O, et al. Modafinil and methylphenidate for neuroenhancement in healthy individuals: A systematic review. Pharmacol Res. 2010;62(3):187-206. PMID:20416377.

63. Sibley BA, Etnier JL. The relationship between physical activity and cognition in children: A meta-analysis. Pediatr Exerc Sci. 2003;15(3):243-56.

64. Valentini E, Ferrara M, Presaghi F, et al. Systematic review and meta-analysis of psychomotor effects of mobile phone electromagnetic fields. J Occup Environ Med. 2010;67(10):708-16. PMID:20837651.

65. Voss MW, Kramer AF, Basak C, et al. Are expert athletes 'expert' in the cognitive laboratory? A meta-analytic review of cognition and sport expertise. Appl Cognit Psychol. 2010;24(6):812-26.

66. Wheaton P, Mathias JL, Vink R. Impact of early pharmacological treatment on cognitive and behavioral outcome after traumatic brain injury in adults: A meta-analysis. J Clin Psychopharmacol. 2009;29(5):468-77. PMID:19745647.

67. Woodward ND, Purdon SE, Meltzer HY, et al. A meta-analysis of cognitive change with haloperidol in clinical trials of atypical antipsychotics: Dose effects and comparison to practice effects. Schizophr Res. 2007;89(1-3):211-24. PMID:17059880.

68. Zhang JP, Burdick KE, Lencz T, et al. Meta-analysis of genetic variation in DTNBP1 and general cognitive ability. Biol Psychiatry. 2010;68(12):1126-33. PMID:21130223.

69. Granda P, Wolf C, Hadorn R. Harmonizing survey data. In: Harkness JA, Braun M, Edwards B, et al, eds. Survey methods in multinational, multicultural and multiregional contexts. Hoboken, NJ: John Wiley & Sons; 2010: 315-32.

70. Bauer DJ, Hussong AM. Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. Psychol Methods. 2009;14(2):101-25. PMID:19485624.

71. Burns RA, Butterworth P, Kiely KM, et al. Multiple imputation was an efficient method for harmonizing the Mini-Mental State Examination with missing item-level data. J Clin Epidemiol. 2011;64(7):787-93. PMID:21292440.

72. Gorsuch R. New procedure for extension analysis in exploratory factor analysis. EPM. 1997;57(5):725-40.

73. McArdle JJ, Grimm KJ, Hamagami F, et al. Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. Psychol Methods. 2009;14(2):126-49. PMID:19485625.

74. Minicuci N, Noale M, Bardage C, et al. Cross-national determinants of quality of life from six longitudinal studies on aging: The CLESA project. Aging Clin Exp Res. 2003;15(3):187-202. PMID:14582681.

75. Gross AL, Inouye SK, Rebok GW, et al. Parallel but not equivalent: Challenges and solutions for repeated assessment of cognition over time. J Clin Exp Neuropsychol. 2012;34(7):758-72. PMID:22540849.

76. van Buuren S, Eyres S, Tennant A, et al. Improving comparability of existing data by response conversion. J Off Stat. 2005;21(1):53-72. PMID:14533743.

77. Blettner M, Sauerbrei W, Schlehofer B, et al. Traditional reviews, meta-analyses and pooled analyses in epidemiology. Int J Epidemiol. 1999;28(1):1-9. PMID:10195657.

78. Cooper H, Patall EA. The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. Psychol Methods. 2009;14(2):165-76. PMID:19485627.

79. Curran PJ, Hussong AM. Integrative data analysis: the simultaneous analysis of multiple data sets. Psychol Methods. 2009;14(2):81-100. PMID:19485623.

80. Ioannidis JPA, Rosenberg PS, Goedert JJ, et al. Commentary: Meta-analysis of individual participants' data in genetic epidemiology. Am J Epidemiol. 2002;156(3):204-10. PMID:12142254.

81. Schmid CH, Landa M, Jafar TH, et al. Constructing a database of individual clinical trials for longitudinal analysis. Control Clin Trials. 2003;24(3):324-40. PMID:12757997.

82. Simmonds MC, Higgins JP, Stewart LA, et al. Meta-analysis of individual patient data from randomized trials: A review of methods used in practice. Clin Trials. 2005;2(3):209-17. PMID:16279144.

83. van der Steen JT, Kruse RL, Szafara KL, et al. Benefits and pitfalls of pooling datasets from comparable observational studies: Combining US and Dutch nursing home studies. Palliat Med. 2008;22(6):750-9. PMID:18715975.

84. van WC. Individual patient meta-analysis—rewards and challenges. J Clin Epidemiol. 2010;63(3):235-7. PMID:19595573.

85. Bennett DA. Review of analytical methods for prospective cohort studies using time to event data: Single studies and implications for meta-analysis. Stat Methods Med Res. 2003;12(4):297-319. PMID:12939098.

86. Friedenreich CM. Methods for pooled analyses of epidemiologic studies. Epidemiol. 1993;4(4):295-302. PMID:8347739.

87. Hofer SM, Piccinin AM. Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. Psychol Methods. 2009;14(2):150-64. PMID:19485626.

88. Jones AP, Riley RD, Williamson PR, et al. Meta-analysis of individual patient data versus aggregate data from longitudinal clinical trials. Clin Trials. 2009;6(1):16-27. PMID:19254930.

89. Mathew T, Nordstrom K. Comparison of one-step and two-step meta-analysis models using individual patient data. Biom J. 2010;52(2):271-87. PMID:20349448.

90. Burgess S, Seaman S, Lawlor DA, et al. Missing data methods in Mendelian randomization studies with multiple instruments. Am J Epidemiol. 2011;174(9):1069-76. PMID:21965185.

91. Donegan S, Williamson P, Gamble C, et al. Indirect comparisons: A review of reporting and methodological quality. Plos One. 2010;5(11):e11054. PMID:21085712.

92. Peyre H, Leplege A, Coste J. Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. Qual Life Res. 2011;20(2):287-300. PMID:20882358.

93. Siddique J, Crespi CM, Gibbons RD, et al. Using latent variable modeling and multiple imputation to calibrate rater bias in diagnosis assessment. Stat Med. 2011;30(2):160-74. PMID:21204122.

94. Spratt M, Carpenter J, Sterne JA, et al. Strategies for multiple imputation in longitudinal studies. Am J Epidemiol. 2010;172(4):478-87. PMID:20616200.

95. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. BMJ. 2009;338:b2393. PMID:19564179.

96. Crane PK, Narasimhalu K, Gibbons LE, et al. Composite scores for executive function items: demographic heterogeneity and relationships with quantitative magnetic resonance imaging. J Int Neuropsychol Soc. 2008;14(5):746-59. PMID:18764970.

97. Teresi JA, Ocepek-Welikson K, Kleinman M, et al. Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): Applications (with illustrations) to measures of physical functioning ability and general distress. Qual Life Res. 2007;16(Suppl 1):43-68. PMID:17484039.

98. Anstey KJ, Byles JE, Luszcz MA, et al. Cohort profile: The Dynamic Analyses to Optimize Ageing (DYNOPTA) project. Int J Epidemiol. 2010;39(1):44-51. PMID:19151373.

99. Bath P. The harmonisation of longitudinal data: A case study using data from cohort studies in The Netherlands and the United Kingdom. Ageing Soc. 2010;30:1419-37.

100. Beer-Borst S, Morabia A, Hercberg S, et al. Obesity and other health determinants across Europe: the EURALIM project. J Epidemiol Community Health. 2000;54(6):424-30. PMID:10818117.

101. Beer-Borst S, Hercberg S, Morabia A, et al. Dietary patterns in six european populations: Results from EURALIM, a collaborative European data harmonization and information campaign. Eur J Clin Nutr. 2000;54(3):253-62. PMID:10713749.

102. Crane PK, Narasimhalu K, Gibbons LE, et al. Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. J Clin Epidemiol. 2008;61(10):1018-27. PMID:18455909.

103. Curran PJ, Hussong AM, Cai L, et al. Pooling data from multiple longitudinal studies: the role of item response theory in integrative data analysis. Dev Psychol. 2008;44(2):365-80. PMID:18331129.

104. Darby S, Hill D, Deo H, et al. Residential radon and lung cancer—detailed results of a collaborative analysis of individual data on 7148 persons with lung cancer and 14 208 persons without lung cancer from 13 epidemiologic studies in Europe. Scand J Psychol. 2006;32 Suppl(1):1-83. PMID:16538937.

105. Fibrinogen Studies Collaboration. Correcting for multivariate measurement error by regression calibration in meta-analyses of epidemiological studies. Stat Med. 2009;28(7):1067-92. PMID:1922208.

106. Grimm KJ, Steele JS, Mashburn AJ, et al. Early behavioral associations of achievement trajectories. Dev Psychol. 2010;46(5):976-83. PMID:20822216.

107. Duncan GJ, Dowsett CJ, Claessens A, et al. School readiness and later achievement. Dev Psychol. 2007;43(6):1428-46. PMID:18020822.

108. Khachaturian AS, Sabbagh M. Commentary on "Developing a national strategy to prevent dementia: Leon Thal Symposium 2009." Creating a national database for successful aging. Alzheimers Dement. 2010;6(2):132-4. PMID:20298973.

109. McArdle JJ, Prescott CA, Hamagami F, et al. A contemporary method for developmental-genetic analyses of age changes in intellectual abilities. Dev Neuropsychol. 1998;14(1):69-114.

110. McArdle JJ, Nesselroade JR. Using multivariate data to structure developmental change. In: Cohen SH, Reese HW, eds. Life-span developmental psychology: Methodological contributions, Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc; 1994: 223-67.

111. McArdle JJ, Prescott CA. Contemporary models for the biometric genetic analysis of intellectual abilities. In: Flanagan DP, Genshaft JL, Harrison PL, eds. Contemporary intellectual assessment: Theories, tests, and issues, New York, NY, US: Guilford Press; 1997: 403-36.

112. Minicuci N, Noale M, Leon Diaz EM, et al. Disability-free life expectancy: A cross-national comparison among Bulgarian, Italian, and Latin American older population. J Aging Health. 2011;23(4):629-81. PMID:21220352.

113. Pluijm SM, Bardage C, Nikula S, et al. A harmonized measure of activities of daily living was a reliable and valid instrument for comparing disability in older people across countries. J Clin Epidemiol. 2005;58(10):1015-23. PMID:16168347.

114. Ruggles S, King ML, Levison D, et al. IPUMS-International. Hist Meth. 2003;36:60-5.

115. Esteve A, Sobek M. Challenges and methods of international census harmonization. Hist Meth. 2003;36:66-79.

116. Schenker N, Raghunathan TE. Combining information from multiple surveys to enhance estimation of measures of health. Stat Med. 2007;26(8):1802-11. PMID:17278184.

117. Slimani N, Kaaks R, Ferrari P, et al. European Prospective Investigation into Cancer and Nutrition (EPIC) calibration study: Rationale, design and population characteristics. Public Health Nutr. 2002;5(6B):1125-45. PMID:12639223.

118. van Buuren S, Eyres S, Tennant A, et al. Assessing comparability of dressing disability in different countries by response conversion. Eur J Public Health. 2003;13(3 Suppl):15-9. PMID:14533743.

119. Hopman-Rock M, van BS, De Kleijn-De VM. Polytomous Rasch analysis as a tool for revision of the severity of disability code of the ICIDH. Disabil Rehabil. 2000;22(8):363-71. PMID:10896097.

120. Dorans NJ, Moses TP, Eignor DR. Principles and Practices of Test Score Equating. ETS RR-10-29. Princeton, New Jersey: Educational Testing Service; 2010.

121. Griffith LE, Shannon HS, Wells RP, et al. Individual participant data meta-analysis of mechanical workplace risk factors and low back pain. Am J Public Health. 2012;102(2):309-18. PMID:22390445.

122. Canadian Study of Health and Aging Working Group. Canadian study of health and aging: study methods and prevalence of dementia. CMAJ. 1994;150(6):899-913. PMID:8131123.

123. Tuokko H, Kristjansson E, Miller J. Neuropsychological detection of dementia: An overview of the neuropsychological component of the Canadian Study of Health and Aging. J Clin Exp Neuropsychol. 1995;17(3):352-73. PMID:7650099.

124. Taylor EM. The appraisal of children with cerebral deficits. Cambridge: Harvard University Press; 1959.

125. Buschke H. Cued recall in amnesia. J Clin Neuropsychol. 1984;6(4):433-40. PMID:6501581.

126. Spreen O, Benton A. Neurosensory centre comprehensive examination for aphasia. Victoria, B.C.: University of Victoria; 1977.

127. Spreen O, Strauss E. Compendium of neuropsychological tests: Administration, norms, and commentary. 2nd ed. New York: Oxford University Press; 1998.

128. Wechsler D, Stone CP. Wechsler memory scale. New York: Psychological Corporation; 1974.

129. Benton A. Revised visual retention test: Clinical and experimental applications. 3rd ed. New York: Psychological Corporation; 1974.

130. Tuokko H, Woodward TS. Development and validation of a demographic correction system for neuropsychological measures used in the Canadian Study of Health and Aging. J Clin Exp Neuropsychol. 1996;18(4):479-616. PMID:8877629.

131. Van der Linden M, Adam S, Baisset-Mouly C, et al. L'evaluation de troubles de la memoire: presentation de quatre tests de memoire episodique (avec etalonnage). Marseille: Solal; 2004.

132. Belleville S, Chatelois J, Fontaine F, et al. Memoria: Batterie informatisee d'evaluation de la memoire pour Mac et PC. Montrea: Institut Universitaire de Geriatire de Montreal.; 2002.

133. Bherer L, Belleville S, Peretz I. Education, age, and the Brown-Peterson technique. Dev Neuropsychol. 2001;19(3):237-51. PMID:11758667.

134. Rey A. Test de copie d'une figure complexe: Manuel. Paris: Les editions de centre de psychologie appliquee; 1959.

135. Regard,M. Cognitive rigidity and flexibility: A neuropsychological study. University of Victoria, Canada. 1981.

136. Statistics Canada. Canadian Community Health Survey (CCHS)—Healthy aging 2009. http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5146&lang=en&db=imdb&adm=8&dis=2. .

137. McComb E, Tuokko H, Brewster P, et al. Mental alternation test: Administration mode, age, and practice effects. J Clin Exp Neuropsychol. 2011;33(2):234-41. PMID:20865619.

138. Read DE. Neuropsychological assessment of memory in the elderly. Can J Psychol. 1987;41(2):158-74. PMID:3502894.

139. Findlay L, Bernier J, Tuokko H, et al. Validation of cognitive functioning categories in the Canadian Community Health Survey-Healthy Aging. Health Rep. 2010;21(4):1-16. PMID:21269015.

140. Butler M, Retzlaff P, Vanderploeg R. Neuropsychological test usage. Prof Psychol. 1991;14:60-76.

141. Schmidt M. Auditory verbal learning test: A handbook. Western Psychological Services; 1996.

142. Lezak MD, Howlesonn DB, Loring DW. Neuropsychological assessment. 4th ed. New York: Oxford University Press; 2004..

143. Mackler K. Test review of the Rey Auditory Verbal Learning Test: A handbook. In: Plake BS, Impara JC, editors. The fourteenth mental measurements yearbook, Lincoln, Nebraska: Buros Institute; 2003.

144. Backman L, Jones S, Berger AK, et al. Cognitive impairment in preclinical Alzheimer's disease: A meta-analysis. Neuropsychology. 2005;19(4):520-31. PMID:16060827.

145. Tierney MC, Yao C, Kiss A. Neuropsychological tests accurately predict incident Alzheimer disease after 5 and 10 years. Neurol. 2005;64(11):1853-59. PMID:15955933.

146. Petersen RC, Smith GE, Waring SC, et al. Aging, memory, and mild cognitive impairment. Int Psychogeriatr. 1997;9(Suppl 1):65-9. PMID:9447429.

147. Schoenberg MR, Dawson KA, Duff K, et al. Test performance and classification statistics for the Rey Auditory Verbal Learning Test in selected clinical samples. Arch Clin Neuropsychol. 2006;21(7):693-703. PMID:16987634.

148. Feeny D, Furlong W, Boyle M, et al. Multi-attribute health status classification systems. Health Utilities Index 5. Pharmacoeconom. 1995;7(6):490-502. PMID:10155335.

149. Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system 7. Med Care. 2002;40(2):113-28. PMID:11802084.

150. Feng Y, Bernier J, McIntosh C, et al. Validation of disability categories derived from Health Utilities Index Mark 3 scores 1. Health Rep. 2009;20(2):43-50. PMID:19728585.

151. Horsman J, Furlong W, Feeny D, et al. The Health Utilities Index (HUI): Concepts, measurement properties and applications. Health Qual Life Outcomes. 2003;1(54): PMID:14613568.

152. Kavirajan H, Hays RD, Vassar S, et al. Responsiveness and construct validity of the health utilities index in patients with dementia 3. Med Care. 2009;47(6):651-61. PMID:19433998.

153. Mulhern RK. Correlation of the Health Utilities Index Mark 2 cognition scale and neuropsychological functioning among survivors of childhood medulloblastoma. Int J Cancer Suppl. 1999;12:91-4. PMID:10679878.

154. Grober E, Buschke H. Genuine memory deficits in dementia. Dev Neuropsychol. 1987;3:13-36.

155. Macht ML, Buschke H. Age differences in cognitive effort in recall. J Gerontol. 1983;38(6):695-700. PMID:6630904.

156. Tuokko H, Crockett D. Cued recall and memory disorders in dementia. J Clin Exp Neuropsychol. 1989;11(2):278-94. PMID:2925836.

157. O'Connell ME, Tuokko H. The 12-item Buschke memory test: Appropriate for use across levels of impairment. Appl Neuropsychol. 2002;9(4):226-33. PMID:12584076.

158. Sarazin M, Berr C, De RJ, et al. Amnestic syndrome of the medial temporal type identifies prodromal AD: A longitudinal study. Neurol. 2007;69(19):1859-67. PMID:17984454.

159. Carlesimo GA, Perri R, Caltagirone C. Category cued recall following controlled encoding as a neuropsychological tool in the diagnosis of Alzheimer's disease: A review of the evidence. Neuropsychol Rev. 2011;21(1):54-65. PMID:21086049.

160. Coley N, Andrieu S, Gardette V, et al. Dementia prevention: Methodological explanations for inconsistent results. Epidemiol Rev. 2008;30:35-66. PMID:18779228.

161. Fortier I, Burton PR, Robson PJ, et al. Quality, quantity and harmony: The DataSHaPER approach to integrating data across bioclinical studies. Int J Epidemiol. 2010;39(5):1383-93. PMID:20813861.

162. Fortier I, Doiron D, Little J, et al. Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. Int J Epidemiol. 2011;40(5):1314-28. PMID:21804097.

163. DerSimonian R, Laird N. Metaanalysis in clinical trials. Control Clin Trials. 1986;7(3):177-88. PMID:3802833.

164. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: An update. Contemp Clin Trials. 2007;28(2):105-14. PMID:16807131.

165. Wallace BC, Schmid CH, Lau J, et al. Meta-Analyst: Software for meta-analysis of binary, continuous and diagnostic data. BMC Med Res Methodol. 2009;9:80. PMID:19961608.

166. Fleiss JL. Statistical methods for rates and proportions. 2nd. New York: John Wiley & Sons; 1981.

167. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539-58. PMID:12111919.

168. Deeks JJ, Higgins JPT, Altman DG. Analyzing and presenting results. In: Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions 4.2.5, Chichester, U.K.: Wiley; 2005. Ch.8 .

169. Williamson P, Altman D, Blazeby J, et al. Driving up the quality and relevance of research through the use of agreed core outcomes. J Health Serv Res Pol. 2012;17(1):1-2. PMID:22294719.

170. Steenhuis RE, Ostbye T. Neuropsychological test performance of specific diagnostic groups in the Canadian Study of Health and Aging (CSHA). J Clin Exp Neuropsychol. 1995;17(5):773-85. PMID:8557817.

171. Bartels C, Wegrzyn M, Wiedl A, et al. Practice effects in healthy adults: A longitudinal study on frequent repetitive cognitive testing. BMC Neurosci. 2010;11:118. PMID:20846444.

172. Duff K. Evidence-based indicators of neuropsychological change in the individual patient: Relevant concepts and methods. Arch Clin Neuropsychol. 2012;27(3):248-61. PMID:22382384.

173. Salthouse TA. Influence of age on practice effects in longitudinal neurocognitive change. Neuropsychology. 2010;24(5):563-72. PMID:20804244.

174. Beglinger LJ, Gaydos B, Tangphao-Daniels O, et al. Practice effects and the use of alternate forms in serial neuropsychological testing. Arch Clin Neuropsychol. 2005;20(4):517-29. PMID:15896564.

175. Niero M, Martin M, Finger T, et al. A new approach to multicultural item generation in the development of two obesity-specific measures: the Obesity and Weight Loss Quality of Life (OWLQOL) questionnaire and the Weight-Related Symptom Measure (WRSM). Clin Ther. 2002;24(4):690-700. PMID:12017412.

176. Serra-Majem L, MacLean D, Ribas L, et al. Comparative analysis of nutrition data from national, household, and individual levels: Results from a WHO-CINDI collaborative project in Canada, Finland, Poland, and Spain. J Epidemiol Community Health. 2003;57(1):74-80. PMID:12490653.

177. Fox JP, Glas CAW. Bayesian estimation of a multilevel IRT model using Gibbs sampling. [References]. Psychometrika. 2001;(2):271-88.

Glossary of Statistical Terms

Attenuation—where observations on bivariate material are subject to errors of measurement the true correlation between the variates will be obscured, usually being underestimated. The correlation is then said to be attenuated.

Binomial distribution—is the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success with probability *p*.

Cocalibrate—consists of the simultaneous analysis of the responses of persons in a single sample to items in more than 2 instruments.

Cohen's d—Cohen's d is defined as the difference between two means divided by a standard deviation for the data.

Configural invariance—configural invariance is achieved if the model of interest fits across the groups. Although the model is the same across groups, the unknown parameters of the model are assumed to be different across the groups.

Confounding factors—in studies there are often unsuspected systematic differences in the way groups were treated in addition to the intended treatment conditions. Statisticians describe systematic differences of this sort as confounding factors or confounding variables.

Continuous variable—a variable that can take on any number of possible values. Practically speaking, when a variable can take on at least 10 values, it can be treated as a continuous variable. For example, it can be plotted on a scatterplot and certain meaningful calculations can be made using the variable.

Convergence—this describes whether the maximum-likehood algorithm has converged or not.

Covariates—variables not controlled for in the experiment that still affect the dependent variable.

C-score—C-scores are similar to ratios in that they both are measures of relative size. C-scores are calculated as the difference between the Z-score of a single measurement for a given individual and the mean Z-score of that individual for all the measurements used in the analysis.

Differential treatment effects—are indicated by the regression lines of the group crossing and having opposite signs. If only one of these criteria is met only a partially differential effect can be assumed.

Dummy variables—a dummy variable (also known as an indicator variable) is one that takes the values 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome.

Extension analysis—determining the relationship of common factors to variables that were not included in the factor analysis.

Factorial invariance—factorial invariance is present if item responses are associated with the same constructs and if the factor parameter estimates are not significantly different.

Generalized linear mixed model—a model for linear and nonlinear effects of continuous and categorical predictor variables on a discrete or continuous but not necessarily normally distributed dependent (outcome) variable.

Goodness of fit—assessment of the agreement of the data with either a hypothesized pattern (e.g., independence of row and column factors in a contengency table or the form of a regression relationship) or a hypothesized distribution (e.g., comparing a histogram with expected frequencies from the normal distribution).

Harmonization—procedures aimed at achieving and improving the comparability of different surveys.

Heterogeneity—in meta-analysis refers to the variation in study outcomes between studies.

Inferential equivalence—the potential for harmonization of selected information from individual studies.

Item Response Theory—provides a model-based linkage between item responses and the latent characteristics measured by a test of scale.

Latent variable—a variable which is unobservable but is supposed to enter into the structure of a system under study, such as demand in economics or the “general” factor in psychology. Unobservable quantities such as errors are not usually described as latent.

Least square means—a method of fitting a straight line or curve based on minimization of the sum of squared differences (residuals) between the predicted and the observed points.

Linear structural equations modeling—a statistical technique for testing and estimating causal relations using a combination of statistical data and qualitative causal assumptions.

Likelihood ratio test—a general purpose test of hypothesis Ho against an alternative H1 based on the ratio of two likelihood functions, one derived from each of Ho and H1. The statistics l is given by l = -2 ln (LH0 / LH1) and? has approximately a C2 distribution with df equal to the difference in the number of parameters in the two hypotheses.

Linear or z-tranformations—a change to a variable characterized by one or more of the following operations: adding a constant to the variable, subtracting a constant from the variable, multiplying the variable by a constant, and/or dividing the variable by a constant.

Measurement invariance—attempts to verify that the factors are measuring the same underlying latent construct within each group.

Meta-regression—an extension to subgroup analyses that allows the effect of continuous, as well as categorical, characteristics to be investigated, and allows the effects of multiple factors to be investigated simultaneously.

Multiple imputation—missing values for any variable are predicted using existing values from other variables. The predicted values, called "imputes", are substituted for the missing values, resulting in a full data set called an "imputed data set."

Nonlinear factor analysis—maximizes shared information among nonoverlapping sources, producing higher-order features of the data which uncover hidden causal factors controlling the observed phenomena.

Point estimate—endeavours to give the best single estimated value of a parameter, as compared with interval estimation, which proceeds by specifying a range of values. Since the point estimate is surrounded by a band of error, the distinction between two methods is sometimes blurred and in interpretation they often amount to the same thing.

Polytomous Rasch model—a generalization of the dichotomous model which can be applied in contexts in which successive integer scores represent categories of increasing level or magnitude of a latent trait, such as increasing ability, motor function, endorsement of a statement, etc.

Repeated measures—in this design, the same experimental unit is subjected to the different treatments under consideration at different points in time. Each unit, therefore, serves as a block. If for example, two different treatments and placebo treatment are applied to the same patient sequentially, this is a repeated measures design.

Sensitivity analysis—a "what-if" type of analysis to determine the sensitivity of the outcomes to changes in parameters. If a small change in a parameter results in relatively large changes in the outcomes, the outcomes are said to be sensitive to that parameter.

Stratum—when data are stratified according to its characteristics, each subgroup is a stratum.

T-score—a ratio of the departure of an estimated parameter from its notional value and its standard error.

Z-score—the Z score or value expresses the number of standard errors by which a sample mean lies above or below the true population mean. The Z-statistic is defined as difference of sample proportions divided by standard error of difference of sample proportions.

Appendix A. Excluded Studies

Baujat B, Mahe C, Pignon JP, et al. A graphical method for exploring heterogeneity in meta-analyses: application to a meta-analysis of 65 trials. Stat Med. 2002;21(18):2641-52. PMID:12228882

Cook KF, Teal CR, Bjorner JB, et al. IRT health outcomes data analysis project: an overview and summary. Qual Life Res. 2007;16 Suppl 1:121-32. PMID:17351824

Kaplan RF, Trevino RP, Johnson GM, et al. Cognitive function in post-treatment Lyme disease: do additional antibiotics help? Neurology. 2003;60(12):1916-22. PMID:12821733

Koopman L, van der Heijden GJ, Glasziou PP, et al. A systematic review of analytical methods used to study subgroups in (individual patient data) meta-analyses. J Clin Epidemiol. 2007;60(10):1002-9. PMID:17884593

Koopman L, van der Heijden GJ, Hoes AW, et al. Empirical comparison of subgroup effects in conventional and individual patient data meta-analyses. Int J Technol Assess Health Care. 2008;24(3):358-61. PMID:18601805

Marston L, Carpenter JR, Walters KR, et al. Issues in multiple imputation of missing data for large general practice clinical databases. Pharmacoepidemiol Drug Saf. 2010;19(6):618-26. PMID:20306452

Riley RD, Simmonds MC, Look MP. Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. J Clin Epidemiol. 2007;60(5):431-9. PMID:17419953

Riley RD. Commentary: like it and lump it? Meta-analysis using individual participant data. Int J Epidemiol. 2010;39(5):1359-61. PMID:20660642

Thompson S, Kaptoge S, White I, et al. Statistical methods for the time-to-event analysis of individual participant data from multiple epidemiological studies. Int J Epidemiol. 2010;39(5):1345-59. PMID:20439481

Zucker DR, Ruthazer R, Schmid CH. Individual (N-of-1) trials can be combined to give population comparative treatment effect estimates: methodologic considerations. J Clin Epidemiol. 2010;63(12):1312-23. PMID:20863658

Appendix B. Variable Description and Categories for the Studies CSHA, CCHS-CLSA, and NuAge

Candidate Variable	Variable Description	Categories	CSHA Variable Format	CCHS-CLSA Variable Format	NuAge Variable Format
Age	Participant's age at recruitment self-reported by the participant.		1. Birthday Day__ Month__ Year__	ANDB_Q01 What is [respondent name]'s age? ___Age in years (MIN:0) (MAX:130) (DK, RF are not allowed)	When were you born? Date: year___ month___ day__
Sex	Gender of the participant.	Male Female	2. Sex Male_ Female_	SEX_Q01 Interviewer: Enter [respondent name]'s sex. If necessary, ask: (Is [respondent name] male or female?) 1. Male 2. Female (DK, RF are not allowed)	Socio-demographic data Response: Male__ Female__
Highest Level of Education	Highest level of education completed by the participant.	None Primary High School Post-Secondary Prefer Not to Answer Don't Know	5. How many years of education did you complete? EDUYEAR Years 88 DK So that means that you (completed primary school, completed part of high school, all of high school some university)? (Select a suitable category) 1 No formal schooling 2 Some primary school 3 Finished primary school 4 Some secondary or high school 5 Completed secondary or high school 6 Some community or technical college, CEGEP, or nursing program 7 Completed community college, technical college, CEGEP, or nursing program 8 Some University 9 Bachelor's degree 10 Master's degree 11 PhD 12 Other 88 DK 99 Didn't Ask EDULEVEL	ED_Q01 What is the highest grade of elementary or high school [respondent name] ever EDU_1 completed? 1 Grade 8 or lower (Québec: Secondary II or lower) (Go to EDU_Q03) 2 Grade 9 – 10 (Québec: Secondary III or IV, Newfoundland and Labrador: 1st year of secondary) (Go to EDU_Q03) 3 Grade 11 – 13 (Québec: Secondary V, Newfoundland and Labrador: 2nd to 4th year of secondary) DK, RF (Go to EDU_Q03) ED_Q02 Did [respondent name] graduate from high school (secondary school)? EDU_2 1 Yes 2 No DK, RF ED_Q03 Has [respondent name] received any other education that could be counted towards EDU_3 a degree, certificate or diploma	Impossible

Appendix B. Variable Description and Categories for the studies CSHA, CCHS-CLSA, and NuAge (cont'd)

Candidate Variable	Variable Description	Categories	CSHA Variable Format	CCHS-CLSA Variable Format	NuAge Variable Format
Highest Level of Education (cont'd)				from an educational institution? 1 Yes 2 No (Go to EDU_END) DK, RF (Go to EDU_END) ED_Q04 What is the highest degree, certificate or diploma [respondent name] has obtained? EDU_4 01 No post-secondary degree, certificate or diploma 02 Trade certificate or diploma from a vocational school or apprenticeship training 03 Non-university certificate or diploma from a community college, CEGEP, school of nursing, etc. 04 University certificate below bachelor's level 05 Bachelor's degree 06 University degree or certificate above bachelor's degree DK, RF	
Number of Years Education	Number of years of education		5. How many years of education did you complete? EDUYEAR Years 88 DK	Impossible	35 SCOLART1 Nb années de scolarité T1 How many years of education have you completed:_______

Appendix B. Variable Description and Categories for the studies CSHA, CCHS-CLSA, and NuAge (cont'd)

Candidate Variable	Variable Description	Categories	CSHA Variable Format	CCHS-CLSA Variable Format	NuAge Variable Format
Household Income	Average current annual income, before taxes, of the participant's entire household.		Partial	INC_Q03A IN2_03A What is your best estimate of the total household income received by all household members, from all sources, before taxes and deductions, in the past 12 months? |_|_|_|_|_|_|_| Income (MIN: 0) (MAX: 500,000) DK, RF (Go to INC_D03B) INC_Q03B IN2_03B INTERVIEWER: Read categories to respondent. What is your best estimate of the total household income received by all household members, from all sources, before taxes and deductions, in the past 12 months? Was it: 1 … less than $50,000 (include income loss)? 2 … $50,000 and more? (Go to INC_Q03H) DK, RF (Go to INC_C07) INC_Q03C IN2_03C INTERVIEWER: Read categories to respondent. Go to INC_C04 Please stop me when I have read the category which applies to ^YOUR1 household. 1 Less than $5,000 2 $5,000 or more but less than $10,000 3 $10,000 or more but less than $15,000	41 REVACCT1 Accepte de donner revenu familial T1 0: non 1: oui 42 REVFAMT1 Revenus familial sur échelle de 0 à 100 000$ T1

Appendix B. Variable Description and Categories for the studies CSHA, CCHS-CLSA, and NuAge (cont'd)

Candidate Variable	Variable Description	Categories	CSHA Variable Format	CCHS-CLSA Variable Format	NuAge Variable Format
Household Income (cont'd)				4 $15,000 or more but less than $20,000 5 $20,000 or more but less than $30,000 6 $30,000 or more but less than $40,000 7 $40,000 or more but less than $50,000 DK, RF INC_Q03H IN2_03H INTERVIEWER: Read categories to respondent. Please stop me when I have read the category which applies to ^YOUR1 household. 1 $50,000 or more but less than $60,000 2 $60,000 or more but less than $70,000 3 $70,000 or more but less than $80,000 4 $80,000 or more but less than $90,000 5 $90,000 or more but less than $100,000 6 $100,000 or more but less than $150,000 7 $150,000 and over DK, RF	

Appendix B. Variable Description and Categories for the studies CSHA, CCHS-CLSA, and NuAge (cont'd)

Candidate Variable	Variable Description	Categories	CSHA Variable Format	CCHS-CLSA Variable Format	NuAge Variable Format
Household Income Categorical	Average current annual income, before taxes, of the participant's entire household based on CSHA categories.	>$10k $10k to $14,999 $15k to $19,999 $20k to $24,999 $25k to $29,999 $30k to $34,999 $35k to $39,999 $40k to $44,999 $45k to $49,999 $50k to $59,999 $60k to $69,999 $70k and more Prefer not to answer Don't know	Complete	Complete	Complete
Country of Birth	Country where the participant was born.		24. Where were you born? Town Province (or country)	SDC_Q1 SDC_1 Go to SDC_Q2 In what country ^WERE ^YOU2 born? 01 Canada (Go to SDC_Q4) 02 China 03 France Plus a list of other countries.	12 pays Pays de naissance T1 13 autpays Si né à l'extérieur du pays, nb années vécues à l'étranger T1
Ever Smoked Cigarettes	Indicator of whether the participant has ever smoked cigarettes.	Never smoked cigarettes Ever smoked cigarettes Prefer not to answer Don't know	Has he/she ever smoked cigarettes regularly (nearly every day)? __Yes for how many years?____years SMOKE, SMOKEYR On average, how many per day?__ SMOKEDAY __Less than 1 pack __One pack __More than 1 pack __No __Don't know Please go to the next question.	SMK_Q201B SMK_01B ^HAVE_C ^YOU1 ever smoked a whole cigarette? 1 Yes (Go to SMK_Q201C) 2 No (Go to SMK_Q202) DK (Go to SMK_Q202) RF SMK_Q201A SMK_01A In ^YOUR1 lifetime, ^HAVE ^YOU2 smoked a total of 100 or more cigarettes (about 4 packs)? 1 Yes (Go to SMK_Q201C) 2 No DK, RF	Impossible

Appendix B. Variable Description and Categories for the studies CSHA, CCHS-CLSA, and NuAge (cont'd)

Candidate Variable	Variable Description	Categories	CSHA Variable Format	CCHS-CLSA Variable Format	NuAge Variable Format
Current Cigarette Smoker	Indicator of whether the participant currently smokes cigarettes.	Non-cigarette smoker Cigarette smoker Prefer not to answer Don't know	Impossible	SMK_Q202 SMK_202 Note: Daily smoker (current) At the present time, ^DOVERB ^YOU2 smoke cigarettes daily, occasionally or not at all? 1 Daily 2 Occasionally (Go to SMK_Q205B) 3 Not at all (Go to SMK_C205D) DK, RF (Go to SMK_END)	44 FUMERT1 Fumez-vous actuellement? T1 0: Non 1: Oui, occasionnellement 2: Oui, régulièrement 3: Non mais j'ai déjà fumé 45 NBFUMET1 Nb/jour cigarettes ou tabac T1
Current Quantity of Cigarettes Smoked	Current average number of cigarettes smoked per week.		Impossible	SMK_Q204 SMK_204 How many cigarettes ^DOVERB ^YOU1 smoke each day now? |_|_| Cigarettes (MIN: 1) (MAX: 99Warning after 60) DK, RF Go to SMK_END Note: Occasional smoker (current) SMK_Q205B SMK_05B On the days that ^YOU2 ^DOVERB smoke, how many cigarettes ^DOVERB ^YOU1 usually smoke? |_|_| Cigarettes DK, RF (MIN: 1) (MAX: 99; warning after 60) SMK_Q205C SMK_05C In the past month, on how many days ^HAVE ^YOU1 smoked 1 or more cigarettes? |_|_| Days (MIN: 0) (MAX: 30) DK, RF Note: Occasional smoker or non-smoker (current)	44 FUMERT1 Fumez-vous actuellement? T1 0: Non 1: Oui, occasionnellement 2: Oui, régulièrement 3: Non mais j'ai déjà fumé 45 NBFUMET1 Nb/jour cigarettes ou tabac T1

Appendix B. Variable Description and Categories for the studies CSHA, CCHS-CLSA, and NuAge (cont'd)

Candidate Variable	Variable Description	Categories	CSHA Variable Format	CCHS-CLSA Variable Format	NuAge Variable Format
Ever Alcohol Consumption	Indicator of whether the participant has ever consumed alcohol.	Never consumed alcohol Ever consumed alcohol Prefer not to answer Don't know	39. Has he/she ever been a regular beer drinker? (at least once a week) __Yes for how many years? BEER, BEERYR __No __Don't know, please go to the next question. 40. Has he/she ever been a regular wine drinker? (at least once a week) __Yes for how many years?___years WINE, WINEYR __No __Don't know Please go to the next question.	ALC_Q01 ALC_1 During the past 12 months, that is, from ^YEARAGO to yesterday, ^HAVE ^YOU2 had a drink of beer, wine, liquor or any other alcoholic beverage? 1 Yes 2 No (Go to ALC_Q05B) DK, RF (Go to ALC_END) ALC_Q05B ALC_5B ^HAVE_C ^YOU1 ever had a drink? 1 Yes 2 No DK, RF	Impossible
Current Use Alcohol	Indicator of whether the participant currently consumes alcohol	Does not consume alcohol Consumes alcohol Prefer not to answer Don't know	Impossible	ALC_Q01 ALC_1 During the past 12 months, that is, from ^YEARAGO to yesterday, ^HAVE ^YOU2 had a drink of beer, wine, liquor or any other alcoholic beverage? 1 Yes 2 No (Go to ALC_Q05B) DK, RF (Go to ALC_END)	When you had beer or ale(in the last month), how many cans or bottles did you usually have at one times? 1 or lees 2 3-3+ Not answered Have you had any wine in the past year? Yes No Not answered Have you had any wine in the past month? Yes No Not answered Over the past months hoe often have you had wine? 01-29 30-30+ Not answered

Appendix B. Variable Description and Categories for the studies CSHA, CCHS-CLSA, and NuAge (cont'd)

Candidate Variable	Variable Description	Categories	CSHA Variable Format	CCHS-CLSA Variable Format	NuAge Variable Format
Current Use Alcohol (cont'd)					When you had wine (in the last month), how many glasses did you usually have at one times? 1 or lees 2 3-3+ Not answered Have you had any liquor in the past month? Yes No Not answered Over the past months hoe often have you had beer or ale? 01-29 30-30+ Not answered When you had beer or ale(in the last month), how many cans or bottles did you usually have at one times? 1 or lees 2 3-3+ Not answered

Appendix B. Variable Description and Categories for the studies CSHA, CCHS-CLSA, and NuAge (cont'd)

Candidate Variable	Variable Description	Categories	CSHA Variable Format	CCHS-CLSA Variable Format	NuAge Variable Format
Standing Height	The vertical measurement or distance from the foot to the head of the participant when he/she is standing.		From doctor 1: Height HEIGHT _______m, _______ft. Continuous 888 Don't know 999 Missing	HWT_2 How tall ^ARE ^YOU2 without shoes on? .. 4 4'0" to 4'11" / 48" to 59" (120.7 to 151.0 cm.) (Go to HWT_N2D) 5 5'0" to 5'11" (151.1 to 181.5 cm.) (Go to HWT_N2E) 6 6'0" to 6'11" (181.6 to 212.0 cm.) (Go to HWT_N2F) 7 7'0" and over (212.1 cm. and over) (Go to HWT_Q3) HWT_N2D HWT_2D INTERVIEWER: Select the exact height. Go to HWT_Q3 00 4'0" / 48" (120.7 to 123.1 cm.) 01 4'1" / 49" (123.2 to 125.6 cm.) 02 4'2" / 50" (125.7 to 128.2 cm.) 03 4'3" / 51" (128.3 to 130.7 cm.) 04 4'4" / 52" (130.8 to 133.3 cm.) 05 4'5" / 53" (133.4 to 135.8 cm.) 06 4'6" / 54" (135.9 to 138.3 cm.) 07 4'7" / 55" (138.4 to 140.9 cm.) 08 4'8" / 56" (141.0 to 143.4 cm.) 09 4'9" / 57" (143.5 to 146.0 cm.) 10 4'10" / 58" (146.1 to 148.5 cm.) 11 4'11" / 59" (148.6 to 151.0 cm.) DK, RF	469 TAIMEST1 Taille mesurée (m)

Appendix B. Variable Description and Categories for the studies CSHA, CCHS-CLSA, and NuAge (cont'd)

Candidate Variable	Variable Description	Categories	CSHA Variable Format	CCHS-CLSA Variable Format	NuAge Variable Format
Weight	Weight of the participant.		From doctor 1: Weight WEIGHT _______kg, _______lbs. Continuous 888 Don't know 999 Missing	HWT_Q3 HWT_3 How much ^DOVERB ^YOU1 weigh? INTERVIEWER Enter amount only. (MIN: 1) (MAX: 575) |_|_|_| Weight DK, RF (Go to HWT_END) HWT_N4 INTERVIEWER Was that in pounds or kilograms? 1 Pounds 2 Kilograms (DK, RF are not allowed)	470 POIMEMT1 poids mémoire (lbs) T1 471 POIMEST1 poids mesuré (kg) T1
Body Mass Index	Weight (in kg) divided by height (in m) squared. Body mass index = (Weight) / (Standing height * 0.01)^2		Complete	Complete	Complete
Hip Circumference	Measured distance around hips.		Impossible	Impossible	481 CIRCHAT1 circonférence hanche (cm) T1
Waist Circumference	Measured distance around waist.		Impossible	Impossible	480 CIRCTAT1 circonférence taille (cm) T1
Heart Rate at Rest	Number of heart beats per minute measured at rest.		Impossible	Impossible	621 RYTHMET1 Rythme cardiaque T1
Diastolic Blood Pressure at Rest	Diastolic blood pressure measured at rest.		DBPSUPIN Continuous 777 Skipped 888 Don't know 999 Missing 6666 NA/Skipped	Impossible	617 TAMINT1 Tension artérielle diastolique (assis) T1
Systolic Blood Pressure at Rest	Systolic blood pressure measured at rest.		SBPSUPIN Continuous 777 Skipped 888 Don't know 999 Missing 6666 NA/Skipped	Impossible	616 TAMAXT1 Tension artérielle systolique (assis) T1

Appendix B. Variable Description and Categories for the studies CSHA, CCHS-CLSA, and NuAge (cont'd)

Candidate Variable	Variable Description	Categories	CSHA Variable Format	CCHS-CLSA Variable Format	NuAge Variable Format
Occurrence of High Blood Pressure	Occurrence of high blood pressure at any point during the life of the participant.	Never had high blood pressure Ever had high blood pressure Prefer not to answer Don't know	From Screen: Now I will read a list of health problems that people often have. For each problem that I read, please tell me if you have had it in the past year. If the problem began longer ago and symptoms lasted into the past year, check "yes". Do not read the examples in parentheses unless the respondent asks for clarification.) Yes No a) High blood pressure (whether controlled 1 2 HBP by medication or not) From proxy 1: 21a) Has he/she suffered from any of the following health problems? Health problem or condition No Don't know Yes How long Has he/she had this condition? High blood pressure PROXHBP HBPYR 1 Yes 2 No / chart slashed 8 Don't know 9 Missing From doctor1 Rate as 1 = yes, describe and indicate duration 2 = questionable 3 = no 4 = not relevant, describe why 5 = subject does not know 6 = subject could not answer 9 = not asked 28. Arterial hypertension (H) 1 2 3 4 5 6 9 ARTERIAL	We are interested in "long-term conditions" which are expected to last, or have already lasted, 6 months or more and that have been diagnosed by a health professional. CCC_071 (^DOVERB_C ^YOU2 have:) .. high blood pressure? 1 Yes (Go to CCC_Q073) 2 No DK, RF CCC_Q072 CCC_072 ^HAVE_C ^YOU2 ever been diagnosed with high blood pressure? 1 Yes 2 No (Go to CCC_Q081) DK, RF (Go to CCC_Q081) CCC_Q073 CCC_073 In the past month, ^HAVE ^YOU2 taken any medication for high blood pressure? 1 Yes 2 No DK, RF CCC_C073A If RESPGENDER = 2, go to CCC_Q073A. Otherwise, go to CCC_Q081. CCC_Q073A CCC_073A ^WERE_C ^YOU2 pregnant when ^YOU1 ^WERE diagnosed with high blood pressure? 1 Yes 2 No (Go to CCC_Q081) DK, RF (Go to CCC_Q081) CCC_Q073B CCC_073B	93 HTEPRET1 Haute pression T1 "Do you have .. At the present time" 0: non 1: oui, n'empêche pas du tout 2: oui, empêche un peu 3: oui, empêche beaucoup

Appendix B. Variable Description and Categories for the studies CSHA, CCHS-CLSA, and NuAge (cont'd)

Candidate Variable	Variable Description	Categories	CSHA Variable Format	CCHS-CLSA Variable Format	NuAge Variable Format
Occurrence of High Blood Pressure (cont'd)			From Nurse1: Variable: C295 Label: CAMDEX 295. High BP? Value labels 0 No 1 Yes 8 Don't know 9 Missing	Other than when ^YOU1 ^WERE pregnant, was there any other time when ^YOU1 ^WERE diagnosed with high blood pressure? 1 Yes 2 No DK, RF	
Current Treatment for High Blood Pressure	Indicator of whether the participant is currently treated for high blood pressure.	Not currently under treatment for high blood pressure Currently under treatment for high blood pressure Prefer not to answer Don't know	From Proxy: Please list all medication he/she is currently taking for any of the above conditions. Don't Know Name of the medication How long .. From Doctor Ant-hypertensive (Yes, No, can't tell, Missing)	CCC_R001 INTERVIEWER: Press <Enter> to continue. Now I'd like to ask about certain chronic health conditions which ^YOU2 may have. We are interested in "long-term conditions" which are expected to last, or have already lasted, 6 months or more and that have been diagnosed by a health professional. CCC_071 (^DOVERB_C ^YOU2 have:) .. high blood pressure? 1 Yes (Go to CCC_Q073) 2 No DK, RF CCC_Q072 CCC_072 ^HAVE_C ^YOU2 ever been diagnosed with high blood pressure? 1 Yes 2 No (Go to CCC_Q081) DK, RF (Go to CCC_Q081) CCC_Q073 CCC_073 In the past month, ^HAVE ^YOU2 taken any medication for high blood pressure? 1 Yes 2 No DK, RF	Impossible

Appendix B. Variable Description and Categories for the studies CSHA, CCHS-CLSA, and NuAge (cont'd)

Candidate Variable	Variable Description	Categories	CSHA Variable Format	CCHS-CLSA Variable Format	NuAge Variable Format
Current Treatment for High Blood Pressure (cont'd)				CCC_C073A If RESPGENDER = 2, go to CCC_Q073A. Otherwise, go to CCC_Q081. CCC_Q073A CCC_073A ^WERE_C ^YOU2 pregnant when ^YOU1 ^WERE diagnosed with high blood pressure? 1 Yes 2 No (Go to CCC_Q081) DK, RF (Go to CCC_Q081) CCC_Q073B CCC_073B Other than when ^YOU1 ^WERE pregnant, was there any other time when ^YOU1 ^WERE diagnosed with high blood pressure? 1 Yes 2 No DK, RF	

Appendix B. Variable Description and Categories for the studies CSHA, CCHS-CLSA, and NuAge (cont'd)

Candidate Variable	Variable Description	Categories	CSHA Variable Format	CCHS-CLSA Variable Format	NuAge Variable Format
Occurrence of Stroke	Occurrence of stroke at any point during the life of the participant.	Never had stroke Ever had stroke Prefer not to answer Don't know	1 Yes 8 Don't know 9 Missing From doctor1 "Did the subject or does the subject have.." 17. history of stroke (H) 1 2 3 4 5 6 9 HXSTROKE From proxy1 21a) Has he/she suffered from any of the following health problems? Health problem or condition No Don't know Yes How long Has he/she had this condition? Stroke PRSTROKE STRYR 1 Yes 2 No / chart slashed 8 Don't know 9 Missing From screen1 27. Now, I will read a list of health problems that people often have. For each problem that I read, please tell me if you have had it in the past year. You can just answer Yes or No. (Note to Interviewer: if the problem began longer ago and symptoms lasted into the past year, check "yes". Do not read the examples in parentheses unless the respondent asks for clarification.) c) Stroke, or effects of stroke 1 2 STROKE	CCC_R001 INTERVIEWER: Press <Enter> to continue. Now I'd like to ask about certain chronic health conditions which ^YOU2 may have. We are interested in "long-term conditions" which are expected to last, or have already lasted, 6 months or more and that have been diagnosed by a health professional. CCC_Q151 CCC_151 ^DOVERB_C ^YOU2 suffer from the effects of a stroke? 1 Yes 2 No DK, RF	103 ACVT1 Thrombose, hémorragie cérébrale, avc T1 "Do you have .. At the present time" 0: non 1: oui, n'empêche pas du tout 2: oui, empêche un peu 3: oui, empêche beaucoup

Appendix B. Variable Description and Categories for the studies CSHA, CCHS-CLSA, and NuAge (cont'd)

Candidate Variable	Variable Description	Categories	CSHA Variable Format	CCHS-CLSA Variable Format	NuAge Variable Format
Occurrence of Diabetes	Occurrence of diabetes at any point during the life of the participant.	Never had diabetes Ever had diabetes Prefer not to answer Don't know	Variable: C297 Label: CAMDEX 297. Diabetic? 0 No 1 Yes 8 Don't know 9 Missing From doctor1 Did the subject or does the subject have.. 36. History of diabetes mellitus 1 2 3 4 5 6 9 From proxy1 21a) Has he/she suffered from any of the following health problems? Health problem or condition No Don't know Yes How long Has he/she had this condition? Diabetes PROXDIAB DIAYR 1 Yes 2 No / chart slashed 8 Don't know 9 Missing From screen1 27. Now, I will read a list of health problems that people often have. For each problem that I read, please tell me if you have had it in the past year. You can just answer Yes or No. n) Diabetes 1 2 DIABETES	CCC_R001 INTERVIEWER: Press <Enter> to continue. Now I'd like to ask about certain chronic health conditions which ^YOU2 may have. We are interested in "long-term conditions" which are expected to last, or have already lasted, 6 months or more and that have been diagnosed by a health professional. CCC_Q101 CCC_101 (^DOVERB_C ^YOU2 have:) .. diabetes? 1 Yes 2 No DK, RF	96 DIABETT1 Diabète T1. "Do you have .. At the present time" 0: non 1: oui, n'empêche pas du tout 2: oui, empêche un peu 3: oui, empêche beaucoup

Appendix B. Variable Description and Categories for the studies CSHA, CCHS-CLSA, and NuAge (cont'd)

Candidate Variable	Variable Description	Categories	CSHA Variable Format	CCHS-CLSA Variable Format	NuAge Variable Format
Occurrence of Myocardial Infarction	Occurrence of myocardial infarction at any point during the life of the participant	Never had myocardial infarction Ever had myocardial infarction Prefer not to answer Don't know	From nurse1: Variable: C296 Label: CAMDEX 296. Heart attack? Value labels 0 No 1 One 2 More than one 8 Don't Know 9 Missing	CCC_Q120 CCC_120 ^HAVE_C ^YOU1 ever had a heart attack? 1 Yes 2 No DK, RF CCC_C121 If CCC_Q119 = 1 or CCC_Q120 = 1, go to CCC_Q131. Otherwise, go to CCC_Q121. CCC_Q121 CCC_121 INTERVIEWER Include congestive heart failure. At the time of data processing, if the respondent reported having either angina (CCC_Q119 = 1) or a heart attack (CCC_Q120 = 1), then the variable for heart disease will be set to "Yes" (CCC_Q121 = 1).	108 AUTRMAT1 Autres maladies T1 0: non 1: oui, n'empêche pas du tout 2: oui, empêche un peu 3: oui, empêche beaucoup 109 TYPMALT1 Type maladie autre T1
Family History of High Blood Pressure	Occurrence of high blood pressure amongst members of the biological family of the participant (mother, father, siblings and children).	No family history of high blood pressure Family history of high blood pressure Prefer not to answer Don't know	Impossible	Impossible	Impossible
Family History of Stroke	Occurrence of stroke amongst members of the biological family of the participant (mother, father, siblings and children).	No family history of stroke Family history of stroke Prefer not to answer Don't know	Impossible	Impossible	Impossible

Appendix B. Variable Description and Categories for the studies CSHA, CCHS-CLSA, and NuAge (cont'd)

Candidate Variable	Variable Description	Categories	CSHA Variable Format	CCHS-CLSA Variable Format	NuAge Variable Format
Family History of Diabetes	Occurrence of diabetes amongst members of the biological family of the participant (mother, father, siblings and children).	No family history of diabetes Family history of diabetes Prefer not to answer Don't know	Impossible	Impossible	Impossible
Family History of Myocardial Infarction	Occurrence of myocardial infarction amongst members of the biological family of the participant (mother, father, siblings and children).	No family history of myocardial infarction Family history of myocardial infarction Prefer not to answer Don't know	Impossible	Impossible	Impossible
Level of Physical Activity	Categorical indicator of the participant's level of physical activity. Based on IPAQ scoring protocol (https://sites.google.com/site/theipaq/scoring-protocol).	Low level of physical activity Moderate level of physical activity High level of physical activity Not applicable	Impossible	SAPE (Physical Activity Scale for the Elderly)	SAPE (Physical Activity Scale for the Elderly)
Total physical activity	Quantitative indicator of global physical activity in metabolic equivalent (MET)-minutes per week. Based on IPAQ scoring protocol (https://sites.google.com/site/theipaq/scoring-protocol).		Impossible	SAPE (Physical Activity Scale for the Elderly)	SAPE (Physical Activity Scale for the Elderly)

Appendix B. Variable Description and Categories for the studies CSHA, CCHS-CLSA, and NuAge (cont'd)

Candidate Variable	Variable Description	Categories	CSHA Variable Format	CCHS-CLSA Variable Format	NuAge Variable Format
Level of Physical Activity (CSHA-based - ordinal)	Categorical indicator of the participant's level of physical activity using CSHA categories (ordinal; 3 categories)	Never [0] Low level of physical activity [1] Moderate level of physical activity [2] High level of physical activity [3] Prefer not to answer [8] Don't know [9]	Complete	Complete	Complete
Level of Physical Activity (CSHA-based - binary)	Categorical indicator of the participant's level of physical activity using CSHA categories (binary)	Never [0] Low level of physical activity [1] Moderate or High level of physical activity [2] Prefer not to answer [8] Don't know [9]	Complete	Complete	Complete

Abbreviations: CCHS = Canadian Community Health Survey; CLSA = Canadian Longitudinal Study on Aging; cm = centimeters; CSHA = Canadian Study of Health and Aging; DK = don't know; ft. = feet; HBP = high blood pressure; IPAQ = International Physical Activity Questionnaire; kgs. = kilograms; lbs = pounds; m = meters; MAX = maximum; MET = metabolic equivalent; MIN = minimum; NA = not applicable; nb = number; NuAge = Quebec Longitudinal Study on Nutrition and Aging; RF = refused; SAPE = Physical Activity Scale for the Elderly

Appendix C. Distribution of Age, Sex, and Education Level for CCHS-CLSA, CSHA, and NuAge

Appendix C. Table 1. Age, Sex and Education Level Distribution of CCHS-CLSA, CSHA and NuAge Studies

Age group	Sex	Education level	CCHS-CLSA		CSHA		NuAge	
			Freq	%	Freq	%	Freq	%
65-69	Male	Low (0-8)	144	0.79	43	0.23	1	0.01
65-69	Male	Medium (9-13)	377	2.06	30	0.16	19	0.1
65-69	Male	High (14+)	578	3.15	5	0.03	22	0.12
65-69	Female	Low (0-8)	153	0.83	55	0.3	9	0.05
65-69	Female	Medium (9-13)	467	2.55	23	0.13	19	0.1
65-69	Female	High (14+)	658	3.59	7	0.04	15	0.08
70-74	Male	Low (0-8)	152	0.83	45	0.25	11	0.06
70-74	Male	Medium (9-13)	292	1.59	30	0.16	33	0.18
70-74	Male	High (14+)	353	1.93	9	0.05	49	0.27
70-74	Female	Low (0-8)	167	0.91	56	0.31	14	0.08
70-74	Female	Medium (9-13)	388	2.12	52	0.28	48	0.26
70-74	Female	High (14+)	402	2.19	10	0.05	25	0.14
75-79	Male	Low (0-8)	151	0.82	91	0.5	8	0.04
75-79	Male	Medium (9-13)	239	1.3	84	0.46	19	0.1
75-79	Male	High (14+)	259	1.41	30	0.16	21	0.11
75-79	Female	Low (0-8)	231	1.26	127	0.69	19	0.1
75-79	Female	Medium (9-13)	381	2.08	111	0.61	41	0.22
75-79	Female	High (14+)	368	2.01	35	0.19	21	0.11
80-84	Male	Low (0-8)	104	0.57	78	0.43	3	0.02
80-84	Male	Medium (9-13)	153	0.83	37	0.2	7	0.04
80-84	Male	High (14+)	121	0.66	13	0.07	7	0.04
80-84	Female	Low (0-8)	198	1.08	115	0.63	1	0.01
80-84	Female	Medium (9-13)	304	1.66	119	0.65	11	0.06
80-84	Female	High (14+)	260	1.42	31	0.17	9	0.05
85-89	Male	Low (0-8)	10	0.05	79	0.43		
85-89	Male	Medium (9-13)	21	0.11	32	0.17		
85-89	Male	High (14+)	24	0.13	13	0.07		
85-89	Female	Low (0-8)	32	0.17	99	0.54		
85-89	Female	Medium (9-13)	42	0.23	112	0.61		
85-89	Female	High (14+)	32	0.17	28	0.15		
90-94	Male	Low (0-8)			13	0.07		
90-94	Male	Medium (9-13)			6	0.03		
90-94	Male	High (14+)			4	0.02		
90-94	Female	Low (0-8)			32	0.17		
90-94	Female	Medium (9-13)			41	0.22		
90-94	Female	High (14+)			5	0.03		

Appendix C. Table 1. Age, Sex and Education Level Distribution of CCHS-CLSA, CSHA and NuAge Studies (cont'd)

Age group	Sex	Education level	CCHS-CLSA		CSHA		NuAge	
			Freq	%	Freq	%	Freq	%
95-99	Male	Low (0-8)			1	0.01		
95-99	Male	High (14+)			1	0.01		
95-99	Female	Low (0-8)			6	0.03		
95-99	Female	Medium (9-13)			7	0.04		
95-99	Female	High (14+)			3	0.02		
100-104	Female	Low (0-8)			1	0.01		

Abbreviations: CCHS = Canadian Community Health Survey; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study on Health and Aging; Freq = frequency; NuAge = Quebec Longitudinal Study on Nutrition and Aging

Appendix C Table 2. Standard Deviations of Cognitive Measures for Females Aged 70-74 and 75-79 with 9-13 Years of Education in CCHS-CLSA, CSHA and NuAge Studies

Study	Cognitive Measure	Standard Deviation 70-74	Standard Deviation 75-79	Min	Group
CCHS-CLSA	Rey	1.85	1.78	1.78	75-79
CSHA	Rey	1.94	2.05	1.94	70-74
	Buschke	2.76	2.78	2.76	70-74
	Buschke Total	2.16	1.72	1.72	75-79
NuAge	Buschke	2.23	2.59	2.23	70-74
	Buschke Total	1.37	2.18	1.37	70-74

Abbreviations: CCHS = Canadian Community Health Survey; CLSA = Canadian Longitudinal Study on Aging; CSHA = Canadian Study on Health and Aging; min = minimum; NuAge = Quebec Longitudinal Study on Nutrition and Aging

Appendix D. Meta-Analysis Results Weighted Mean Difference Analysis

Appendix D. Table 1. Summary of Traditional Meta-Analysis Results for Rey from CCHS and CSHA and Buschke Free from NuAge

Method of Statistical Harmonization	Adjustment	Summary Effect Estimate (95% CI)	I^2	Q^*	p-value
Raw Data – Using Hedge's G	Unadjusted	0.16 (0.01, 0.30)	0.78	8.96	0.01
	Unadjusted (Including Participants with Complete Data for Covariates)	0.16 (0.01, 0.30)	0.75	7.84	0.02
	Study-Specific Statistically Significant Covariates	0.11 (-0.02, 0.24)	0.72	7.18	0.03
	Common Set of Covariates	0.11 (-0.02, 0.23)	0.66	5.81	0.06
T-Score – Weighted Mean Difference	Unadjusted†	1.18 (0.05, 2.32)	0.64	5.45	0.06
	Unadjusted† (Including Participants with Complete Data for Covariates)	1.19 (0.11, 2.27)	0.55	4.47	0.11
	Common Set of Covariates	1.09 (-0.20, 2.39)	0.67	5.98	0.05
C-Score – Weighted Mean Difference	Unadjusted	0.16 (0.03, 0.29)	0.73	7.46	0.02
	Unadjusted (Including Participants with Complete Data for Covariates)	0.17 (0.04, 0.30)	0.70	6.69	0.04
	Common Set of Covariates	0.11 (-0.01, 0.22)	0.63	5.33	0.07

*All Q Statistics have 2 degrees of freedom

†T-Scores minimally adjusted for age, sex and education level

Abbreviations: CI = confidence interval

Appendix D. Table 2. Summary of traditional meta-analysis results for rey from CCHS and CSHA and Buschke total from NuAge

Method of Statistical Harmonization	Adjustment	Summary Effect Estimate (95% CI)	I^2	Q^*	p-value
Raw Data – Using Hedge's G	Unadjusted	0.13 (-0.04, 0.30)	0.84	12.20	0.002
	Unadjusted (Including Participants with Complete Data for Covariates)	0.13 (-0.04, 0.30)	0.82	10.99	0.004
	Study-Specific Statistically Significant Covariates	0.10 (-0.04, 0.24)	0.75	8.14	0.02
	Common Set of Covariates	0.10 (-0.04, 0.23)	0.72	7.06	0.03
T-Score – Weighted Mean Difference	Unadjusted†	0.89 (-0.55, 2.34)	0.78	9.10	0.01
	Unadjusted† (Including Participants with Complete Data for Covariates)	0.89 (-0.55, 2.33)	0.75	7.86	0.02
	Common Set of Covariates	0.98 (-0.45, 2.41)	0.72	7.21	0.03
C-Score – Weighted Mean Difference	Unadjusted	0.15 (-0.001, 0.30)	0.78	9.10	0.01
	Unadjusted (Including Participants with Complete Data for Covariates)	0.15 (-0.002, 0.31)	0.76	8.18	0.02
	Common Set of Covariates	0.12 (0.001, 0.23)	0.60	4.95	0.08

*All Q Statistics have 2 degrees of freedom

†T-Scores minimally adjusted for age, sex and education level

Abbreviations: CI = confidence interval

Appendix D. Table 3. Summary of traditional meta-analysis results for rey from CCHS and Buschke free from CSHA and NuAge

Method of Statistical Harmonization	Adjustment	Summary Effect Estimate (95% CI)	I^2	Q^*	p-value
Raw Data – Using Hedge's G	Unadjusted	0.18 (-0.01, 0.36)	0.88	16.46	<0.001
	Unadjusted (Including Participants with Complete Data for Covariates)	0.18 (0.0001, 0.36)	0.85	13.37	0.001
	Study-Specific Statistically Significant Covariates	0.12 (-0.04, 0.29)	0.85	13.0	0.002
	Common Set of Covariates	0.13 (-0.05, 0.31)	0.85	12.9	0.002
T-Score – Weighted Mean Difference	Unadjusted†	1.40 (-0.33, 3.13)	0.85	13.51	0.001
	Unadjusted† (Including Participants with Complete Data for Covariates)	1.44 (-0.25, 3.13)	0.83	11.56	0.003
	Common Set of Covariates	1.38 (-0.53, 3.28)	0.85	13.43	0.001
C-Score – Weighted Mean Difference	Unadjusted	0.18 (0.01, 0.36)	0.86	14.45	0.001
	Unadjusted (Including Participants with Complete Data for Covariates)	0.19 (0.03, 0.35)	0.83	11.5	0.003
	Common Set of Covariates	0.13 (-0.04, 0.30)	0.84	12.22	0.002

*All Q Statistics have 2 degrees of freedom

†T-Scores minimally adjusted for age, sex and education level

Abbreviations: CI = confidence interval

Appendix D. Table 4. Summary of traditional meta-analysis results for rey from CCHS, Buschke free from CSHA and Buschke total from NuAge

Method of Statistical Harmonization	**Adjustment**	**Summary Effect Estimate (95% CI)**	**I^2**	**Q***	**p-value**
Raw Data – Using Hedge's G	Unadjusted	0.16 (-0.04, 0.35)	0.90	20.32	<0.001
	Unadjusted (Including Participants with Complete Data for Covariates)	0.16 (-0.04, 0.36)	0.88	16.67	<0.001
	Study-Specific Statistically Significant Covariates	0.12 (-0.06, 0.29)	0.86	13.98	0.001
	Common Set of Covariates	0.12 (-0.06, 0.31)	0.86	14.29	0.001
T-Score – Weighted Mean Difference	Unadjusted†	1.14 (-0.79, 3.06)	0.89	17.42	<0.001
	Unadjusted† (Including Participants with Complete Data for Covariates)	1.18 (-0.72, 3.08)	0.87	15.26	<0.001
	Common Set of Covariates	1.28 (-0.72, 3.27)	0.86	14.74	0.001
C-Score – Weighted Mean Difference	Unadjusted	0.17 (-0.02, 0.36)	0.87	15.90	<0.001
	Unadjusted (Including Participants with Complete Data for Covariates)	0.18 (-0.007, 0.36)	0.84	12.85	0.002
	Common Set of Covariates	0.14 (-0.04, 0.31)	0.83	11.56	0.002

*All Q Statistics have 2 degrees of freedom

†T-Scores minimally adjusted for age, sex and education level

Abbreviations: CI = confidence interval

Appendix D. Table 5. Summary of traditional meta-analysis results for rey from CCHS, Buschke total from CSHA and Buschke free from NuAge

Method of Statistical Harmonization	**Adjustment**	**Summary Effect Estimate (95% CI)**	**I^2**	**Q[*]**	**p-value**
Raw Data – Using Hedge's G	Unadjusted	0.15 (0.02, 0.27)	0.73	7.35	0.03
	Unadjusted (Including Participants with Complete Data for Covariates)	0.15 (0.02, 0.28)	0.71	6.80	0.03
	Study-Specific Statistically Significant Covariates	0.09 (-0.01, 0.20)	0.63	5.47	0.07
	Common Set of Covariates	0.10 (-0.008, 0.21)	0.61	5.18	0.08
T-Score – Weighted Mean Difference	Unadjusted[†]	1.08 (0.27, 1.88)	0.43	3.48	0.18
	Unadjusted[†] (Including Participants with Complete Data for Covariates)	1.12 (0.31, 1.94)	0.41	3.37	0.19
	Common Set of Covariates	1.06 (-0.12, 2.23)	0.65	5.35	0.07
C-Score – Weighted Mean Difference	Unadjusted	0.15 (0.05, 0.26)	0.70	6.61	0.04
	Unadjusted (Including Participants with Complete Data for Covariates)	0.16 (0.05, 0.27)	0.68	6.27	0.04
	Common Set of Covariates	0.10 (0.006, 0.20)	0.57	4.68	0.10

[*] All Q Statistics have 2 degrees of freedom

[†] T-Scores minimally adjusted for age, sex and education level

Abbreviations: CI = confidence interval

Appendix D. Table 6. Summary of traditional meta-analysis results for rey from CCHS and Buschke total from CSHA and NuAge

Method of Statistical Harmonization	Adjustment	Summary Effect Estimate (95% CI)	I^2	Q*	p-value
Raw Data – Using Hedge's G	Unadjusted	0.12 (-0.03, 0.27)	0.81	10.51	0.005
	Unadjusted (Including Participants with Complete Data for Covariates)	0.12 (-0.03, 0.28)	0.80	9.89	0.007
	Study-Specific Statistically Significant Covariates	0.09 (-0.03, 0.20)	0.69	6.39	0.04
	Common Set of Covariates	0.09 (-0.03, 0.22)	0.69	6.42	0.04
T-Score – Weighted Mean Difference	Unadjusted†	0.77 (-0.42, 1.95)	0.71	6.93	0.03
	Unadjusted† (Including Participants with Complete Data for Covariates)	0.81 (-0.39, 2.01)	0.70	6.67	0.04
	Common Set of Covariates	0.94 (-0.37, 2.26)	0.70	6.57	0.04
C-Score – Weighted Mean Difference	Unadjusted	0.14 (0.02, 0.27)	0.76	8.31	0.02
	Unadjusted (Including Participants with Complete Data for Covariates)	0.15 (0.02, 0.27)	0.75	7.83	0.02
	Common Set of Covariates	0.11 (0.01, 0.21)	0.54	4.35	0.11

*All Q Statistics have 2 degrees of freedom

†T-Scores minimally adjusted for age, sex and education level

Abbreviations: CI = confidence interval

Appendix D. Table 7. Summary of traditional meta-analysis results for HUI cognition from CCHS-CLSA, rey from CSHA and Buschke free from NuAge

Method of Statistical Harmonization	**Adjustment**	**Summary Effect Estimate (95% CI)**	**I^2**	**Q[*]**	**p-value**
Raw Data – Using Hedge’s G	Unadjusted	0.13 (-0.02, 0.28)	0.80	9.86	0.007
	Unadjusted (Including Participants with Complete Data for Covariates)	0.13 (-0.02, 0.27)	0.75	8.10	0.02
	Study-Specific Statistically Significant Covariates	0.09 (-0.04, 0.22)	0.74	7.58	0.02
	Common Set of Covariates	0.09 (-0.03, 0.22)	0.66	5.83	0.054
T-Score – Weighted Mean Difference	Unadjusted[†]	1.07 (-0.11, 2.26)	0.66	5.9	0.053
	Unadjusted[†] (Including Participants with Complete Data for Covariates)	1.05 (-0.06, 2.16)	0.57	4.61	0.10
	Common Set of Covariates	0.95 (-0.35, 2.25)	0.67	6.02	0.049
C-Score – Weighted Mean Difference	Unadjusted	0.13 (-0.002, 0.27)	0.75	7.95	0.02
	Unadjusted (Including Participants with Complete Data for Covariates)	0.13 (0.002, 0.27)	0.70	6.64	0.04
	Common Set of Covariates	0.10 (-0.02, 0.21)	0.62	5.30	0.07

[*] All Q Statistics have 2 degrees of freedom

[†] T-Scores minimally adjusted for age, sex and education level

Abbreviations: CI = confidence interval

Appendix D. Table 8. Summary of traditional meta-analysis results for HUI cognition from CCHS-CLSA, rey from CSHA and Buschke total from NuAge

Method of Statistical Harmonization	Adjustment	Summary Effect Estimate (95% CI)	I^2	Q[*]	p-value
Raw Data – Using Hedge's G	Unadjusted	0.11 (-0.06, 0.28)	0.84	12.67	0.002
	Unadjusted (Including Participants with Complete Data for Covariates)	0.11 (-0.06, 0.28)	0.81	10.77	0.005
	Study-Specific Statistically Significant Covariates	0.08 (-0.06, 0.22)	0.77	8.61	0.01
	Common Set of Covariates	0.08 (-0.05, 0.22)	0.71	6.98	0.03
T-Score – Weighted Mean Difference	Unadjusted[†]	0.80 (-0.67, 2.27)	0.78	9.21	0.01
	Unadjusted[†] (Including Participants with Complete Data for Covariates)	0.80 (-0.65, 2.22)	0.74	7.66	0.02
	Common Set of Covariates	0.85 (-0.58, 2.27)	0.72	7.14	0.03
C-Score – Weighted Mean Difference	Unadjusted	0.12 (-0.04, 0.28)	0.80	9.84	0.007
	Unadjusted (Including Participants with Complete Data for Covariates)	0.12 (-0.04, 0.28)	0.76	8.39	0.02
	Common Set of Covariates	0.11 (-0.02, 0.22)	0.61	5.10	0.08

[*] All Q Statistics have 2 degrees of freedom

[†] T-Scores minimally adjusted for age, sex and education level

Abbreviations: CI = confidence interval

Appendix D. Table 9. Summary of traditional meta-analysis results for HUI cognition from CCHS-CLSA and Buschke free from CSHA and NuAge

Method of Statistical Harmonization	**Adjustment**	**Summary Effect Estimate (95% CI)**	**I^2**	**Q[*]**	**p-value**
Raw Data – Using Hedge's G	Unadjusted	0.15 (-0.06, 0.36)	0.90	20.62	<0.001
	Unadjusted (Including Participants with Complete Data for Covariates)	0.15 (-0.05, 0.35)	0.88	16.54	<0.001
	Study-Specific Statistically Significant Covariates	0.11 (-0.08, 0.30)	0.88	16.56	<0.001
	Common Set of Covariates	0.12 (-0.07, 0.31)	0.86	14.53	0.001
T-Score – Weighted Mean Difference	Unadjusted[†]	1.30 (-0.52, 3.12)	0.87	14.85	0.001
	Unadjusted[†] (Including Participants with Complete Data for Covariates)	1.32 (-0.48, 3.11)	0.85	12.91	0.002
	Common Set of Covariates	1.24 (-0.77, 3.25)	0.87	14.96	0.001
C-Score – Weighted Mean Difference	Unadjusted	0.15 (-0.04, 0.35)	0.89	18.70	<0.001
	Unadjusted (Including Participants with Complete Data for Covariates)	0.15 (-0.03, 0.34)	0.87	14.88	0.001
	Common Set of Covariates	0.12 (-0.06, 0.30)	0.85	13.10	0.001

[*] All Q Statistics have 2 degrees of freedom

[†] T-Scores minimally adjusted for age, sex and education level

Abbreviations: CI = confidence interval

Appendix D. Table 10. Summary of traditional meta-analysis results for HUI cognition from CCHS-CLSA, Buschke free from CSHA and Buschke total from NuAge

Method of Statistical Harmonization	Adjustment	Summary Effect Estimate (95% CI)	I^2	Q*	p-value
Raw Data – Using Hedge's G	Unadjusted	0.13 (-0.09, 0.35)	0.92	23.61	<0.001
	Unadjusted (Including Participants with Complete Data for Covariates)	0.13 (-0.09, 0.35)	0.90	19.36	<0.001
	Study-Specific Statistically Significant Covariates	0.11 (-0.09, 0.30)	0.89	17.66	<0.001
	Common Set of Covariates	0.11 (-0.09, 0.31)	0.87	15.77	<0.001
T-Score – Weighted Mean Difference	Unadjusted†	1.05 (-0.95, 3.04)	0.89	18.51	<0.001
	Unadjusted† (Including Participants with Complete Data for Covariates)	1.07 (-0.91, 3.05)	0.88	16.28	<0.001
	Common Set of Covariates	1.14 (-0.95, 3.23)	0.88	16.17	<0.001
C-Score – Weighted Mean Difference	Unadjusted	0.13 (-0.09, 0.35)	0.90	20.44	<0.001
	Unadjusted (Including Participants with Complete Data for Covariates)	0.14 (-0.07, 0.34)	0.88	16.53	<0.001
	Common Set of Covariates	0.12 (-0.07, 0.31)	0.84	12.59	0.002

*All Q Statistics have 2 degrees of freedom

†T-Scores minimally adjusted for age, sex and education level

Abbreviations: CI = confidence interval

Appendix D. Table 11. Summary of traditional meta-analysis results for HUI cognition from CCHS-CLSA, Buschke total from CSHA and Buschke free from NuAge

Method of Statistical Harmonization	Adjustment	Summary Effect Estimate (95% CI)	I^2	Q[*]	p-value
Raw Data – Using Hedge's G	Unadjusted	0.12 (-0.004, 0.24)	0.72	7.01	0.03
	Unadjusted (Including Participants with Complete Data for Covariates)	0.12 (-0.001, 0.24)	0.70	6.04	0.049
	Study-Specific Statistically Significant Covariates	0.09 (-0.03, 0.21)	0.68	6.30	0.04
	Common Set of Covariates	0.09 (-0.02, 0.20)	0.60	4.97	0.08
T-Score – Weighted Mean Difference	Unadjusted[†]	0.97 (0.18, 1.76)	0.40	3.32	0.19
	Unadjusted[†] (Including Participants with Complete Data for Covariates)	0.99 (0.21, 1.77)	0.36	3.11	0.21
	Common Set of Covariates	0.91 (-0.24, 2.07)	0.61	5.18	0.08
C-Score – Weighted Mean Difference	Unadjusted	0.13 (0.03, 0.23)	0.66	5.81	0.06
	Unadjusted (Including Participants with Complete Data for Covariates)	0.13 (0.03, 0.23)	0.59	4.85	0.088
	Common Set of Covariates	0.09 (-0.002, 0.19)	0.55	4.43	0.11

[*] All Q Statistics have 2 degrees of freedom

[†] T-Scores minimally adjusted for age, sex and education level

Abbreviations: CI = confidence interval

Appendix D. Table 12. Summary of traditional meta-analysis results for HUI cognition from CCHS-CLSA and Buschke total from CSHA and NuAge

Method of Statistical Harmonization	Adjustment	Summary Effect Estimate (95% CI)	I^2	Q*	p-value
Raw Data – Using Hedge's G	Unadjusted	0.10 (-0.05, 0.24)	0.80	9.76	0.008
	Unadjusted (Including Participants with Complete Data for Covariates)	0.10 (-0.05, 0.24)	0.77	8.65	0.01
	Study-Specific Statistically Significant Covariates	0.08 (-0.05, 0.21)	0.73	7.32	0.03
	Common Set of Covariates	0.08 (-0.04, 0.20)	0.67	6.10	0.047
T-Score – Weighted Mean Difference	Unadjusted†	0.69 (-0.47, 1.84)	0.69	6.49	0.04
	Unadjusted† (Including Participants with Complete Data for Covariates)	0.71 (-0.44, 1.87)	0.67	6.07	0.048
	Common Set of Covariates	0.81 (-0.48, 2.10)	0.68	6.30	0.04
C-Score – Weighted Mean Difference	Unadjusted	0.12 (-0.006, 0.24)	0.74	7.73	0.02
	Unadjusted (Including Participants with Complete Data for Covariates)	0.12 (0.0001, 0.24)	0.70	6.62	0.04
	Common Set of Covariates	0.10 (0.001, 0.20)	0.53	4.28	0.12

*All Q Statistics have 2 degrees of freedom

†T-Scores minimally adjusted for age, sex and education level

Abbreviations: CI = confidence interval

Appendix E. Additional Regression Results

Appendix E. Table 1. Summary of relationship of T scores for cognitive measures with variables of interest

Variable	Description	CCHS		CSHA n=1730			NuAge	
		REY	HUI	REY	BUSC1	BUSC_T	BUSC1	BUSC_T
G_Age	Age in years (n)	7,060	7,058	1,488	1,488	1,488	432	432
	Intercept	50.0000	50.0000	50.0000	50.0000	50.0000	50.0000	50.0000
	Beta	0.0000	0.0000	-0.0000	-0.0000	0.0000	0.0000	0.0000
	R-square	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	p-value	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
G_129	Gender (n)	7,060	7,058	1,488	1,488	1,488	432	432
	Intercept	50.0000	50.0000	50.0000	50.0000	50.0000	50.0000	50.0000
	Beta (Ref: male)	-0.0000	-0.0000	0.0000	0.0000	0.0000	-0.0000	-0.00000
	R-square	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	p-value	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
G_161_C	Highest level of education (n)	7,060	7,060	1,488	1,488	1,488	432	432
	Intercept	50.0000	50.0000	49.9816	50.0795	49.9984	50.0940	50.2683
	Beta (Ref: Low (0-8)							
	Moderate (9-13)	0.0000	0.0000	0.0588	-0.2545	0.0053	-0.1570	-0.4481
	High (14+)	0.0000	0.0000	-0.0548	0.2372	-0.0049	-0.0573	-0.1635
	R-square	0.0000	0.0000	0.0000	0.0003	0.0000	0.0000	0.0003
	p-value	1.0000	1.0000	0.9892	0.8153	0.9999	0.9921	0.9375
G_297	Country of birth (n)	7,058	7,056	1,488	1,488	1,488	432	432
	Intercept	48.7473	49.6602	49.0965	49.8638	49.5637	50.8868	51.1381
	Beta (Ref: other country)	1.5457	0.4194	1.3566	0.2045	0.6551	-0.9899	-1.2705
	R-square	0.0036	0.0003	0.0041	0.0001	0.0010	0.0009	0.0015
	p-value	< 0.0001	0.1697	0.0135	0.7100	0.2334	0.5303	0.4205
G_940	Ever/current smoked cigarettes (n)	7,058	7,056	1,301*	1,301*	1,301	432	432
	Intercept	49.9776	50.4173	50.5273	50.5487	50.2103	49.7371	49.7214
	Beta (Ref: never)	0.0308	-0.6094	-0.7712	-0.9033	-0.2868	0.5824	0.6171
	R-square	0.000002	0.0008	0.0015	0.0020	0.0002	0.0008	0.0009
	p-value	0.9039	0.0170	0.1657	0.1071	0.6029	0.5475	0.5239
G_2526	Ever/current alcohol consumption (n)	7,060	7,058	1,289	1,289	1,289	432	432
	Intercept	48.2874	50.3179	50.2092	50.5104	50.3687	44.1332	47.9754
	Beta (Ref: never)	1.8590	-0.3451	0.2516	-1.1290	-1.1986	6.1665	2.1281
	R-square	0.0025	0.00009	0.0001	0.0022	0.0025	0.0176	0.0021
	p-value	<0.0001	0.4349	0.7083	0.0911	0.0726	0.0057	0.3421
G_355	Standing height in centimeter (n)	7,024	7,022	1,377	1,377	1,377	432	432
	Intercept	47.3332	50.5118	46.5496	47.9305	53.4555	41.2094	34.9620
	Beta	0.0161	-0.0031	0.0223	0.0136	-0.0204	0.0544	0.0931
	R-square	0.0002	0.00001	0.0005	0.0002	0.0004	0.0025	0.0074
	p-value	0.1895	0.8024	0.4151	0.6181	0.4534	0.2973	0.0742

Appendix E. Table 1. Summary of relationship of T scores for cognitive measures with variables of interest (cont'd)

Variable	Description	CCHS		CSHA n=1730			NuAge	
		REY	HUI	REY	BUSC1	BUSC_T	BUSC1	BUSC_T
G_217	Weight in Kg	6,956	6,954	1,420	1,420	1,420	432	432
	Intercept	49.1409	50.2085	49.2215	46.6716	49.0394	52.5501	47.3578
	Beta	0.0118	-0.0026	0.0123	0.0526	0.0158	-0.0351	0.0363
	R-square	0.0003	0.00002	0.0003	0.0053	0.0005	0.0026	0.0028
	p-value	0.1220	0.7346	0.5177	0.0058	0.4100	0.2907	0.2736
G_162	Body mass index (N)	6,924	6,922	1,357	1,357	1,357	432	432
	Intercept	49.5068	50.3061	50.3508	46.5906	48.8816	55.4833	49.9240
	Beta	0.0200	-0.0108	-0.0061	0.1481	0.0520	-0.1973	0.0027
	R-square	0.00009	0.00003	0.0000	0.0046	0.0006	0.0082	0.0000
	p-value	0.4228	0.6637	0.9186	0.0127	0.3816	0.0603	0.9793
G_326	Occurrence of high blood pressure(N)	7,058	7,055	1,349	1,349	1,349	432	432
	Intercept	50.1377	49.99	50.2537	49.7083	49.6610	50.429	49.7764
	Beta (Ref: never)	-0.2403	0.0318	-0.5059	1.0007	1.1146	-0.8869	0.4690
	R-square	0.0001	0.000002	0.0006	0.0024	0.0031	0.0020	0.0006
	p-value	0.3166	0.8945	0.3655	0.0734	0.0406	0.3578	0.6269
G_290	Occurrence of diabetes (N)	7,058	7,056	1,455	1,455	1,455	432	432
	Intercept	50.2658	50.1048	50.0894	49.9876	49.8359	49.8600	49.7602
	Beta (Ref: never)	-1.5012	-0.5819	-0.8708	0.1272	0.9099	1.5125	2.5900
	R-square	0.0033	0.0005	0.0009	0.0000	0.0009	0.0019	0.0056
	p-value	<0.0001	0.0621	0.2621	0.8702	0.2451	0.3628	0.1188
G_388	Occurrence of myocardial infarction (N)	7,046	7,043	1,431	1,431	1,431	424	424
	Intercept	50.1058	50.1037	50.1462	50.0846	49.9405	49.8041	50.0110
	Beta (Ref: never)	-0.7705	-0.7546	-0.6850	0.3375	1.3464	1.5824	-0.0490
	R-square	0.0006	0.0006	0.0006	0.0002	0.0026	0.0029	0.0000
	p-value	0.0333	0.0370	0.3447	0.6382	0.0543	0.2690	0.9725
G_245_CB	Categorical indicator of the participants level of physical activity using CSHA categories (3 categories) (N)	7060	7058	1271	1271	1271	432	432
	Intercept	48.9042	49.5623	49.2264	48.5627	49.1926	51.6882	51.1951
	Beta (Ref: none)							
	Low level	0.4616	-0.7936	1.5737	2.3309	1.7180	-1.9179	0.2914
	Moderate or high level	1.4369	0.7204	2.4643	3.6444	1.8030	-1.8856	-1.5525
	R-square	0.0032	0.0025	0.0127	0.0279	0.0077	0.0034	0.0048
	p-value	< 0.0001	0.0001	0.0003	< 0.0001	0.0074	0.4811	0.3531

Appendix E. Table 1. Summary of relationship of T scores for cognitive measures with variables of interest (cont'd)

Variable	Description	CCHS		CSHA n=1730			NuAge	
		REY	HUI	REY	BUSC1	BUSC_T	BUSC1	BUSC_T
G_245_CO	Categorical indicator of the participants level of physical activity using CSHA categories (4 categories) (N)	7,060	7,058	1,271	1,271	1,271	432	432
	Intercept	48.9042	49.5623	49.2264	48.5627	49.1926	51.6882	51.1951
	Beta (Ref: none)							
	Low level	0.4616	-0.7936	1.5737	2.3309	1.7180	-1.9179	0.2914
	Moderate level	1.2255	0.6285	2.2189	3.3708	1.5214	-1.3698	-1.4205
	High level	2.1000	1.0085	3.2841	4.5582	2.7435	-2.8115	-1.7894
	R-square	0.0043	0.0027	0.0134	0.0288	0.0086	0.0072	0.0051
	p-value	< 0.0001	0.0002	0.0007	< 0.0001	0.0118	0.3781	0.5346

Appendix E. Table 2. Summary of relationship of centered scores for cognitive measures with variables of interest

Variable	**Description**	**CCHS**		**CSHA n=1730**			**NuAge**	
		REY	**HUI**	**REY**	**BUSC1**	**BUSC_T**	**BUSC1**	**BUSC_T**
G_Age	Age in years (n)	7,106	7,103	1,497	1,730	1,730	432	432
	Intercept	3.6550	1.086	1.6300	2.5670	2.0236	2.3318	2.7491
	Beta	-0.0531	-0.0155	-0.0239	-0.0340	-0.0251	-0.0368	-0.0423
	R-square	0.0869	0.0073	0.0332	0.0615	0.0391	0.0184	0.0113
	p-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.0048	0.0269
G_129	Gender (n)	7,106	7,104	1,497	1,730	1,730	432	432
	Intercept	-0.4398	-0.0477	-0.3303	-0.1055	0.1260	-0.6811	-0.6629
	Beta (Ref: male)	0.3522	-0.0092	0.0967	-0.0590	-0.1701	0.5660	0.5537
	R-square	0.0272	0.00002	0.0027	0.0009	0.0087	0.0693	0.0310
	p-value	<0.0001	0.7203	0.0442	0.2133	0.0001	<0.0001	0.0002
G_161_C	Highest level of education (n)	7,060	7,058	1,488	1,719	1,719	432	432
	Intercept	-0.7151	-0.1701	-0.4317	-0.1169	0.0177	-0.5738	-0.6508
	Beta (Ref: Low (0-8)							
	Moderate (9-13)	0.4248	0.0912	0.2440	-0.0541	-0.0007	0.1798	0.2251
	High (14+)	0.7367	0.1883	0.4610	-0.0057	0.0342	0.2931	0.4668
	R-square	0.0660	0.0044	0.0313	0.0007	0.0002	0.0084	0.0110
	p-value	<0.0001	<0.0001	<0.0001	0.5271	0.8773	0.1632	0.0939
G_297	Country of birth (n)	7,104	7,102	1,497	1,730	1,730	432	432
	Intercept	-0.3347	-0.0770	-0.3763	-0.1980	-0.0388	-0.3535	-0.2498
	Beta (Ref: other country)	0.1211	0.0297	0.1596	0.0824	0.0863	-0.0264	-0.1292
	R-square	0.0020	0.00012	0.0070	0.0016	0.0021	0.0001	0.0006
	p-value	0.0002	0.3598	0.0012	0.0922	0.0571	0.8759	0.6018
G_940	Ever/current smoked cigarettes (n)	7,104	7,102	1,309*	1,521*	1,521*	432	432
	Intercept	-0.2212	-0.0226	-0.2327	-0.1505	-0.0302	-0.2887	-0.2912
	Beta (Ref: never)	-0.0227	-0.0441	-0.0421	0.0427	0.1244	-0.1960	-0.1649
	R-square	0.0001	0.0004	0.0005	0.0005	0.0050	0.0083	0.0027
	p-value	0.3987	0.1034	0.4024	0.3902	0.0059	0.0589	0.2779
G_2526	Ever/current alcohol consumption (n)	7,106	7,104	1,297	1,507	1,507	432	432
	Intercept	-0.4540	-0.0589	-0.2543	-0.1232	0.0145	-0.9818	-0.7094
	Beta (Ref: never)	0.2360	0.0065	0.0331	0.0020	0.0564	0.6356	0.3614
	R-square	0.0036	0.000003	0.0002	0.0000	0.0007	0.0162	0.0025
	p-value	<0.0001	0.8907	0.5853	0.9736	0.2982	0.0080	0.3041
G_355	Standing height in centimeter (n)	7,070	7,068	1,386	1,600	1,600	432	432
	Intercept	0.9935	-0.2234	-0.5207	-0.9502	-0.9520	1.7744	0.2221
	Beta	-0.0074	0.0010	0.0017	0.0051	0.0061	-0.0133	-0.0036
	R-square	0.0046	0.00009	0.0003	0.0028	0.0048	0.0131	0.0005
	p-value	<0.0001	0.4292	0.4994	0.0341	0.0055	0.0172	0.6574

Appendix E. Table 2. Summary of relationship of centered scores for cognitive measures with variables of interest (cont’d)

Variable	**Description**	**CCHS**		**CSHA n=1730**			**NuAge**	
		REY	**HUI**	**REY**	**BUSC1**	**BUSC_T**	**BUSC1**	**BUSC_T**
G_217	Weight in Kg	7,000	6,998	1,428	1,654	1,654	432	432
	Intercept	-0.1766	-0.1044	-0.3992	-0.6849	-0.4315	0.4269	-0.2449
	Beta	-0.0008	0.0007	0.0020	0.0085	0.0070	-0.0111	-0.0017
	R-square	0.0001	0.0001	0.0010	0.0154	0.0121	0.0224	0.0002
	p-value	0.3389	0.3780	0.2332	<0.0001	<0.0001	0.0018	0.7502
G_162	Body mass index (N)	6,968	6,966	1,365	1,576	1,576	432	432
	Intercept	-0.3991	-0.0702	-0.3295	-0.6260	-0.2715	0.2293	-0.3264
	Beta	0.0063	0.0007	0.0032	0.0208	0.0127	-0.0218	-0.0014
	R-square	0.0008	0.00001	0.0003	0.0101	0.0044	0.0087	0.0000
	p-value	0.0165	0.7784	0.5477	<0.0001	0.0081	0.0529	0.9320
G_326	Occurrence of high blood pressure(N)	7,104	7,102	1,358	1,561	1,561	432	432
	Intercept	-0.1928	-0.0414	-0.2530	-0.1601	0.0055	-0.3125	-0.3702
	Beta (Ref: never)	-0.0773	-0.0197	-0.0290	0.0616	0.0556	-0.1355	0.0096
	R-square	0.0013	0.00008	0.0002	0.0010	0.0010	0.0040	0.0000
	p-value	0.0022	0.4397	0.5650	0.2144	0.2168	0.1905	0.9494
G_290	Occurrence of diabetes (N)	7,104	7,102	1,464	1,691	1,691	432	432
	Intercept	-0.1962	-0.0391	-0.2633	-0.1452	0.0026	-0.3818	-0.3905
	Beta (Ref: never)	-0.2258	-0.0769	-0.0590	0.0118	0.0918	0.0508	0.2686
	R-square	0.0067	0.0008	0.0005	0.0000	0.0012	0.0002	0.0025
	p-value	<0.0001	0.0202	0.3994	0.8624	0.1484	0.7758	0.3032
G_388	Occurrence of myocardial infarction (N)	7,092	7,090	1,440	1,665	1,665	424	424
	Intercept	-0.1993	-0.0374	-0.2535	-0.1390	0.0116	-0.4033	-0.3721
	Beta (Ref: never)	-0.2927	-0.1172	-0.0768	0.0456	0.1173	0.2247	0.0787
	R-square	0.0083	0.0013	0.0010	0.0003	0.0025	0.0051	0.0003
	p-value	<0.0001	0.0023	0.2391	0.4740	0.0434	0.1436	0.7234
G_245_CB	Categorical indicator of the participants level of physical activity using CSHA categories (3 categories) (N)	7,106	7,104	1,279	1,484	1,484	432	432
	Intercept	-0.4286	-0.1310	-0.3772	-0.2937	-0.0717	-0.2695	-0.2689
	Beta (Ref: never)							0.2790
	Low level	0.1079	-0.0670	0.2251	0.2348	0.1629	-0.0041	-0.1593
	Moderate or high level	0.2480	0.1174	0.2946	0.3989	0.2256	-0.1358	
								0.0076
	R-square	0.0082	0.0039	0.0231	0.0357	0.0141	0.0026	0.1928
	p-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.5737	

Appendix E. Table 2. Summary of relationship of centered scores for cognitive measures with variables of interest (cont'd)

Variable	Description	CCHS		CSHA n=1730			NuAge	
		REY	HUI	REY	BUSC1	BUSC_T	BUSC1	BUSC_T
G_245_CO	Categorical indicator of the participants level of physical activity using CSHA categories (4 categories) (N)	7,106	7,104	1,279	1,484	1,484	432	432
	Intercept	-0.4286	-0.1310	-0.3772	-0.2937	-0.0717	-0.2695	-0.2689
	Beta (Ref: none)							
	Low level	0.1079	-0.0670	0.2251	0.2348	0.1629	-0.0041	0.2790
	Moderate level	0.2077	0.1005	0.2619	0.3530	0.1810	-0.0728	-0.1338
	High level	0.3742	0.1706	0.4044	0.5512	0.3736	-0.2487	-0.2051
	R-square	0.0115	0.0045	0.0246	0.0382	0.0169	0.0075	0.0080
	p-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.3602	0.3274

Appendix F. Summary of Two-Stage IPD Meta-Analysis

Appendix F. Summary of two stage IPD meta-analysis

Selection of memory tests	**Type of outcome**	**Ranges of values from the analysis with different selections of covariates**			
		Effect size	**I^2**	**Q**	**p-value**
CCHS:REY **CSHA:REY** **NuAge: Buschke Free**	Raw	1.71 – 2.12	0.72 – 0.74	7.13 – 7.62	0.022 – 0.028
	T-score	1.09-1.20	0.64-0.71	5.58-6.91	0.032-0.061
	C-score	0.13-0.16	0.74-0.75	7.75-8.16	0.017-0.021
CCHS:REY **CSHA:REY** **NuAge: Buschke Total**	Raw	1.69-1.72	0.78-0.82	8.95-11.33	0.003-0.011
	T-score	0.96-1.03	0.74-0.76	7.62-8.46	0.015-0.022
	C-score	0.15-0.16	0.69-0.75	6.49-8.10	0.017-0.039
CCHS:REY **CSHA:Buschke Free** **NuAge: Buschke Free**	Raw	2.44-3.05	0.90-0.92	20.96-26.76	<0.001
	T-score	1.36-1.40	0.85-0.87	12.97-15.48	<0.001-0.002
	C-score	0.15-0.18	0.84-0.86	12.46-14.34	0.001-0.002
CCHS:REY **CSHA:Buschke Free** **NuAge: Buschke Total**	Raw	2.37-2.78	0.91-0.94	22.37-31.14	<0.001
	T-score	1.30-1.74	0.88-0.88	16.03-16.23	<0.001
	C-score	0.16-0.18	0.81-0.86	10.50-14.28	0.001-0.005
CCHS:REY **CSHA:Buschke Total** **NuAge: Buschke Free**	Raw	1.46-2.04	0.68-0.73	6.32-7.34	0.025-0.043
	T-score	0.95-1.06	0.43-0.63	3.48-3.57	0.068-0.175
	C-score	0.11-0.15	0.71-0.71	6.84-6.93	0.031-0.033
CCHS:REY **CSHA:Buschke Total** **NuAge: Buschke Total**	Raw	1.32-1.63	0.74-0.82	7.58-11.38	0.023-0.003
	T-score	0.79-0.88	0.67-0.68	6.06-6.24	0.044-0.048
	C-score	0.12-0.15	0.64-0.73	5.62-7.31	0.026-0.060
CCHS:HUI **CSHA:REY** **NuAge: Buschke Free**	Raw	1.48-1.82	0.74-0.75	7.59-7.96	0.019-0.023
	T-score	0.94-1.08	0.67-0.72	6.03-7.12	0.029-0.049
	C-score	0.11-0.13	0.76-0.77	8.29-8.54	0.014-0.016
CCHS:HUI **CSHA:REY** **NuAge: Buschke Total**	Raw	1.36-1.45	0.78-0.82	9.13-11.38	0.003-0.010
	T-score	0.85-0.89	0.74-0.77	7.77-8.71	0.013-0.021
	C-score	0.12-0.13	0.73-0.78	7.51-9.28	0.010-0.023
CCHS:HUI **CSHA:Buschke Free** **NuAge: Buschke Free**	Raw	2.28-2.85	0.91-0.93	22.13-28.99	<0.001
	T-score	1.22-1.30	0.86-0.88	14.47-17.30	<0.001-0.001
	C-score	0.12-0.15	0.86-0.89	14.72-18.36	<0.001-0.001
CCHS:HUI **CSHA:Buschke Free** **NuAge: Buschke Total**	Raw	2.22-2.59	0.91-0.94	23.38-32.89	<0.001
	T-score	1.07-1.17	0.89-0.89	17.33-18.00	<0.001
	C-score	0.13-0.14	0.85-0.89	13.40-18.89	<0.001-0.001
CCHS:HUI **CSHA:Buschke Total** **NuAge: Buschke Free**	Raw	1.28-1.77	0.65-0.73	5.77-7.53	0.023-0.056
	T-score	0.83-0.95	0.39-0.59	3.25-4.84	0.089-0.197
	C-score	0.09-0.13	0.61-0.66	5.11-5.81	0.055-0.078
CCHS:HIU **CSHA:Buschke Total** **NuAge: Buschke Total**	Raw	1.15-1.41	0.71-0.82	6.84-11.06	0.004-0.033
	T-score	0.71-0.77	0.64-0.66	5.48-5.81	0.055-0.065
	C-score	0.10-0.12	0.54-0.70	4.36-6.67	0.036-0.113
CCHS:HIU **CSHA:Buschke Total** **NuAge: Buschke Total**	Latent variable	0.08-0.08	.89-.89	17.76-17.83	<0.001

Appendix G. Summary of One-Stage IDP Meta-Analysis

Appendix G. Summary of one stage IPD meta-analysis

Variable in equation	Unified raw score		t-score		c-score	
	Unadjusted	Adjusted*	Selected	Adjusted*	Selected	Adjusted*
	Coeff. (95% CI)	Coeff. (95% CI)	Coeff. (95% CI)	Coeff. (95% CI)	Coeff. (95% CI)	Coeff. (95% CI)
CCHS-Rey CSHA-Rey NuAge-Buschke free	2.48 (1.88, 3.09)	2.20 (1.61, 2.78)	1.23 (0.77, 1.69)	1.36 (0.88, 1.84)	0.20 (0.15, 0.24)	0.18 (0.13, 0.22)
CCHS-Rey CSHA-Rey NuAge-Buschke total	2.45 (1.85, 3.05)	2.17 (1.59, 2.76)	1.20 (0.74, 1.65)	1.34 (0.86, 1.82)	0.19 (0.14, 0.24)	0.17 (0.12, 0.22)
CCHS-Rey CSHA-Buschke free NuAge-Buschke free	3.34 (2.65, 4.03)	2.96 (2.29, 3.63)	1.38 (0.93, 1.84)	1.56 (1.08, 2.04)	0.14 (0.17, 0.26)	0.19 (0.15, 0.24)
CCHS-Rey CSHA-Buschke free NuAge-Buschke total	3.32 (2.63, 4.01)	2.94 (2.27, 3.61)	1.35 (0.89, 1.80)	1.54 (1.06, 2.02)	0.21 (0.16, 0.26)	0.19 (0.14, 0.24)
CCHS-Rey CSHA-Buschke total NuAge-Buschke free	2.53 (1.90, 3.16)	2.17 (1.55, 2.78)	1.09 (0.64, 1.53)	1.24 (0.76, 1.72)	0.19 (0.14, 0.24)	0.16 (0.12, 0.21)
CCHS-Rey CSHA-Buschke total NuAge-Buschke total	2.51 (1.88, 3.13)	2.15 (1.54, 2.76)	1.06 (0.61, 1.51)	1.22 (0.74, 1.70)	0.18 (0.13, 0.23)	0.16 (0.11, 0.21)
CCHS-HUI CSHA-Rey NuAge-Buschke free	2.04 (1.36, 2.71)	1.77 (1.08, 2.46)	1.05 (0.59, 1.50)	1.06 (0.58, 1.53)	0.15 (0.10, 0.20)	0.13 (0.08, 0.18)
CCHS-HUI CSHA-Rey NuAge-Buschke total	2.01 (1.34, 2.68)	1.75 (1.07, 2.44)	1.01 (0.56, 1.47)	1.04 (0.57, 1.52)	0.14 (0.09, 0.19)	0.12 (0.07, 0.17)
CCHS-HUI CSHA-Buschke free NuAge-Buschke free	2.91 (2.16, 3.66)	2.54 (1.78, 3.30)	1.16 (0.72, 1.61)	1.29 (0.81, 1.76)	0.17 (0.12, 0.22)	0.15 (0.10, 0.20)
CCHS-HUI CSHA-Buschke free NuAge-Buschke total	2.89 (2.14, 3.63)	2.53 (1.77, 3.28)	1.13 (0.69, 1.58)	1.27 (0.80, 1.75)	0.16 (0.11, 0.21)	0.14 (0.09, 0.19)
CCHS-HUI CSHA-Buschke total NuAge-Buschke free	2.09 (1.40, 2.79)	1.75 (1.04, 2.46)	0.93 (0.48, 1.37)	1.01 (0.53, 1.48)	0.14 (0.09, 0.19)	0.12 (0.07, 0.17)
CCHS-HUI CSHA-Buschke total NuAge-Buschke total	2.06 (1.38, 2.75)	1.73 (1.02, 2.43)	0.90 (0.45, 1.35)	1.00 (0.52, 1.47)	0.13 (0.08, 0.18)	0.11 (0.06, 0.16)

Latent variable: unadjusted: 0.088 (0.063, 0.113)
selected: 0.088 (0.063, 0.113)
adjusted: 0.095 (0.069, 0.121)

* Adjusted for Age, Gender, weight, height and alcohol