

ABSTRACT

ABSTRACT: *Objective:* Recognizing the discrete dimensions that underlie negative symptoms in schizophrenia and how these dimensions are understood across localities might result in better understanding and treatment of these symptoms. To this end, the objectives of this study were to 1) identify the Positive and Negative Syndrome Scale negative symptom dimensions of expressive deficits and experiential deficits and 2) analyze performance on these dimensions over 15 geographical regions to determine whether the items defining them manifest similar reliability across these regions. *Design:* Data were obtained for the baseline Positive and Negative Syndrome Scale visits of 6,889 subjects across 15 geographical regions. Using confirmatory factor analysis, we examined whether a two-factor negative symptom structure that is found in schizophrenia (experiential deficits and expressive deficits) would be replicated in our sample, and using differential item functioning, we tested the degree to which specific items from each negative symptom subfactor performed across geographical regions in comparison with the United States. *Results:* The two-factor negative symptom solution was replicated in this sample. Most geographical regions showed moderate-to-large differential item functioning for Positive and Negative Syndrome Scale expressive deficit items, especially N3 Poor Rapport, as compared with Positive and Negative Syndrome Scale experiential deficit items, showing that these items might be interpreted or scored differently in different regions. Across countries, except for India, the differential item functioning values did not favor raters in the United States. *Conclusion:* These results suggest that the Positive and Negative Syndrome Scale negative symptom factor can be better represented by a two-factor model than by a single-factor model. Additionally, the results show significant differences in responses to items representing the Positive and Negative Syndrome Scale expressive factors, but not the experiential factors, across regions. This could be due to a lack of equivalence between the original and translated versions, cultural differences with the interpretation of items, dissimilarities in rater training, or diversity in the understanding of scoring anchors. Knowing which items are challenging for raters across regions can help to guide Positive and Negative Syndrome Scale training and improve the results of international clinical trials aimed at negative symptoms.

KEYWORDS: Schizophrenia, Positive and Negative Syndrome Scale, PANSS, negative symptoms, expressive deficits, experiential deficits

Negative Symptom Dimensions of the Positive and Negative Syndrome Scale Across Geographical Regions: Implications for Social, Linguistic, and Cultural Consistency

by ANZALEE KHAN, PhD; LORA LIHARSKA, MS; PHILIP D. HARVEY, PhD; ALEXANDRA ATKINS, PhD; DANIEL ULSHEN, BA; and RICHARD S.E. KEEFE, PhD

Dr. Khan is Senior Biostatistician at NeuroCog Trials and holds an appointment in the Psychopharmacology Research Program at the Nathan S. Kline Institute for Psychiatric Research in Orangeburg, New York. Ms. Liharska is a researcher at Columbia University Medical Center in New York, New York. Dr. Harvey is Professor of Psychiatry and Behavioral Sciences at the University of Miami School of Medicine in Miami, Florida. Dr. Atkins and Mr. Ulshe are employees of NeuroCog Trials. Dr. Keefe is an employee of NeuroCog Trials and Professor of Psychiatry and Behavioral Sciences at Duke University Institute for Brain Sciences in Chapel Hill, North Carolina.

Innov Clin Neurosci. 2017;14(11–12):30–40

Emerging consensus in recent literature posits a two-factor structure of negative symptoms in schizophrenia: experiential deficits and expressive deficits.¹ This presents a departure from the previous one-factor model by differentiating negative symptoms into two potentially correlated but distinct dimensions.² Experiential deficits comprise avolition (decreased motivation and lack of interest in daily activities), asociality (social withdrawal and reduced value and interest in social contact), and anhedonia (decreased anticipation and experience of pleasure); expressive

deficits comprise blunted affect (decreased emotional expressivity and diminished facial expression) and alogia (poverty of speech).^{3–6} The implications of this two-dimensional model are far-reaching: recent research suggests that each dimension might have its own distinct underlying causes and clinical and functional correlates.^{6–8}

The expressive-experiential distinction has been shown to have vast importance in relation to functional outcomes in schizophrenia. Specifically, experiential deficits have been shown to be more robust predictors of

FUNDING: This study is a secondary analysis of existing data, so it received no funding from any agency in the public, commercial, or non-profit sectors.

DISCLOSURES: Dr. Khan currently or in the past three years has received funding support, received honoraria, served as a consultant, and served as a speaker for the National Institute of Mental Health, the Research Foundation for Mental Hygiene, Inc., the NY State Office of Mental Health, the Qatar Foundation at Weill Cornell Medicine, and eResearch Technologies. She also holds positions at NeuroCog Trials and at Manhattan Psychiatric Center Nathan S. Kline Institute for Psychiatric Research. Ms. Liharska has no conflicts of interest relevant to the content of this article. Dr. Harvey has served as a consultant to AbbVie, Allergan, Akili, Boehringer Ingelheim, Forum Pharmaceuticals, Genentech, Lundbeck Pharmaceuticals, Otsuka Digital Health, Roche Pharma, Sanofi, Sunovion, and Takeda Pharmaceuticals for the past three years. Dr. Atkins currently or in the past three years has received funding from the National Institute of Mental Health and the National Institute on Aging, and is a full-time employee of NeuroCog Trials. Mr. Ulshe is a full-time employee of NeuroCog Trials. Dr. Keefe currently or in the past three years has received investigator-initiated research funding support from the Department of Veteran's Affairs, the Feinstein Institute for Medical Research, the National Institute of Mental Health, the Research Foundation for Mental Hygiene, Inc., and the Singapore National Medical Research Council. He currently or in the past three years has received honoraria, served as a consultant, speaker, or advisory board member for AbbVie, Acadia, Akebia, Akili, Alkermes, Astellas, Asubio, Avanir, AviNeuro/ChemRar, Axovant, Biogen, BiolineRx, Biomarin, Boehringer-Ingelheim, Cerecor, CoMentis, FORUM, Global Medical Education (GME), GW Pharmaceuticals, Intracellular Therapeutics, Janssen, Lundbeck, MedScape, Merck, Minerva Neurosciences Inc., Mitsubishi, Moscow Research Institute of Psychiatry, Neuralstem, NeuroniX, Novartis, NY State Office of Mental Health, Otsuka, Pfizer, Reviva, Roche, Sanofi, Shire, Sunovion, Takeda, Targacept, University of Moscow, University of Texas Southwest Medical Center, and WebMD. He is also a shareholder in NeuroCog Trials, Inc., and in Sengenix.

CORRESPONDENCE: Anzalee Khan, PhD; Email: anzalee.khan@nki.rfmh.org

functional outcome than expressive deficits.^{6,9–11} This finding has also been observed in clinical high risk populations.¹² Additionally, the magnitude of impairment in experiential deficits, but not in expressive deficits, has been shown to be associated with employment outcomes of fewer hours worked and lesser wages earned.¹³

Recent work has suggested that experiential deficits and expressive deficits might identify different clusters of patients, meaning that there might be distinct subgroups of patients with primary impairments in emotional experience and expression.¹⁴ These findings suggest that an individual assessment of both negative symptom dimensions might increase the sensitivity of treatment outcome.¹⁵ Given the marked heterogeneity in schizophrenia symptoms, the ability to more precisely differentiate the pathological mechanisms of the two negative-symptom dimensions across symptom severity levels and geographical regions is critically important for the development of effective treatment interventions.

Many psychometric measurements, especially in psychiatry, have items that might perform differently with diverse groups.¹⁶ The Positive and Negative Syndrome Scale (PANSS), which estimates the latent trait of the intensity of symptoms in schizophrenia, was developed and validated 30 years ago on a hospitalized sample of subjects with chronic schizophrenia, without accounting for group differences across highly important domains.¹⁷ Therefore, evidence for the validity of inferences made from symptom scores includes an evaluation of group differences (e.g., geographical regions) in symptom presentation. The presence of bias in ratings and the impact of symptoms on overall functioning is of interest. Scores that perform in markedly different ways across demographic, regional, cultural, or clinical severity characteristics might not be valid representations of the target construct. Our previous studies have shown that the PANSS might not be equivalently rated across country-specific and cultural disparities, not only with regard to symptom expression but also with respect to rater judgment of symptom severity scores.¹⁸ When looking at comparisons of reliability estimates across geo-cultural groups, Khan et al¹⁸ found increased variability among those scores of the PANSS Negative Symptoms

subscale that had the lowest reliability. The investigators further observed differential item functioning (DIF) for the PANSS Negative Symptoms subscale as compared with the PANSS Positive Symptoms and General Psychopathology subscales. The DIF method estimates the difference in the probability of raters from different countries scoring symptoms similarly when assessing subjects of the same severity level. Theoretically, if an item is free of construct-irrelevant variance, then subjects with the same severity level—even if scored by different raters, and regardless of geographical location—should have the same probability of a similar symptom presentation on that item. When a statistically significant difference in probability is observed, the following might have contributed to the difference: ambiguity of the description of the item/symptom being measured; issues with rater training, rater difficulty in comprehending the construct being measured, subject's severity level; language or translation validity; and influence due to geographical characteristics.

Despite the extensive psychometric work done on the PANSS for the past 30 years, only within the past 10 years have more modern psychometric techniques (such as item response analysis and DIF) been applied to the PANSS. In general, the use of these techniques is done during scale development as opposed to *post hoc*. However, as negative symptoms remain difficult for raters to assess reliably in international clinical trials¹⁹ and across cultures,^{18,20} using these techniques *post hoc* can help to further elucidate the psychometric properties of negative symptom assessments by identifying their validity and reliability across international trials.

Existing research has already shown that a two-factor model of the PANSS fits negative symptoms data significantly better than a one-factor model.¹⁵ The purpose of this article is to 1) assess the replicability of the two-factor solution of negative symptoms (expressive deficits and experiential deficits) commonly found in people with schizophrenia or schizoaffective disorder, and people at clinical high risk of psychosis^{1,5,12,15,21} and 2) compare items with differential functioning across geographical regions. The results of this study can be used to customize and guide protocol development and rater training.

METHODS

Sample. Data were provided for 7,348 subjects who were enrolled between 1992 and 2005 in one of 16 randomized, double-blind clinical trials comparing risperidone, risperidone depot, or paliperidone to other antipsychotic drugs (e.g., haloperidol, olanzapine) or a placebo (information on these trials is presented in Appendix 1). As these were comparative open-label and double-blind trials examining the safety and efficacy of antipsychotics, subjects included in these studies were selected based on overall symptom presentation, rather than primarily on the severity of negative symptoms. The data used in the current analysis comprise the baseline (pre-treatment) data collected in these trials, but can be seen as representative of individuals who enter multicenter, international clinical trials. All studies were carried out in accordance with the latest version of the Declaration of Helsinki. Study procedures were reviewed by the respective ethics committees, and informed consent was obtained from all subjects after the procedures were fully explained.

Data analysis included baseline PANSS item scores from 6,889 out of the 7,348 subjects for whom data was provided. A total of 459 subjects (6.25%) were removed from the analysis. Of these 459 subjects, 92 (20.04%) were removed due to having diagnoses other than schizophrenia or schizoaffective disorder. An additional three subjects (0.65%) with no reported diagnosis and two subjects (0.44%) with missing PANSS item scores were removed. Lastly, 362 subjects were removed from geographical regions that did not have an adequate sample size (at least 100 subjects per group as required for the DIF analysis).

Data source. Data for the analysis were provided by Ortho-McNeil Janssen Pharmaceuticals, Inc. (Raritan, New Jersey, USA). The data for each subject included a study identifier, de-identified subject number, sex, age at the time of study entry, age at the time of onset of illness, medication to which subject was randomized, country of residence during the time of study participation, and scores for each of the 30 PANSS items at the baseline visit. To maintain confidentiality, no treatment code information was included in the data, nor did any exchange of information occur that could have identified either the subjects or the investigative sites participating in the studies.

Measures. The PANSS¹⁷ is a 30-item rating instrument comprising three subscales: the seven-item Positive Symptoms subscale (P-P7), the seven-item Negative Symptoms subscale (N1-N7), and the 16-item General Psychopathology subscale (G1-G16). All 30 items are rated on a seven-point scale (1=absent to 7=extreme).

Currently, there are over 40 official language versions of the PANSS. Translations have been carried out per international guidelines, through collaborations between specific sponsors and translation agencies in the geo-cultural groups concerned. Translation standards for the PANSS follow internationally recognized guidelines with the objective to achieve semantic equivalence as outlined by the Multi-Health Systems translation policy.

All raters participating in the 16 clinical trials received rater training and certification on the PANSS prior to conducting PANSS assessments. Processes for rater training differed across studies, but all raters received didactic training overseen by a PANSS subject matter expert. Didactic training on the PANSS consisted of a detailed overview of each PANSS item and its anchor points. Following the overview, all raters were required to view and score a PANSS “Gold Score” video, which is a recorded interview of a rater conducting a structured clinical interview with either a patient with schizophrenia or an actor trained to portray a patient with schizophrenia. The rater’s scores on the interview were then compared to the consensus scores of two or more expert raters. In order to receive certification, a rater was expected to have an intra-class correlation (ICC) of at least 0.80 with the Gold Score ratings. It is expected that for some studies there might have been exceptions to the ICC \geq 0.80 requirement, based on rater qualifications and experience. Specific inter-rater reliability values within and across studies were not available.

The categorization of data was based on country, culture, and language. Because a minimum of 100 subjects per group is recommended for performing DIF analysis,²² attention was placed on the number of subjects per country. Additionally, an attempt was made to match groups to raters who were more likely to share their language and culture. To the extent possible, based on the available sample size, an attempt was made to maintain individual countries as individual categories.

The resulting categories and rationales for combining multiple countries into single categories are presented in Table 1. Despite its heterogeneity of language and culture, the United States (US) is identified as a separate category for several reasons. First, Gören’s study²³ examining the most culturally diverse countries in the world places the United States near the middle of all countries. Although New York and San Francisco are within the top 10 most culturally diverse cities, the only Western country ranked in the top 20 most diverse countries is Canada.²³ Second, the original scale development of the PANSS occurred in the United States, and its psychometric properties were validated based on the country’s population by diverse United States raters.¹⁷ Additionally, our team used the United States as a reference group in a previous DIF analysis of the PANSS.¹⁸

Statistical analysis. We first conducted an exploratory factor analysis (EFA) to determine if the dataset adhered to the seven PANSS negative symptom factor (NSF) items (i.e., N1 Blunted Affect, N2 Emotional Withdrawal, N3 Poor Rapport, N4 Passive Social Withdrawal, N6 Lack of Spontaneity and Flow of Conversation, G7 Motor Retardation, and G16 Active Social Avoidance). Next, we performed a confirmatory factor analysis (CFA) on the seven NSF items for the entire dataset. For the CFA, the Kaiser-Meyer-Olkin (KMO) measure evaluates whether the responses given by the sample are adequate; Kaiser²⁴ recommends a 0.50 value for KMO as the minimum (barely accepted), with values between 0.70 and 0.80 determined to be acceptable and values above 0.90 determined to be excellent.²⁴ We also assessed Bartlett’s test for significance (<0.05), which indicated a rejection of the null hypothesis.

The following indices of goodness-of-fit were computed and used for model evaluation: the chi-square difference test, comparative fit index (CFI; values >0.90 represent acceptable fit), Tucker-Lewis index (TLI; values >0.90 represent acceptable fit), root mean square error of approximation (RMSEA; values <0.05 represent acceptable fit), and goodness-of-fit index (GFI; values >0.90 represent satisfactory fit).²⁵⁻²⁷ CFA and chi-square difference tests were conducted using SPSS 23.0²⁸ and R.²⁹

We investigated the validity of the PANSS expressive-experiential distinction across 15 countries or geographical regions—South

America-Mexico; Austria-Germany; Belgium-Netherlands; Brazil; Canada; the Nordic region (Denmark, Finland, Norway, and Sweden); France; Great Britain; India; Italy; Poland; Eastern Europe (Romania, Slovakia, Ukraine, Croatia, Estonia, and Czech Republic); Russia; South Africa; and Spain—as compared with the United States. The Mantel-Haenszel statistic was used in the analysis of DIF, as it creates meaningful comparisons of item performance for different geographical regions by comparing raters assessing subjects in similar countries, rather than by comparing overall group performance on an item. For DIF of the expressive and experiential deficit items, the expectation is that two individual item responses have a probability of $p \leq 0.05$ in accordance with the Rasch model; “ α ” is the type I error for a single test (incorrectly rejecting a true null hypothesis). Thus, when the data fit the model, the probability of a correct finding is $(1-\alpha)$ for one item and $(1-\alpha)^n$ for “ n ” items. Consequently, the type I error for n independent items is $1-(1-\alpha)^n$. Therefore, the level for each single test is α/n . For example, in order to reject the hypothesis that “the entire set of items fits the Rasch model” in a finding of $p \leq 0.05$ for four items on the expressive factor and three items on the experiential factor, at least one item would need to be reported with $p \leq 0.013$ and $p \leq 0.017$, respectively.

Subjects were matched by severity level on the PANSS and grouped by geographic region; since DIF only allows for two groups per analysis, each region was compared with the United States. DIF testing was based on the chi-square statistic and is highly sensitive to sample size.³⁰ If the sample size is large, statistical significance can emerge even when DIF is quite small. DIF effect sizes can be investigated to alleviate this concern, because even though statistical significance is necessary for an item to demonstrate DIF, it is not sufficient. Zumbo et al³¹ note that an item only demonstrates DIF if the significant difference in chi-square has at least a moderate effect size (0.30–0.79). Therefore, three criteria were used to flag items as differentially functioning: 1) statistically significant chi-square test statistic ($p \leq 0.05$), 2) effect size (ES), and 3) Educational Testing Services (ETS) DIF classification criteria. Since the statistically significant test statistic does not indicate that the magnitude of the DIF is significant,³² a review of both the effect size

TABLE 1. Sample characteristics of geographical groupings

CATEGORY NAME	COUNTRY	NUMBER	JUSTIFICATION FOR GROUPINGS
South America-Mexico (Spanish-speaking)	Argentina, Chile, Mexico	264	Grouping was based primarily on language of the PANSS administration (i.e., Spanish)
Austria-Germany	Austria, Germany	360	Grouping was based primarily on language of the PANSS administration (i.e., German) and geographic location
Belgium-Netherlands	Belgium, Netherlands	210	Grouping was based primarily on language of administration and geo-cultural similarities, according to the realities and challenges of the society
Brazil	Brazil	135	Brazil was identified as separate from the Americas based on language (Brazilian Portuguese versus Spanish) and geo-cultural differences
Canada	Canada	473	Canada was not selected to be part of another category based on the language use of French-Canadian. It should be noted that some PANSS were also administered in English
Nordic	Denmark, Finland, Norway, Sweden	541	Denmark, Finland, Norway and Sweden were combined, as they are Nordic countries with geographical and cultural regions in Northern Europe and the North Atlantic. These four countries have much in common with respect to their lifestyle, history, religion, Scandinavian languages, and social structure
France	France	272	France was not selected to be part of another category based on language (i.e., French)
Great Britain	Great Britain	158	Great Britain was not selected to be part of another category based on language and within-country cultural heterogeneity
India	India	314	India was not selected to be part of another category based on language, geographical location, and within-country cultural heterogeneity
Italy	Italy	115	Italy was not selected to be part of another category based on language (i.e., Italian) and within-country cultural heterogeneity
Poland	Poland	267	Poland was not selected to be part of another category based on language, (i.e., Polish)
Eastern Europe	Romania, Slovakia, Ukraine, Croatia, Estonia, Czech Republic	441	The six countries identified as Eastern Europe were categorized based on location and Slavic language origin
Russia	Russia	161	Russia was not selected to be part of another category based on language (i.e., Russian) and culture
South Africa	South Africa	311	South Africa was not selected to be part of another category based on language and geographic location
The United States	The United States	2737	The United States was not selected to be part of another category based on language, geographic location, and within-country cultural heterogeneity
Spain	Spain	130	Spain was not selected to be part of another category based on language (i.e., Spain-Spanish) and within-country cultural heterogeneity

TABLE 2. Sample demographic and clinical characteristics of study population

TOTAL NUMBER OF SUBJECTS	6,889
Male (%)	4493 (65.22%)
Female (%)	2396 (34.78%)
AGE	
Mean	40.12 years
SD	12.35 years
AGE AT ONSET	
Mean	25.05 years
SD	8.56 years
PANSS TOTAL SCORE	
Mean	83.11
SD	19.09
PANSS POSITIVE SYMPTOM SUBSCALE SCORE	
Mean	19.49
SD	6.70
PANSS NEGATIVE SYMPTOM SUBSCALE SCORE	7–44
Mean	22.70
SD	6.93
Range (min–max)	7–48
PANSS NEGATIVE SYMPTOM FACTOR SCORE	
Mean	21.45
SD	6.98
Range (min–max)	7–48

SD: standard deviation; PANSS: Positive and Negative Syndrome Scale

and the ETS criteria make for a more robust finding. Additionally, the ETS categories were created only for identifying items that display statistically significant DIF, which cannot be represented by effect size alone. A description of the ETS DIF used in this analysis can be found in Appendix 2.

In DIF analyses, the focal group refers to the group of interest, whereas the reference group refers to the group with which the focal group is being compared.³³ In the current study, all distinct geographical regions except for the United States were chosen as focal groups and the United States was chosen as the reference group. As the PANSS was developed in the United States and initially validated on a United States population sample, the authors chose to compare each country grouping with that of the United States. The Mantel-Haenszel procedure is performed in jMetrik™ and produces effect size computation and ETS DIF item classifications

as follows: AA (negligible DIF), BB (moderate DIF), and CC (large DIF) levels.³⁴ Additional classifications include the following:

- AA =negligible DIF
- BB+ =moderate DIF favoring the focal group (indicating the item appears more uniformly and reliably scored for the severity level vs. the United States)
- BB- =moderate DIF favoring the reference group (indicating the item appears more uniformly and reliably scored for the severity level vs. the comparison region)
- CC+ =large DIF favoring the focal group (indicating the item appears more uniformly scored for the severity level vs. the United States)
- CC- =large DIF favoring the reference group (indicating the item appears more uniformly and reliably scored for the severity level vs. the comparison region).

RESULTS

Demographic and clinical data are reported in Table 2 and include subject age, age of onset, sex, and PANSS total scores.

The KMO measure of sampling adequacy ranged from 0.95 to 0.99, and Bartlett's test of sphericity was significant ($p \leq 0.001$), indicating that a factor analysis was appropriate for these data. Consistent with findings in people with schizophrenia,^{1,5,21} we identified the two-factor solution of negative symptoms in our sample (Figure 1).

Fit indices for the two models are presented in Table 3. Chi-square difference tests comparing

the one-factor and two-factor models found that the two-factor model exhibited a significantly better fit than did the one-factor model ($\chi^2=68.127$, degrees of freedom=1, $p \leq 0.001$); therefore, the two-factor model was selected as the final model. The two-factor model shows good model fit (Table 3).^{25–27}

DIF analysis was performed for items on the PANSS expressive deficit factor and the PANSS experiential deficit factor for 15 geographical regions as compared with the United States. Results are presented in Table 4. All significant differences in chi-square also reported moderate effect sizes, thereby confirming the DIF for all items with moderate (BB) to large (CC) ETS classifications.

With respect to the expressive deficit factor, more DIF was observed for items in the expressive deficit factor category than for items in the experiential deficit factor category. Across countries, there were 16 cases of large DIF and 21 cases of moderate DIF for expressive deficit items (out of 60 total item-by-region comparisons), as compared with four cases of large DIF and 10 cases of moderate DIF for experiential deficit. The following regions showed moderate-to-large DIF for all items of the expressive deficit factor: Austria-Germany, Nordic, France, and Poland. Similarly, Austria-Germany, Brazil, and South Africa showed large DIF (CC) for three items of the expressive factor. France and Spain showed large DIF (CC) for N3 Poor Rapport and N6 Lack of Spontaneity and Flow of Conversation as compared with the United States. India was the only country

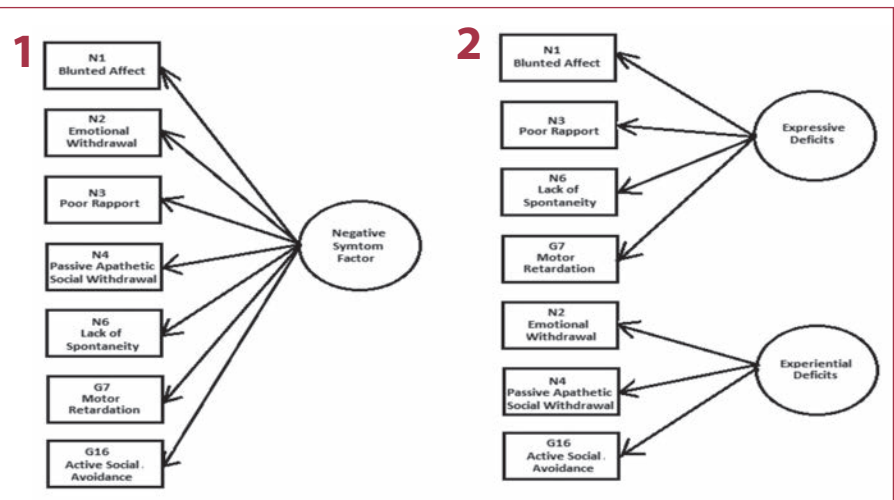


FIGURE 1. Two models of the PANSS negative symptoms factor: 1) a one-factor model; 2) a two-factor model

that showed DIF favoring the United States (as evidenced by CC-), indicating the item is more reliably and uniformly scored for the severity level in the United States than in India. Every moderate-to-large DIF was in favor of the non-United States geographical region under investigation. Of all the items of the NSF, N3 Poor Rapport showed the most moderate and large DIF ($n=13$; 86.67%) across countries, with seven countries reporting large DIF. Similarly, N6 Lack of Spontaneity and Flow of Conversation showed moderate and large DIF for 66.67% countries ($n=10$). Canada was the only country to report no DIF across all items of the expressive deficit domain.

With respect to the experiential deficit factor, there were no geographical regions that showed moderate-to-large DIF for all items. Out of all the factors of the NSF, item G16 Active Social Avoidance reported negligible DIF for 14 of the 15 countries investigated (93.33%). Large DIF was observed for N2 Emotional Withdrawal and N4 Passive Apathetic Social Withdrawal for Brazil and India. Brazil demonstrated the largest DIF classifications (CC) across all countries (i.e., five of the seven NSF items). Seven regions demonstrated no DIF across all items of the PANSS experiential deficit factor (South America-Mexico, Belgium-Netherlands, Nordic, Great Britain, Eastern Europe, Russia, and Spain), as compared with only one region (Canada) that showed no DIF for the PANSS expressive deficit domain. Similar to the findings observed in the PANSS expressive deficit domain, India was the only country to show large DIF in favor of the United States, indicating that the item is more reliably and uniformly scored for the severity level within the United States. Overall, there were many fewer observed items with large DIF for PANSS experiential deficits (only 14 of 45 observed cases of moderate-to-large DIF).

DISCUSSION

Despite the multiple psychometric analyses of the PANSS over the past 30 years, this study is the first to assess performance on the PANSS expressive and experiential deficit factors across varying levels of symptom severity (borderline to extremely mentally ill) and across multiple (i.e., 15) geographical regions. Our first aim was to assess whether the items attributed to the PANSS expressive and experiential deficit factors were observed within our dataset. To this end,

the expressive and experiential items of the PANSS NSF show good model fit and distinct deficits in these two domains, indicating that the PANSS expressive deficit and experiential deficit factors can be reliably used as distinct efficacy endpoints to further characterize negative symptoms. Our second aim was to assess performance of the items attributed to the PANSS expressive and experiential deficits to identify DIF across geographic regions. Our findings showed that, similar to previous studies,^{18,19} negative symptom items show increased variability in scores across raters. Specifically, we showed DIF across multiple countries for most items of the PANSS expressive deficit.

Having determined that negative symptoms of schizophrenia are particularly difficult to assess due to the challenge of eliciting reliable information from a potentially disengaged subject during a clinical interview, this study evaluated differences in PANSS expressive and experiential deficits across 15 geographic regions. The observed differences can help inform protocol development, elucidate ways to customize rater training and data review, and determine endpoints for clinical trials that will subsequently affect accuracy of symptom presentation and trial outcome.

The present study found that all items of the PANSS experiential and expressive deficit factors showed evidence of DIF across geographic regions after matching subjects on PANSS total score. Items of the PANSS expressive deficit factor showed more DIF with the United States compared with other geographical regions. Specifically, for N3 Poor Rapport, 13 of the 15 geographic regions had moderate-to-large DIF with the United States than any other item. N3 Poor Rapport assesses the rater's opinion of the subject's interpersonal empathy, openness in conversation, and sense of closeness, interest, or involvement with the rater. With the available data from our sample, it is unclear how to interpret these findings, given that we do not have information on subject disposition (i.e., hospitalized or non-hospitalized, length of hospitalization); good rapport could be expected with longer inpatient stays as symptoms improve and subjects become more familiar with raters, but such differences cannot be assessed with the available data. We therefore hypothesize that cultural differences between raters might

TABLE 3. Factor loadings and model fit

SEVERITY LEVELS	PANSS NSF
PANSS expressive deficits, % variance	55.12%
N1 Blunted Affect	0.749
N3 Poor Rapport	0.696
N6 Lack of Spontaneity	0.783
G7 Motor Retardation	0.744
PANSS experiential deficits, % variance	14.65%
N2 Emotional Withdrawal	0.666
N4 Passive Apathetic Social Withdrawal	0.779
G16 Active Social Avoidance	0.886
Chi-squared	99.215
p^*	≤ 0.001
Root mean square error of approximation (RMSEA)**	0.045
Comparative fit index (CFI)***	0.923
Tucker-Lewis index (TLI)****	0.913
Goodness-of-fit index (GFI)*****	0.906

PANSS: Positive and Negative Syndrome Scale; NSF: negative symptom factor; *if significance level set at $P < 0.05$; **RMSEA ≤ 0.05 is "good fit," ***CFI ≥ 0.90 is "satisfactory fit," ****TLI ≥ 0.90 is "satisfactory fit," *****GFI ≥ 0.90 is "satisfactory fit"

have affected the scoring of N3 Poor Rapport in the geographical regions examined due to distinctive interpersonal norms. While assessing the influence of language and culture on the PANSS across seven geographic regions, Yavorsky et al³⁵ found differences in the rating of negative symptoms, particularly poor rapport, due to the ways in which different groups conceptualize N3 Poor Rapport. Moreover, N3 Poor Rapport characterizes behaviors that are perhaps less accepted in the United States, Canada, and India, as compared with in other geographic regions where they might be more tolerated. Item N1 Blunted Affect showed DIF favoring seven geographic regions compared with the United States; item N6 Lack of Spontaneity and Flow of Conversation showed DIF favoring nine geographic regions as compared with the United States, with India being the only region that showed DIF favoring the United States. N1 Blunted Affect is scored solely on the rater's observation of the subject's physical manifestations. However, Mesquita and Frijda³⁶ reviewed evidence that there are cultural similarities and differences in all components of the affect process, including in

TABLE 4. Differential item functioning of PANSS expressive and experiential deficits with the United States as the reference geographical region

GEOGRAPHICAL REGION	PANSS EXPRESSIVE DEFICITS				PANSS EXPERIENTIAL DEFICITS		
	N1 BLUNTED AFFECT	N3 POOR RAPPORT	N6 LACK OF SPONTANEITY	G7 MOTOR RETARDATION	N2 EMOTIONAL WITHDRAWAL	N4 PASSIVE/APATHETIC SOCIAL WITHDRAWAL	G16 ACTIVE SOCIAL AVOIDANCE
South America-Mexico							
<i>p</i>	0.44	<0.001	<0.001	0.15	0.33	0.07	0.03
ES	0.03	0.54	0.35	0.09	0.04	0.13	0.16
ETS Class	AA	BB+	BB+	AA	AA	AA	AA
Austria-Germany							
<i>p</i>	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.01
ES	0.51	0.8	0.78	0.68	0.41	0.34	0.17
ETS Class	BB+	CC+	CC+	CC+	BB+	BB+	AA
Belgium-Netherlands							
<i>p</i>	<0.001	<0.001	0.03	<0.001	0.04	0.28	0.25
ES	0.23	0.48	0.2	0.31	0.13	0.07	-0.09
ETS Class	AA	BB+	AA	BB+	AA	AA	AA
Brazil							
<i>p</i>	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.01
ES	0.78	0.99	0.68	-0.29	0.75	0.66	-0.25
ETS Class	CC+	CC+	CC+	AA	CC+	CC+	AA
Canada							
<i>p</i>	<0.001	<0.001	0.02	0.27	<0.001	<0.001	0.02
ES	0.2	0.17	0.14	-0.06	0.24	0.31	0.13
ETS Class	AA	AA	AA	AA	AA	BB+	AA
Nordic							
<i>p</i>	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.37
ES	0.31	0.54	0.38	0.32	0.26	0.27	-0.04
ETS Class	BB+	BB+	BB+	BB+	AA	AA	AA
France							
<i>p</i>	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.14
ES	0.44	0.91	0.67	0.53	0.42	0.53	0.1
ETS Class	BB+	CC+	CC+	BB+	BB+	BB+	AA
Great Britain							
<i>p</i>	0.89	<0.001	0.02	0.42	0.04	<0.001	0.21
ES	-0.02	0.33	0.23	0.07	0.17	0.27	0.13
ETS Class	AA	BB+	AA	AA	AA	AA	AA
India							
<i>p</i>	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.55
ES	-0.87	-0.09	-0.57	-0.47	-0.75	-0.7	0.06
ETS Class	CC-	AA	BB-	BB-	CC-	CC-	AA
Italy							
<i>p</i>	0.27	<0.001	<0.001	0.01	0.07	<0.001	<0.001
ES	0.13	1	0.45	0.29	0.17	0.59	0.36
ETS Class	AA	CC+	BB+	AA	AA	BB+	BB+
Poland							
<i>p</i>	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.01
ES	0.47	0.49	0.41	0.41	0.43	0.29	0.18
ETS Class	BB+	BB+	BB+	BB+	BB+	AA	AA

TABLE 4, continued. Differential item functioning of PANSS expressive and experiential deficits with the United States as the reference geographical region

GEOGRAPHICAL REGION	PANSS EXPRESSIVE DEFICITS				PANSS EXPERIENTIAL DEFICITS		
	N1 BLUNTED AFFECT	N3 POOR RAPPORT	N6 LACK OF SPONTANEITY	G7 MOTOR RETARDATION	N2 EMOTIONAL WITHDRAWAL	N4 PASSIVE/APATHETIC SOCIAL WITHDRAWAL	G16 ACTIVE SOCIAL AVOIDANCE
Eastern Europe							
<i>p</i>	0.26	<0.001	0.02	0.01	0.64	0.86	0.11
ES	0.08	0.34	0.17	0.16	0.03	0.001	-0.09
ETS Class	AA	BB+	AA	AA	AA	AA	AA
Russia							
<i>p</i>	<0.001	<0.001	0.03	0.54	<0.001	0.01	0.06
ES	0.41	0.63	0.24	-0.05	0.27	0.23	-0.19
ETS Class	BB+	CC+	AA	AA	AA	AA	AA
South Africa							
<i>p</i>	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.16
ES	0.79	0.76	0.63	0.23	0.31	0.38	-0.07
ETS Class	CC+	CC+	CC+	AA	BB+	BB+	AA
Spain							
<i>p</i>	0.01	<0.001	0.01	0.18	0.06	0.41	0.31
ES	0.27	0.62	0.32	0.13	0.18	0.08	-0.11
ETS Class	AA	CC+	CC+	AA	AA	AA	AA
Differential Item Functioning (DIF)							
Number of large DIF	3	7	5	1	2	2	0
Number of moderate DIF	5	6	5	5	4	5	1
Number of negligible DIF	7	2	5	9	9	8	14
PANSS=Positive and Negative Syndrome Scale; ES=effect size; ETS=Education Testing Services; AA=negligible DIF; BB+=moderate DIF favoring the focal group; BB-=moderate DIF favoring the reference group; CC+=large DIF favoring the focal group; CC-=large DIF favoring the reference group							

emotion regulation and display. In explaining geographical differences in negative symptoms, not only are the norms for expression of emotions and flow of conversation relevant, but so are the norms for experience of behavior. Examinations of emotion intensity perception have confirmed that facial expressions with varying levels of intensity of positive and negative affect are perceived and categorized differently across cultures.³⁷ Intrinsic biological factors, such as genes and the central nervous system, are substantially shaped by cultural and social contexts during development. These relations between biology and context subscribe to observed behavioral patterns of individuals (e.g., PANSS raters) and cultural agreement/disagreement in identifying expressions of affect (N1) and verbal interaction (N6).^{38–40} For this study, we observed large DIF for Brazil, India, and South Africa as compared with the United States, and moderate DIF for Austria-Germany,

Nordic, France, Poland, and Russia as compared with the United States. It has been shown that individualistic cultures (e.g., the United States, South Africa, Austria-Germany, France) tend to endorse physical display of expression and conversation, while collectivistic cultures (e.g., Brazil, India, Nordic, Poland, Russia) encourage the control of expressions of affect to maintain group harmony.^{41–43} Thus, the role of display rules in regulating and interpreting affect and flow of conversation in a variety of contexts has been well-documented and varies across cultures. Although the United States is considered an individualistic culture, it also comprises a heterogeneous community of raters and subjects from other geographic locations, such that it would be difficult to assess pure cultural differences in presentation and interpretation of affect and flow of conversation. However, since it is more difficult to regulate the variability of subjects with schizophrenia, each

rater should have a clear understanding of the presentation of negative symptoms within and across their specific cultural contexts.

It is important to look at items N4 Passive/Apathetic Social Withdrawal and G16 Active Social Avoidance, as these items of the PANSS experiential deficit are scored exclusively by reports from the subject's caregiver. These have previously been shown to be the two best items for predicting everyday social outcomes in people with schizophrenia.⁴⁴ Both of these items showed the fewest amount of DIF between the 15 geographic regions as compared with the United States. Moreover, only one country—Italy—showed DIF (at a moderate level) for G16 Active Social Avoidance, with no countries showing large DIF. Additionally, Brazil and India, highly heterogeneous countries with multiple subcultures and distinct languages, showed large DIF for N4 Passive/Apathetic Social Withdrawal. With the exception of Brazil

and India, the small number of DIF identified for these two items might be due to less subjectivity in interpretation on the part of the rater or to less variability in the presentation of these core, unmistakable features of the illness across most geographical regions.

There are several likely explanations for DIF among raters across and within diverse geographical locations.⁴⁵ One key reason might be variation across raters in measurement procedures and variability in interpretation of measurement result. This variability in measurement procedures (e.g., PANSS administration, interview skill, interview environment) and variability in interpretation (i.e., scoring the PANSS NSF) implies that when differences occur once raters have agreed upon criteria for administering and scoring a symptom, they are the result of decision-making differences in the scoring of the item.¹⁸ Since cultural differences cannot be standardized, the development of a standardized international PANSS training curriculum is not possible. However, training can be culturally adapted to manage these differences by supplementing the standard PANSS training with additional culture-specific training. Akin to the linguistic and cultural validation processes employed in the translation of rating scales, rater training could also include linguistic and cultural methodologies based on findings from cultural analysis of rating scales. For countries for which normative data are not available, this can be achieved by providing “cultural translations” of specific PANSS items, concepts, and symptoms. Such “cultural translation” could involve the employment of native culture-specific experts to provide detailed guidance on how specific items and concepts on the PANSS are manifested in their cultures. For instance, when deploying rater training for negative symptom trials, training should be customized for geographical location, cultural and language norms, and expectations of what constitutes endorsing each anchor point for the items with large DIF. It should be especially ensured that the training received for raters in the United States, India, Brazil, and other heterogeneous regions, captures within-region variability in language, culture, and social constructs. Our study identified significant moderate-to-large DIF for items of the PANSS expressive deficit across geographical locations as compared with the United States. Dissimilar social interpretations

due to geographical and cultural influences might lead to different ratings of social and emotional behaviors present in the PANSS expressive deficit and should be subjected to interim item analysis throughout a clinical trial.

Despite social, linguistic, and cultural differences between sites, large international clinical trials will continue to be conducted, and data from these trials will be combined to assess efficacy. For this reason, it is important to underscore that DIF does not denote that the scores provided by the raters are not appropriate for the culture, but that the interpretation of the anchor points as outlined in the PANSS can be further explored to lessen large scoring discrepancies among regions. For example, in a previous DIF study conducted by our group in which all subjects viewed the same PANSS interview video, we also found differences in the interpretation of anchors across geo-cultural regions.¹⁸ The expectation of supplemental training is not to homogenize the understanding of a symptom, but rather to clearly define that symptom within a social and cultural context.

Limitations. The present study has some limitations. First, we examined subjects with chronic schizophrenia who were screened for enrollment in various clinical trials and who were taking one or more antipsychotic medications. Consequently, this study's results are not generalizable to subjects in different illness courses, such as first-episode subjects or subjects who are not on an antipsychotic medication. Second, the data used in the analysis comprises data collected in 16 clinical trials that did not specifically focus on negative symptoms, although the overall Negative Symptom subscale score and the NSF score were higher than the overall Positive Symptom subscale score for this sample. Additionally, scores on the NSF ranged from 7 (lowest possible score) to 48 (highest possible score is 49). The baseline data from these 16 trials are also representative of individuals who enter multicenter international clinical trials. Third, this analysis focuses on raters from 15 geographical locations with varying levels of proficiency and experience in scoring the PANSS. Although all raters received rater training and certification prior to conducting PANSS assessments, training and certification processes differed across the 16 studies, and specific interrater reliability values were not

available. Fourth, although this is a very diverse sample, it does not include every area in which clinical trials are commonly conducted (e.g., the Philippines). Additionally, some could argue that our groupings are themselves heterogeneous (e.g., Finland among the Nordic countries has language differences; grouping Mexico and South America together was not based on geographic location, but rather on language similarities). Fifth, as this study examines PANSS scores at baseline only and not longitudinally, treatment change was not addressed. Sixth, our dataset did not contain the language in which the PANSS was administered, the specific site location within the geographical region, or rater information (e.g., experience level, qualifications). We recognize that these could influence the differences in scoring responses. Seventh, rater training could not be examined using the currently available data and should be addressed in future studies assessing cross-regional comparisons. Finally, we acknowledge that individuals with negative symptoms might not provide accurate information or enough information for adequate assessment of a symptom.

CONCLUSIONS

Following research conducted over the past 30 years, this study addressed how items of the PANSS expressive and experiential deficits function across cultures. Items of the PANSS expressive deficit show more DIF across 15 geographic regions, as compared with the items of the experiential deficit. These differences among geographical regions might be related to rater cultural interpretations, language differences, social experiences, probability of the subject endorsing negative symptoms, rater training, and/or subject geo-cultural variability. The results of this study could be useful in protocol development, rater training practices across geographical regions, and decision-making among clinicians and researchers. Furthermore, these results might highlight subtle phenomenological differences between expressive and experiential deficits that can be used to guide future research. Future efforts to develop scales assessing negative symptoms would benefit from examining whether a scale functions in the same way across regions, cultures, languages, severity levels, and in relationship to functional outcomes. Harvey et al⁴⁶ use these factor structures to examine their

predictions of multiple aspects of everyday functioning in an independent sample of people with schizophrenia by comparing expressive and experiential deficits to the PANSS NSF.

APPENDIX 1

Clinical trials information. Clinical trials included in the analysis dataset are presented below. Some international studies or studies conducted prior to the year 2001 that do not have clinicaltrials.gov identifiers are listed as “Data on File” with the pharmaceutical company or include a link to the relevant publication in which the data were previously presented.

- RIS-SCH-401 (NCT00297388)
- RIS-SCP-402 (NCT00061802)
- RIS-INT-2 (Peuskens J. Risperidone in the treatment of chronic schizophrenic patients: an international multi-center double blind parallel-group comparative study versus haloperidol. Jan 1992. Janssen Clinical Research Report no.: RIS-INT-2)
- RIS-INT-57 (NCT00558298)
- 076477-SCH-305 (NCT00668837)
- R076477-SCH-303 (NCT00650793)
- RIS-INT-61 (NCT00558298)
- RIS-INT-57 (NCT00558298)
- RIS-INT-3 (Marder SR. Risperidone versus haloperidol versus placebo in the treatment of chronic schizophrenia. Nov 1991. Janssen Clinical Research Report no.: RIS-INT-3)
- R076477-SCH-304 (NCT00077714)
- RIS-INT-50 (Data on File: RIS-INT-50. Janssen Pharmaceutical Products, L.P., Titusville, NJ; 2000)
- RIS-USA-112 (Conley RR, Mahmoud R. A randomized double-blind study of risperidone and olanzapine in the treatment of schizophrenia or schizoaffective disorder. *Am J Psychiatry*. 2001;158(5):765–774)
- RIS-USA-121 (NCT00253136)
- RIS-USA-250 (NCT00378183)
- RIS-USA-305 (NCT00236353)
- RIS-USA-79 (NCT00253110)

APPENDIX 2

Educational Testing Service DIF classification system. The ETS system for DIF classification has been in place for nearly 25

years. The ETS DIF criteria combine the Mantel-Haenszel procedure with the contrast between the Rasch-based item difficulty estimates for the different groups. As described by Zieky,⁴⁷ statistical analyses are used to designate items into three ETS DIF categories according to the direction, size, and significance of the DIF statistics.^{48,49} These categories were created to “avoid identifying items that display practically trivial but statistically significant DIF.”⁵⁰ The three categories are as follows:

- A=negligible or nonsignificant DIF
- B=slight to moderate DIF
- C=moderate to large DIF.

The rules currently used by the ETS to classify items as A, B, or C are based on the magnitude of the Mantel-Haenszel delta difference (MH D-DIF) statistic and its statistical significance. The Mantel-Haenszel approach⁵¹ to DIF analysis, developed by Holland and Thayer,⁵² involves the creation of “k” two-by-two tables, where k is the number of score categories on the matching criterion.^{51,52} For the kth score level, the data can be summarized as follows: NF1k denote the numbers of examinees in the reference and focal groups, respectively, who answered correctly; and NROk and NFOk are the numbers of examinees in the reference and focal groups who answered incorrectly. Nk is the total number of examinees. In developing the MH D-DIF index, Holland and Thayer⁵³ elected to express the statistic on the ETS delta scale of item difficulty. An MH D-DIF value of -1, for example, means that the item is estimated to be more difficult for the focal group than for the reference group by an average of one delta point, conditional on ability.⁵³ Expressing the amount of DIF in this way was intended to make the MH D-DIF statistic more useful for test development.

Therefore, an A item is one for which either the Mantel-Haenszel (MH) chi-square statistic is not significant at the 0.05 level, or MH D-DIF is smaller than 1 in absolute value. A C item is one for which the MH D-DIF statistic is significantly greater than 1 in absolute value at the 0.05 level and has an absolute value of 1.5 or more. Items that do not meet the definition for either A or C items are considered B items. More explicitly, an item is declared a B item if it does not meet the qualifications for a C item and if 1) MH CHISQ is greater than 3.84, and 2) if [MH D-DIF] is 1 or greater.

REFERENCES

1. Blanchard JJ, Cohen AS. The structure of negative symptoms within schizophrenia: implications for assessment. *Schizophr Bull*. 2006;32(2):238–245.
2. Kirkpatrick B, Fenton WS, Carpenter WT, Marder SR. The NIMH-MATRICS consensus statement on negative symptoms. *Schizophr Bull*. 2006;32(2):214–219.
3. Messinger JW, Trémeau F, Antonius D, et al. Avolition and expressive deficits capture negative symptom phenomenology: implications for DSM-5 and schizophrenia research. *Clin Psychol Rev*. 2011;31(1):161–168.
4. Strauss GP, Keller WR, Buchanan RW, et al. Next-generation negative symptom assessment for clinical trials: validation of the brief negative symptom scale. *Schizophr Res*. 2012;142(1–3):88–92.
5. Kring AM, Gur RE, Blanchard JJ, et al. The clinical assessment interview for negative symptoms (CAINS): final development and validation. *Am J Psychiatry*. 2013;170(2):165–172.
6. Galderisi S, Bucci P, Mucci A, et al. Categorical and dimensional approaches to negative symptoms of schizophrenia: focus on long-term stability and functional outcome. *Schizophr Res*. 2013;147(1):157–162.
7. Lyne J, Renwick L, Madigan K, et al. Do psychosis prodrome onset negative symptoms predict first presentation negative symptoms? *Eur Psychiatry*. 2014;29(3):153–159.
8. Quinlan T, Roesch S, Granholm E. The role of dysfunctional attitudes in models of negative symptoms and functioning in schizophrenia. *Schizophr Res*. 2014;157(1–3):182–189.
9. Green MF, Bearden CE, Cannon TD, et al. Social cognition in schizophrenia, part 1: performance across phase of illness. *Schizophr Bull*. 2012;38(4):854–864.
10. Rassovsky Y, Horan WP, Lee J, et al. Pathways between early visual processing and functional outcome in schizophrenia. *Psychol Med*. 2011;241(3):487–497.
11. Ventura J, Subotnik KL, Gitlin MJ, et al. Negative symptoms and functioning during the first year after a recent onset of schizophrenia and eight years later. *Schizophr Res*. 2015;161(2–3):407–413.
12. Schlosser DA, Campellone TR, Biagianni B, et al. Modeling the role of negative symptoms in determining social functioning in individuals at clinical high risk of psychosis. *Schizophr Res*. 2015;169(1–3):204–208.
13. Llerena K, Reddy LF, Kern RS. The role of experiential and expressive negative symptoms on

- job obtainment and work outcome in individuals with schizophrenia. *Schizophr Res*. 2017 Jun 7.
14. Strauss GP, Horan WP, Kirkpatrick B, et al. Deconstructing negative symptoms of schizophrenia: avolition-apathy and diminished expression clusters predict clinical presentation and functional outcome. *J Psychiatry Res*. 2013;47(6):783–790.
 15. Jang SK, Choi HI, Park S, et al. A two-factor model better explains heterogeneity in negative symptoms: evidence from the Positive and Negative Syndrome Scale. *Front Psychol*. 2016;7:707.
 16. Castro SM, Curi M, Torman VB, Riboldi J. Differential item functioning in the Beck depression inventory. *Rev Bras Epidemiol*. 2001;18(1):54–67.
 17. Kay SR, Fiszbein A, Opler LA. The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophr Bull*. 1974;13(2):261–276.
 18. Khan A, Yarovsky Y, Liechti S, et al. A rasch model to test the cross-cultural validity in the positive and negative syndrome scale (PANSS) across six geo-cultural groups. *BMC Psychol*. 2013;1(1):5.
 19. Daniel DG, Alphas L, Cazorla P, et al. Training for assessment of negative symptoms of schizophrenia across languages and cultures: comparison of the NSA-16 with the PANSS Negative Subscale and Negative Symptom factor. *Clin Schizophr Relat Psychoses*. 2011;5(2):87–94.
 20. Brekke JS, Barrio C. Cross-ethnic symptom differences in schizophrenia: the influence of culture and minority status. *Schizophr Bull*. 1997;23(2):305–316.
 21. Foussias G, Remington G. Negative symptoms in schizophrenia: avolition and Occam's razor. *Schizophr Bull*. 2010;36(2):359–369.
 22. Zumbo BD. *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (ordinal) Item Scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense; 1999.
 23. Goren E. Economic Effects of Domestic and Neighbouring Countries' Cultural Diversity. Working Papers V-352-13. Oldenburg, Germany: University of Oldenburg Department of Economics; 2013. Retrieved from <http://www.etsg.org/ETSG2013/Papers/042.pdf>.
 24. Kaiser HF. An index of factorial simplicity. *Psychometrika*. 1974;39(1):31–36.
 25. Hair JF, Black WC, Babin BJ, Anderson RE. *Multivariate Data Analysis, a Global Perspective*. 7th ed. Upper Saddle River, NJ: Pearson Prentice Hall; 2010.
 26. Forza C, Filippini R. TQM impact on quality conformance and customer satisfaction: A causal model. *Int J Prod Econ*. 1998;55(1):1–20.
 27. Awang Z. *Structural Equation Modeling Using Amos Graphic*. Malaysia: UiTM Press; 2012.
 28. IBM Corp. Released 2015. IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY: IBM Corp.
 29. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria; 2013.
 30. Kim SH, Cohen AS, Alagoz, C, Kim S. DIF detection and effect size measures for polytomous scored items. *J Edu Meas*. 2007;44(2):93–116.
 31. Gelin MN, Zumbo BD. Differential item functioning results may change depending on how an item is scored: An illustration with the Center for Epidemiologic Studies Depression Scale. *Edu Psychol Meas*. 2003;63(1):65–74.
 32. Monahan PO, McHorney CA, Stump TE, Perkins AJ. Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *J Edu Behav Stat*. 2007;32(1):92–109.
 33. Angoff WH. Perspectives on differential item functioning methodology. In: Holland PW, Wainer H, eds. *Differential Item Functioning*. Hillsdale, NJ: Erlbaum; 1993: 3–23.
 34. Longford NT, Holland PW, Thayer DT. Stability of the MH D-DIF statistics across populations. In: Holland PW, Wainer H, eds. *Differential Item Functioning*. Hillsdale, NJ: Erlbaum; 1993: 171–196.
 35. Yavorsky C, Liechti S, Opler M. The impact of language and culture on the delivery of standardized rater training for the PANSS across seven countries. *Eur Psychiatry*. 2010;25(1):1555.
 36. Mesquita B, Frijda NH. Cultural variations in emotions: a review. *Psychol Bull*. 1992;112(2):179–204.
 37. Engelmann JB, Pogossyan M. Emotion perception across cultures: the role of cognitive mechanisms. *Front Psychol*. 2013;4:118.
 38. McCrae RR, Costa PT, Ostendorf F, et al. Nature over nurture: temperament, personality, and lifespan development. *J Pers Soc Psychol*. 2000;78(1):173–186.
 39. Adolphs R, Jansari A, Tranel D. Hemispheric perception of emotional valence from facial expressions. *Neuropsychol*. 2001;15(4):516–524.
 40. Elfenbein HA, Ambady N. Cultural similarity's consequences: A distance perspective on cross-cultural differences in emotion recognition. *J Cross-Cult Psychol*. 2003;34(1):92–110.
 41. Markus HR, Kitayama S. Culture and the self: Implications for cognition, emotion, and motivation. *Psychol Rev*. 1991;98(2):224–253.
 42. Heine SJ, Lehman DR, Markus HR, Kitayama S. Is there a universal need for positive self-regard? *Psychol Rev*. 1999;106(4):766–794.
 43. Matsumoto D, Yoo SH, Fontaine J. Mapping expressive differences around the world: the relationship between emotional display rules and individualism versus collectivism. *Jo Cross-Cult Psychol*. 2008;39(1):55–74.
 44. Robertson BR, Prestia D, Twamley EW, et al. Social competence versus negative symptoms as predictors of real world social functioning in schizophrenia. *Schizophr Res*. 2014;160(1-3):136–141.
 45. van Herk H, Poortinga YH, Verhallen TM. Response styles in rating scales: Evidence of method bias in data from six EU countries. *J Cross-Cult Psychol*. 2004;35(3):346–360.
 46. Harvey PD, Khan A, Keefe RSE. Using the PANSS to define different domains of negative symptoms: prediction of everyday functioning by impairments in emotional expression and emotional experience. *Innov Clin Neurosci*. 2017;14(11-12):18–22.
 47. Zieky M. DIF statistics in test development. In: Holland PW, Wainer H, eds. *Differential Item Functioning*. Hillsdale, NJ: Erlbaum; 1993: 337–347.
 48. Linacre JM. WINSTEPS, Version 3.81. Chicago, IL: Winsteps; 2014.
 49. Camilli G, Shepard LA. Methods for identifying biased test items. Thousand Oaks, CA: Sage Publications, Inc.; 1994.
 50. Clauser BE, Mazor KM. Using statistical procedures to identify differentially functioning test items. *Educ Meas*. 1998;17(1):31–44.
 51. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst*. 1959;22(4):719–748.
 52. Holland PW, Thayer DT. Differential item functioning and the Mantel-Haenszel procedure. In: Wainer H, Braun HI, eds. *Test Validity*. Hillsdale, NJ: Erlbaum; 1988: 129–145.
 53. Holland PW, Thayer DT. An alternate definition of the ETS delta scale of item difficulty (ETS Program Statistics Research Technical Report No. 85-64). Princeton, NJ: Educational Testing Service; 1985.