Schizophrenia Bulletin vol. 38 no. 3 pp. 630–641, 2012
doi:10.1093/schbul/sbq136
Advance Access publication on December 6, 2010

The Cultural Adaptability of Intermediate Measures of Functional Outcome in Schizophrenia*

Dawn I. Velligan[1], Maureen Rubin[2], Megan M. Fredrick[1], Jim Mintz[1], Keith H. Nuechterlein[3], Nina R. Schooler[4], Judith Jaeger[5], Nancy M. Peters[6], Raimund Buller[7], Stephen R. Marder[3], and Sanjay Dube[8]

[1]Department of Psychiatry, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, TX 78229-3900; [2]Department of Social Work, University of Texas, San Antonio; [3]Semel Institute for Neuroscience, University of California, Los Angeles; VA Desert Pacific Mental Illness Research, Education and Clinical Center; [4]Department of Psychiatry, State University of New York Downstate Medical Center; [5]Astra Zeneca Pharmaceuticals, Albert Einstein College of Medicine, and Zucker Hillside Hospital, Long Island, New York; [6]Sanofi-Aventis Pharmaceuticals; [7]Lundbeck Pharmaceuticals; [8]School of Medicine, University of Pittsburgh

*To whom correspondence should be addressed; tel: 210-567-5508, fax: 210-567-129, e-mail: velligand@uthscsa.edu

The Measurement and Treatment Research to Improve Cognition in Schizophrenia initiative was designed to encourage the development of cognitive enhancing agents for schizophrenia. For a medication to receive this indication, regulatory agencies require evidence of improvement in both cognition and functional outcome. Because medication trials are conducted across multiple countries, we examined ratings of the cross-cultural adaptability of 4 intermediate measures of functional outcome (Independent Living Scales, UCSD Performance-based Skills Assessment, Test of Adaptive Behavior in Schizophrenia, Cognitive Assessment Interview [CAI]) made by experienced clinical researchers at 31 sites in 8 countries. English-speaking research staff familiar with conducting medication trials rated the extent to which each subscale of each intermediate measure could be applied to their culture and to subgroups within their culture based on gender, geographic region, ethnicity, and socioeconomic status on the Cultural Adaptation Rating Scale. Ratings suggested that the CAI would be easiest to adapt across cultures. However, in a recent study, the CAI was found to have weaker psychometric properties than some of the other measures. Problems were identified for specific subscales on all the performance-based assessments across multiple countries. India, China, and Mexico presented the greatest challenges in adaptation. For international clinical trials, it would be important to use the measures that are most adaptable, to adapt subscales that are problematic for specific countries or regions, or to develop a battery composed of the subscales from different instruments that may be most acceptable across multiple cultures with minimal adaptation.

Key words: schizophrenia/co-primary measures/intermediate measures/international clinical trials/functional capacity measures/cognitive impairments

Introduction

Schizophrenia is an illness characterized by cognitive deficits in the areas of attention, memory, and executive functions.[1–3] These cognitive deficits have been found to be related to impairments in role functioning in individuals with schizophrenia and are considered a core feature of this disorder.[2,3] Efforts to improve outcomes in schizophrenia have increasingly focused on ways to address cognitive impairments, with the ultimate goal of improving functional outcomes.[4–8]

The Measurement and Treatment Research to Improve Cognition in Schizophrenia (MATRICS) initiative was designed to encourage the development of cognitive-enhancing pharmaceutical agents for schizophrenia by developing a process by which a medication could receive an indication for the treatment of cognitive dysfunction in schizophrenia.[9–13] This initiative was a collaboration among academicians, industry partners, and government agencies and resulted in recommendations for study design and the development of a consensus cognitive battery—the MATRICS Consensus Cognitive Battery (MCCB)—to assess cognition in studies of novel compounds seeking this indication.[12,13] Representatives of the U.S. Food and Drug Administration (FDA) indicated that improvement in performance on neuropsychological tests was not sufficient to establish an indication for improving cognition in schizophrenia.[11] The FDA indicated that a compound would also need to demonstrate that it improved a co-primary measure of functional outcome that had more face validity for everyday functioning than cognitive testing.[11,14] This model is similar to that used to get approval for cognitive enhancing medications in dementia, which requires evidence of improvement or slower decline in cognitive functioning and everyday living skills.

As part of the MATRICS initiative, the MATRICS Psychometric and Standardization study was conducted to examine the psychometric properties of a number of co-primary measures.[14] While measures generally had acceptable psychometric properties, it was decided not to endorse a single co-primary measure but to conduct further evaluation of these intermediate assessments of functional outcome.[14] With this purpose in mind, the Validation of Intermediate Measures (VIM) study was developed to assess the reliability, validity, and utility of a number of intermediate measures that have face validity for assessing functional outcome in schizophrenia. Because longer term functional outcomes, such as employment or changes in marital status, are not likely to be improved during the course of a typical clinical trial, the study focused on intermediate measures of functional capacity or everyday functioning that are thought be more amenable to change over this time period.[14] The goal of the VIM study was to identify the measure or measures with the best psychometric properties for use in clinical trials of compounds designed to improve cognition in schizophrenia.[15] Measures selected for investigation in the VIM study were those rated highest by a RAND panel method as having the most promise for use in this context.[15] The measures included 3 performance-based or functional capacity measures in which the subject must perform specific tasks that are rated by an examiner for accuracy and one interview-based measure assessing cognitive problems in everyday life. Functional capacity measures included the Independent Living Scales,[16] the UCSD Performance-based Skills Assessment (UPSA),[17] and the Test of Adaptive Behavior in Schizophrenia (TABS).[18] The interview measure was the Cognitive Assessment Interview.[19] Both the UPSA and TABS have brief versions that were also evaluated for psychometric properties in the VIM study.

Examining the psychometric properties of these measures in a US sample is an important step in finding the most appropriate intermediate measure for use in clinical trials of novel compounds designed to improve cognition. However, many efficacy studies of novel compounds are now conducted as multisite international trials. This necessitates that these measures be applicable in culturally distinct locations. However, the types of everyday activities in the intermediate measures assessed in the VIM study were developed for study participants in the United States, and several of the items in these measures may not reflect everyday activities around the world. Different cultures may influence the degree to which the everyday activities investigated are familiar to and/or are comfortable for a given subgroup of individuals in other countries. Moreover, there may be larger differences with respect to familiarity and comfort in less developed countries among socioeconomic, gender, or ethnic groups, than are found in the US population. Lack of familiarity or comfort with the form or content of the test items could influence the validity of an item as a measure of a person's skills. In fact, there is evidence from cross-cultural studies that societal contexts and environmental differences can influence measurement.[20,21]

We developed the Cross-Cultural Adaptability of Intermediate Measures (CIM) Study to examine which of the intermediate measures examined in the VIM study would be rated as most appropriate for use across cultures, by expert investigators conducting clinical trials. We obtained their ratings of the overall adaptability of each intermediate measure and its applicability across genders, socioeconomic strata, ethnicity, and geographic region (rural vs urban) for patients typically seen at their site. We were also interested in opinions regarding which subtests of each intermediate measure were most adaptable to their country and cultural context. The study was designed and carried out by the MATRICS cross-cultural subcommittee and ran concurrently with the VIM study. Our goal was to identify the measures that would be rated as most applicable for assessing the broadest range of individuals. This study represents only an initial step to guide the selection of co-primary measures for international clinical trials investigating treatments for cognition and functional outcome. Follow-up studies will be needed to validate whether the measures rated most adaptable across multiple cultural contexts perform well in groups of individuals diagnosed with schizophrenia.

Method

A Measure of Cross-cultural Adaptability

The MATRICS cross-cultural subcommittee reviewed the literature on guidelines and methods for adapting tests across cultures including the International Testing Commission Guidelines.[22–28] Typical approaches to ensuring cross-cultural adaptability include a series of meetings among test developers and individuals who are highly knowledgeable about the population to be assessed in the target culture. All aspects of an instrument are discussed with respect to how well or poorly they fit into the cultural context and how they should be modified to make the instrument more appropriate for a given culture. Based on review of the literature, we designed a survey instrument known as the Cross-Cultural Adaptation Rating Scale (C-CARS). The C-CARS was designed to be completed by investigators and frontline staff conducting clinical trials across the world. The C-CARS asks these raters to assess the degree to which each intermediate measure would be appropriate for use in their culture as a way of assessing everyday functioning. The measure contains nine 7-point rating scale items assessing raters' opinions about the overall adaptability of the measure and its use with specific subgroups in the culture based upon gender, socioeconomic status, ethnicity,

and area of residence (rural vs urban). Higher scores on the C-CARS items reflect ratings of better cross-cultural adaptability. A section for detailed comments to describe problems identified in cross-cultural adaptation is included for each question.

Expert Raters

Raters participating in the study were investigators and frontline research assistants at clinical trial sites in the United States, Argentina, China, Germany, India, Mexico, Russia, and Spain. Potential sites were identified by industry partners, who were participants in the MATRICS initiative. Countries chosen were those in which the MCCB was being translated and normative data on the battery was being collected as part of the MATRICS initiative. Through established contacts in each of 8 countries, the industry partners identified investigators in each country with experience in conducting clinical trials in schizophrenia. Raters were not selected for expertise in psychometrics or the assessment of functional outcome but rather for their knowledge of everyday activities in the culture and their experience with the target population for the intermediate measures. Sites were contacted in the order in which their information was received, except that in India sites were chosen in primarily Hindi-speaking regions because this was the language of translation for the MCCB. In the United States and countries thought to be more similar culturally to the United States (ie, Spain and Germany), fewer sites were recruited for participation (2 sites per country), while in less westernized countries thought to be most dissimilar culturally to the United States (ie, Russia, India, China, Mexico, and Argentina), more sites were recruited (5 sites per country). At each site, the goal was to have each intermediate measure independently assessed by a minimum of 2 bilingual (English/language of MCCB translation) raters; ie, one principal investigator and one frontline research assistant. A total of 31 sites from 8 countries were recruited for participation.

Procedures

Once sites agreed to participate in the study, all materials were mailed beginning in May 2009. Materials included copies of the C-CARS for each section of each intermediate measure, copies of each intermediate measure, and a DVD demonstrating administration of each intermediate measure in English. Individual site initiation visits were conducted by telephone/video conferencing to review the study protocol, to provide clarity regarding expectations and procedures, and to address any questions that might arise. All sites were asked to have each rater review all the instruments, manuals, and scoring sheets, view the DVD of the administration of each instrument, and practice administering the sections to peers to get a feel for the intermediate measure. They were then asked to independently complete all C-CARS questions for each subsection of each intermediate measure. Subsections were reviewed rather than the intermediate measures as a whole because it was believed that the subsections of each measure would differ with respect to the extent to which they could be adapted across cultures. In addition, this approach allowed us to have the brief measures rated (ie, Brief UPSA, Brief TABS) separately from their corresponding full measures. Raters were asked to provide comments about any problems they saw with the cross-cultural adaptability of each section and were asked specifically to comment on all sections that were rated as having less than "good = 5" cultural adaptability on C-CARS. All data were faxed, mailed, or emailed to the coordinating site (The University of Texas Health Science Center, San Antonio) for data entry. Because an investigator (D.I.V.) at the coordinating site developed one of the intermediate measures under review (TABS), all data entry was supervised by another investigator from a different institution (M.R.). All sites had email contact with the project staff and were encouraged to ask questions throughout the process. Once sites returned the scored C-CARS forms, key project personnel (M.F., M.R.) reviewed the responses for consistency and flagged comments that needed clarification. Email messages were sent to sites to clarify comments or scores that were not understood. All sites responded to these queries before data were locked for analysis on January 11, 2010.

Intermediate Measures

CAI—The Cognitive Assessment Interview[19] is a semi-structured interview developed from the CGI-CogS[29] and the SCoRS[30] using classical test theory methods and statistical approaches to select the "best" items. The CAI contains 10 items that assess the domains of the MCCB through a clinical interview with the patient alone or patient and caregiver. The rater uses all available information to rate cognition on a series of Likert scales, with higher scores reflecting more severe cognitive impairment. Questions include: "Do you have difficulty keeping figures in mind while paying bills?" "Do you have trouble learning or remembering instructions or other important information?" "Do you have trouble coming up with alternatives when your plans are disturbed?"

ILS—The Independent Living Scales[16] is a performance-based test of competence in instrumental activities of daily living. The items require the examinee to do problem solving, to demonstrate knowledge, or to perform a task. The ILS comprised the following 5 subscales (total of 70 items): Memory/orientation—eg, person is asked to remember the name of a new doctor and the time of an appointment when asked later; Managing money—eg,

person is asked to make out a check/money order to a utility company; Managing home and transportation—eg, person is asked how they would go about getting repairs made to their home; Health and safety—eg, the person is asked what they would do if they cut their hand and it was bleeding badly, and Social adjustment—eg, person is asked to name 2 reasons why it is important to have relationships. The ILS yields a total score and 2 factor scores (1) Problem Solving and (2) Performance/Information. Higher scores reflect better functional performance.

The UPSA[17] is designed to assess an individual's ability to perform functional tasks. The UPSA version 2.0 used in the VIM study assesses 5 skill areas that are considered essential to functioning in the community: General Comprehension—the person is asked to read a newspaper article about the opening of a water park, to remember information, and plan a trip there; Finance—the person is asked to pay a bill, make change, etc. Social/Communications—the person is asked to read a letter from their doctor about an appointment and what to bring with them, to call and reschedule their appointment, and to remember the information when the letter is removed; Transportation—the person is asked to plan a bus route to specific destinations and to answer questions about it; Household Chores—the person is asked to write a shopping list for items needed to prepare a specific dish based on what is present and what is missing from a mock pantry. A summary score is calculated for each subscale as well as a total score. Higher scores reflect better performance.

The TABS[18] was designed to assess underlying abilities needed to complete goal-directed adaptive behavior, such as initiation, planning and sequencing, and problem identification. The TABS comprised 6 test areas: Medication Management—the person is asked to fill a medication container based upon instructions on pill bottles and to remember to call for a refill at a specific time; Empty Bathroom—the person is asked what would be needed to stock an empty bathroom to get ready every day; Shopping Skills—the person is asked how he or she would get to the store by using a map, to remember a grocery list, and to pay for items with a set amount of money; Clothes Closet—the person is asked to select appropriate clothing for specific activities; Work and Productivity—the person is asked to make packets of flyers and stack them for mailing, and Social Skills—basic skills such as voice volume and eye contact are rated during the assessment. Scores for each subtest and the total score are the percent correct. Higher scores indicate better adaptive functioning.

Brief Scales. The UPSA and the TABS have brief versions available. The UPSA Brief is composed of subtests assessing Finance and Communication. The TABS Brief is composed of subtests assessing Medication Management and Work.

Data Analysis

Data analyses were designed to answer the following research questions.

1. To what extent are the intermediate measures rated as being adaptable to different cultural contexts?
2. Are measures rated differently with respect to cultural acceptability/adaptability based upon gender, rural vs urban residence, socioeconomic status (SES), or ethnic minority status?
3. Are there specific subscales on the functional measures that are rated as likely to be more adaptable than others across countries?

Interrater reliability was examined by subscale within scale and country. We report a percentage agreement statistic and a weighted kappa statistic that corrects for chance agreement. Reliability was calculated within site and averaged across sites and countries. The index of agreement was calculated by summing the number of C-CARS item ratings of the same subscale made by different raters that were within one rating point of each other and dividing by the number of paired ratings. Chance agreement was calculated using the marginal frequencies of the C-CARS rating points based on the entire data set. The complete factorial design of the study is complex. The statistical design has Raters nested within countries. These factors are crossed with the within-rater, repeated measures factors, Tests, and Subscales nested within Tests. We simplified the analyses examining ratings of cultural adaptability overall (Item 1 on the C-CARS) and with respect to gender, minority status, SES, and rural/urban residence by collapsing across subscales of the tests. We also calculated differences scores for ratings of adaptability between genders (gender sensitivity), the majority and minority population (ethnic sensitivity), high and low socioeconomic groups (SES sensitivity), and rural and urban dwellers (region sensitivity). Comparisons of means were done using mixed effects regression models. We focused on planned comparisons between the United States (where the scales were all developed) and each of the other countries. Computing all pairwise differences between countries would have resulted in a large number of post hoc comparisons that would not be particularly useful. To correct for experiment-wise error, we used the Holms–Bonferroni approach. This is a step-down method in which comparisons are ordered from most to least significant. For each research question, the most significant pairwise comparison was examined at the corrected alpha for the total number of comparisons and each subsequent test was examined at alpha divided by the number of remaining comparisons (SAS Institute, 2002–2008). In addition to the questions addressed using the data analytic model, we also report on the comments provided by raters to identify specific problems in adaptation.

Table 1. Expert Raters Requested/Participating in the Cross-Cultural Adaptability of Intermediate Measures Study

Country	Requested Number of Raters	Obtained Number of Raters	Degrees	Comment
Argentina	10	8	MD/PhD = 7, BS = 1	One site withdrew completely; one site had no research assistant and one site had 2 research assistants.
China	10	7	MD/PhD = 7	One site withdrew; one principal investigator sent only comments and no ratings.
Germany	4	4	MD/PhD = 3, RN = 1	
India	10	8	MD/PhD = 5, MA/MS = 3	One site sent ratings for only one intermediate measure and did not respond to contact attempts.
Mexico	10	8	MD/PhD = 6, MA/MS = 3	Five sites participated, but 1 site had no research assistant and the principal investigator at 1 site did not speak English well enough to participate.
Russia	10	11	MD/PhD = 11	One site had no research assistant and 2 sites had 2 research assistants.
Spain	4	4	MD/PhD = 4	
United States	4	5	MD/PhD = 5	One site had 2 research assistants.

Results

Study Overview

Table 1 lists the countries participating, the number of raters requested per country, and the number of raters completing the study. In some cases, a site had more than one research assistant. In some cases in which a principal investigator did not speak English well enough, only a research assistant rated the intermediate measures. Overall, 56 of a minimum target number of 62 individuals were expert raters for the CIM study. Raters had on average 12.47 (SD = 7.96) years of clinical trial experience in schizophrenia. The majority of raters (85.7%; $n = 48$) had advanced degrees (MD, MD/PhD, PhD). These data appear in table 1.

Interrater Reliability of the C-CARS. Median percent interrater agreement varied from 74.1% in Mexico to 98.4% in Germany. The overall kappa statistic for interrater agreement corrected for chance averaged across countries and tests was 0.70. This is considered in the acceptable range. Kappa was greater than 0.77 for all countries with the exception of China ($\kappa = 0.21$) and Mexico ($\kappa = 0.35$), which fell well below the acceptable range.

To What Extent Are the Intermediate Measures Rated as Being Adaptable to Different Cultural Contexts?

In an effort to determine in general how culturally appropriate the intermediate measures were rated across different countries, we examined item 1 of the C-CARS in a two-way (scale × country) factorial analysis of variance. Item 1 asks clinical researchers to rate how each subscale of each intermediate measure would work with typical patients at their sites. Data were averaged across subscales for each test. Results appear in table 2. The interview measure, the CAI, was rated significantly higher than all other measures. Differences between the TABS and ILS were not significant but both tests were rated significantly higher than the UPSA. Cultural adaptability was rated significantly lower in India than in the United States for the TABS, ILS, and UPSA, significantly lower in China for the ILS, and significantly lower in both Mexico and China for the UPSA. Results for the UPSA Brief and TABS Brief scales on C-CARS item were similar to those for the full scales. Our a priori cutoff of 5 coincides with significant

Table 2. CCARS1—Global Rating

	Full Scales					Brief Scales	
	CAI	TABS	ILS	UPSA	Mean	TABS	UPSA
United States	6.03	6.07	6.12	6.00	6.06	5.80	6.10
Germany	6.92	6.29	5.45	5.95	6.15	6.38	6.25
Argentina	6.60	6.04	5.95	5.30	5.97	5.63	5.81
Spain	6.46	5.88	5.45	5.45	5.81	4.88	5.13
Russia	6.39	6.01	5.55	5.22	5.79	5.59	5.10
Mexico	6.31	5.39	5.70	4.84*	5.56	4.89	4.61**
China	5.81	5.71	5.17*	4.63*	5.33	5.50	4.50**
India	6.10	4.60**	4.98**	3.98**	4.91	4.13*	3.19**
Mean	6.33	5.75	5.54	5.17	5.65	5.35	5.09

Note: CAI, Cognitive Assessment Interview; TABS, Test of Adaptive Behavior in Schizophrenia; ILS, Independent Living Scales; UPSA, UCSD Performance-based Skills Assessment. Estimates from mixed effects regression. Asterisks indicate significantly poorer adaptability than in the United States value (*$P < .05$, **$P < .01$ Holm–Bonferroni adjusted). Root mean squared SE = 0.26 (range: 0.17–0.36). Overall scale means: CAI > TABS > ILS > UPSA, all pairwise differences significant at Holm–Bonferroni $P = .001$ except TABS vs ILS, $P = .017$.

Table 3. CCARS1-Global Rating Report Card

	Full Scales				Brief Scales	
	CAI	TABS	ILS	UPSA	TABS	UPSA
United States	A	A	A	A	A	A
Germany	A	A	A	A	A	A
Argentina	A	A	A	C	B	A
Spain	A	A	A	F	F	F
Russia	A	A	A	D	A	B
Mexico	A	D	B	F	C	F
China	B	B	C	F	C	F
India	A	F	F	F	F	F
GPA	3.9	3.0	3.1	1.4	2.4	1.9

Note: Letter grades assigned conventionally (90–100 = A, 80–89 = B, 70–79 = C, 60–69 = D, <60% = F). Grade Point Average based on A = 4, B = 3, C = 2, D = 1. Abbreviations are explained in the first footnote to table 2.

differences between the United States and other countries, providing some validation of its use as a metric.

To give readers a quick way to examine the adaptability of the intermediate measures across countries, we calculated the percentage of ratings that met the a priori cutoff score of 5 or higher (good or better) for each intermediate measure for all subscales and raters and translated this percentage into a letter grade where 90–100 = A, 80–89 = B, 70–79 = C, 60–69 = D, <60% = F. These results appear in table 3.

Are Measures Rated Differently With Respect to Cultural Acceptability/Adaptability Based Upon Gender, Rural vs Urban Residence, SES, or Ethnic Minority Status?

The mean C-CARS scores for cultural adaptability based upon gender by scale and country appear in table 4. Adaptability for females was rated significantly lower in India than the United States for all measures but the CAI. Adaptability for males was rated as lower in India than the United States for the TABS and UPSA and lower in China than the United States for the UPSA. Gender sensitivity scores were calculated by subtracting C-CARS ratings for females from those of males. Results indicate that in every country but India and China the TABS was rated as more adaptable for females. The UPSA was rated as more adaptable for females in Russia, Mexico, and India. All tests but the UPSA were rated as more adaptable to males ONLY in India. Only the gender sensitivity differences between the United States and India on the CAI, TABS, and ILS were statistically significant.

Table 5 presents cultural adaptability ratings for different socioeconomic strata. Adaptability for low SES individuals was rated as significantly lower in India than the United States for the TABS, ILS, and UPSA and significantly lower in China and Mexico than the United States for the UPSA. These data suggest that differences for the overall level of adaptability reflect problems particularly in adaptation to lower SES groups. An SES sensitivity scale was calculated by taking the C-CARS score for the highest SES and subtracting the score for the lowest SES. Comparison of means suggested that the TABS, ILS, and UPSA were rated as less adaptable to the lower SES individuals in Mexico and India than in the United States.

Table 6 presents the C-CARS ratings for ethnic minority status by country and scale. Adaptability for individuals of ethnic minority status was rated lower for India than the United States with respect to the ILS and lower for Mexico than the United States with respect to the UPSA .A score for sensitivity to ethnic minority status was created by subtracting the ethnic minority score from the score for the overall cultural adaptability (C-CARS1). While all scales were rated as less adaptable

Table 4. Adaptability by Respondent Gender

	Full Scales				Brief Scales	
	CAI	TABS	ILS	UPSA	TABS	UPSA
A. Adaptability for use with FEMALES						
United States	6.03	6.17	6.12	6.08	5.80	6.10
Germany	6.92	6.46	5.40	5.95	6.38	6.13
Argentina	6.56	6.23	6.06	5.60	5.81	6.0
Spain	6.46	5.88	5.45	5.40	4.88	5.00
Russia	6.52	6.08	5.76	5.90	5.59	5.55
Mexico	6.31	5.89	5.91	5.31	5.44	5.33
China	5.81	5.88	5.29	5.06	5.79	5.14
India	6.06	4.71**	4.93**	4.05**	4.13*	3.19**
Mean	6.33	5.91	5.61	5.42	5.48	5.3
B. Adaptability for use with MALES						
United States	6.03	5.97	6.12	6.04	5.80	6.10
Germany	6.92	6.29	5.40	5.75	6.38	6.25
Argentina	6.58	6.02	6.10	5.48	5.81	6.00
Spain	6.46	5.67	5.50	5.35	4.88	5.13
Russia	6.50	5.98	5.67	5.52	5.55	5.40
Mexico	6.28	5.76	5.98	5.09	5.33	5.17
China	5.81	5.83	5.29*	4.94	5.79	5.14
India	6.4	4.98*	5.58	4.28**	4.44	3.81**
Mean	6.37	5.81	5.70	5.30	5.50	5.37
C. Gender sensitivity (Section A minus Section B)						
United States	0.00	**0.20**	0.00	0.04	0.00	0.00
Germany	0.00	**0.17**	0.00	0.20	0.00	−0.13
Argentina	−0.02	**0.21**	−0.04	0.13	0.00	0.00
Spain	0.00	**0.21**	−0.05	0.05	0.00	−0.13
Russia	0.02	**0.11**	−0.09	**0.38**	0.05	0.15
Mexico	0.04	**0.13**	−0.09	**0.22**	0.11	0.17
China	0.00	0.05	0.00	0.11	0.00	0.00
India	**−0.33***	**−0.27****	**−0.65****	**−0.23**	**−0.31***	**−0.63****
Mean	−0.04	**0.10**	**−0.09**	**0.11**	−0.02	−0.07

Note: In Sections A-C, entries marked with * differ significantly from the United States value. In Section C, bold entries differ significantly from zero, indicating Gender sensitivity. Positive values indicate greater adaptability for females. Root mean squared standard error of means: Section A = 0.26 (range = 0.16–0.39), Section B = 0.26 (range = 0.14–0.41). Abbreviations are explained in the first footnote to table 2.

Table 5. Adaptability by Respondent SES

	Full Scales				Brief Scales	
	CAI	TABS	ILS	UPSA	TABS	UPSA
A. Adaptability for use with LOWER SES						
United States	5.93	5.93	5.92	5.92	5.70	6.00
Germany	6.54	5.75	4.80	5.60	5.63	5.75
Argentina	6.38	5.56	5.23	5.08	5.13	5.06
Spain	6.33	5.67	5.2	5.15	4.63	4.75
Russia	6.24	5.87	5.33	5.26	5.55	5.1
Mexico	5.37	4.94	4.98	4.16*	4.83	3.94**
China	5.74	5.12	4.91	4.14*	4.93	4.07**
India	5.90	4.46**	4.35**	3.45**	4.06*	2.88**
Mean	6.05	5.41	5.09	4.84	5.06	4.69
B. Adaptability for use with MIDDLE SES						
United States	6.03	6.10	6.16	6.04	5.90	6.10
Germany	6.92	6.21	5.40	6.05	6.13	6.25
Argentina	6.65	6.17	6.05	5.65	5.75	6.00
Spain	6.46	5.88	5.50	5.55	4.88	5.13
Russia	6.47	6.08	5.76	5.89	5.55	5.60
Mexico	6.31	5.87	5.87	5.38	5.61	5.39
China	5.81	5.86	5.37	5.03	5.71	5.00
India	6.54	5.75	5.93	5.28	5.13	4.75*
Mean	6.40	5.99	5.76	5.61	5.58	5.53
C. Adaptability for use with HIGHER SES						
United States	6.03	6.10	6.16	5.96	5.90	6.10
Germany	7.00	6.42	5.70	6.40	6.50	6.75
Argentina	6.65	6.27	6.18	5.85	5.75	6.31
Spain	6.46	5.88	5.60	5.60	4.88	5.25
Russia	6.38	6.11	5.87	5.78	5.55	5.65
Mexico	6.33	6.09	6.29	5.91	5.72	6.06
China	5.90	6.14	5.57	*5.29*	5.93	5.29
India	6.75	*6.33*	6.42	*5.95*	5.63	5.69
Mean	6.44	6.17	5.97	5.84	5.73	5.89
D. SES sensitivity (Section C minus Section A)						
United States	0.10	0.17	0.24	0.04	0.20	0.10
Germany	0.46	**0.67**	**0.90**	0.80	**0.88**	**1.00**
Argentina	0.27	**0.71**	**0.95**	**0.78**	**0.63**	**1.25**
Spain	0.13	0.21	0.40	0.45	0.25	0.50
Russia	0.14	0.24	**0.55**	**0.54**	0.00	0.55
Mexico	**0.96**	**1.15***	**1.36***	**1.76****	**0.89**	**2.11****
China	0.17	**1.02**	**0.66**	**1.14**	**1.00**	**1.21**
India	**0.85**	**1.88****	**2.07****	**2.50****	**1.56****	**2.81****
Mean	**0.38**	**0.75**	**0.89**	**1.00**	0.68	1.19

Note: In Sections A–D, marked with * differ significantly from the United States value. In Section D, bold entries differ significantly from zero, indicating SES sensitivity. Positive values indicate greater adaptability for higher SES. Root mean squared standard error of means: Sections A = 0.37 (range = 0.22–0.55), Section B = 0.24 (range = 0.16–0.37), Section C = 0.22 (range = 0.14–0.31). Abbreviations are explained in the first footnote to table 2.

Table 6. Adaptability for ETHNIC MINORITIES

	Full Scales				Brief Scales	
	CAI	TABS	ILS	UPSA	TABS	UPSA
A. Adaptability for use with ETHNIC MINORITY						
United States	5.87	5.73	5.92	5.84	5.50	6.00
Germany	6.50	5.88	4.90	5.40	5.63	5.75
Argentina	6.48	5.67	5.60	5.11	5.44	5.25
Spain	6.33	5.42	5.05	5.15	4.63	4.75
Russia	6.15	5.64	5.38	5.39	5.36	5.06
Mexico	5.52	4.76	4.86	4.04*	4.67	3.67**
China	5.60	5.43	5.06	4.43	5.43	4.50
India	6.06	4.71	4.69*	4.40	4.19	3.94**
Mean	6.06	5.40	5.18	4.97	5.10	4.86
B. ETHNIC MINORITY SENSITIVITY (Section A minus table 1)						
United States	−0.17	−0.33	−0.20	−0.16	−0.30	−0.10
Germany	**−0.42**	−0.42	−0.55	−0.55	**−0.75**	−0.50
Argentina	−0.13	**−0.38**	−0.35	−0.19	−0.19	−0.56
Spain	−0.13	**0.46**	−0.40	−0.30	−0.25	−0.38
Russia	−0.24	**0.37**	−0.16	0.16	−0.23	−0.04
Mexico	**0.80**	**0.63**	**0.73**	**0.80**	−0.22	**0.94**
China	−0.21	−0.29	−0.11	−0.20	−0.07	0.00
India	−0.04	0.10	−0.28	0.43	0.06	**0.75**
Mean	**0.27**	**0.35**	**0.35**	**0.20**	−0.24	−0.22

Note: In Section A–B, entries marked with * differ significantly from the United States value. In Section B, bold entries differ significantly from zero, indicating ETHNIC MINORITY sensitivity. Negative values indicate poorer adaptability with ethnic minorities. Root mean squared standard error of means in Section A = 0.33 (range = 0.22–0.48). Abbreviations are explained in the first footnote to table 2.

to ethnic minorities in Mexico, the CAI was rated as less adaptable for ethnic minorities in Germany and the TABS was rated as less adaptable for ethnic minorities in multiple countries, none of these sensitivity measures was significantly different from the ethnic minority/general population difference in the United States.

Table 7 presents mean C-CARS ratings for rural and urban residence by scale and country. The TABS, ILS, and UPSA were rated as being less adaptable to rural residents in India and Mexico, and the ILS and UPSA were rated as being less adaptable to rural residents in China. All scales were more sensitive to urban/rural differences in India and Mexico than in the United States. All performance-based scales were more sensitive to urban/rural differences in China than in the United States.

Are There Specific Subscales on the Functional Measures That Are Rated as Likely to be More Adaptable Than Others Across Countries?

We also examined the individual subscales of the intermediate measures with respect to ratings of adaptability across cultures. All CAI subscales were rated close to the mean for all countries. Therefore, these data are not presented in figure form. Figures 1–3 present the data for the UPSA, ILS, and TABS. For the TABS and the UPSA which have Brief versions—subscales included in the brief versions are underlined.

UPSA

With respect to the a priori C-CARS cutoff score for acceptable performance, Household Management and Comprehension which are not included in the UPSA

Table 7. Adaptability by URBAN-RURAL Residence

	Full Scales				Brief Scales	
	CAI	TABS	ILS	UPSA	TABS	UPSA
A. Adaptability for use in URBAN areas						
United States	6.07	6.10	6.16	6.04	5.90	6.10
Germany	6.92	6.29	5.40	6.05	6.13	6.25
Argentina	6.65	6.21	6.06	5.70	5.81	6.25
Spain	6.46	5.88	5.50	5.55	4.88	5.13
Russia	6.45	6.18	5.73	5.86	5.73	5.69
Mexico	6.37	5.85	5.87	5.42	5.28	5.50
China	5.86	5.93	5.37	5.06	5.79	5.00
India	6.63	5.85	6.16	5.58	5.31	5.38
Mean	6.42	6.04	5.78	5.66	5.6	5.66
B. Adaptability for use in RURAL areas						
United States	5.97	6.03	6.04	5.84	5.80	6.00
Germany	6.88	6.25	5.10	5.85	6.13	6.13
Argentina	6.21	5.75	5.38	5.05	5.44	5.31
Spain	6.38	5.79	5.45	5.20	4.63	5.13
Russia	6.18	5.67	5.24	5.12	5.45	5.05
Mexico	5.31	4.91*	4.80*	3.87**	4.72	3.94**
China	5.36	5.02	4.51**	3.97**	4.93	3.79**
India	5.71	4.65**	4.46**	3.43**	4.38	3.06**
Mean	6.00	5.51	5.12	4.79	5.13	4.80
C. URBAN-RURAL sensitivity (Section A minus Section B)						
United States	0.10	0.07	0.12	0.20	0.10	0.10
Germany	0.04	0.04	0.30	0.20	0.00	0.13
Argentina	0.44	**0.46**	**0.68**	**0.65**	0.38	**0.94**
Spain	0.08	0.08	0.05	0.35	0.25	0.00
Russia	0.27	**0.51**	**0.49**	**0.74**	0.27	**0.63**
Mexico	1.06	**0.94****	**1.17****	**1.56****	**0.56**	**1.56****
China	**0.5**	**0.90****	0.86	1.09	**0.86**	**1.21**
India	0.92	**1.21****	**1.71****	**2.15****	**0.94**	**2.31****
Mean	**0.43**	**0.53**	**0.67**	**0.87**	0.42	0.86

Note: In Sections A-C, entries in bold italics differ significantly from the US value. In Section C, bold entries differ significantly from zero, indicating URBAN-RURAL sensitivity. Positive values indicate greater adaptability for URBAN residents. Root mean squared standard error of means: Section A = 0.25 (0.15–0.37), Section B = 0.36 (0.21–0.52). Abbreviations are explained in the first footnote to table 2.

brief, were rated as most adaptable across all countries; both failing to meet the cutoff of "5" in only one country. With respect to the UPSA brief, both the Communication and Finance subtests were rated below the acceptable level in 4 countries.

ILS

With respect to the C-CARS cutoff score, the Money Subscale failed to reach the criterion in every country, and the Home subscale failed to meet the criterion in India. Otherwise, the subscales were rated as likely to perform well across countries.

TABS

With respect to the a priori cutoff score on the C-CARS, the Medication subtest failed to meet the criterion in all but one country, and the bathroom subtest failed to meet the criterion in India. While there was variability, all other subtests were rated as in the acceptable range in all countries.

What Specific Problems in Cross-cultural Adaptation Were Identified?

While the specifics of problems identified with each scale and suggestions for better cross-cultural adaptation are presented in detail in another manuscript; here, we briefly mention the types of issues encountered. The problems identified with the specific subtests typically involved the context of the test, specific props used, or what the person would be asked to do. For example, raters indicated that residents would likely be unfamiliar with specific locations, such as a water park (UPSA) or with the specific type of store or bathroom pictured (TABS). Moreover, raters indicated that residents would likely be unfamiliar with specific items, such as medication bottles that are individualized (TABS) or insurance cards (UPSA). There were also specific activities that raters indicated would not regularly be performed by residents, such as paying a bill (UPSA, ILS), paying taxes (ILS), or filling a medication container (TABS). While the specific content of items such as these can be altered to better fit a specific cultural context, there is a question as to whether this alteration would change the nature and the demands of the test.

Discussion

Overall, the interview measure (CAI) was rated as most easily adapted to other countries. The CAI requires the rater to make a judgment about the extent to which everyday activities (eg, reading a newspaper) are negatively impacted by cognitive problems (eg, concentration problems). As in rating symptoms, raters in other countries are used to making judgments based upon the report of the patient alone or the patient and their caregiver. Unfortunately, of the 4 measures examined in the VIM study, the CAI was found to have the weakest relationships to the MCCB and to an interview-based measure of functional outcome.[15]

The ILS and TABS were not rated as significantly different from one another, but both were rated as more culturally acceptable than the UPSA. In general, performance-based measures were rated as being less culturally adaptable in India, China, and Mexico than in other countries. Raters also identified difficulties in adapting the different subtests of the performance-based tests to specific countries. There were also problems identified by raters when considering adaptation for rural dwellers, lower SES, and ethnic minority subgroups. This was particularly true of the UPSA. While it would make sense to eliminate very rural residents from clinical

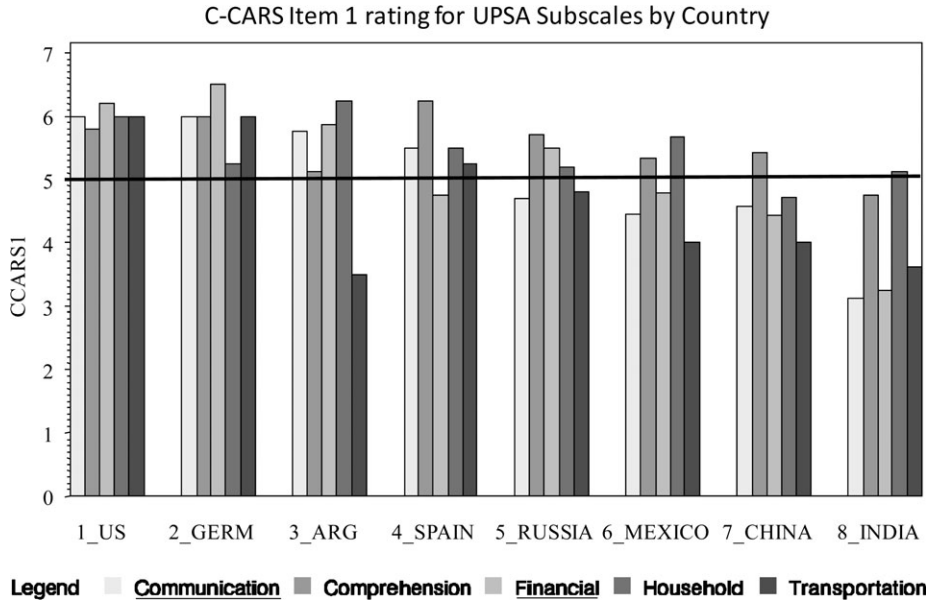

Fig. 1. Cross-cultural Adaptation Rating Scale Item 1 Rating for UCSD Performance-Based Skills Assessment Subscales by Country.

trials across countries, industry partners involved in the MATRICS initiative indicated that eliminating lower SES groups and ethnic minorities would likely create substantial problems in both recruitment and generalization of findings for large-scale pharmaceutical trials.

These data must be judged in the context of methodological limitations. Countries selected and numbers of sites invited to participate in the CIM study were based upon the impression of the cross-cultural subcommittee of the MATRICS group regarding which countries were likely to be less or more similar to the United States (5 vs 2 sites invited; respectively). These impressions may not have been accurate. The C-CARS was developed specifically for this trial. While interrater reliability was generally acceptable, agreement among raters regarding the cultural acceptability of the intermediate measures was poor in Mexico and China. Moreover, no patients were assessed using the intermediate measures. Rather, we asked expert clinical assessors to judge the degree to which each measure would apply to typical patients in their culture. Ultimately, the validity of these ratings must be based on experience using the scales in the field. Future studies need to examine the relationships among intermediate measures, community functioning, and cognitive performance in patient samples across countries.

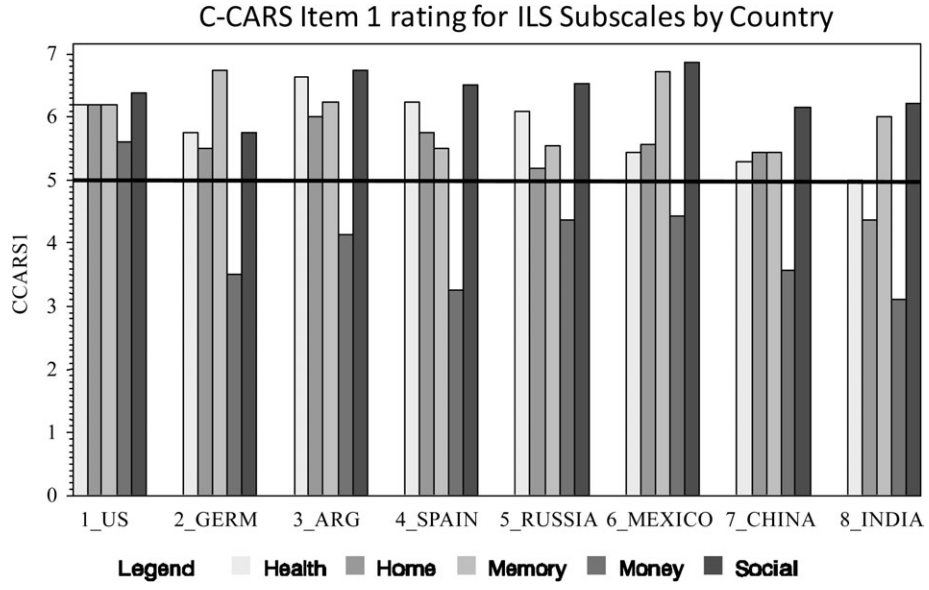

Fig. 2. Cross-cultural Adaptation Rating Scale Item 1 Rating for Independent Living Scales Subscales by Country.

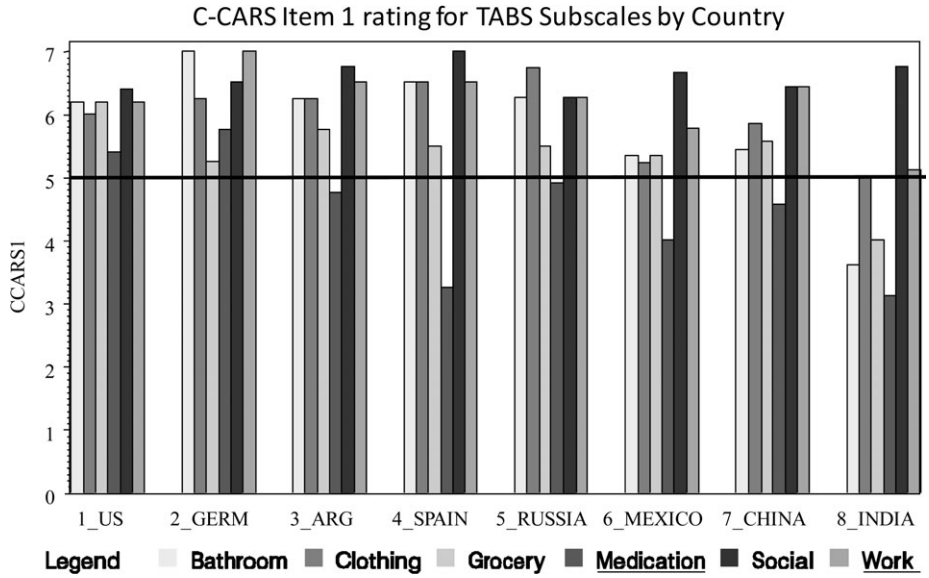

Fig. 3. Cross-cultural Adaptation Rating Scale Item 1 Rating for Test of Adaptive Behavior in Schizophrenia Subscales by Country.

Despite these limitations, these results represent an important first step in identifying and addressing problems in the cross-cultural adaptation of intermediate measures of functional outcome for use in international clinical trials. The goal of identifying or developing a co-primary measure of functional outcome that will be applicable across cultures is an important one. The World Health Organization has been working to identify functional measures that are applicable to multiple countries. For example, while not useful as a measure of change in a short-term intervention trial, the reliability and validity of the WHO Disability Assessment Schedule across cultures has led investigators to conclude that "universal measures of disability are feasible for use in patients with long-term physical and psychiatric illnesses."[31] The MATRICS initiative will investigate whether a universal measure of functional outcome for application to clinical trials in schizophrenia is feasible.

There are several ways to move forward given these data. One approach is to modify the subtests of the intermediate measures for use in each country where problems in adapting the measure to the culture were identified. For example, with respect to the UPSA, each time a request is made to use the scale in a specific country, the scale is adapted collaboratively by the author and investigators in that country to better reflect everyday functioning in that country or specific region of that country. This is done while attempting to keep the cognitive demands of the test as similar as possible to the United States/English version of the test.[20] Empirical evidence is needed to determine the success of this approach and whether the adapted intermediate measures would maintain their test-retest reliability and concurrent validity. This approach has been used in altering the UPSA for different cultures.[20] However, data to support the validity and reliability of the UPSA are not available in many of the countries in which the test has been adapted.

A second approach would be to go back to the data generated by the VIM Study. These data could be examined to identify other subtest combinations with acceptable psychometric properties for inclusion in a new intermediate measure or battery of intermediate measures (including subtests from more than one intermediate measure) that could then be tested in samples of patients in the United States and other countries. For example, the majority of subtests from the ILS and TABS, as well as the Household Management and Comprehension items of the UPSA were judged as working well across multiple cultures. Using existing data from the VIM study, the investigators may be able to pick subtests for a new measure that are likely to have the best psychometrics and be easily adaptable across cultures. The reliability and validity of any identified subtests would then need to be tested across cultures in actual patients in this new combination. This 2-step method of identifying tests with good psychometric properties in the United States and those rated as acceptable in other cultures would have been helpful in the development of the MCCB. While several tests for verbal memory with acceptable psychometric properties were identified, the measure chosen cannot be administered in China due to different structural components of the Chinese language.

Neither of the 2 approaches outlined above may work in countries that are very culturally dissimilar to the United States. In such locales, the existing measures may simply be too inappropriate for adaptation or reconfiguring. If so, an entirely new functional outcome measure could be developed in a country with large differences from the United States, such as India.

Attempts could be made to choose relevant functional behaviors that would also translate to more westernized cultures. It is critical that any co-primary measure of functional outcome utilized in multisite trials of novel medications for schizophrenia be relevant for individuals across countries. Future research will be needed to examine the options discussed above.

Funding

This work was supported by an extension to contract HHSN 278 2004 41003C from the National Institute of Mental Health (to S.M., Principal Investigator).

Acknowledgments

The authors wish to thank the members of the MATRICS-CT (Co-primary and Translation) Scientific Board and the Cross-cultural Subcommittee each consisting of representatives from academia, pharmaceutical industry, National Institute of Mental Health, and the Foundation at NIH (FNIH). This Board and subcommittee provided excellent input and guidance for the study described in this article. We also wish to thank the investigators and research assistants, listed below, at sites across the world for their time and effort on this project. The Authors have declared that there are no conflicts of interest in relation to the subject of this study.

References

1. Gold JM, Harvey PD. Cognitive deficits in schizophrenia. *Psychiatr Clin North Am*. 1993;16:295–312.
2. Green MF. What are the functional consequences of neurocognitive deficits in schizophrenia. *Am J Psychiatry*. 1996;153:321–330.
3. Velligan DI, Mahurin RK, Diamond PL, et al. The functional significance of symptomatology and cognitive function in schizophrenia. *Schizophr Res*. 1997;25:21–31.
4. Velligan DI, Diamond P, Mintz J, et al. The use of individually tailored environmental supports to improve medication adherence and outcomes in schizophrenia. *Schizophr Bull*. 2008;34:483–493.
5. Velligan DI, Diamond P, Maples N, et al. Comparing the efficacy of interventions that use environmental supports to improve outcomes in patients with schizophrenia. *Schizophr Res*. 2008;102:1–3 312–319.
6. Velligan DI, Kern RS, Gold JM. Cognitive rehabilitation for schizophrenia and the putative role of motivation and expectancies. *Schizophr Bull*. 2006;32:474–485.
7. Bell M, Wayne Z, Tamasine G, Wexler B. Neurocognitive enhancement therapy with vocational services: work outcomes at a two year follow up. *Schizophr Res*. 2008;105:1–3 18–29.
8. Wexler BE, Bell MD. Cognitive remediation and vocational rehabilitation for schizophrenia. *Schizophr Bull*. 2005;31: 931–941.
9. Green MF, Nuechterlein KH, Gold JM, et al. Approaching a consensus cognitive battery for clinical trials in schizophrenia: the NIMH-MATRICS conference to select cognitive domains and test criteria. *Biol Psychiatry*. 2004;56:301–307.
10. Nuechterlein KH, Robbins TW, Einat H. Distinguishing separable domains of cognition in human and animal studies: what separations are optimal for targeting interventions? A summary of recommendations from breakout group 2 at the measurement and treatment research to improve cognition in schizophrenia new approaches conference. *Schizophr Bull*. 2005;31:870–874.
11. Buchanan RW, Davis M, Goff D, et al. A Summary of the FDA-NIMH-MATRICS workshop on clinical trial design for neurocognitive drugs for schizophrenia. *Schizophr Bull*. 2005;31(1):5–19.
12. Nuechterlein KH, Green MF, Kern RS, et al. The MATRICS consensus cognitive battery, part 1: test selection, reliability, and validity. *Am J Psychiatry*. 2008;65:203–213.
13. Kern RS, Nuechterlein KH, Green MF, et al. The MATRICS consensus cognitive battery, part 2: co-norming and standardization. *Am J Psychiatry*. 2008;165:214–220.
14. Green MF, Nuechterlein KH, Kern RS, et al. Functional co-primary measures for clinical trials in schizophrenia: results from the MATRICS psychometric and standardization study. *Am J Psychiatry*. 2008;165:2.
15. Green MF, Schooler NR, Kern RS, et al. Evaluation of Co-Primary Measures for Clinical Trials of Cognition Enhancement in Schizophrenia. *Am J Psychiatry*. In press.
16. Loeb PA. *Independent Living Scales Manual.* San Antonio, TX: Psychological Corporation; 1996.
17. Patterson TL, Goldman S, McKibbin CL, et al. UCSD performance-based skills assessment: development of a new measure of everyday functioning for severely mentally ill adults. *Schizophr Bull*. 2001;27:235–245.
18. Velligan DI, Diamond P, Glahn DC, et al. The reliability and validity of the test of adaptive behavior in schizophrenia (TABS). *Psychiatry Res*. 2007;151:1–2 55–66.
19. Bilder R, Ventura J, Reise S, et al. *Cognitive Assessment Interview (CAI). Interviewer's Manual: Definition and Rating Guidelines.* CA: Neuropsychiatric Institute, UCLA; 2008.
20. Harvey PD, Helldin L, Bowie CR, et al. Performance-based measurement of functional disability in schizophrenia: a cross-national study in the United States and Sweden. *Am J Psychiatry*. 2009;166:821–827.
21. Harvey PD, Velligan DI, Bellack AS. Performance-based measures of functional skills: usefulness in clinical treatment studies. *Schizophr Bull*. 2007;33:1138–1148.
22. Meyer K, Sprott H, Mannion AF. Cross cultural adaptation, reliability and validity of the German version of the pain catastrophizing scale. *J Psychosom Res*. 2008;64:469–478.
23. Rahman MBA, Indran SK. Disability in schizophrenia and mood disorders in a developing country. *Soc Psychiatry Psychiatr Epidemiol*. 1997;32:387–390.
24. Chavez LM, Canino G, Negron G, et al. Psychometric properties of the Spanish version of two mental health outcome measures: World Health Organization Disability Assessment Schedule II and Lehman's Quality of Life Interview. *Ment Health Serv Res*. 2005;7(3):145–159.
25. Cook L, Schmitt-Cascalliar AP, Brown C. Adapting achievement and aptitude tests. A review of methodological issues. In: Hambleton RK, Merenda PF, Spielberger CD, eds. *Adapting Educational and Psychological Tests for Cross-Cultural Assessment.* London, UK: Lawrence Erlbaum Associates; 2005:171–192.

26. Tanzer NK, Sim CQE. Adapting instruments for use in multiple languages and cultures: a review of the ITC guidelines for test adaptations. *Eur J Psychol Assess*. 1999;15:258–269.
27. Matias-Carrelo LE, Chavez LM, Negron G, et al. The Spanish translation and cultural adaptation of five mental health outcome measures. *Cult Med Psychiatry*. 2003;27:291–313.
28. Van Widenfelt BM, Treffers PDA, Beurs E, et al. Translation and cross-cultural adaptation of assessment instruments used in psychological research with children and families. *Clin Child Fam Psychol Rev*. 2005;8(2):135–147.
29. Bilder R, Ventura J, Cienfuegos A. *Clinical Global Impression of Cognition in Schizophrenia, Version 3.2: Interviewer's Manual: Definitions and Guidelines.* University of California Los Angeles; 2003.
30. Keefe RSE, Poe M, Walker TM, et al. The schizophrenia cognition rating scale: an interview-based assessment and its relationship to cognition, real-world functioning, and functional capacity. *Am J Psychiatry*. 2006;163: 426–432.
31. Chopra P, Herrman H, Kennedy G. Comparison of disability and quality of life measures in patients with long-term psychotic disorders and patients with multiple sclerosis: an application of the WHO disability assessment schedule II and WHO quality of life-BREF. *Int J Rehabil Res*. 2008;31(2): 141–149.