Published in final edited form as:
Int Psychogeriatr. 2009 February ; 21(1): 129–137. doi:10.1017/S1041610208007862.

Japanese-English language equivalence of the Cognitive Abilities Screening Instrument among Japanese-Americans

Laura E. Gibbons, PhD[1], **Susan McCurry, PhD**[2], **Kristoffer Rhoads, PhD**[3], **Kamal Masaki, MD**[4], **Lon White, MD MPH**[5], **Amy R. Borenstein, PhD, MPH**[6], **Eric B. Larson, MD MPH**[7], and **Paul K. Crane, MD MPH**[8]

[1]Department of Medicine, University of Washington, Box 359780, Harborview Medical Center, 325 Ninth Avenue, Seattle, WA 98104, USA.

[2]Department of Psychosocial and Community Health, University of Washington, 9709 3rd Ave NE, Suite 507, Seattle WA, 98115-2053, USA

[3]Department of Psychiatry and Behavioral Sciences, University of Washington, Department of Physical Medicine and Rehabilitation, Virginia Mason Medical Center, 1100 Ninth Avenue, Seattle, 98101

[4]Kuakini Medical Center and the Pacific Health Research Institute, 347 N. Kuakini St. HPM9, Honolulu, HI 96817

[5]Kuakini Medical Center and the Pacific Health Research Institute, 347 N. Kuakini St. HPM9, Honolulu, HI 96817

[6]Department of Epidemiology and Biostatistics, College of Public Health MDC-56, University of South Florida, 13201 Bruce B. Downs Blvd, Tampa, FL, 33612, U.S.A.

[7]Center for Health Studies, Group Health Cooperative and Department of Medicine, University of Washington, 1730 Minor Avenue, Suite 1600, Seattle, WA, 98101.

[8]Center for Health Studies, Group Health Cooperative and Department of Medicine, University of Washington, Box 359780, Harborview Medical Center, 325 Ninth Avenue, Seattle, WA 98104, USA.

Abstract

Background—The Cognitive Abilities Screening Instrument (CASI) was designed for use in cross-cultural studies of Japanese and Japanese-American elderly in Japan and the United States. The measurement equivalence in Japanese and English has not been confirmed in prior studies.

Methods—We analyzed the 40 CASI items for differential item functioning (DIF) related to test language, as well as self-reported proficiency with written Japanese, age, and educational attainment in two large epidemiologic studies of Japanese-American elderly: the *Kame* Project (n=1,708) and the Honolulu-Asia Aging Study (HAAS; n=3,148). DIF was present if the demographic groups differed in the probability of success on an item, controlling for their underlying cognitive functioning ability.

Results—While 7 CASI items had DIF related to language of testing in *Kame* (registration of one item; recall of one item; similes; judgment; repeating a phrase; reading and performing a command; and following a 3-step instruction), the impact of DIF on participants' scores was minimal. Mean scores for Japanese and English speakers in *Kame* changed by < 0.1 SD after accounting for DIF related to test language. In HAAS, there were not enough participants tested in Japanese to assess

Corresponding author: Laura E. Gibbons, PhD, Box 359780, Harborview Medical Center, 325 Ninth Avenue, Seattle, WA 98104, USA. gibbonsl@u.washington.edu. (206) 744-184 (phone); (206) 744-9917 (fax).

DIF related to test language. In both studies, DIF related to written Japanese proficiency, age, and educational attainment had minimal impact.

Conclusions—To the extent that DIF could be assessed, the CASI appeared to meet the goal of measuring cognitive function equivalently in Japanese and English. Stratified data collection would be needed to confirm this conclusion. DIF assessment should be used in other studies with multiple language groups to confirm that measures function equivalently or if not, to form scores that account for DIF.

Keywords

cognitive testing; cross-cultural; dementia; differential item functioning; item response theory; test bias

INTRODUCTION

The Cognitive Abilities Screening Instrument (CASI) was developed to facilitate cross-cultural studies of cognitive aging and dementia in Japanese and Japanese-American elderly individuals (Teng et al., 1994). It combined elements from the most commonly used tests of global cognitive functioning in the United States and in Japan. It has been used in numerous studies in the Pacific Rim; among these are the *Kame* Project (Graves et al., 1996) and the Honolulu-Asia Aging Study (HAAS) (White et al., 1996).

CASI development involved careful consideration of linguistic and neuropsychological content to ensure that Japanese and English versions tapped the same cognitive domains (Teng et al., 1994). These are essential steps toward effective cross-cultural comparisons, analogous to exercises carried out across different sites to standardize dementia diagnosis methodology and to harmonize instruments across cultures (Larson et al., 1998). Additional steps need to be taken to ensure these preparatory steps were successful (Teresi et al., 2006). Here we analyze CASI data from two large studies of Japanese-American elderly individuals to investigate how well the CASI performs across language groups. We focus on differential item functioning (DIF).

DIF is present in a test item if people from different demographic groups have unequal probabilities of item success when controlling for the ability measured by the test (Camilli and Shepard, 1994). DIF assessment is a first step in determining whether test items may be biased (Camilli and Shepard, 1994). Next, the impact of DIF should be considered. Epidemiologists may not care how many items have DIF as much as they do about DIF's impact on groups. Could group differences be caused (or masked) by items with DIF? For clinicians, the primary interest may be the DIF's impact on individual scores. If DIF has a large impact, scores that account for DIF should be used. See (Crane et al., 2007) p. 82 for further discussion.

Previous studies have found DIF related to language in the Mini-Mental State Examination (Folstein et al., 1975) when comparing Spanish vs. English test-takers (Crane *et al.*, 2006; Dorans and Kulick, 2006; Edelen *et al.*, 2006; Jones, 2006; Marshall *et al.*, 1997; Morales *et al.*, 2006; Teresi *et al.*, 1995). In addition, tests of cognitive function are often found to have DIF related to education and other demographic factors (Crane, 2006; Crane *et al.*, 2006; Crane *et al.*, 2004; Jones and Gallo, 2001; Jones and Gallo, 2002; Teresi *et al.*, 1995; Teresi *et al.*, 2000). None of these studies evaluated DIF's impact on individuals or groups.

The primary goal of this study was to examine CASI items for DIF in Japanese Americans to determine how successful the test developers were in developing an instrument that has equivalent measurement properties in Japanese and English speakers. Secondary goals were

to further investigate the CASI for DIF related to other covariates, including age, sex, and education. We were interested in DIF's presence and its impact on groups and individuals.

METHODS

Overview

We used item-level CASI data from the *Kame* project and HAAS to evaluate items for DIF within each study. We used a hybrid ordinal logistic regression – item response theory (IRT) approach for DIF detection. We analyzed data from the *Kame* Project regarding language of testing; there were too few Japanese test-takers in HAAS (n=120) to address this question. We analyzed CASI items from both studies for DIF related to age, sex and educational attainment. We determined DIF presence and its impact on groups and individuals.

Data source: The *Kame* Project

A schematic representation of the samples analyzed in this study is shown in Figure 1. A cohort of 3,045 eligible individuals aged 65 and older, 96% of whom were of 100% Japanese origin, was identified in a November 1991 census of Japanese Americans in King County. Study census details are described elsewhere (Graves et al., 1996). Of those eligible, 1,985 (65.2%) participated in the baseline evaluation (1992–1994).

The CASI was administered to all participants. Participants were stratified by CASI score and age and sampled into the clinical and neuropsychological evaluation phase of the prevalence study. Of 1,985 individuals participating in CASI screening, 450 were sampled from three cognitive strata and five age strata, with individuals scoring <81 or age ≥80 sampled with 1.0 probability, and younger and higher scoring individuals sampled at lower frequencies. Results of the clinical and neuropsychologic evaluations were reviewed in consensus conferences, and dementia diagnoses were made using DSM-III-R criteria (American Psychiatric Association, 1987). Data analyzed here come from the baseline examination 1992–1994 for the 1,836 non-demented participants. Trained interviewers rated CASI scores as invalid for participants with limitations in hearing, eyesight, or motor control, leaving 1,708 evaluated here.

Data source: HAAS

The Honolulu Heart Program (HHP) cohort included 8,006 Japanese-American men born 1900–1919 living on the island of Oahu, Hawaii, at enrollment in 1965 (Syme et al., 1975). Three midlife examinations conducted 1965–1968, 1968–1970, and 1971–1974 included collection of clinical and demographic information. At the fourth examination (1991–1993), HAAS began as an extension of the HHP. At this examination, 3,734 HHP cohort members (80% of survivors) participated and took the CASI. A multi-step procedure was used to identify individuals with dementia, detailed in (White et al., 1996). Dementia diagnoses were made using DSM-III-R criteria (American Psychiatric Association, 1987). At HHP Exam 4 (the baseline HAAS exam), there were 3,508 dementia-free participants, of whom 3,148 had valid CASI scores; these data are analyzed here.

HAAS wives of Japanese ancestry—A probability sampling based on age and CASI score determined which men were selected for further evaluation in the second stage of the dementia assessment. Those selected were asked to identify a caregiver or potential caregiver; this person was invited for a proxy interview. In most cases, this was the participant's wife. CASI scores were available for 489 wives of HAAS participants, 477 of whom were of Japanese ancestry and were included in our analyses of DIF related to sex. We thus examined HAAS CASI items for DIF related to sex on 477+3,148 = 3,625 participants.

Measure: CASI

The CASI is a 40-item test of global cognitive functioning (Teng et al., 1994). Items were identical or similar to ones used in the Hasegawa Dementia Screening Scale (Hasegawa, 1983), the MMSE (Folstein et al., 1975) and the Modified MMSE (Teng and Chui, 1987). A new item on judgment was added. The CASI was developed in parallel in English and Japanese at three workshops in which items were scrutinized for cultural equivalence, back-translated and pilot-tested.

Item response theory

We scored CASI items using IRT, so that scores accounting for DIF could be formed, as explained below. We used PARSCALE (Muraki and Bock, 2003), with Samejima's graded response model (Samejima, 1969), a polytomous extension of the 2-parameter logistic model. The mean score was zero with a standard deviation of one.

DIF detection methods

We have developed an approach to DIF assessment that combines ordinal logistic regression and IRT. Details of this approach have been previously outlined (Crane, 2006; Crane *et al.*, 2006; Crane *et al.*, 2004). We used the Stata (StataCorp, 2007) program **difwithpar** (Crane et al., 2006) to detect DIF and obtain IRT scores accounting for DIF. The program is available by typing "ssc install difwithpar" at the Stata command prompt.

We examined three ordinal logistic regression models for each item:

$$f(\text{item response}) = \text{cut} + \beta_1{}^{*}\theta \tag{1}$$

$$f(\text{item response}) = \text{cut} + \beta_1{}^{*}\theta + \beta_2{}^{*}\ \text{group} \tag{2}$$

$$f(\text{item response}) = \text{cut} + \beta_1{}^{*}\theta + \beta_2{}^{*}\ \text{group} + \beta_3{}^{*}\theta{}^{*}\text{group} \tag{3}$$

In these models, *cut* represents the cutpoint(s) for each level in the proportional odds ordinal logistic regression model, θ (theta) is the IRT estimate of cognitive ability, and "group" is the indicator for the demographic covariate we are assessing. In model 3, β_3 is the coefficient for the ability-group interaction term.

Two types of DIF are identified (Crane et al., 2006). In items with *non-uniform DIF*, demographic interference between ability and item responses differs at varying ability levels. In items with *uniform DIF*, the interference is the same across all ability levels. To detect non-uniform DIF, we compared the log likelihoods of models 2 and 3 to test the significance of the interaction term. We used an alpha level of 0.002, based on Bonferroni adjustment for the 27 items with enough discordance to analyze in *Kame*. For comparability, we used 0.002 for all other non-uniform DIF assessment. To detect uniform DIF, we determined the relative difference between parameters associated with θ [β_1 from equation 1 and equation 2] using the formula $|(\beta_{1(\text{equation 2})}-\beta_{1(\text{equation 1})})/\beta_{1(\text{equation 1})}|$. If the relative difference was >10%, group membership interfered with the expected relationship between θ and item responses.

We accounted for DIF by using items free of DIF as anchors and estimated group-specific item parameters for items with DIF. The need for group-specific item parameters required that

continuous covariates be categorized. The resulting ability estimates (θ) were generated from all of the items, but only DIF-free items were used in common across groups.

Spurious false-positive and false-negative results may occur if ability scores used for DIF detection includes many items with DIF, see (Millsap, 2006). We used the θ score which accounted for DIF as the ability level for DIF detection, and re-ran models 1–3. We compared the lists of items found with DIF using original and modified θs. If the same items were identified, we concluded our findings were not related to spurious DIF. If different items were identified, we used the most recent findings to generate new θ estimates and repeated these steps until the same items were identified on successive runs. The final θ values account for DIF related to that covariate.

Individual and group level DIF impact

We determined DIF's impact for each covariate for individuals by subtracting unadjusted IRT scores from IRT scores accounting for DIF. If DIF made no impact the difference would be 0. We needed a reference point for understanding DIF's impact. Since minimal clinically important score differences (Guyatt et al., 2002) have not been specified for the CASI, we used the median standard error of measurement; individual differences related to DIF larger than this value have a *salient* difference related to DIF (Crane et al., 2006). We performed analogous calculations to compare group mean scores when accounting for and when ignoring DIF to determine group-level DIF impact.

Covariates assessed for DIF

A schematic representation of covariates assessed for DIF is shown in Figure 1. We analyzed Japanese vs. English language testing in the *Kame* project; too few HAAS participants took the test in Japanese to analyze DIF related to language. We analyzed data from the studies separately for DIF related to self-reported proficiency reading or writing Japanese, age, sex, and self-reported educational attainment. Categories are as in Table 1, except the two lowest educational groups were combined in *Kame*.

Other statistics

Demographic features were compared using Kruskal-Wallis tests for continuous and chi-squared tests for categorical variables.

Institutional review

The *Kame* Project was approved by the University of Washington institutional review board. HAAS was approved by institutional review boards of Kuakini Medical Center (Honolulu, Hawaii) and the Honolulu Department of Veterans Affairs. All participants in both studies gave written informed consent.

RESULTS

Demographic characteristics

Study cohort demographic characteristics are shown in Table 1. On average, *Kame* participants were more likely to be tested in and read or write Japanese, were younger, and had more years of education than HAAS participants (all p values < 0.001). HAAS wives were slightly younger (53% were ≤ 75; $p<0.001$) and had fewer years of education (43% had ≤ 8 years; $p< 0.001$) than HAAS men.

DIF

In these non-demented, community-dwelling participants, some of the items were answered correctly by almost everyone (meaning they were too easy for those participants) and hence did not provide enough variability to be analyzed for DIF. This was particularly the case in *Kame*, where only 27 of the 40 items had enough variability to be assessed for DIF. HAAS had more participants, and all items had enough variability to be analyzed for DIF related to at least one covariate.

Language—In *Kame*, seven CASI items had DIF related to language of testing (Table 2), and four of these had DIF related to self-reported proficiency reading or writing Japanese (see Table S1 published as supplementary material online attached to the electronic version of this paper at www.journals.cambridge.org/jid_IPG). The group level impact of DIF related to language of testing was minimal, with mean group changes close to zero (See Figure S1 published as supplementary material). On the individual level, minimal changes due to DIF occurred in both the positive and negative directions in both Japanese and English speakers, with only one participant experiencing salient DIF. In HAAS, no items had DIF related to self-reported proficiency reading or writing Japanese.

Age, sex, education—In both studies, few items had DIF related to age (Table 2), and no items had DIF related to sex. There were more items with DIF related to educational attainment identified in *Kame* (13 of 27 items) than in HAAS (3 of 35 items) (Table 2). However, the impact of this DIF was minimal. Figure 2 shows the impact of DIF related to education in *Kame*. Only 27 *Kame* participants (1.5%) and 1 (0.2%) HAAS participant had salient DIF related to education. Other DIF impacts were smaller (see Figure S1 published as supplementary material). Mean group changes were 0.1 point or less.

DISCUSSION

The CASI was designed to facilitate measurement of cognitive functioning in Japanese and Japanese-American elderly participants, whether they were more comfortable in Japanese or in English. Careful traditional methods were employed to increase comparability of CASI scores across languages. This paper represents the first attempt to quantify how successful these efforts were in producing a test that is valid for assessing cognitive functioning in Japanese and English speaking Japanese-American participants.

The CASI was better targeted to participants in the HAAS study than the *Kame* Project. Many CASI items were not possible to analyze for DIF in *Kame* due to a high proportion of correct scores. Few HAAS participants took the test in Japanese. These facts made a thorough investigation of the CASI for DIF related to language of testing impossible. The evaluation we were able to perform found a few items with DIF in *Kame*, though that DIF appeared to have minimal impact on individuals or groups. This suggests that effort directed at the development of the CASI (Teng et al., 1994) was worthwhile; the test appears to have minimal DIF related to language.

The Japanese-English language analyses were unique to this paper, but findings were similar to prior studies comparing Spanish versus English in the MMSE (Crane *et al.*, 2006; Dorans and Kulick, 2006; Edelen *et al.*, 2006; Jones, 2006; Marshall *et al.*, 1997; Morales *et al.*, 2006; Teresi *et al.*, 1995), which also found several items with DIF related to language of testing. We also examined DIF related to self-reported proficiency reading or writing Japanese. There was no DIF in HAAS and little in *Kame*, and DIF's impact on individuals or groups was minimal in *Kame*.

DIF related to other demographic covariates has been examined for several global cognitive tests (Crane, 2006; Crane *et al.*, 2006; Crane *et al.*, 2004; Jones and Gallo, 2001; Jones and Gallo, 2002; Teresi *et al.*, 2000). Consistent with this literature, we found few items had DIF related to sex or age. A few items had DIF related to education, though this DIF had minimal impact on scores of individuals or groups. This finding is different from the extant literature on DIF's impact related to education in global cognitive tests; generally individual and group impact of DIF related to educational attainment is much greater than that related to other covariates.

Several limitations should be kept in mind when interpreting our results. While we performed sensitivity evaluations of our DIF findings using other criteria, there is no universal agreement on DIF detection techniques. Different methods may have found different results – though in broad terms most approaches seem to have similar findings when applied to the same data sets (Millsap, 2006). It should also be noted that the women included in the HAAS analysis of DIF related to sex were wives of HAAS participants who were more likely to have dementia and are not representative of all women of Japanese ancestry on the island of Oahu. In *Kame* the study census enumerated >90% of the Japanese American population living in King County in 1990; therefore, the *Kame* population can be said to represent the Japanese-American population of King County, Washington (Graves et al., 1996).

It is of some interest that despite item-level data on some 3,700 Japanese-American participants across two studies we were unable to fully evaluate the CASI for DIF related to language of testing. There are several reasons for this. First, an extremely high proportion of HAAS participants took the CASI in English, so despite having a sample roughly twice the size of *Kame*, roughly 1/3 as many HAAS participants took the CASI in Japanese. Second, many CASI items had 100% correct responses in *Kame*. While methods for DIF analyses of small sample sizes exist (Lai et al., 2005), these methods fail when there is a discrepancy between ability levels and item difficulties.

A further complication of any analyses of DIF related to language of test administration is differences in covariates across language groups. In HAAS, for example, 9% of the Japanese speakers had a high school education, compared with 51% of their English-speaking counterparts. A naïve analysis of DIF related to language of test administration may uncover DIF that is in fact due to education. Methods exist for assessing DIF related to language when other covariates (such as education) are very different across language groups, but for the reasons outlined above we were unable to perform such analyses.

A more powerful way to determine whether there was DIF in the CASI related to language of test administration is to perform targeted, stratified data collection. Hundreds of Japanese-speaking (and English-speaking) elderly Japanese-Americans would need to be sought and examined with the CASI. Sample sizes for DIF analyses are discussed in (Crane *et al.*, 2006). Ideally both English-speaking and Japanese-speaking cohorts would have broad and overlapping distributions of educational attainment and other demographic characteristics. While this stratified sample would not represent the general population, IRT has important invariance properties such that item parameters are invariant across samples within a linear transformation if model assumptions are met.

To the extent that we could assess DIF, the CASI appeared to meet the goal of measuring cognitive function equivalently in Japanese and English test takers. We found many items with DIF related to the covariates we analyzed, but DIF impact was minimal – producing small and clinically insignificant differences in scores. Targeted data collection would be needed to confirm this conclusion. Methods such as these should be used in other studies with multiple

language groups, to confirm that measures function equivalently or, if not, to form scores that account for DIF.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

These analyses were supported by an Alzheimer's Association Investigator Initiated Research Grant (P Crane). Data collection was supported by NIA N01-AG-4-2149, NIA U01-AG-1-9349-01, NIA 1RO1 AG19349-01, NHLBI N01-HC-05102, and NIH R01 09769. Programming to detect differential item functioning was supported by NIH P50 AG05136 (Raskind). Dr. Larson was supported by NIH U01 AG06781. The funding sources had no role in how the study was conducted or in the decision to publish the results.

Conflict of interest declaration

These analyses were funded by an Alzheimer's Association Investigator Initiated Research Grant and the NIH. The funding sources had no role in how the study was conducted or in the decision to publish the results.

Description of authors' roles

L. Gibbons contributed to the conception and design of the paper and the drafting of the manuscript, and conducted all the statistical analysis. S. McCurry and K. Rhoads made critical revisions of the manuscript for important intellectual content. K.Masaki contributed to data collection and made critical revisions of the manuscript for important intellectual content. L. White, A. Borenstein and E. Larson obtained funding, contributed to data collection and made critical revisions of the manuscript for important intellectual content. P.Crane obtained funding and contributed to the conception and design, interpretation of data, drafting of the manuscript.

References

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. Washington, D.C.: American Psychiatric Association; 1987.

Camilli, G.; Shepard, LA. Methods for Identifying Biased Test Items. Thousand Oaks: Sage; 1994.

Crane PK. Commentary on comparing translations of the EORTC QLQ-C30 using differential item functioning analyses. Quality of Life Research 2006;15:1117–1118.

Crane PK, Gibbons LE, Jolley L, van Belle G. Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. Medical Care 2006;44:S115–S123. [PubMed: 17060818]

Crane PK, et al. A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. Quality of Life Research 2007;16:69–84. [PubMed: 17554640]

Crane PK, van Belle G, Larson EB. Test bias in a cognitive test: differential item functioning in the CASI. Statistics in Medicine 2004;23:241–256. [PubMed: 14716726]

Dorans NJ, Kulick E. Differential item functioning on the Mini-Mental State Examination: an application of the Mantel-Haenszel and standardization procedures. Medical Care 2006;44:S107–S114. [PubMed: 17060817]

Edelen MO, Thissen D, Teresi JA, Kleinman M, Ocepek-Welikson K. Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: application to the Mini-Mental State Examination. Medical Care 2006;44:S134–S142. [PubMed: 17060820]

Folstein MF, Folstein SE, McHugh PR. "Mini-Mental State". A practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research 1975;12:189–198. [PubMed: 1202204]

Graves AB, et al. Prevalence of dementia and its subtypes in the Japanese American population of King County, Washington state. The Kame Project. American Journal of Epidemiology 1996;144:760–771. [PubMed: 8857825]

Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR. Methods to explain the clinical significance of health status measures. Mayo Clinic Proceedings 2002;77:371–383. [PubMed: 11936935]

Hasegawa, K. The clinical assessment of dementia in the aged: A dementia screening scale for psychogeriatric patients. In: Bergener, M.; Lehr, U.; Lang, E.; Schmitz-Scherzer, R., editors. Aging in the eighties and beyond. New York: Springer; 1983. p. 207-218.

Jones RN. Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination: detecting differential item functioning using MIMIC modeling. Medical Care 2006;44:S124–S133. [PubMed: 17060819]

Jones RN, Gallo JJ. Education bias in the Mini-Mental State Examination. International Psychogeriatrics 2001;13:299–310. [PubMed: 11768377]

Jones RN, Gallo JJ. Education and sex differences in the Mini-Mental State Examination: effects of differential item functioning. Journals of Gerontology. Series B. Psychological Sciences and Social Sciences 2002;57B:P548–P558.

Lai JS, Teresi J, Gershon R. Procedures for the analysis of differential item functioning (DIF) for small sample sizes. Evaluation in the Health Professions 2005;28:283–294.

Larson EB, et al. Standardization of the clinical diagnosis of the dementia syndrome and its subtypes in a cross-national study: The Ni-Hon-Sea experience. Journals of Gerontology. Series A. Medical Sciences 1998;53A:M313–M319.

Marshall SC, Mungas D, Weldon M, Reed B, Haan M. Differential item functioning in the Mini-Mental State Examination in English- and Spanish-speaking older adults. Psychology and Aging 1997;12:718–725. [PubMed: 9416639]

Millsap RE. Comments on methods for the investigation of measurement bias in the Mini-Mental State Examination. Medical Care 2006;44:S171–S175. [PubMed: 17060824]

Morales LS, Flowers C, Gutierrez P, Kleinman M, Teresi JA. Item and scale differential functioning of the Mini-Mental State Exam assessed using the differential item and test functioning (DFIT) Framework. Medical Care 2006;44:S143–S151. [PubMed: 17060821]

Muraki, E.; Bock, D. PARSCALE for Windows. Chicago: Scientific Software International; 2003.

Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph. 1969No. 17

StataCorp. Stata Statistical Software: release 10. College Station, TX: StataCorp LP; 2007.

Syme SL, Marmot MG, Kagan A, Kato H, Rhoads G. Epidemiologic studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii and California: introduction. American Journal of Epidemiology 1975;102:477–480. [PubMed: 1202949]

Teng EL, Chui HC. The Modified Mini-Mental State (3MS) examination. Journal of Clinical Psychiatry 1987;48:314–318. [PubMed: 3611032]

Teng EL, et al. The Cognitive Abilities Screening Instrument (CASI): a practical test for cross-cultural epidemiological studies of dementia. International Psychogeriatrics 1994;6:45–58. [PubMed: 8054493]discussion 62.

Teresi JA, Golden RR, Cross P, Gurland B, Kleinman M, Wilder D. Item bias in cognitive screening measures: comparisons of elderly white, Afro-American, Hispanic and high and low education subgroups. Journal of Clinical Epidemiology 1995;48:473–483. [PubMed: 7722601]

Teresi JA, Kleinman M, Ocepek-Welikson K. Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. Statistics in Medicine 2000;19:1651–1683. [PubMed: 10844726]

Teresi JA, Stewart AL, Morales LS, Stahl SM. Measurement in a multiethnic society: overview to the special issue. Medical Care 2006;44:S3–S4. [PubMed: 17060831]

White L, et al. Prevalence of dementia in older Japanese-American men in Hawaii: The Honolulu-Asia Aging Study. Journal of the American Medical Association (JAMA) 1996;276:955–960.

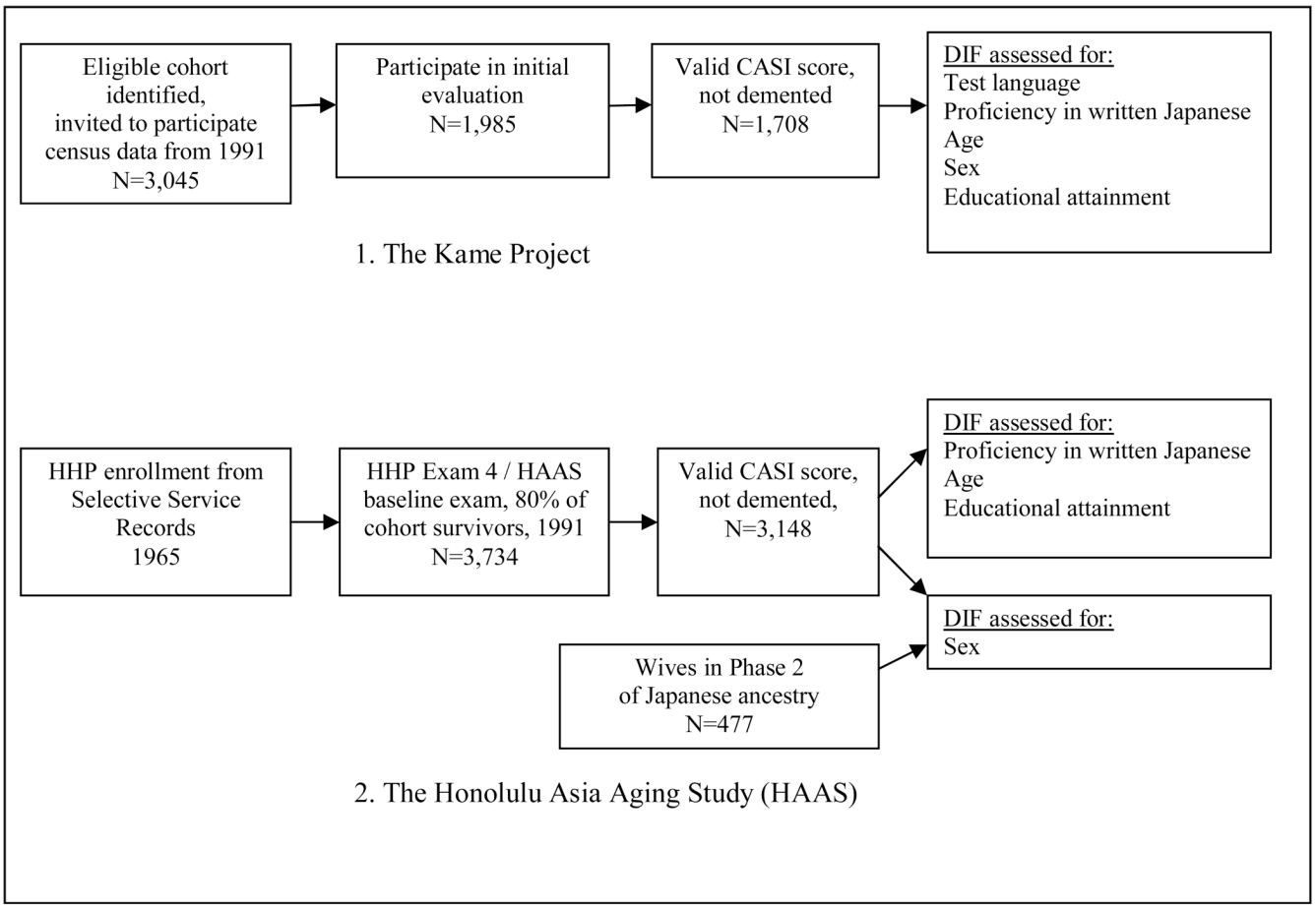

Figure 1.
Schematic representation of study designs and covariates examined for DIF.

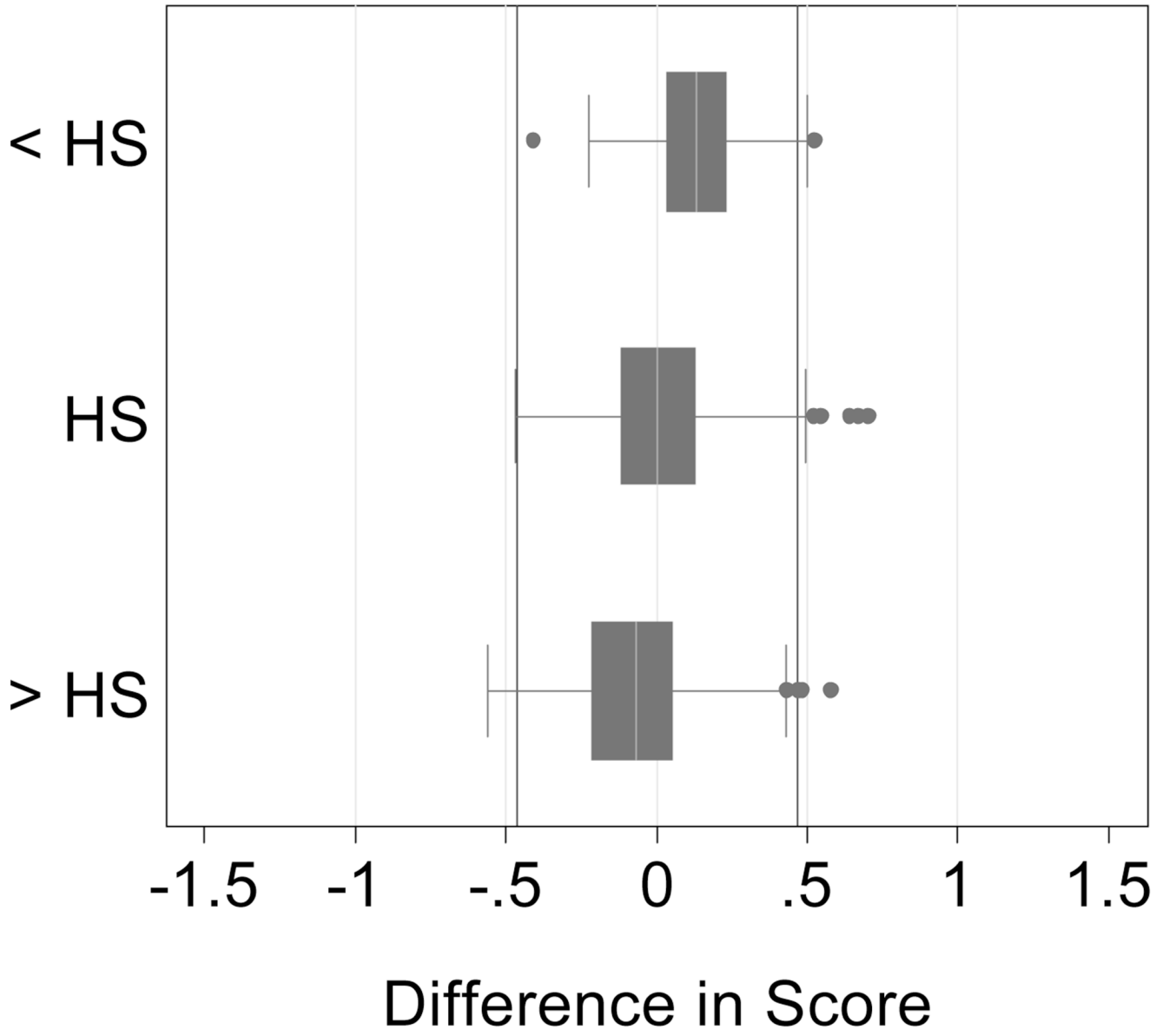

Figure 2.
This figure demonstrates the minimal impact of DIF with respect to educational attainment in *Kame*, operationalized as the difference between the IRT CASI score accounting for DIF and the naïve IRT CASI score. In the box-and-whiskers plots, the box spans the 25th to 75th percentiles, with the median indicated. The whiskers define 1 ½ times the inter-quartile range; individual observations more extreme than this are indicated with dots. The bold vertical lines indicate the median value of the standard error of measurement for the population, the threshold for *salient* DIF.

Table 1

Demographic characteristics of participants in the *Kame* Project and HAAS

	Kame		HAAS	
Characteristic	**n**	**%**	**n**	**%**
Language of testing				
English	1348	79	2439	95
Japanese or both	352	21	120	5
Do you read or write Japanese now?				
No	864	53	1761	58
Yes, but with a lot of difficulty	221	14	424	14
Yes, but with some difficulty	201	12	504	17
Yes, no difficulty	340	21	344	11
Age, years				
64–74	698	41	1018	32
75–79	590	35	1363	43
80–98	420	25	767	24
Sex				
Female	964	56	0*	0
Male	744	44	3148	100
Education, years				
0–8	124	7	898	29
9–11	203	12	692	22
12	729	43	926	29
13+	649	38	632	20

*477 wives of Japanese descent who completed the CASI were included in the analyses of DIF related to sex.

Table 2

Item-level findings of differential item functioning related to language of testing (Kame), age, and education.*

	Language of testing		Age**				Education***			
	Kame		Kame		HAAS		Kame		HAAS	
Item	**non-uniform**	**uniform**	**non-uniform**	**uniform**	**non-uniform**	**uniform**	**non-uniform**	**uniform**	**non-uniform**	**uniform**
Birth place	-	-	-	-	0.667	−0.9	-	-	0.415	−4.6
Year of birth	-	-	-	-	-	-	-	-	-	0.3
Date of birth	-	-	-	-	0.727	1.9	-	-	0.062	-
Age	0.200	4.7	0.354	2.3	**< 0.001**	−2.0	0.206	−4.6	0.477	2.7
Minutes in an hour	-	-	-	-	0.203	−0.3	-	-	0.740	−1.6
Direction of sunset	0.425	9.0	0.166	**18.2**	0.430	4.5	0.104	1.8	0.039	−5.4
Instant recall 1	**0.002**	**−11.2**	0.316	**−17.1**	**< 0.001**	**−17.8**	**< 0.001**	**−22.9**	0.661	**−15.5**
Instant recall 2	-	-	-	-	0.003	**−11.7**	-	-	0.734	−1.8
Digits backwards 1	-	-	-	-	0.353	−2.4	-	-	0.587	−2.0
Digits backwards 2	0.692	1.8	0.429	2.2	0.747	1.9	0.415	−0.7	0.759	−3.7
Digits backwards 3	0.027	1.3	0.561	−1.7	0.039	3.5	**< 0.001**	−2.8	0.902	−3.7
Recall 1a	0.298	5.9	0.244	3.8	0.045	0.6	0.146	3.5	0.034	5.1
Recall 1b	**< 0.001**	4.6	0.017	2.9	0.147	2.1	**< 0.001**	4.3	0.586	5.1
Recall 1c	0.157	1.1	0.507	3.9	0.550	2.0	**0.002**	1.9	0.954	4.6
Serial 3 subtraction 1	0.271	3.6	0.059	**21.9**	0.191	6.8	0.273	−5.9	0.240	−2.6
Serial 3 subtraction 2	0.208	−1.0	0.201	−3.6	0.846	5.7	0.293	−7.7	0.803	−4.8
Serial 3 subtraction 3	0.146	5.0	0.516	1.5	0.156	1.2	0.154	−0.7	0.212	−2.9
Year	-	-	-	-	0.285	0.4	-	-	0.590	2.0
Month	-	-	-	-	0.287	0.9	-	-	0.464	2.2
Date	0.015	−0.3	0.028	−3.2	0.184	−0.6	**< 0.001**	−0.5	0.827	3.6
Day of the week	0.329	−5.5	0.803	−5.2	0.205	1.0	0.251	−8.9	0.667	5.3

	Language of testing		Age**				Education***			
	Kame		Kame		HAAS		Kame		HAAS	
Item	non-uniform	uniform	non-uniform	uniform	non-uniform	uniform	non-uniform	uniform	non-uniform	uniform
Season	-	-	-	-	0.864	3.8	-	-	0.494	1.3
Spatial orientation A	-	-	-	-	0.152	0.8	-	-	0.615	−3.0
Spatial orientation B	-	-	-	-	0.923	−3.3	-	-	0.423	−6.5
Animal fluency	0.072	−2.7	0.234	−5.0	0.431	−4.7	0.127	−9.4	0.953	−1.3
Similes	**< 0.001**	−2.3	0.839	−5.6	**< 0.001**	0.3	**< 0.001**	−7.4	0.012	−9.7
Judgment	**< 0.001**	4.3	**< 0.001**	−5.5	**< 0.001**	−0.4	**< 0.001**	−2.9	0.403	−3.4
Repeat phrase 1	0.751	**−13.9**	0.488	−3.6	0.040	−6.3	0.003	**−20.3**	0.119	−7.2
Repeat phrase 2	0.012	−4.4	0.651	−6.5	0.386	−3.1	0.018	−7.3	0.723	−7.6
Read / follow command	0.051	**−10.1**	0.080	−7.1	0.150	−7.6	0.949	**−18.9**	0.285	**−13.1**
Write a specific sentence	0.532	−6.9	0.557	−0.4	0.582	2.8	0.017	**−14.4**	0.634	**−15.8**
Copy interlocked pentagons	0.006	0.5	0.716	−4.8	0.229	−4.4	**< 0.001**	−9.8	0.946	−3.8
Three step command	**0.001**	5.6	0.739	−5.6	0.072	−3.9	**< 0.001**	0.9	0.243	0.2
Recall 2a	0.528	0.8	0.632	3.4	0.116	0.8	0.343	−0.1	0.835	3.2
Recall 2b	0.026	−0.3	0.338	0.6	0.775	−0.6	**< 0.001**	0.9	0.768	2.9
Recall 2c	0.166	0.0	0.724	−1.6	0.020	0.4	0.353	−0.1	0.081	0.5
Name body parts	0.587	−4.2	0.963	1.4	0.056	−0.6	0.165	−4.7	0.869	−2.6
Object naming 1	-	-	-	-	**< 0.001**	−3.2	-	-	0.373	2.8
Object naming 2	-	-	-	-	-	-	-	-	0.456	0.3
Object recall	0.811	3.3	0.621	−5.3	0.293	−2.1	0.513	7.1	0.498	5.5

* Non-uniform DIF is present (bold) if p <0.002. Uniform DIF is present (bold) if the % change in β_1 is > 10.

** Age categories were 64–74, 75–79 and 80–98 years.

*** Education categories were 0–11, 12, and 13+ years in Kame and 0–8, 9–11, 12, and 13+ years in HAAS.

- Indicates an item did not have enough discordance to analyze with these subgroups