ORIGINAL ARTICLE

Developing an international scoring system for a consensus-based social cognition measure: MSCEIT-managing emotions

G. S. Hellemann[1,2], M. F. Green[1,2,3]*, R. S. Kern[1,2,3], G. Sitarenios[4] and K. H. Nuechterlein[1,5]

[1] *Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, California 90095, USA*
[2] *VISN 22 Mental Health Mental Illness Research, Education, and Clinical Center (MIRECC), Los Angeles CA, USA*
[3] *VA Research Enhancement Award Program (REAP) on Enhancing Community Integration for Homeless Veterans, Los Angeles CA, USA*
[4] *Multi-Health Systems Inc., Toronto, Canada*
[5] *Department of Psychology, University of California, Los Angeles, California 90095, USA*

Background. Measures of social cognition are increasingly being applied to psychopathology, including studies of schizophrenia and other psychotic disorders. Tests of social cognition present unique challenges for international adaptations. The Mayer–Salovey–Caruso Emotional Intelligence Test, Managing Emotions Branch (MSCEIT-ME) is a commonly-used social cognition test that involves the evaluation of social scenarios presented in vignettes.

Method. This paper presents evaluations of translations of this test in six different languages based on representative samples from the relevant countries. The goal was to identify items from the MSCEIT-ME that show different response patterns across countries using indices of discrepancy and content validity criteria. An international version of the MSCEIT-ME scoring was developed that excludes items that showed undesirable properties across countries.

Results. We then confirmed that this new version had better performance (i.e. less discrepancy across regions) in international samples than the version based on the original norms. Additionally, it provides scores that are comparable to ratings based on local norms.

Conclusions. This paper shows that it is possible to adapt complex social cognitive tasks so they can provide valid data across different cultural contexts.

Introduction

Cognition enhancement in schizophrenia is a major public health goal and it presents a vexing challenge for drug development and psychosocial treatment. To increase the focus on cognition-enhancing drugs for schizophrenia, the National Institute of Mental Health (NIMH) sponsored the Measurement and Treatment Research to Improve Cognition in Schizophrenia (MATRICS) Initiative (Green *et al.* 2004; Marder & Fenton, 2004; Buchanan *et al.* 2005). A key product of this Initiative was a standard cognitive battery that can be used in clinical trials of schizophrenia to measure cognitive improvement: the MATRICS Consensus Cognitive Battery (MCCB)

(Nuechterlein & Green, 2006; Kern *et al.* 2008; Nuechterlein *et al.* 2008).

An initial decision regarding the composition of the MCCB was to include social cognition as one of the domains. Managing Emotions (Branch 4) of the Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT-ME) was selected as the representative measure of social cognition (Mayer *et al.* 2002). The MSCEIT is a standardized and fairly comprehensive measure of emotion processing, including the abilities of perceiving emotions, facilitating thought, and understanding emotions, in addition to managing emotions. The Managing Emotions branch includes two subtests, Emotion Management and Emotional Relations and its selection was based on a review of published findings for candidate social cognition tests, as well as the results from a psychometric study with schizophrenia patients (Nuechterlein *et al.* 2008). The MSCEIT-ME is designed to asses 'the ability to be open to feelings, and to modulate them in oneself and others so as to promote personal understanding

* Address for corresponding author: M. F. Green, Ph.D., UCLA Semel Institute for Neuroscience and Human Behavior, 760 Westwood Plaza, 27-462, Los Angeles, California 90095, USA.

(Email: mgreen@ucla.edu)

and growth' (Mayer *et al.* 2002). Although management of emotions is similar to emotion regulation, this measure differs from other measures of emotion regulation that have been used in schizophrenia in that the MSCEIT-ME involves knowledge about how to effectively regulate one's own emotions. As such, it is not a direct physiological measure of regulation in the way that electroencephalography or functional magnetic resonance imaging paradigms are.

At the time the decision was made to include social cognition as a domain in the MCCB (Green *et al.* 2004), there were relatively few studies of social cognition in schizophrenia. Since then, social cognition has emerged as a major focus for schizophrenia research, as well as for psychopathology research more generally. Social cognition refers to those cognitive functions that are employed in socially-relevant situations (Harvey & Penn, 2010) and includes emotion perception, emotion regulation, social perception, mentalizing, and attributional style, as well as complex social abilities like empathy (Pinkham *et al.* 2016). It is critically important for social and vocational outcomes in schizophrenia and it has somewhat higher correlations than non-social cognitive deficits with functional outcomes (Fett *et al.* 2011). Social cognition has become a treatment target in schizophrenia, as evidenced by energetic development of training interventions (Wolwer *et al.* 2005; Roberts & Penn, 2009; Horan *et al.* 2011), and promising pharmacological treatments (Pedersen *et al.* 2011; Davis *et al.* 2013; Fischer-Shofty *et al.* 2013).

Clinical trials, especially large trials that are intended for approval by the U.S. Food and Drug Administration and regulatory agencies in other countries, are often conducted internationally. Similarly, advancements in psychosocial treatment are quickly internationally adapted and their treatment outcomes studied. The MCCB was initially published only in English. To be useful for large multi-site trials, the MCCB needed to be translated into other languages. Extending a cognitive battery to a new language involves two distinct steps: (1) translation and cultural adaptation of tests, and (2) norming the tests on community samples so that scores from different countries have the same performance metric and can be combined. The MCCB has now been translated into over 20 languages (Green *et al.* 2014).

For many non-social cognition tests (e.g. speed of processing, memory, and vigilance) the procedures for translation, adaptation, and norming are fairly well understood (Hambleton *et al.* 2005). The process is also quite clear for social cognition tests that involve basic processes such as identifying emotions in facial images. However, the situation for social cognition assessments that involve inference and knowledge of cultural norms is much more complex. This complexity

is reflected in tests of social cognition such as the MSCEIT-ME, where participants are asked to evaluate the usefulness of different strategies of emotion management across a wide range of situations that are presented in vignettes. For example, a vignette might describe a person experiencing negative emotions at their workplace due to being passed over for promotion in favor of a less qualified colleague, and ask about the appropriateness and effectiveness of different emotion regulation strategies in this situation. The participant is asked to rate each strategy on a 5-point rating scale on how effective it would be, with the scale anchored at 1 as 'Very Ineffective' and at 5 as 'Very Effective'. The most appropriate and effective solutions to social cognition problems like the one above will presumably depend on social context, and there may not be a 'correct' answer.

The MSCEIT-ME assessment uses a 'consensus-based' scoring method in which the value or score assigned to a response to a social problem is based on the proportion of a population that selects that response. For example, if 75% of the population selects a particular response as the preferred choice, then that response is subsequently assigned an equivalent proportion score for that item (e.g. 0.75 in a range of 0–1.00). The original norms of the MSCEIT-ME that generate the scores were established using a large (N = 2112) sample of English-speaking participants drawn from seven countries. Most participants were from the USA (1240), with others from South Africa (231), India (194), the Philippines (170), the UK (115), Scotland (122), and Canada (37).

Indeed, in early use of the MCCB outside of the USA, questions about cross-cultural validity of the MCCB focused mainly on the MSCEIT-ME as opposed to the other non-social cognitive tests. This concern is similar to what was seen when clinicians were surveyed about the cross-cultural adaptability of intermediate measures of functioning (Velligan *et al.* 2012). These intermediate measures, such as the MSCEIT, are based on participants' evaluation and response to situations that are presented verbally or as text.

Because of the importance of social cognition deficits in schizophrenia and other psychiatric disorders, and their increasing role as treatment targets, development of a valid international scoring system for the MSCEIT-ME test was viewed as a crucial step in multinational adoption of the MCCB for clinical trials.

In this paper, we describe a data-driven approach to address questions about the degree to which scoring on the MSCEIT-ME is influenced by differences in cultural norms that influence the judgment of the appropriateness of social behavior. Our goal was to develop a single scoring system that can be used

internationally for the MSCEIT-ME. This work was done using six different published translations of the MSCEIT-ME (Hindi, Japanese, Simplified Chinese, Russian, Spanish for Spain, Spanish for Central and South America) that were generated using the same established procedures for translation and cultural adaptation as the other tests in the MCCB. We obtained normative samples from six corresponding regions (India, Japan, China, Russia, Spain, and Central and South America), each of which was stratified by age, sex, and education according to each region's population. This broad sample of different languages and different cultural contexts provided a solid basis for evaluating the consequences of language translation and differing cultural contexts.

Methods

These analyses are based on an international sample of 1597 participants from six selected regions ranging from 204 to 350 participants per region. In each of the regions recruitment was stratified to ensure a representative sample of the age, gender and educational distribution. In all regions exclusion criteria were a diagnosis of a major psychiatric disorder, a history or current alcohol or drug abuse, a psychotic disorder in first degree relatives, and inability to understand the language of the test sufficiently to comprehend testing instructions. All participants were compensated small amounts (less than US$40) for their time and transportation costs. Local IRBs approved all protocols. There were some differences in recruitment strategies, and in the number and selection of the recruitment sites. All regions included urban and rural areas in their recruitment, with some regions specifically selecting recruitment sites in rural and urban communities to ensure that they were represented in the final sample. Most sites recruited using local advertisements as well as other strategies. The specific recruitment strategies for each region and sample sizes are provided in Table 1.

To create an international scoring system we needed to select items from the original MSCEIT-ME that would create comparable scores regardless of language and country of origin. Consistent with this goal, we evaluated how discrepant the consensus ratings were across the different regions and excluded items that showed large differences in consensus ratings.

To quantify the degree of discrepancy between consensus ratings in different regions we decided to use a measure of discrepancy known as L1 (Manhattan norm), which is defined as the sum of the absolute differences between the percentages of choices for each of the response options. This distance metric has two key advantages: First, it directly mirrors the scoring

Table 1. Sample and sampling characteristics.

	China	India	Japan	Latin America	Russia	Spain
Sample size	286	220	338	350	204	210
Age M (S.D.)	38.8 (11.6)	42.1 (11.3)	39.8 (11.4)	39.0 (11.6)	38.8 (11.8)	42.7 (11.3)
Highest grade M (S.D.)	10.9 (3.2)	12.2 (3.5)	14.9 (2.2)	9.6 (4.9)	12.77 (2.5)	10.6 (3.8)
Gender (male)	143 (50%)	109 (49%)	167 (49%)	175 (50%)	102 (50%)	107 (51%)
Number of sites	6	13	9	6	5	3
Rural/urban recruited separately	N (sites represent geographic regions of China)	Y	N	Y	N	Y
Reimbursement	Transportation (50 RMB)	Transportation + Time (NA)	Transportation + Time (NA)	Transportation + Time ($10)	Time ($10)	Time (40 EUR)
Recruitment	Referrals and advertisements in the community	Mostly families and friends of patients	Advertisements in the community; Recruitment from existing pool of volunteers	Advertisements at universities; Advertisements for parents at local schools; Recruitment from existing volunteer pools	Advertisements within large companies; Friends and relatives of participants in earlier studies.	Personal recruitment of likely volunteers; Advertisements

algorithm of the MSCEIT-ME. A respondent's score on an item in the MSCEIT-ME is the percentage rating of the consensus scoring of the response option. For example, if the selected response option was also chosen by 17% of the participants that created the norms for consensus scoring, it is scored as 0.17. Hence, the L1 norm of the difference between two sets of consensus ratings is proportional to the average difference in a respondent's score if their performance on that item were to be scored based on the consensus rating of one region *v.* the other consensus rating of another region. Second, the L1 norm is less influenced than other discrepancy metrics by the skew of the frequency distribution of the consensus ratings within each country. That is an important feature because consensus ratings of behaviors generally show a large degree of skew in that one choice is often identified as the best by the majority of participants, and other choices are considered inappropriate and are infrequently endorsed. Third, as long as the consensus ratings are unimodal and have the same overall pattern the L1 norm also measures the variability of participants' responses and how much any individual response on average deviates from the population mode. We computed the L1 value for each possible pairwise comparison between regions and then averaged those values. Items with the highest average discrepancies were considered for exclusion.

To capture another important aspect of the differences in response pattern between regions that is not measured by the L1 norm, we also identified and excluded items that were considered to be 'reversed.' We determined for each item the median response option based on the five-point rating scale from 'very ineffective' to 'very effective', i.e. 50% of the respondents considered the presented strategy that or more effective, and 50% considered it effective to that degree or less. Items were considered reversed in which, based on the median response option, one region rated the action as effective and another region rated the action as ineffective. Reversal/non-reversal can also be viewed as general measure of agreement between the shapes of the response patterns, as they are typically unimodal and their general shape can be described by the location of the mode.

Although some situations presented in the vignettes might not be appropriate across all regions, we expected that most of the items would be minimally affected by cultural and language differences based on how the original vignettes were selected. The situations presented in the MSCEIT-ME were designed to cover a wide range of social situations while still being sufficiently generic so that most people could relate to them. In addition, the original norms were established on an English speaking sample that was culturally heterogeneous – it included English speakers mainly from North American, but also from other countries (e.g. South Africa, India), who might not consider English their primary language and who were likely to have a range of ethnic identities. Thus, heavily culturally-loaded items would tend to be dropped before inclusion due to undesirable psychometric properties in the original MSCEIT-ME development process.

For scoring the items included in the international scoring system of the MSCEIT-ME, we decided to use the original consensus norms instead of establishing a new set of norms based on the seven regional norming samples for two reasons. First, the regional samples are large enough ($n > 200$) to determine if the consensus ratings of the different language versions differ from the original ratings, but they are small compared to the sample for the original consensus norms. Second, using the original norms and modifying the items included provides a more straightforward path to future expansion of the international MSCEIT-ME scoring method to other countries. If the international consensus ratings were to be based on an aggregate of the regional norming samples available now, adding new regions would necessitate calculation of new consensus ratings that would change the overall scoring system. Hence, it seemed desirable to anchor the international MSCEIT-ME on norms with the largest norming sample available and choose items that were as culturally neutral as possible. Finally, to confirm that the international scoring system provides a valid representation of participants' responses, we scored the MSCEIT-ME using region-specific norms to determine whether the international version and the region-specific versions of the MSCEIT-ME resulted in similar scores for participants.

Results

When examining pairs of regions for discrepancy, no one region or pair of regions stood out as particularly distinctive. The average discrepancies ranged from 51.6 (India *v.* Japan) to 27.1 (India *v.* Latin America), with a mean of 43.22 and a standard deviation of 9.88 (Table 2). These discrepancies represent the sum absolute differences for all response options for each items, i.e. the discrepancy score of 43.22 represents an average difference of 8.64% per response option (43.22/5) between the consensus ratings for that item for that pair of regions. No pair of regions showed a discrepancy that was more than one standard deviation larger than the mean. The absence of regions that were extremely dissimilar from one another based on the original norms supports our decision to use the original norms as the basis for the international

Table 2. *Average distances across all items.*

	Russia	Latin America	China	Spain	Japan	India
Original norms	41.5	39.5	44.4	36.2	39.2	50.1
Russia		41.8	44.0	50.9	42.6	49.4
Latin America			34.2	30.4	45.2	27.1
China				45.6	48.8	36.6
Spain					45.4	42.3
Japan						51.6

scoring system. The consensus ratings of all regions that these analyses are based on are available in online Supplement 1.

A small number of items (3 of 29) showed a reversal. Since a reversal indicates a major conceptual difference in the evaluation of effectiveness between regions, these items were dropped from scoring for the international version without further discussion. Three additional items were dropped in a second phase, meaning a total of 6 out of 29 items were excluded in the international version of MSCEIT-ME scoring. The final decision for dropping each item was made using a consensus-based decision process based on the statistical properties of the item (the discrepancy score), and case-by-case considerations. One of the goals of this process was to generate an international scoring system that maintained the overall structure of the MSCEIT as much as possible, including all of the original vignettes. Thus, when two items from the same vignette showed some discrepancies between regions, the poorer item was considered for omission. There was only one vignette in which two items were dropped. In that case, one item was reversed and the other had an average discrepancy score which was in the top 10% of discrepancies. The final set of items omitted from scoring for the international MSCEIT-ME consists of D13, D22, D43, D44, H12, and H32.

When evaluating the performance of the international MSCEIT-ME scoring method after the removal of these six items compared with region specific norms, we found very strong associations between the performances on these two scoring methods (most correlations were $r > 0.8$). The only region that did not reach this threshold was Russia ($r = 0.725$), but we consider this degree of convergence to be still acceptable (Table 3). This pattern of strong associations indicates that the items selected for the international scoring system are sufficiently homogeneous across cultures to be used for valid comparison of scores.

We did not identify common elements in the content of the six excluded items. The content of the excluded items was not specific to scenarios based on gender

Table 3. *Associations between international MSCEIT-ME scoring method and region specific norms*

	International MSCEIT: Original *v.* Regional consensus norms
Russia	0.725
Latin America	0.867
China	0.929
Spain	0.877
Japan	0.860
India	0.808

(two feature female protagonists, four male protagonists), valence of the emotion (two feature managing positive emotions, four feature managing negative emotions), or setting (two are set in a work environment, one in a family setting, and four deal with personal emotions). Details are provided in online Supplement 1.

We compared the performance of the international norms excluding the six items with the original norms and found that for the vast majority of cases the international norms reduced overall discrepancies (compare Tables 2 and 4). The only exception was for some dyads including Russia.

As a further analysis we evaluated whether the relationship between other non-social cognitive measures that are part of the MCCB and social cognition measured with the MSCEIT-ME changed based on which scoring algorithm is used. While in some cases the changes are statistically significant, the effect sizes of these changes are very small. In no case does the effect size reach the threshold Cohen defines as small ($q = 0.10$). Details are available in online Supplement 2.

Discussion

Overall, these results show that participants evaluated strategies for the management of emotions in social situations very similarly across diverse geographical regions. Only a small proportion of items showed

Table 4. *Average distances on the items included in the international norms.*

Overall	Russia	Latin America	China	Spain	Japan	India
Original norms	40.2	33.7	33.6	31.7	32.2	43.3
Russia		43.15	39.7	54.4	40.8	50.5
Latin America			31.9	29.3	43.6	25.7
China				43.3	47.0	36.9
Spain					44.1	40.8
Japan						49.3

evidence of either major conceptual differences or differences in the degree of consensus for a response option. The international scoring version, which was created by excluding items that contributed the most variability between countries, produced scores that correlated highly with scores based on region-specific data, and provides a reliable measure of social cognition that is comparable across different populations.

The items that showed strong response variability across cultures did not show a consistent or predictable content. Several items that appeared to be culturally-loaded (e.g. normative behavior at the workplace, gender roles) did not result in strongly discrepant response patterns. This suggests greater international similarity than differences in people's ability to assess strategies for successful emotion management.

As mentioned earlier, one potential reason for the relatively culture-neutral nature of the MSCEIT-ME might be that the original English language norms of the MSCEIT-ME are based on English speaking participants from seven countries (US, Canada, UK, Scotland, India, Philippines, and South Africa) representing a broad range of different cultural backgrounds, including participants who would not consider English their first language. Thus, the items on the MSCEIT-ME were already to some extent pre-selected for absence of strong cultural effects. Additionally one of the limitations of this study is that we do not have detailed information about each participant's primary language, or more fine-grained cultural and ethnical self-identification. Similarly, the vignettes of the MSCEIT-ME explicitly ask the participant to evaluate the behavior of somebody else. Evaluating behavior from somebody else's perspective, even if those individuals are engaged in scenarios that are not commonly observed in one's own socio-cultural environment, requires the same social cognitive processes across cultures. If the vignettes are implicitly read as 'Evaluate the behavior of this person from their perspective, even if the situation is not something you would be likely experience yourself', then respondents can formulate their response using perspective-taking (seeing things from someone else's point of view). This ability

is reinforced by the global reach of media and the internet, and at least partial familiarity with other cultures. Thus, the relative robustness of the original MSCEIT-ME, with only a limited number of items showing strong cultural variability might be partly attributable to respondents' perspective-taking.

While we consider the MSCEIT-ME with the international scoring system to be a relatively culture-independent measurement of emotion management, certain aspects of social cognition are culture dependent. Thus, while there are some social situations and emotion management strategies that are relatively invariant across cultures, the details of appropriate social behavior in some social situations depend on the specific cultural context, and the ability to evaluate behavior for its adherence (or violation) of social norms. Thus, one could develop a social cognition measure to detect the degree to which an individual is skilled in social situations that require nuanced navigation of social norms within a specific culture.

The purpose of the international MSCEIT-ME scoring system, and the international versions of the MCCB of which it is a part, is to provide researchers with reliable and validated measurement instruments that can be used across the widest possible range of geographical locations, and yield comparable results. The international scoring system for the MSCEIT-ME is now available and included with the MCCB test kits. The results of this adaptation process suggest that versions of social cognition tests can be developed that are valid and reasonably culture independent.

Supplementary material

For supplementary material accompanying this paper visit https://doi.org/10.1017/S0033291717001052

Acknowledgements

conducting the research on the performance of the MCCB in each included region (in alphabetical order):

China: Robert K. Heaton[1], Hua Jin [1], Chuan Shi[2,3,4] and Xin Yu[2,3,4]

[1]Department of Psychiatry, University of California at San Diego, CA, USA; [2]Peking University Institute of Mental Health (Sixth Hospital), Beijing, China; [3]National Clinical Research Center for Mental Disorders (Peking University Sixth Hospital) Beijing, China; [4]Key Laboratory of Health (Peking University) Beijing China.

India: Mark Marsico[1].

[1]Center for Observational & Real-World Research, Merck & Co., Inc., Kenilworth, USA.

Japan: Kunimasa Arima[1], Yasuhiro Kaneda[2], Crystal Lin[3], Ichiro Sora[4] and Kiyohisa Takahasi[5]

[1]National Komorokogen Hospital; [2]Iwaki Clinic; [3]Sakata Clinic; [4]Department of Psychiatry Kobe University Graduate School of Medicine; [5]Japan Foundation for Neuroscience and Mental Health.

Latin America: Delfina de Achával[1,2], Salvador Rogelio Apiquian[3], Rodrigo Nel Cordoba Rojas[4,5], Elsa Y. Costanzo[1,2] and M. Guinjoan[1,2].

[1]Adult Psychiatry Service, FLENI Foundation, Argentina; [2]Department of Psychiatry & Mental Health, University of Buenos Aires School of Medicine, Argentina; [3]Arete Proyectos, Mexico City, Mexico; [4]Grupo Cisne, Bogotá, Colombia; [5]University of Rosario, Bogotá, Colombia.

Russia: Andrey Alexeev[1], Taissia Lepilkina[1], Margarita Morozova[1], Yana Morozova[1] and Victoria Irzhevsky[1]

[1]Moscow Research Institute of Psychiatry, Moscow, Russia.

Spain: R. Rodriguez-Jimenez[1,2], A. Bagney[1,2], J.L. Santos[3], A. Lobo[2,4] and T. Palomo[2,5].

[1]Department of Psychiatry, Instituto de Investigación Hospital 12 de Octubre (imas12), Madrid, Spain; [2]Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Spain; [3]Department of Psychiatry, Hospital Virgen de La Luz, Cuenca, Spain; [4]Department of Psychiatry, Universidad de Zaragoza, Instituto de Investigación Sanitaria Aragón (IIS Aragón), Spain; [5]Department of Psychiatry, Universidad Complutense de Madrid, Spain.

Disclosures

References

Buchanan RW, Davis M, Goff D, Green MF, Keefe RSE, Leon AC, Nuechterlein KH, Laughren T, Levin R, Stover E, Fenton W, Marder SR (2005). A summary of the FDA-NIMH-MATRICS workshop on clinical trial design for neurocognitive drugs for schizophrenia. *Schizophrenia Bulletin* **31**, 5–19.

Davis MC, Lee J, Horan WP, Clarke AD, McGee MR, Green MF, Marder SR (2013). Effects of single dose intranasal oxytocin on social cognition in schizophrenia. *Schizophrenia Research* **147**, 393–397.

Fett AK, Viechtbauer W, Dominguez MD, Penn DL, van Os J, Krabbendam L (2011). The relationship between neurocognition and social cognition with functional outcomes in schizophrenia: a meta-analysis. *Neuroscience and Biobehavioral Reviews* **35**, 573–588.

Fischer-Shofty M, Brune M, Ebert A, Shefet D, Levkovitz Y, Shamay-Tsoory SG (2013). Improving social perception in schizophrenia: the role of oxytocin. *Schizophrenia Research* **146**, 357–362.

Green MF, Harris JG, Nuechterlein KH (2014). The MATRICS consensus cognitive battery: what we know 6 years later. *American Journal of Psychiatry* **171**, 1151–1154.

Green MF, Nuechterlein KH, Gold JM, Barch DM, Cohen J, Essock S, Fenton WS, Frese F, Goldberg TE, Heaton RK, Keefe RSE, Kern RS, Kraemer H, Stover E, Weinberger DR, Zalcman S, Marder SR (2004). Approaching a consensus cognitive battery for clinical trials in schizophrenia: The NIMH-MATRICS conference to select cognitive domains and test criteria. *Biological Psychiatry* **56**, 301–307.

Hambleton RK, Merenda PF, Spielberger CD (2005). *Adapting Educational and Psychological Tests for Cross-cultural Assessment*. Lawrence Erlbaum: Mahwah, NJ.

Harvey PD, Penn DL (2010). Social cognition: The key factor predicting social outcome in people with schizophrenia? *Psychiatry* **7**, 41–44.

Horan WP, Kern RS, Tripp C, Hellemann G, Wynn JK, Bell M, Marder SR, Green MF (2011). Efficacy and specificity of social cognitive skills training for outpatients with psychotic disorders. *Journal of Psychiatric Research* **45**, 1113–1122.

Kern RS, Nuechterlein KH, Green MF, Baade LE, Fenton WS, Gold JM, Keefe RSE, Mesholam-Gately R, Mintz J, Seidman LJ, Stover E, Marder SR (2008). The MATRICS Consensus Cognitive Battery. Part 2. Co-norming and standardization. *American Journal of Psychiatry* **165**, 214–220.

Marder SR, Fenton WS (2004). Measurement and treatment research to improve cognition in schizophrenia: NIMH MATRICS Initiative to support the development of agents for improving cognition in schizophrenia. *Schizophrenia Research* **72**, 5–10.

Mayer JD, Salovey P, Caruso DR (2002). *Mayer-Salovey-Caruso Emotional Intelligence Test* (*MSCEIT*) *User's Manual*. MHS Publishers: Toronto.

Nuechterlein KH, Green MF (2006). *MATRICS Consensus Cognitive Battery*. MATRICS Assessment, Inc.: Los Angeles.

Nuechterlein KH, Green MF, Kern RS, Baade LE, Barch D, Cohen J, Essock S, Fenton WS, Frese FJ, Gold JM, Goldberg T, Heaton R, Keefe RSE, Kraemer H, Mesholam-Gately R, Seidman LJ, Stover E, Weinberger D, Young AS, Zalcman S, Marder SR (2008). The MATRICS Consensus Cognitive Battery. Part 1. Test selection, reliability, and validity. *American Journal of Psychiatry* **165**, 203–213.

Pedersen CA, Gibson CM, Rau SW, Salimi K, Smedley KL, Casey RL, Leserman J, Jarskog LF, Penn DL (2011). Intranasal oxytocin reduces psychotic symptoms and improves Theory of Mind and social perception in schizophrenia. *Schizophrenia Research* **132**, 50–53.

Pinkham AE, Penn DL, Green MF, Harvey PD (2016). Social cognition psychometric evaluation: results of the initial psychometric study. *Schizophrenia Bulletin* **42**, 494–504.

Roberts DL, Penn DL (2009). Social cognition and interaction training (SCIT) for outpatients with schizophrenia: a preliminary study. *Psychiatry Research* **166**, 141–147.

Velligan DI, Rubin M, Fredrick MM, Mintz J, Nuechterlein KH, Schooler NR, Jaeger J, Peters NM, Buller R, Marder SR, Dube S (2012). The cultural adaptability of intermediate measures of functional outcome in schizophrenia. *Schizophrenia Bulletin* **38**, 630–641.

Wolwer W, Frommann N, Haufmann S, Piaszek A, Streit M, Gaebel W (2005). Remediation of impairments in facial affect recognition in schizophrenia: efficacy and specificity of a new training program. *Schizophrenia Research* **80**, 295–303.