PNAS

Main Manuscript for

Bridging Big Data: Procedures for Combining Non-equivalent Cognitive Measures from the ENIGMA Consortium

Eamonn Kennedy[1-3]*, Shashank Vadlamani[1], Hannah M Lindsey[1,3], Pui-Wa Lei[4], Mary Jo-Pugh[1,2], Maheen Adamson[5,6], Martin Alda[7], Silvia Alonso-Lana[8,9], Sonia Ambrogi[10], Tim J Anderson[11-13], Celso Arango[14], Robert F Asarnow[15-17], Mihai Avram[18], Rosa Ayesa-Arriola[9,19], Talin Babikian[15,20], Nerisa Banaj[10], Laura J Bird[21], Stefan Borgwardt[18,22], Amy Brodtmann[23,24], Katharina Brosch[25], Karen Caeyenberghs[26], Vince D Calhoun[27], Nancy D Chiaravalloti[28,29], David X Cifu[30], Benedicto Crespo-Facorro[9,31], John C Dalrymple-Alford[11,12,32], Kristen Dams-O'Connor[33,34], Udo Dannlowski[35], David Darby[36-38], Nicholas Davenport[39,40], John DeLuca[29,41], Covadonga M Diaz-Caneja[14], Seth G Disner[39,40], Ekaterina Dobryakova[29,42], Stefan Ehrlich[43,44], Carrie Esopenko[33], Fabio Ferrarell[45]i, Lea E Frank[46], Carol Franz[47,48], Paola Fuentes-Claramonte[8,9], Helen Genova[29,49], Christopher C Giza[20,50,51], Janik Goltermann[35], Dominik Grotegerd[35], Marius Gruber[35,52], Alfonso Gutierrez-Zotes[9,53,54], Minji Ha[55], Jan Haavik[56,57], Charles Hinkin[15], Kristen R Hoskinson[58,59], Daniela Hubl[60], Andrei Irimia[61-63], Andreas Jansen[25], Michael Kaess[64,65], Xiaojian Kang[5], Kimbra Kenney[66,67], Barbora Keřková, Mohamed Salah Khlif[23], Minah Kim[68,69], Jochen Kindler[64], Tilo Kircher[25], Karolina Knížková[70,71], Knut K Kolskår[72-74], Denise Krch[29,42], William S Kremen[47,48], Taylor Kuhn[15], Veena Kumari[75], Jun Soo Kwon[55,68,69], Roberto Langella[10], Sarah Laskowitz[76], Jungha Lee[55], Jean Lengenfelder[29,42], Spencer W Liebel[1,3], Victoria Liou-Johnson[5], Sara M Lippa[67,77], Marianne Løvstad[73,74], Astri Lundervold[78], Cassandra Marotta[36,37], Craig A Marquardt[39,40], Paulo Mattos[79], Ahmad Mayeli[45], Carrie R McDonald[80,81], Susanne Meinert[35,82], Tracy R Melzer[11,12,32], Jessica Merchán-Naranjo[14], Chantal Michel[64], Rajendra A Morey[76,83], Benson Mwangi[84], Daniel J Myall[12], Igor Nenadić[25], Mary R Newsome[85,86], Abraham Nunes[7,87], Terence O'Brien[88,89], Viola Oertel[90], John Ollinger[67], Alexander Olsen[91-93], Victor Ortiz García de la Foz[19], Mustafa Ozmen[94], Heath Pardoe[38], Marise Parent[95], Fabrizio Piras[10], Federica Piras[10], Edith Pomarol-Clotet[9], Jonathan Repple[35,52], Geneviève Richard[72], Jonathan Rodriguez[47], Mabel Rodriguez[70], Kelly Rootes-Murdy[27], Jared Rowland[96-98], Nicholas P Ryan[26,99], Raymond Salvador[9], Anne-Marthe Sanders[72-74], Andre Schmidt[100], Jair C Soares[84], Gianfranco Spalleta[10], Filip Španiel[70,101], Alena Stasenko[47,81], Frederike Stein[25], Benjamin Straube[25], April Thames[15], Florian Thomas-Odenthal[25], Sophia I Thomopoulos[102], Erin Tone[103], Ivan Torres[104,105], Maya Troyanskaya[85,86], Jessica A Turner[106], Kristine M Ulrichsen[72-74], Guillermo Umpierrez[107], Elisabet Vilella[9,53,54], Lucy Vivash[36,37], William C Walker[108,109], Emilio Werden[38], Lars T Westlye[72,73,110], Krista Wild[111], Adrian Wroblewski[25], Mon-Ju Wu[84], Glenn R Wylie[29,112], Lakshmi N Yatham[104], Giovana B Zunta-Soares[84], Paul M Thompson[102,113], David F Tate[1,3], Frank G Hillary[114-116], Emily L Dennis[1,3#], Elisabeth A Wilde[1,3#]

[#] joint contribution

1 Department of Neurology, University of Utah School of Medicine, Salt Lake City, UT, 84132
2 Division of Epidemiology, University of Utah, Salt Lake City, UT, 84132
3 George E. Wahlen Veterans Affairs Medical Center, Salt Lake City, UT, 84132
4 Department of Educational Psychology, Counseling, and Special Education, Pennsylvania State University, University Park, PA, 16801
5 WRIISC-WOMEN & Rehabilitation Department, VA Palo Alto, Palo Alto, CA, USA.
6 Neurosurgery, Stanford School of Medicine, Stanford, CA, USA
7 Department of Psychiatry, Dalhousie University, Halifax, Canada.
8 FIDMAG Research Foundation, Barcelona, Spain.
9 Centro Investigación Biomédica en Red Salud Mental (CIBERSAM), Madrid, Spain.
10 Laboratory of Neuropsychiatry, Santa Lucia Foundation IRCCS, Rome, Italy.
11 Department of Medicine, University of Otago, Christchurch, New Zealand.
12 New Zealand Brain Research Institute, Christchurch, New Zealand.
13 Department of Neurology, Te Whatu Ora – Health New Zealand Waitaha Canterbury, Christchurch, New Zealand.
14 Department of Child and Adolescent Psychiatry, Institute of Psychiatry and Mental Health, Hospital General Universitario Gregorio Marañón, IiSGM, CIBERSAM, School of Medicine, Universidad Complutense, Madrid, Spain.
15 Department of Psychiatry and Biobehavioral Sciences, Semel Institute for Neuroscience and Human Behavior, UCLA, Los Angeles, CA, USA
16 Brain Research Institute, UCLA, Los Angeles, CA, USA
17 Department of Psychology, UCLA, Los Angeles, CA, USA
18 Translational Psychiatry, Department of Psychiatry and Psychotherapy, University of Lübeck, Lübeck, Germany.
19 Department of Psychiatry, Marqués de Valdecilla University Hospital, IDIVAL, School of Medicine, University of Cantabria, Santander, Spain.
20 UCLA Steve Tisch BrainSPORT Program, Los Angeles, CA, USA
21 Turner Institute for Brain and Mental Health, School of Psychological Sciences, Monash University, Clayton, Melbourne, VIC, Australia.
22 Center of Brain, Behaviour and Metabolism (CBBM), University of Lübeck, Lübeck, Germany.
23 Cognitive Health Initiative, Central Clinical School, Monash University, Melbourne, VIC, Australia.
24 Department of Medicine, Royal Melbourne Hospital, Melbourne, VIC, Australia.
25 Department of Psychiatry and Psychotherapy, University of Marburg, Marburg, Germany.
26 Cognitive Neuroscience Unit, School of Psychology, Deakin University, Geelong, Australia
27 Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State University, Georgia Institute of Technology, and Emory University University, Atlanta, GA, USA.
28 Centers for Neuropsychology, Neuroscience & Traumatic Brain Injury Research, Kessler Foundation, East Hanover, NJ, USA.
29 Department of Physical Medicine & Rehabilitation, Rutgers, New Jersey Medical School, Newark, NJ, USA.
30 Rehabilitation Medicine Department, National Institutes of Health Clinical Center, Bethesda, MD.
31 Department of Psychiatry, Virgen del Rocio University Hospital, School of Medicine, University of Seville, IBIS, Seville, Spain.

32 School of Psychology, Speech and Hearing, University of Canterbury, Christchurch, New Zealand.
33 Department of Rehabilitation and Human Performance, Icahn School of Medicine at Mount Sinai, New York, NY.
34 Department of Neurology, Icahn School of Medicine at Mount Sinai, New York, NY.
35 Institute for Translational Psychiatry, University of Münster, Germany.
36 Department of Neuroscience, Monash University, Melbourne, Australia.
37 Department of Neurology, Alfred Health, Melbourne, Australia.
38 The Florey Institute of Neuroscience and Mental Health, Melbourne, Australia.
39 Department of Psychiatry and Behavioral Sciences, University of Minnesota Medical School, Minneapolis, MN.
40 Minneapolis VA Health Care System, Minneapolis, MN.
41 Kessler Foundation, East Hanover, NJ, USA.
42 Center for Traumatic Brain Injury, Kessler Foundation, East Hanover, NJ.
43 Translational Developmental Neuroscience Section, Division of Psychological and Social Medicine and Developmental Neurosciences, Faculty of Medicine, TU Dresden, Germany.
44 Eating Disorders Research and Treatment Center, Department of Child and Adolescent Psychiatry, Faculty of Medicine, TU Dresden, Germany.
45 Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA.
46 Department of Psychology, University of Oregon, Eugene, OR, USA.
47 Department of Psychiatry, University of California, San Diego, La Jolla, CA, USA
48 Center for Behavior Genetics of Aging, University of California, San Diego, La Jolla, CA, USA
49 Center for Autism Research, Kessler Foundation, East Hanover, NJ, USA.
50 Department of Pediatrics, Division of Neurology, UCLA Mattel Children's Hospital, Los Angeles, CA, USA
51 Department of Neurosurgery, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA
52 Department of Psychiatry, Psychosomatic Medicine and Psychotherapy, University Hospital Frankfurt, Goethe University, Frankfurt, Germany.
53 Hospital Universitari Institut Pere Mata, Tarragona, Spain.
54 Institut d'Investiació Sanitària Pere Virgili-CERCA, Universitat Rovira i Virgili, Tarragona, Spain.
55 Department of Brain and Cognitive Sciences, Seoul National University College of Natural Sciences, Seoul, South Korea.
56 Department of Biomedicine, University of Bergen, Bergen, Norway.
57 Division of Psychiatry, Haukeland University Hospital, Bergen, Norway.
58 Center for Biobehavioral Health, The Abigail Wexner Research Institute at Nationwide Children's Hospital, Columbus, OH.
59 Section of Pediatrics, The Ohio State University College of Medicine, Columbus, Ohio, USA.
60 Translational Research Centre, University Hospital of Psychiatry and Psychotherapy, University of Bern, Bern, Switzerland.
61 Ethel Percy Andrus Gerontology Center, Leonard Davis School of Gerontology, University of Southern California, Los Angeles, CA, USA
62 Department of Biomedical Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, CA, USA

63 Department of Quantitative & Computational Biology, Dornsife College of Arts & Sciences, University of Southern California, Los Angeles CA, USA.
64 University Hospital of Child and Adolescent Psychiatry and Psychotherapy, University of Bern, Bern, Switzerland.
65 Clinic of Child and Adolescent Psychiatry, Centre of Psychosocial Medicine, University of Heidelberg, Heidelberg, Germany.
66 Department of Neurology, Uniformed Services University, Bethesda, MD.
67 National Intrepid Center of Excellence, Walter Reed National Military Medical Center, Bethesda, MD.
68 Department of Neuropsychiatry, Seoul National University Hospital, Seoul, South Korea.
69 Department of Psychiatry, Seoul National University College of Medicine, Seoul, South Korea.
70 National Institute of Mental Health, Klecany, Czech Republic.
71 Department of Psychiatry, First Faculty of Medicine, Charles University and General University Hospital, Prague, Czech Republic.
72 NORMENT, Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway.
73 Department of Psychology, University of Oslo, Oslo, Norway.
74 Sunnaas Rehabilitation Hospital, Nesodden, Norway.
75 Department of Life Sciences, College of Health, Medicine and Life Sciences, Brunel University London, Uxbridge, United Kingdom.
76 Brain Imaging and Analysis Center, Duke University, Durham, NC.
77 Department of Neuroscience, Uniformed Services University of the Health Sciences, Bethesda, MD, USA.
78 Department of Biological and Medical Psychology, University of Bergen, Bergen, Norway.
79 Institute D'Or for Research and Education (IDOR), São Paulo, Brazil.
80 Department of Radiation Medicine and Applied Sciences and Psychiatry, UC San Diego, La Jolla, CA, USA.
81 Center for Multimodal Imaging and Genetics, UC San Diego, San Diego, CA, USA.
82 Institute for Translational Neuroscience, University of Münster, Münster, Germany
83 VISN 6 MIRECC, Durham VA, Durham, NC.
84 Center of Excellence on Mood Disorders, Louis A. Faillace, MD, Department of Psychiatry and Behavioral Sciences, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX, USA.
85 Michael E. DeBakey Veterans Affairs Medical Center, Houston, TX.
86 H. Ben Taub Department of Physical Medicine and Rehabilitation, Baylor College of Medicine, Houston, TX.
87 Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada.
88 Department of Medicine, The Royal Melbourne Hospital, The University of Melbourne, Parkville, VIC, Australia.
89 Department of Neuroscience, The Central Clinical School, Alfred Health, Monash University, Melbourne, VIC, Australia.
90 Department of Psychiatry, Psychosomatic Medicine and Psychotherapiey, Frankfurt University, Frankfurt, Germany.
91 Department of Psychology, Norwegian University of Science and Technology, Trondheim, Norway.
92 Department of Physical Medicine and Rehabilitation, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway.
93 NorHEAD - Norwegian Centre for Headache Research, Trondheim, Norway.

94 Department of Electrical and Electronics Engineering, Antalya Bilim University, Antalya, Turkey.
95 Neuroscience Institute & Department of Psychology, Georgia State University, Atlanta, GA, USA.
96 W.G. (Bill) Hefner VA Medical Center, Salisbury, NC.
97 Department of Neurobiology & Anatomy, Wake Forest School of Medicine, Winston-Salem, NC.
98 VA Mid-Atlantic Mental Illness Research Education and Clinical Center (MA-MIRECC), Durham, NC.
99 Department of Paediatrics, The University of Melbourne, Parkville, VIC, Australia.
100 University of Basel, Department of Psychiatry (UPK), Basel, Switzerland.
101 3rd Faculty of Medicine Charles University, Prague, Czech Republic.
102 Imaging Genetics Center, Stevens Neuroimaging & Informatics Institute, Keck School of Medicine of USC, Marina del Rey, CA, 90007
103 Department of Psychology, Georgia State University, Atlanta, GA.
104 Department of Psychiatry, University of British Columbia, Vancouver, BC, Canada.
105 British Columbia Mental Health and Substance Use Services Research Institute, Vancouver, BC, Canada.
106 Psychiatry and Behavioral Health, Ohio State Wexner Medical Center, Columbus, OH, USA.
107 Division of Endocrinology, Emory University School of Medicine, Atlanta, GA, USA.
108 Department of Physical Medicine and Rehabilitation, Virginia Commonwealth University, Richmond, VA.
109 Richmond Veterans Affairs (VA) Medical Center, Central Virginia VA Health Care System , Richmond, VA.
110 KG Jebsen Center for Neurodevelopmental Disorders, University of Oslo, Oslo, Norway.
111 Department of Psychology, Phoenix VA Health Care System, Phoenix, AZ, USA.
112 Rocco Ortenzio Neuroimaging Center, Kessler Foundation, East Hanover, NJ, USA.
113 Departments of Neurology, Pediatrics, Psychiatry, Radiology, Engineering, and Ophthalmology, USC, Los Angeles, CA, 90007
114 Department of Psychology, Penn State University, State College, PA, 16801
115 Department of Neurology, Hershey Medical Center, State College, PA, 16801
116 Social Life and Engineering Science Imaging Center, Penn State University, State College, PA, 16801

Corresponding Author: Eamonn Kennedy

Email: eamonn.kennedy@utah.edu

Author Contributions:

Designed research: Eamonn Kennedy, Hannah Lindsey, Pui-Wa Lei, Paul M Thompson, David F Tate, Frank G Hillary, Emily L Dennis, Elisabeth A Wilde

Performed research: Eamonn Kennedy, Shashank Vadlamani, Hannah M Lindsey, Emily L Dennis, Elisabeth A Wilde

Contributed data: Maheen Adamson, Martin Alda, Silvia Alonso-Lana, Sonia Ambrogi, Tim J Anderson, Celso Arango, Robert F Asarnow, Mihai Avram, Rosa Ayesa-Arriola, Talin Babikian,

Nerisa Banaj, Laura J Bird, Stefan Borgwardt, Amy Brodtmann, Katharina Brosch, Karen Caeyenberghs, Vince D Calhoun, Nancy D Chiaravalloti, David X Cifu, Benedicto Crespo-Facorro, John C Dalrymple-Alford, Kristen Dams-O'Connor, Udo Dannlowski, David Darby, Nicholas Davenport, John DeLuca, Covadonga M Diaz-Caneja, Seth G Disner, Ekaterina Dobryakova, Stefan Ehrlich, Carrie Esopenko, Fabio Ferrarelli, Lea E Frank, Carol Franz, Paola Fuentes-Claramonte, Helen Genova, Christopher C Giza, Janik Goltermann, Dominik Grotegerd, Marius Gruber, Alfonso Gutierrez-Zotes, Minji Ha, Jan Haavik, Charles Hinkin, Kristen R Hoskinson, Daniela Hubl, Andrei Irimia, Andreas Jansen, Michael Kaess, Xiaojian Kang, Kimbra Kenney, Barbora Keřková, Mohamed Salah Khlif, Minah Kim, Jochen Kindler, Tilo Kircher, Karolina Knížková, Knut K Kolskår, Denise Krch, William S Kremen, Taylor Kuhn, Veena Kumari, Jun Soo Kwon, Roberto Langella, Sarah Laskowitz, Jungha Lee, Jean Lengenfelder, Spencer W Liebel, Victoria Liou-Johnson, Sara M Lippa, Marianne Løvstad, Astri Lundervold, Cassandra Marotta, Craig A Marquardt, Paulo Mattos, Ahmad Mayeli, Carrie R McDonald, Susanne Meinert, Tracy R Melzer, Jessica Merchán-Naranjo, Chantal Michel, Rajendra A Morey, Benson Mwangi, Daniel J Myall, Igor Nenadić, Mary R Newsome, Abraham Nunes, Terence O'Brien, Viola Oertel, John Ollinger, Alexander Olsen, Victor Ortiz García de la Foz, Mustafa Ozmen, Heath Pardoe, Marise Parent, Fabrizio Piras, Federica Piras, Edith Pomarol-Clotet, Jonathan Repple, Geneviève Richard, Jonathan Rodriguez, Mabel Rodriguez, Kelly Rootes-Murdy, Jared Rowland, Nicholas P Ryan, Raymond Salvador, Anne-Marthe Sanders, Andre Schmidt, Jair C Soares, Gianfranco Spalleta, Filip Španiel, Alena Stasenko, Frederike Stein, Benjamin Straube, April Thames, Florian Thomas-Odenthal, Sophia I Thomopoulos, Erin Tone, Ivan Torres, Maya Troyanskaya, Jessica A Turner, Kristine M Ulrichsen, Guillermo Umpierrez, Elisabet Vilella, Lucy Vivash, William C Walker, Emilio Werden, Lars T Westlye, Krista Wild, Adrian Wroblewski, Mon-Ju Wu, Glenn R Wylie, Lakshmi N Yatham, Giovana B Zunta-Soares,

Analyzed data: Eamonn Kennedy, Shashank Vadlamani, Pui-Wa Lei

Wrote paper: Eamonn Kennedy

Competing Interest Statement: Dr. Arango has been a consultant to or has received honoraria or grants from Acadia, Angelini, Biogen, Boehringer, Gedeon Richter, Janssen Cilag, Lundbeck, Medscape, Menarini, Minerva, Otsuka, Pfizer, Roche, Sage, Servier, Shire, Schering Plough, Sumitomo Dainippon Pharma, Sunovion and Takeda. Dr. Brodtmann serves on the editorial boards of Neurology and International Journal of Stroke. Dr. Diaz-Caneja has received honoraria from Exeltis and Angelinii. Dr. Giza: consultant for NBA, NFL, NHLPA, Los Angeles Lakers; Advisory Board: Highmark Interactive, Novartis, MLS, NBA, USSF; Medicolegal 1-2 cases annually. Dr. Soares: ALKERMES (Research Grant), ALLERGAN (Research Grant), ASOFARMA (Consultant), ATAI (Stock), BOEHRINGER Ingelheim (Consultant), COMPASS (Research Grant), JOHNSON & JOHNSON (Consultant), LIVANOVA (Consultant), PFIZER (Consultant), PULVINAR NEURO LLC (Consultant), RELMADA (Consultant), SANOFI (Consultant), SUNOVIAN (Consultant). Dr. Thompson received partial research support from Biogen, Inc., for research unrelated to this manuscript. Dr. Yatham has been on speaker or advisory boards for, or has received research grants from, Alkermes, Abbvie, Canadian Institutes of Health Research, Sumitomo Dainippon Pharma, GlaxoSmithKline, Intracellular Therapies, Merck, Sanofi, Sequiris, Servier, and Sunovion, over the past 3 years, all outside this work. The collection of this cohort was partially supported by an investigator-initiated research grant from Biogen (US). Biogen had no role in the analysis or writing of this manuscript. Eisai (JP) and Life Molecular Imaging for research unrelated to this manuscript. Dr. Wylie has received research support from the NJ Commission for brain injury research, from the Dept of Veterans' Affairs, from Biogen, from

Bristol, Myers, Squibb, from Genetech, and has served on advisory boards for the CDMRP and the VA. All of these activities are unrelated to this research. The views expressed in this article are those of the author(s) and do not reflect the official policy of the Department of Army/Navy/Air Force, Department of Defense, or U.S. Government.

Classification: Major: Social. Minor: Psychological and Cognitive Sciences

Keywords Harmonization, Verbal learning, Mega analysis, Tools

Abstract

Investigators in the cognitive neurosciences have turned to Big Data to address persistent replication and reliability issues by increasing sample sizes, statistical power, and representativeness of data. While there is tremendous potential to advance science through open data sharing, these efforts unveil a host of new questions about how to integrate data arising from distinct sources and instruments. We focus on the most frequently assessed area of cognition - memory testing - and demonstrate a process for reliable data harmonization across three common measures. We aggregated raw data from 53 studies from around the world which measured at least one of three distinct verbal learning tasks, totaling N = 10,505 healthy and brain-injured individuals. A mega analysis was conducted using empirical bayes harmonization to isolate and remove site effects, followed by linear models which adjusted for common covariates. After corrections, a continuous item response theory (IRT) model estimated each individual subject's latent verbal learning ability while accounting for item difficulties. Harmonization significantly reduced inter-site variance by 37% while preserving covariate effects. The effects of age, sex, and education on scores were found to be highly consistent across memory tests. IRT methods for equating scores across AVLTs agreed with held-out data of dually-administered tests, and these tools are made available for free online. This work demonstrates that large-scale data sharing and harmonization initiatives can offer opportunities to address reproducibility and integration challenges across the behavioral sciences.

Significance Statement

Data sharing can increase the quality and rigor of scientific claims, but with large scale data sharing comes an increased need to combine non-equivalent measurements across studies. Auditory verbal learning tasks (AVLTs) are one of the most common research and clinical tools to evaluate memory constructs, but numerous distinct AVLTs are in common use, which creates challenges for data aggregation. We report methods to convert raw scores across common verbal learning instruments, constructed using harmonizing data from 53 studies from around the world. This approach can be replicated in other domains to address long standing data compatibility issues for researchers and clinicians.

Main Text

Introduction

Data sharing consortia aim to increase the robustness and statistical power of results by aggregating large and diverse samples (1, 2). While analyses of large datasets can provide unparalleled statistical power, data aggregation without robust harmonization can mask and even introduce flaws and biases (3). This is a critical consideration in the behavioral sciences, where multi-site collaboration often requires the synthesis of non-identical cognitive measures (4–6). For example, verbal memory/recall is a core cognitive function, and deficits in learning and memory

are one of the most common and widely assessed patient complaints (7). However, a wide variety of auditory verbal learning tasks (AVLTs) exist that can be administered to assess verbal memory and recall, and these differ across a range of qualitative and quantitative features (7–9). Such differences in assessment instruments can contribute to inconsistencies in the measurement of neurocognitive performance (10).

Methods to accurately convert scores across AVLTs could facilitate highly powered studies of verbal memory and recall, offering opportunities for new clinical insights (11, 12). However, data from single sites are typically biased by the specific attributes, demographics, and inclusion criteria of the study which can increase variance/error and confound reproducibility (13–15). To address these limitations, emerging data harmonization approaches offer new ways to perform data transformations that remove unwanted influences in aggregated data while preserving meaning (12). Data harmonization of large and heterogeneous AVLT data sources represents an appropriate framework for the development of cross-AVLT score conversion tools, but such efforts come with analytical and organizational challenges (1, 4, 16, 17).

Here, we report a retrospective multisite (n = 53 datasets) mega study analysis of three Common AVLTs: the California Verbal Learning Test (CVLT)(8), the Rey Auditory Verbal Learning Test (RAVLT)(18), and the Hopkins Verbal Learning Test-Revised (HVLT)(19), drawing from international healthy and brain-injured populations across 13 countries and 8 languages. In contrast to meta-analyses which combine summary statistics from several sites, mega-analysis centralizes and pools individual raw data from many sites prior to analysis. This allows for a richer range of experimental designs which can consider subtle single item differences in detail (1, 2). Our primary hypothesis was that conversion performance would be significantly improved by a mega-analytic pipeline combining harmonization and item response theory (IRT) relative to unadjusted conversions. The goal of this study was to establish crosswalks between common memory measures as a means to address long standing data compatibility issues for AVLTs through freely available conversion tools: verbal-learning.chpc.utah.edu

Results

Data Summary: An overview of the key features differentiating the CVLT, RAVLT, and HVLT assessments are provided in Table 1. Supplementary Table S1 shows summary characteristics itemized for each of the 53 aggregated datasets after applying exclusion criteria. Overall, the sample size was N = 10,505 (31.8% female) which included both controls and TBI groups. The median age was 42 years with an interquartile range of 30-55 years. Studies showed significant differences in both item scores and covariates (Fig. 1). Table 2 shows the summary statistics of the full cohort after aggregation. Each instrument was represented by >1000 subjects across >10 studies, indicating good representation of AVLTs. Significant differences in demographic characteristics were evident across measures, indicating that covariate adjustment was required.

Harmonization: The ComBat-GAM algorithm (13) was implemented to correct for site-specific variations such as differences in inclusion/exclusion criteria (Supplementary Table S2) while explicitly preserving real covariate effects (e.g., age) for further analysis as described elsewhere (4). Fig. 1 shows a comparison of single site mean scores, before (gray dots) and after (colored dots) site harmonization. A line is drawn to connect each site from its pre- to post-harmonized mean value. Gray distributions portray the variation in mean scores across sites before harmonization; colored distributions portray site mean score distributions after harmonization (CVLT: Blue, RAVLT: Orange, HVLT: Red). The unadjusted distributions of scores (gray areas) exhibit much higher variance than their post-harmonized equivalents, and overall harmonization reduced total variance by 37% across all items.

Covariate Adjustment: Fig. 2a shows boxplots of sum of raw learning trial scores as percentages stratified by group (traumatic brain injury (TBI) vs. control) and sex/gender. TBI history and male sex/gender were both associated with lower sum scores across all three tests.

Age-related declines were well-fit by quadratics; years of education were well-fit by a straight line. The effects of both age and education were significant and consistent across all tests. Linear models within each measure were used to assess covariates (Table 3) after language and country of origin effects were removed separately prior to harmonization.

Anchor Items and Ability Measures: Anchor items similar in format and nature were identified for each of the three required crosswalks (1. RAVLT ↔ CVLT; 2. RAVLT ↔ HVLT; and 3. CVLT ↔ HVLT). After expert consensus and trials of different anchor combinations, we elected to use immediate free recall learning trials, short delay, and long delay free recall as anchor items, where available. Short delay was used as an anchor item between CVLT and RAVLT only (short delay is not assessed in HVLT). False positive measures were not recorded consistently across sites and were not used. Recognition hits showed inconsistent behavior and were excluded from conversions (see Limitations). Since all site effects and measured covariate effects had been removed prior to IRT analysis, we assumed scores were distributed randomly across measures and ability scores did not require further scaling.

Item Response Theory: A continuous IRT model (Fig. 3) was used to estimate the latent trait of all individuals while accounting for different item difficulties and discriminations across tests. Statistically, individuals with the same verbal learning ability will achieve equivalent scores on different measures. Third degree polynomials estimated the relationship between observed scores and ability scores for each. Cubic polynomial fits of ability vs. score are shown in Fig. 3a for immediate, short, and long delay items. Horizontal lines of equivalent ability connect equivalent scores. Items with longer delay show larger differences in difficulty across AVLTs.

Validation with Dually Administered Tests: We validated the derived conversions on held-out data not used in other analyses. Validation was conducted by comparing the conversion estimates to real data where two verbal learning tests were administered to the same set of individuals (Fig. 3b; n=36). How well conversion lines fit the dually administered test scores is an independent measure of conversion accuracy.

Fig. 3b shows a scatter plot of the raw sum of learning scores for a subset of cases who were administered both the CVLT and RAVLT after harmonization. As a proportion correct, repeated assessment scores vary within subjects by 10.7% on average, presented as a confidence area derived by 100x cross validation assuming a fit through the origin (Fig. 3b, blue shaded area). These data are fitted against the IRT-derived conversion scores (black dotted line) for RAVLT to CVLT. The line falls within the 95% confidence bound for the dually administered tests, indicating agreement. Compared to the same conversion model constructed using unadjusted data, the harmonized conversion exhibited a 9.5% lower root mean squared error against the held-out data, indicating that harmonization moderately improved conversion.

Potential Clinical Application: An individual with a chief complaint of memory loss is undergoing evaluation to address a question of possible neurodegenerative disease. Initial neuropsychological examination revealed a score on the CVLT suggesting memory impairment. However, the individual is seeking a second opinion with clinicians at another institution who elect to repeat memory testing. The second neuropsychologist administers the RAVLT. Using the online conversion tool the RAVLT is converted to an equivalent CVLT score, and these scores can now be directly compared after the same CVLT norms are applied.

Discussion

There have never been more studies published annually in the history of the neurosciences (20). This intensive rate of research offers unparalleled opportunity for data combination and nuanced examination of cognitive and behavioral changes associated with neurological diagnoses. However, “high volume science” lacks coordination between studies, which poses critical challenges for the integration of findings and data harmonization. For example, AVLTs are the most common method for learning and memory assessment, but they were independently developed and without explicit quantitative reference to pre-existing instrumentation. Over the last

70 years, this has led to a scenario where clinicians and researchers routinely use distinct AVLTs with incomparable results (7–9). This is not only a technical inconvenience but is problematic for the interpretation and reproducibility of results and findings .

Constructing reliable standards for converting scores across common AVLTs is challenging, because conversions should be made independent of factors such as language, study group, and instrumental details. For example, given more words to recall, it is more likely that more words will be recalled. Naim et al. found the average number of memory items recalled (R) scales with the root of M items presented (21), and there are other subtle differences between seemingly similar assessments (Table 1). Together, large-sample mega-analysis and harmonization present a promising solution to address these concerns in aggregate, offering new opportunities to explore previously incompatible datasets and examine more interesting clinical features.

Beyond conversions, comparing the difficulty of tests on the same ability scale (Fig. 3a) may assist with the selection of AVLT across different research and clinical contexts. For example, the HVLT was the easiest test overall, while the CVLT was the most challenging test. The HVLT may be most appropriate for the assessment of individuals who are at risk for significant impairment. Conversely, the CVLT has sufficient dynamic range to discriminate within high ability groups, while the RAVLT may be well suited for studies involving a wide range of abilities. However, these are relatively coarse recommendations, and individually-derived measures from the RAVLT and other AVLTs may be more or less suitable given specific scenarios (22).

In the process of converting across AVLTs, site effects such as different settings, inclusions, and procedures were found to have an appreciable impact on verbal learning scores (Fig. 1, gray distributions). This may also be attributable to underlying differences in inclusion/exclusion criteria across studies (Table S2). However, a detailed list of all the ways our sources differed was not necessary to remove these effects in aggregate with a harmonization algorithm. We confirmed our primary hypothesis that conversion error would be reduced by implementing a mega-analytic pipeline combining harmonization and IRT. In the end, simple cubic models effectively equated the instrumental component of scores without reference to underlying factors (Fig. 3a). In time, these conversions may be found to be suitable for clinical utilization at the individual level, although verifying this will require further independent scrutiny (see Limitations).

Appropriate harmonization transforms data in ways that preserve its core relationships (4). For this study, these relationships include the associations between scores and ability, between scores and covariates (e.g., age-related memory decline), and the measurement of the underlying cognitive construct. These relationships can be assessed directly by comparing pre/post harmonized data (Fig. 1). After harmonization, the higher scores associated with younger age, female sex/gender, more education, and controls persisted for all AVLTs, despite a large drop in cross-site variance. Interestingly, despite the large sample size (to our knowledge, the largest AVLT dataset to date), no Race/Ethnicity variable was consistently associated with higher or lower scores (Table 3).

Beyond covariate effects, harmonizing AVLT offers opportunities for new, highly powered mega-analytic investigations of verbal learning and memory. For example, aggregating previously incompatible data sources could yield new estimates of the impact of a range of clinical conditions on the ability to encode, combine, store, and retrieve information. This may be particularly beneficial as a means to functionally characterize the imaging findings from large-scale ENIGMA mega studies, where interesting imaging features are emerging in clinical populations not seen in smaller samples (23) Large-scale harmonization efforts also hold the promise of advancing theory. For example, through common definitions of cognitive constructs, these efforts may influence cognitive theory by advancing the development of cognitive ontologies (24).

Conclusion

To address long standing incompatibilities across commonly used AVLTs, we aggregated a comprehensive dataset of more than ten thousand participants drawn from 53 international datasets that recorded performance on verbal learning tasks. We confirmed our hypothesis, and found that site harmonization and IRT could isolate the instrumental component of scores. This work suggests the specific choice of AVLT has a pronounced effect; Averaged across items, the CVLT was the most challenging test, although it was similar to the RAVLT in difficulty, while the HVLT was the least difficult, as expected due to its lower complexity (25). Free conversion tables and tools (available at verbal-learning.chpc.utah.edu) can assist clinicians to track and compare patient raw scores against large reference groups, regardless of differences in AVLT administration practices. More broadly, this work demonstrates that data harmonization of large data sharing initiatives can offer new tools and a means to address long standing data challenges across the behavioral sciences.

Materials and Methods

Data Sources and Inclusion Criteria: Drawing from a range of international studies of head injury and comparator groups and controls for a variety of conditions (26–30) this secondary multisite (N = 53 datasets) mega study analysis focuses on three AVLTs: the CVLT; (8), HVLT; (18), and RAVLT (17). To mitigate balance issues, we included only comparator and groups with TBI to reduce the mean difference in ability across assessments. As described in prior work (4), we aggregated data contributed by collaborators in the Psychiatric Genomics Consortium (PGC), the Enhancing NeuroImaging Genetics through Meta-Analysis Consortium (ENIGMA) working groups (31), the ENIGMA Brain Injury working group (15), and the Long-term Impact of Military-relevant Brain Injury Consortium – Chronic Effects of Neurotrauma Consortium, LIMBIC-CENC (17). The University of Utah provided overall Institutional Review Board (IRB) study approval. Inclusion criteria, performance validity testing, and other factors varied across studies (see Supplementary Table S2).

To limit sources of variability, we excluded anyone with a known clinically diagnosed mental health or neurological condition other than TBI. Consistent with standard AVLT administration practices, we included only participants aged 16 years or over. In the case of longitudinal or serial measurement designs, only the first measurement of AVLTs per person were included; repeated measurements were dropped.

Verbal Learning Task Contents and Scoring: Table 1 provides a summary overview of the key features of the AVLTs assessed. AVLT scores on each trial denote the number of correct words that are recalled. The maximum score reflects the number of memory items per list. The sum of the total words recalled across all immediate free recall (learning) trials is the immediate free recall summary score (Sum of Learning Trials). These raw scores are often subsequently normed so that the performance of the individual can be contextualized relative to a population of interest. However, in this work we exclusively assess raw scores, and not t-scores, or normative scores. We focus on raw scores because normative values are occasionally updated over time and are based upon distinct normative samples between instruments.

The California Verbal Learning Test: The CVLT (8) refers to a family of instruments that assess verbal learning and memory deficits. The CVLT has been revised twice, and three iterations exist (CVLT-I, CVLT-II, and CVLT-3). Additionally, the CVLT comes in standard, short, and alternate forms. In this work, we estimated crosswalks for the more recent CVLT-II and the CVLT-3. While the CVLT-3 is nominally a revision of the CVLT-II, in practice the target words, their order, and their number are the same for both the CVLT-3 and CVLT-II. Thus, we refer to both CVLT-3 and CVLT-II standard and alternate forms together as ‘CVLT’. Table 1 provides a numerical overview of the key features of the CVLT. The CVLT uses M = 16-word list lengths, which are drawn from 4 semantic categories, and 5 consecutive learning trials. The CVLT is a comprehensive test that includes a distractor list, cued and free recall assessments, short and long delay trials, and a recognition trial with 48 words.

The Hopkins Verbal Learning Test–Revised: The HVLT-R (19) is a relatively short measure of verbal learning and memory deficits. The HVLT exists in two primary forms (original and revised) denoted together as 'HVLT'. Table 1 outlines the key features of the HVLT. The HVLT does not use a distractor list for immediate recall and does not assess cued or a short delay recall performance. The HVLT uses M = 12-word list lengths, which are drawn from 3 semantic categories, and uses a small (N=24) total pool of words for scoring. The HVLT has three consecutive learning trials.

The Rey Auditory Verbal Learning Test: The RAVLT (18) is a measure of verbal learning and memory deficits. Table 1 provides an overview of the key features of the RAVLT. The RAVLT draws from random, semantically unrelated words, and employs a M = 15-word list length, a distractor list, as well as a large (N=50) total pool of words for scoring recognition hits. Alternate forms also exist for the RAVLT.

Covariates: Language, country of origin, age at testing, sex/gender, race/ethnicity, site/study, military/civilian status, TBI history, and education level were included and adjusted for in this study. The exclusion criteria were used to rule out the presence of any other clinically relevant variables, including epilepsy, dementia, and mild or early onset cognitive impairment (32). While some of the studies recorded gender, others recorded biological sex, and these were aggregated into a single variable. Ethnicity was binarized to Hispanic/Latino, or Not Hispanic/Latino. Perspectives on race/ethnicity differ widely according to cultural context (33), and we elected to use broad categories of Black, White, Asian, and Other. Covariate coefficients per AVLT model were converted to percentages, averaged, and then applied back to adjust the full cohort. This means each covariate had the same effect on scores regardless of the instrument used.

Statistical Analysis: Analysis was performed in Python 3 and in R. Kruskal–Wallis *H* tests (omnibus) were used to test for overall significance across groups. Where normality was confirmed, Welch's *t* tests were used for post-hoc pairwise comparisons with additional correction for multiple comparisons. Overall missing values were low (<5%) and any missing data points were imputed with nearest neighbor imputation. After data cleaning and imputation, Empirical Bayes harmonization using the ComBat-GAM algorithm (12) was used to remove unwanted site effects while preserving instrumental effects for further analysis.

Site and Covariate Modeling: Modeling was conducted in three stages: first, the overall dataset was divided into three subsets, one per AVLT instrument. The ComBat-GAM algorithm (13) was applied to each of these three subsets separately to remove site effects within each AVLT. Prior work evaluating the harmonization of cognitive measurements with this approach found that harmonization preserved real covariate effects while removing site effects (4). After site correction, covariate adjustment was performed as follows: the overall dataset was again divided into three distinct subsets by instrument, and ordinary least squares (OLS) linear models were used to estimate and remove covariates averaged across all three AVLT linear models.

Item Response Theory: A Continuous Model IRT family (23, 34–36) was used to estimate each subject's verbal learning ability because item score ranges were different across tests. To calibrate parameters, we used Shojima's (35) simplified expectation maximization (EM) method by assuming non-informative priors for item parameters as implemented in the EstCRM (Continuous Response Model) R package (34). After all data adjustments, samples taking different AVLTs were assumed to be randomly equivalent (see Limitations), such that verbal learning ability estimates were placed on the same scale using a 'random equivalent groups' linking design. The relative difficulty of all items across tests was inferred on a single ability scale, and tables of equivalent raw scores were linked across AVLTs.

Transparency and Openness: Raw data are available upon reasonable request pending study approval and data transfer agreements between participating institutions. Code used for analysis and online tool creation are available online.

Limitations

Individuals who took different tests may have different abilities on average, meaning some residual correlation between ability and choice of instrument may have persisted. However, a post-hoc analysis of converted scores found that the average percentage difference in scores across AVLTs was just 0.052%.

This study considered only a limited binary interpretation of lifetime history of TBI, which exists along a spectrum of severity and has distinct phenotypes (37). For example, late effects of TBI are highly variable with injury severity a major factor (e.g chronic memory/learning deficits are expected with severe injury but are uncommon with mild TBI). A secondary sensitivity analysis conducted using only controls did not significantly alter the conversions.

This study aggregated 53 international datasets, but these were primarily from English-speaking and western hemisphere countries, which limits the international representativeness of the sample. Our IRT conversions were validated against held-out data of dually administered tests (Fig. 3b) s, but no data to independently assess the other two conversions.

We attempted to construct a crosswalk for recognition memory trials, but unlike the other items, we could not establish low error IRT results for the recognition item and the recognition hits memory trial conversions were not included.

Acknowledgements

- Spanish Ministry of Science and Innovation, Instituto de Salud Carlos III
  - PI15-00852
  - PI18-00945
  - JR19-00024
  - PI17-00481
  - PI20-00721
  - Sara Borrell contract (CD19-00149)
- European Union
  - NextGenerationEU (PMP21/00051)
  - PI19/01024
  - Structural Funds
  - Seventh Framework Program
  - H2020 Program under the Innovative Medicines Initiative 2 Joint Undertaking: Project PRISM-2 (Grant agreement No.101034377)
  - Project AIMS-2-TRIALS (Grant agreement No 777394)
  - Horizon Europe
- National Institutes of Health
  - U01MH124639
  - P50MH115846
  - R01MH113827
  - R25MH080663
  - K08MH068540
  - R01NS100973
  - R01EB006841
  - P20GM103472
  - RO1MH083553
  - T32MH019535
  - R01 HD061504
  - RO1MH083553
  - R01AG050595
  - R01AG076838

- R01AG060470
- R01AG064955
- P01AG055367
- K23MH095661
- R01MH094524
- R01MH121246
- T32MH019535
- R01NS124585
- R01NS122827
- R61NS120249
- R01NS122184
- U54EB020403
- R01MH116147
- R56AG058854
- P41EB015922
- R01MH111671
- P41RR14075
- M01RR01066
- R01EB006841
- R01EB005846
- R01 EB000840
- RC1MH089257
- U24 RR021992
- NCRR 5 month-RR001066 (MGH General Clinical Research Center)

· NSF
  - 2112455

· Madrid Regional Government:
  - B2017/BMD-3740 AGES-CM-2

· Dalhousie Medical Research Foundation

· Research Nova Scotia
  - RNS-NHIG-2021-1931

· US Department of Defense:
  - Award # AZ150145

· US Department of Veterans Affairs
  - 1I01RX003444

· NJ Commission on TBI Research Grants
  - CBIR11PJT020
  - CBIR13IRG026

· Department of Psychology, University of Oslo

· Sunnaas Rehabilitation Hospital
  - HF F32NS119285

· Canadian Institutes of Health Research:
  - Grant 166098

· Neurological Foundation of New Zealand

· Canterbury Medical Research Foundation, University of Otago.

· Biogen US
  - Investigator-initiated grant

· Italian Ministry of Health:
  - RF-2019-12370182

- Ricerca Corrente RC 23
- National Institute on Aging:
- National Health and Medical Research Council
  - Investigator Grant APP1176426
- PA Health Research:
  - Grant SAP #4100077082 to Dr. Hillary
- La Caixa Foundation
  - ID: 100010434, fellowship code: LCF/BQ/PR22/11920017
- Research Council of Norway
  - 248238
- Health Research Council of New Zealand
  - Sir Charles Hercus Early Career Development (17/039)
  - 14-440
- South-Eastern Norway Regional Health Authority
  - 2018076
- Norwegian ExtraFoundation for Health and Rehabilitation
  - 2015/ FO5146
  - 2015044
- Stiftelsen K.G. Jebsen
  - SKGJ MED-02
- German Research Foundation:
  - DFG grant FOR2107 to Andreas Jansen, JA 1890/7-1, JA 1890/7-2
  - DFG grant FOR2107 to Igor Nenadić, NE2254/1-2,NE2254/3-1,NE2254/4-1
  - DFG grant FOR2107, KI588/14-1 and FOR2107, KI588/14-2
  - DFG, grant FOR2107 DA1151/5-1 and DA1151/5-2
  - SFB-TRR58, Projects C09 and Z02
- Central Norway Regional Health Authority (RHA) and the Norwegian University of Science and Technology (NTNU)
- National Health and Medical Research Council
  - APP1020526
- Brain Foundation
- Wicking Trust
- Collie Trust
- Sidney and Fiona Myer Family Foundation.
- U.S. Army Medical Research and Materiel Command (USAMRMC)
  - Award #13129004
- Department of Energy
  - DE- FG02-99ER62764
- Mind Research Network
- National Association for Research in Schizophrenia and Affective Disorders
  - Young Investigator Award
- Blowitz Ridgeway and Essel Foundations
- NOW ZonMw TOP 91211021
- UCLAS Easton Clinic for Brain Hellath
- UCLA Brain Injury Research Center
- Stan and Patty Silver
- Clinical and Translational Research Center
  - UL1RR033176
  - UL1TR000124

· Mount Sinai Institute for NeuroAIDS Disparities
· VA Rehab SPIRE
· CDMRP PRAP
· VA RR&D IK2RX002922
· Veski Fellowship
· Femino Foundation grant
· Fundación Familia Alonso
· Fundación Alicia Koplowitz
· CIBERSAM, Madrid Regional Government (B2017/BMD-3740 AGES-CM-2)
· 2019R1C1C1002457, 21-BR-03-01
· 2020M3E5D9079910, 21-BR-03-01
· Interdisciplinary Center for Clinical Research (IZKF) of the medical faculty of Münster

References

1. P. M. Thompson, *et al.*, ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries. *Transl. Psychiatry* **10**, 100 (2020).

2. A. Nagaraj, E. Shears, M. de Vaan, Improving data access democratizes and diversifies science. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 23490–23498 (2020).

3. S. M. Rajtmajer, T. M. Errington, F. G. Hillary, How failure to falsify in high-volume science contributes to the replication crisis. *Elife* **11**, e78830 (2022).

4. E. Kennedy, *et al.*, Harmonizing PTSD Severity Scales Across Instruments and Sites. *Neuropsychology* **Accepted**.

5. F.-F. Pan, L. Huang, K.-L. Chen, Q.-H. Zhao, Q.-H. Guo, A comparative study on the validations of three cognitive screening tests in identifying subtle cognitive decline. *BMC Neurol.* **20**, 78 (2020).

6. D. R. Roalf, *et al.*, Bridging cognitive screening tests in neurologic disorders: A crosswalk between the short Montreal Cognitive Assessment and Mini-Mental State Examination. *Alzheimers. Dement.* **13**, 947–952 (2017).

7. T. J. Guilmette, D. Rasile, Sensitivity, specificity, and diagnostic accuracy of three verbal memory measures in the assessment of mild brain injury. *Neuropsychology* **9**, 338–344 (1995).

8. S. Woods, D. Delis, J. Scott, J. Kramer, J. Holdnack, The California Verbal Learning Test – second edition: Test-retest reliability, practice effects, and reliable change indices for the standard and alternate forms. *Archives of Clinical Neuropsychology* **21**, 413–420 (2006).

9. O. Spreen, E. Strauss, *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary* (Oxford University Press, USA, 1998).

10. S. A. Cardenas, L. Kassem, M. A. Brotman, E. Leibenluft, F. J. McMahon, Neurocognitive functioning in euthymic patients with bipolar disorder and unaffected relatives: A review of the literature. *Neurosci. Biobehav. Rev.* **69**, 193–215 (2016).

11. B. Jennett, G. Teasdale, Wechsler D. A standardized memory scale for clinical use. *J. Psychol.*

12. I. Thiruselvam, J. B. Hoelzle, Refined measurement of verbal learning and memory: Application of item response theory to California Verbal Learning Test - Second Edition (CVLT-II) learning trials. *Arch. Clin. Neuropsychol.* **35**, 90–104 (2020).

13. R. Pomponio, *et al.*, Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *Neuroimage* **208**, 116450 (2020).

14. J. Radua, *et al.*, Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. *Neuroimage* **218**, 116956 (2020).

15. E. A. Wilde, E. L. Dennis, D. F. Tate, The ENIGMA Brain Injury Working Group: Approach, Challenges, and Potential Benefits. *Brain Imaging Behav.* **Under Review** (2019).

16. C. Coryn, K. Hobson, R. McCowen, Meta-analysis as a method of multi-site evaluation: An example from international development. *Evaluation Journal of Australasia* **15**, 4–14 (2015).

17. D. X. Cifu, K. J. Dixon, Chronic effects of neurotrauma consortium. *Brain Inj.* **30**, 1397–1398

(2016).

18. M. R. Schoenberg, *et al.*, Test performance and classification statistics for the Rey Auditory Verbal Learning Test in selected clinical samples. *Arch. Clin. Neuropsychol.* **21**, 693–703 (2006).

19. R. H. B. Benedict, D. Schretlen, L. Groninger, J. Brandt, Hopkins Verbal Learning Test – Revised: Normative Data and Analysis of Inter-Form and Test-Retest Reliability. *The Clinical Neuropsychologist* **12**, 43–55 (1998).

20. D. R. Priestley, *et al.*, Establishing ground truth in the traumatic brain injury literature: if replication is the answer, then what are the questions? *Brain Commun* **5**, fcac322 (2023).

21. M. Naim, M. Katkov, S. Romani, M. Tsodyks, Fundamental Law of Memory Recall. *Phys. Rev. Lett.* **124**, 018101 (2020).

22. K. Barwegen, Z. J. Resch, G. P. Ovsiew, K. J. Jennette, J. R. Soble, A-232 Head-To-Head Comparison of the Rey Auditory Verbal Learning Test Effort Score and Forced Choice Embedded Performance Validity Indicators Among Patients with and without Verbal Memory Impairment. *Arch. Clin. Neuropsychol.* **37**, 1388–1388 (2022).

23. T. Wang, L. Zeng, Item Parameter Estimation for a Continuous Response Model Using an EM Algorithm. *Appl. Psychol. Meas.* **22**, 333–344 (1998).

24. J. Hastings, *et al.*, Interdisciplinary perspectives on the development, integration, and application of cognitive ontologies. *Front. Neuroinform.* **8**, 62 (2014).

25. A. M. Shapiro, R. H. Benedict, D. Schretlen, J. Brandt, Construct and concurrent validity of the Hopkins Verbal Learning Test-revised. *Clin. Neuropsychol.* **13**, 348–358 (1999).

26. A. W. Toga, K. L. Crawford, The Alzheimer's Disease Neuroimaging Initiative informatics core: A decade in review. *Alzheimers. Dement.* **11**, 832–839 (2015).

27. A. Schmidt, *et al.*, Structural Network Disorganization in Subjects at Clinical High Risk for Psychosis. *Schizophrenia Bulletin*, sbw110 (2016).

28. A. J. Lundervold, H. B. Halleland, E. J. Brevik, J. Haavik, L. Sørensen, Verbal Memory Function in Intellectually Well-Functioning Adults With ADHD: Relations to Working Memory and Response Inhibition. *Journal of Attention Disorders* **23**, 1188–1198 (2019).

29. N. Petrovsky, *et al.*, Sensorimotor Gating is Associated with CHRNA3 Polymorphisms in Schizophrenia and Healthy Volunteers. *Neuropsychopharmacology* **35**, 1429–1439 (2010).

30. S. P. Broglio, *et al.*, The Natural History of Sport-Related Concussion in Collegiate Athletes: Findings from the NCAA-DoD CARE Consortium. *Sports Med.* **52**, 403–415 (2022).

31. M. W. Logue, *et al.*, Smaller Hippocampal Volume in Posttraumatic Stress Disorder: A Multisite ENIGMA-PGC Study: Subcortical Volumetry Results From Posttraumatic Stress Disorder Consortia. *Biol. Psychiatry* **83**, 244–253 (2018).

32. E. Kennedy, *et al.*, Traumatic Brain Injury and Early Onset Dementia in Post 9-11 Veterans. *Brain Inj.* **36**, 620–627 (2022).

33. N. S. Nasir, V. M. Hand, Exploring Sociocultural Perspectives on Race, Culture, and Learning. *Rev. Educ. Res.* **76**, 449–475 (2006).

34. C. Zopluoglu, EstCRM: An R package for Samejima's continuous IRT model. *Appl. Psychol. Meas.* **36**, 149 (2012).

35. K. Shojima, A noniterative item parameter solution in each EM cycle of the continuous response model. *Educ. Technol. Res. Dev.* **28**, 11–22 (2005).

36. F. Samejima, Homogeneous case of the continuous response model. *Psychometrika* **38**, 203–219 (1973).

37. M. J. Pugh, *et al.*, Phenotyping the Spectrum of Traumatic Brain Injury: A Review and Pathway to Standardization. *J. Neurotrauma* **38**, 3222–3234 (2021).

Figure 1. Comparing proportions of memory items recalled before and after harmonization. Mean scores for each site (dots) are shown broken out by instrument (color) and item (Top: Trial 1 immediate free recall, Middle: Total sum of all Trials, Bottom: Long-delay free recall scores).

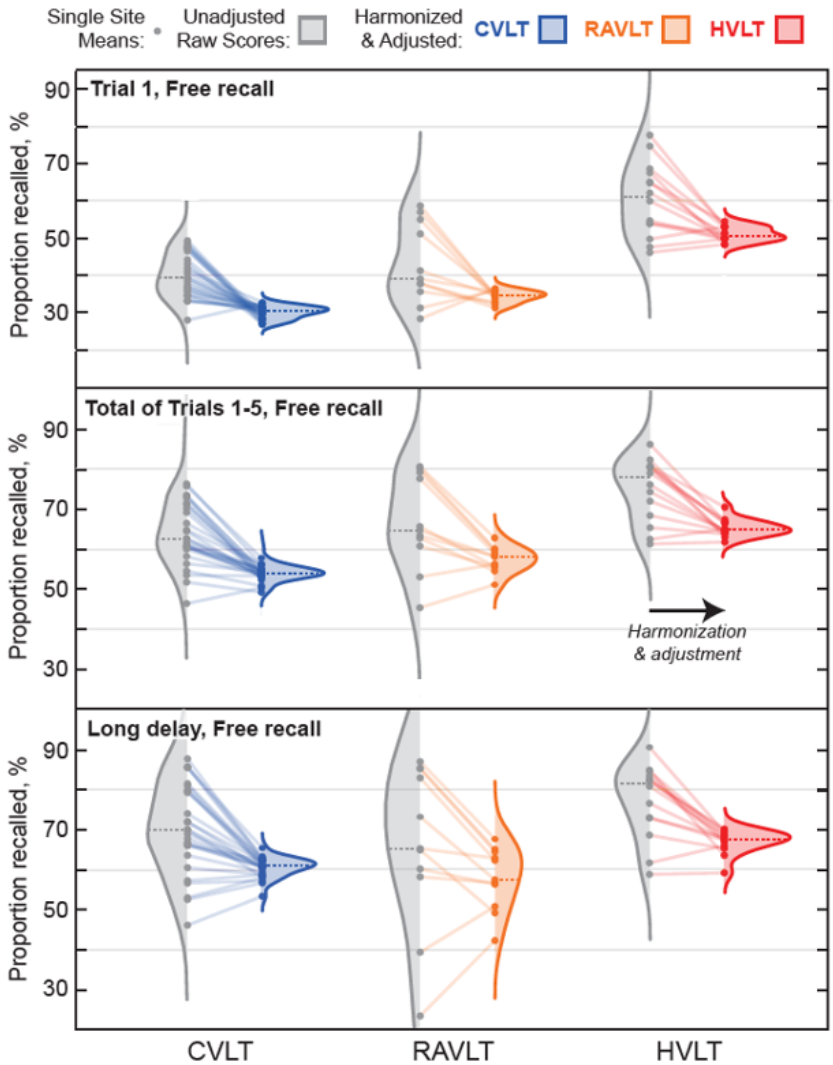

Figure 2. Visualizing covariate effects on scores across AVLTs. **(a)** Boxplots of scores stratified by group (TBI vs. control) and sex/gender indicated that males and those with history of TBI had significantly lower scores on average for all tests. Age-related declines **(b)** and the beneficial effects of education **(c)** on scores were consistent across all AVLTs.

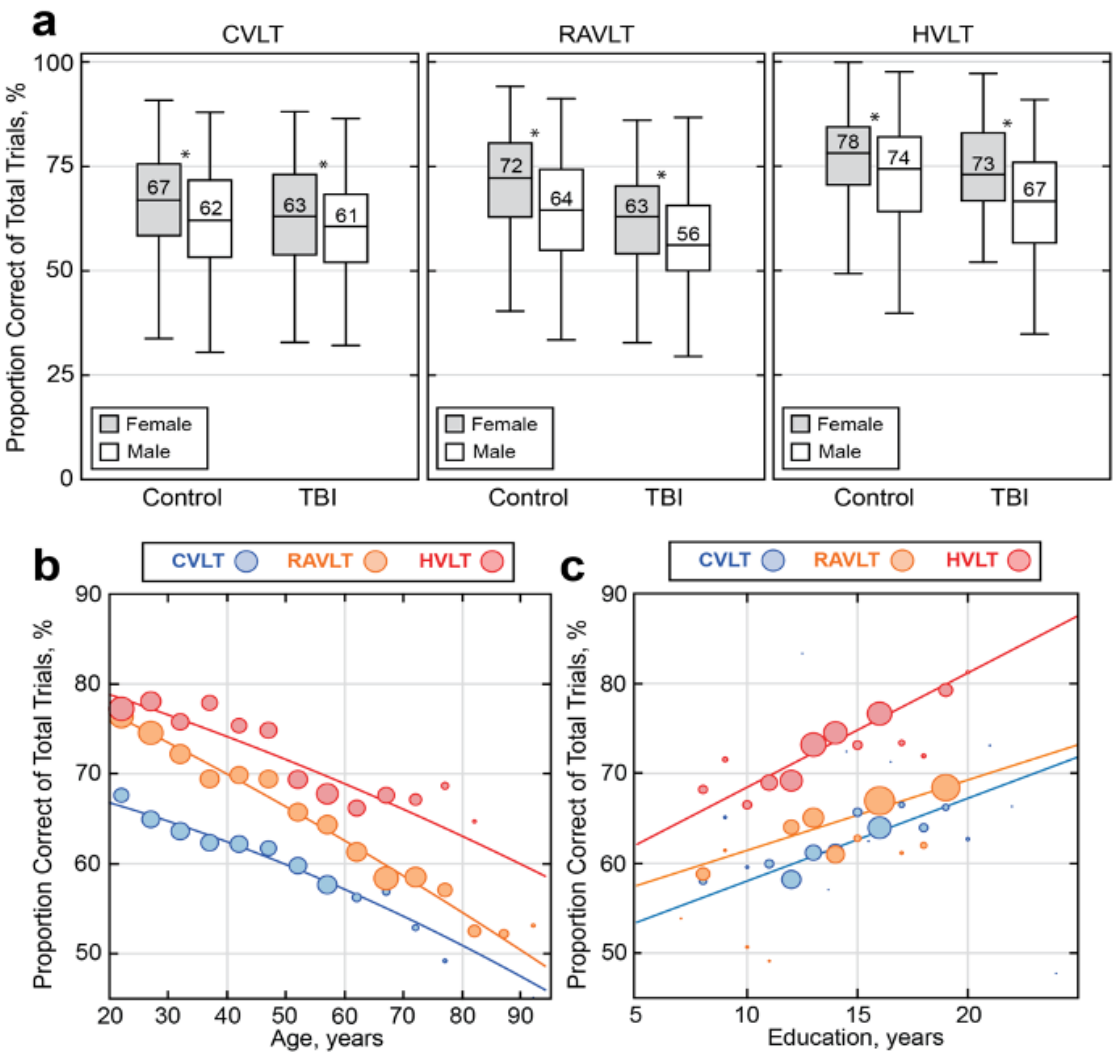

Figure 3. Visualizing and Validating Conversions. (a) Average raw scores as a function of individual ability are shown approximated as cubic polynomial fits for immediate, short, and long delay trials. Horizontal lines of equivalent ability connect equivalent raw scores across tests, which facilitates the construction of crosswalks. **(b)** Scatter plot and fit of the adjusted raw sum of learning Trial scores for a subset of cases who were administered both the CVLT and RAVLT (n=36). The confidence area of the dually assessed data is shown in blue, and agrees with the derived crosswalk for CVLT->RAVLT (n=9,362, black dotted line).

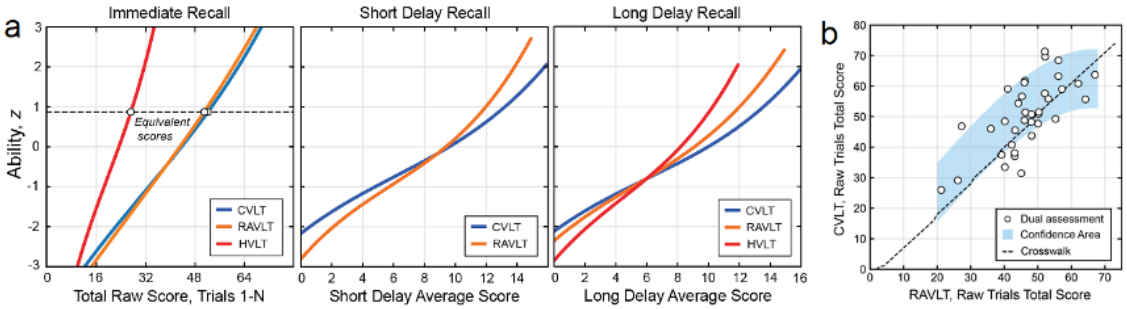

Table 1. Summary of the key features of the three AVLTs.

	CVLT	RAVLT	HVLT
List lengths (words)	16	15	12
Semantic categories	4	-	3
Immediate free recall trials	5	5	3
Recognition trial, total words	48	50	24
Long delay duration (minutes)	20-30	20	20-25
Distractor list (List B)	Yes	Yes	No
Cued recall trials	Yes	No	No
Short Delay Trial	Yes	Yes	No

Table 2. Descriptive characteristics of the total cohort by instrument.

	CVLT	RAVLT	HVLT	*p*
Sample size, N	6,634	2,728	1,143	
# Datasets	28	11	14	
Age: 16-35	34.70%	32.70%	48.20%	<0.001
35-65	61.20%	40.40%	41.90%	<0.001
65+	4.10%	26.90%	9.90%	<0.001
Gender: Male	76.50%	48.60%	66.10%	<0.001
Female	23.50%	51.40%	33.90%	<0.001
Region: Americas	90.70%	44.50%	69.00%	<0.001
Europe	8.30%	55.40%	27.60%	<0.001
Asia Pacific	1.10%	0.10%	3.40%	<0.001
Language: English	90.70%	37.60%	82.20%	<0.001
Spanish	1.90%	7.40%	6.40%	<0.001
German	0.20%	35.3%	11.50%	<0.001
Italian	0.00%	8.80%	0.00%	<0.001
Other	7.20%	10.90%	0.00%	<0.001
Education: Secondary	20.80%	28.30%	15.20%	<0.001
Some College	14.80%	14.00%	19.70%	<0.001
Postgraduate	7.20%	21.40%	5.50%	<0.001
Clinical Population: TBI	64.70%	11.90%	21.90%	<0.001
Controls	35.30%	88.10%	78.10%	<0.001
Mean Raw Scores: Trial 1	4.9 ± 1.9	5.0 ± 1.9	6.1 ± 1.7	<0.001
Total Learning	43.2 ± 9.7	42.9 ± 8.6	23.2 ± 4.2	<0.001
Long Delay	9.6 ± 3.1	8.7 ± 3.0	7.9 ± 2.2	<0.001

Table 3. Blocked linear regressions predicting sum of raw learning scores per instrument. The average percentage effect across all AVLTs are shown, indicating that Age>65 had the largest impact on scores overall (-11.4%). * indicates significance after correction for multiple comparisons.

	CVLT	RAVLT	HVLT		
Sample size	6,634	2,728	1,143		
Explained variance	10%	30%	22%	% effect	Significant
Max score of T1-TN	80	75	36	all	all tests
Age group (Ref: 16-35 years)					
35-65	-2.7*	-6.5*	-1.0*	-4.90%	Yes
65+	-7.8*	-12.2*	-2.9*	-11.40%	Yes
Sex/gender (Ref: Male)					
Female	3.1*	3.7*	1.6*	4.40%	Yes
Race/Ethnicity (Ref: White)					
Multi	-1.3	-3	-0.6	-2.40%	No
Black	-2.7*	-0.8	-2.8*	-4.10%	No
Asian	1.1	0.5	-1.9*	-1.10%	No
Hispanic	0.5	-1.1	-0.1	-0.40%	No
TBI (Ref: No TBI)	-1.3*	-1.6*	-1.6*	-2.80%	Yes
Education (per year)	0.6*	0.6*	0.4*	0.90%	Yes