HHS Public Access

Author manuscript

Compr Psychiatry. Author manuscript; available in PMC 2015 December 25.

Published in final edited form as:

Compr Psychiatry. 2014 November ; 55(8): 1862–1874. doi:10.1016/j.comppsych.2014.08.046.

Auditory Tasks for Assessment of Sensory Function and Affective Prosody in Schizophrenia

Eva Petkova[a,b], **Feihan Lu**[c], **Joshua Kantrowitz**[b,d], **Jamie L. Sanchez**[b], **Jonathan Lehrfeld**[b], **Nayla Scaramello**[b], **Gail Silipo**[b], **Joanna DiCostanza**[b], **Marina Ross**[b], **Zhe Su**[a], **Daniel C. Javitt**[b,d], and **Pamela D. Butler**[b,e]

[a]Department of Child and Adolescent Psychiatry, New York University Langone Medical Center, New York, NY 10016

[b]Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY 10962

[c]Department of Statistics, Columbia University, New York, NY 10027

[d]Department of Psychiatry, Columbia University College of Physicians and Surgeons, New York, NY 10027

[e]Department of Psychiatry, New York University School of Medicine, New York, NY 10016

Abstract

Schizophrenia patients exhibit impairments in auditory-based social cognition, indicated by deficits in detection of prosody, such as affective prosody and basic pitch perception. However, little is known about the psychometric properties of behavioral tests used to assess these functions. The goal of this paper is to characterize the properties of prosody and pitch perception tasks and to investigate whether they can be shortened. The pitch perception test evaluated is a tone-matching task developed by Javitt and colleagues (J-TMT). The prosody test evaluated is the auditory emotion recognition task developed by Juslin and Laukka (JL-AER). The sample includes 124 schizophrenia patients (SZ) and 131 healthy controls (HC). Properties, including facility and discrimination, of each item were assessed. Effects of item characteristics (e.g., emotion) were also evaluated. Shortened versions of the tests are proposed based on facility, discrimination, and/or ability of item characteristics to discriminate between patients and controls. Test-retest reliability is high for patients and controls for both the original and short forms of the J-TMT and JL-AER. Thus, the original as well as short forms of the J-TMT and JL-AER are suggested for inclusion in clinical trials of social cognitive and perceptual treatments. The development of short forms further increases the utility of these auditory tasks in clinical trials and clinical practice. The

Corresponding author: Eva Petkova, PhD; Address: One Park Ave, 7th floor, New York, NY 10016; Phone: 646-754-5143; eva.petkova@nyumc.org.

Disclosures

Within the past 36 month, Dr. Javitt reports receiving honoraria from Sunovion, BMS, Eli Lilly, Takeda, Omeros, Otsuka, Consensus Medical Communications, Guidepoint global, American Capital, Clearpoint communications, Vindico Medical Communication, and Clearview Healthcare; research support from Pfizer and Roche; equity in Glytech, Inc. and AASI; intellectual property rights for use of glycine, D-serine and glycine transport inhibitors in schizophrenia, and serving on scientific advisory board of Promentis. Dr. Kantrowitz has conducted clinical research supported by the NIMH, the Stanley Foundation, Roche-Genetech, EnVivo, Sunovion, Novartis and Pfizer. He reports having received consulting payments within the last 2 years from Otsuka Pharmaceuticals, the Healthcare Advisory Board, Vindico Medical Education, Health Advances, LLC, Strategic Edge Communications. He owns a small number of shares of common stock in GlaxoSmithKline. Other authors report no financial relationships with commercial interests.

large SZ vs. HC differences reported here also highlight the profound nature of auditory deficits and a need for remediation.

Keywords

auditory; emotion; prosody; psychometrics; test-retest reliability

1. INTRODUCTION

Social cognitive dysfunction is a core feature of schizophrenia and is among the strongest predictors of impaired functional outcome[3–7], rendering it a major determinant of long-term disability. A large effort, exemplified for instance by the Cognitive Neuroscience Treatment Research to Improve Cognition in Schizophrenia (CNTRICS) initiative, is currently devoted to task development for assessing cognitive (including social cognitive) and perceptual deficits in schizophrenia that can be used in clinical trials[8–12]. In the realm of verbal communication (e.g., prosody) and the related ability to detect pitch, there are several tests available (e.g.,[1, 2, 13–17]), but their psychometric properties have not been systematically described. The focus of this paper is on reporting the psychometric properties of easy-to-administer behavioral tests for assessment of auditory function.

Basic pitch perception deficits, assessed with tone matching tasks (TMTs), have been studied in schizophrenia since the 1990s[18–21], are well-replicated[2, 16, 22–25], and show large diagnostic differences of ~1.2 SDs across numerous studies. Impaired pitch perception in schizophrenia does not appear to be due to deficits in attention and working memory[26], rather it is related to structural[27] and functional[28, 29] impairment in primary auditory cortex. People with schizophrenia (SZ) have repeatedly shown deficits on the TMT developed by Javitt and colleagues, which we call here J-TMT[2, 16, 23–25].

Emotional prosody, also referred to as auditory emotion recognition (AER), depends on recognition of complex physical characteristics of tones, such as low base frequency and low pitch variability for sadness[1, 23]. Detection of emotion in others based on tone of voice is crucial for social interactions[1] and is impaired across a number of tasks in schizophrenia[2, 14, 16, 22–25, 30–32]. Not surprisingly, deficits in recognizing differences in pitch between tones is related to impairment in AER[2, 16, 23, 25, 33]. However, there is little standardization of characteristics of stimuli in most AER tasks. Our group[2, 16, 23] and others[2, 22] have recently begun using a task developed by Juslin and Laukka[1] (JL-AER) in which physical features of stimuli have been characterized. Deficits with large effect sizes are reported on this task as well as significant relationships with measures of functional outcomes (e.g.,[2]).

Ability to measure auditory functioning both in clinical practice and in schizophrenia research is important for understanding auditory perceptual function and social cognition, as well as how changes in auditory function relate to higher-level deficits, e.g., verbal memory following remediation of auditory function[34, 35]. This paper is a detailed investigation of J-TMT and JL-AER to assess psychometric properties of the tests and allow an informed approach to possibly shortening the tests. The J-TMT consists of pairs of 100-msec tones

with a 500-msec inter-tone interval[2, 16, 25]. There are five pitch differentials (PDs), i.e., differences between tones in the pairs: 2.5%, 5%, 10%, 20% and 50%, with the 2.5% differential being most difficult to detect. Tones are either identical (half) or differ in frequency by the specified percent for each PD. In each PD level, 13 unique pairs of tones are presented once in a fixed sequence and immediately repeated in the same sequence, 26 pairs per PD, total of 5*26=130 pairs. Three average base frequencies (500, 1000, 2000 Hz) are used in each PD to avoid learning effects. Participants verbally indicate whether the pitch is the same or different for each pair. The stimuli in JL-AER consist of audio recordings of two male and two female actors portraying five emotions (anger, disgust, fear, happiness, and sadness) with two levels of intensity (weak and strong) and utterances with no emotional expression ("neutral")[1]. The sentences are semantically neutral and have two forms: a statement and a question (e.g., "It is eleven o'clock," "Is it eleven o'clock?"). This yields 88 stimuli. Participants are asked to identify the emotional expression of each utterance.

In this paper we study how different *characteristics* of the items on the test affect the ability of the items to differentiate between healthy controls (HC) and SZ. The characteristic we consider for the tone pairs in J-TMT are: PD, order of the tones within a pair, average base frequency, and sequence number of tone-pair within each PD. The characteristics we consider for the utterances in JL-AER are: emotion, speaker, form (question vs. statement), and intensity (weak vs. strong). The results inform the construction of tests for assessing tone differentiation and prosody and indicate possible ways of shortening the tests. Repeated assessments with J-TMT and JL-AER are used to estimate the test-retest reliabilities of the original test scores and the scores on the shortened versions in order to make recommendations for the use of these tasks in clinical practice and research.

2. METHODS

2.1. Participants

The total sample consists of 124 patients meeting Diagnostic and Statistical Manual of Mental Disorder (DSM-IV) criteria for schizophrenia (n=99) or schizoaffective disorder (n=25) and 131 HC. Table 1 shows the number of participants for J-TMT and JL-AER tests with demographic and clinical characteristics. Patients were recruited from inpatient and outpatient facilities associated with the Nathan Kline Institute for Psychiatric Research (NKI). Diagnoses were obtained using the Structured Clinical Interview for *DSM-IV* (SCID)[36] and all available clinical information. Controls with a history of SCID-defined Axis I psychiatric disorders were excluded. Patients and controls were excluded if they had any neurological or ophthalmologic disorders that might affect performance or met criteria for alcohol or substance dependence within the last six months or abuse within the last month. The study was approved by NKI/Rockland Psychiatric Center and Rockland County Department of Mental Health Institutional Review Board. All participants provided informed consent according to the Declaration of Helsinki.

The study sample includes subsets of participants whose auditory data have been presented in previous publications[2, 16, 23, 25, 33]. However, evaluation of psychometric characteristics is a new use of the data.

2.2. Analysis

2.2.1. Item-level analysis—We first describe the individual items within the tasks with respect to “facility” and “discrimination”, which are item-level characteristics from the classic test theory framework[37]. “Facility” is the mean value of an item in the population. In both J-TMT and JL-AER tasks, the responses to all items have only two levels – correct and wrong – and thus, “facility” for each item corresponds to the proportion of the population that answered the item correctly. Facility is also sometimes referred to as “difficulty” of an item, though it should be noted that higher proportion correct here indicates lower difficulty. In the framework of classic test theory, items that are too easy or too difficult are considered not very useful for generating a broad range of scores and differentiating between subjects. A range between 0.25 and 0.75 is suggested. The “discrimination” of an item is a measure of how well the item separates individuals with low and high abilities on the construct assessed by the test. “Discrimination” can be estimated in different ways; here we use the correlation between the score on the item and the score on the total test minus the item. Guidelines recommend that good items should have discrimination above 0.2[37]. Facility and discrimination values were computed using only the first assessment from each subject.

2.2.2. Effect of item characteristics—To better understand the auditory tests and to help shorten them, we study how the characteristics of the test items are related to the ability of the item to differentiate between HC and SZ. Information about whether a specific characteristic is related to how different the responses of SZ and HC are helps determine if some items in the original tests can be eliminated. For example, if the type of utterance in the JL-ARE had an effect, such that utterances that were questions resulted in a larger difference between correct responses of SZ and HC than utterances that were statements, then we might consider completely or partially eliminating the utterances that are statements. We study the following features of the J-TMT items: (J-TMT.i) PD – 5 levels, 2.5%, 5%, 10%, 20% and 50%; (J-TMT.ii) average base frequency – 3 levels (500, 1000 and 2000 Hz); (J-TMT.iii) order of tones within a pair – 3 levels (high first, low first and same); (J-TMT.iv) order of pairs in a PD level – 26 levels (1 to 26, the first 13 and the second 13 are identical). The features of the JL-AER items that we investigate are: (JL-AER.i) emotion – 6 levels (anger, disgust, fear, happiness, sadness, and neutral); (JL-AER.ii) speaker – 4 levels (1, 2, 3 and 4); (JL-AER.iii) sentence form – 2 levels (question or statement); (JL-AER.iv) intensity of emotion – 2 levels (strong and weak; the neutral emotion level does not have an intensity of emotion).

The effects of item characteristics are studied in descriptive and inferential ways. Descriptively, we construct “hyper-items” by combining the items having the same level of a given characteristic and computing their facility and discrimination. Hyper-items are the means of responses on all items that have the same level of a characteristic. For example, the mean of responses to all tone pairs with a PD of 2.5% would be one hyper-item, with 5 hyper-items defined by the 5 levels of PD (2.5%, 5%, 10%, 20% and 50%). As another example, there are 6 hyper-items corresponding to the feature “emotion” for JL-AER (anger, disgust, fear, happiness, sadness, and neutral). Hyper-items are computed for all of the characteristics of the tasks, with a total of 37 hyper-item for J-TMT and 14 for JL-AER. The

facility and discrimination of the hyper-items are used to identify poor (on average) performance of items with a given characteristic, using the same guidelines reported above.

Formal inferences regarding the effect of a characteristic on the ability of an item to differentiate between SZ and HC are based on Generalized Linear Mixed Models (GLMMs)[38]. We modeled the probability of a correct answer on the items as a function of a given characteristic, diagnostic group, and their interaction. The models include random subject effects to account for the potential correlation between the responses given by the same subject. A significant interaction between diagnosis and an item characteristic indicate that the effect of diagnosis on the probability of answering an item correctly depends on the value of the item characteristic. We computed the χ^2 tests for independence between correct response and diagnosis individually for all items. For ease of visualization of the results we present box-plots of the χ^2 values by levels of the characteristics; the χ^2-test statistics are on 1 degree of freedom and statistical significance at α=0.05 is achieved when the test statistic is equal or greater than 3.84. We explored whether the results from the "single covariate" analyses about the effect of an item characteristics on the ability to differentiate between diagnoses, would change when we control for the other item characteristics (i.e., the other characteristics are included as main effects in the models). We also tested whether the joint distribution of any two characteristics is associated with the ability of an item to differentiate between SZ and HC. This was accomplished by including in the models the 3-way interactions (diagnosis-by-characteristics1-by-characteristic2) and assessing the significance of those interaction terms. We did not explore higher order interactions between the characteristics, since the interpretation of such high order terms would be difficult and because no 3-way interactions were found statistically significant.

2.2.3. Summary Scores—The recommended way of scoring the auditory tests is to obtain the percent correct responses across all items. The summary scores for the original and the proposed shorter versions were ranked based on area under the curve (AUC) of the receiver operating characteristics[39]. The test-retest reliability of summary scores is measured by the intra-class correlation coefficient (ICC), separately for SZ and HC. The variances required for computing the ICCs are estimated based on GLMMs modeling the individual scores with no fixed effects and only random subject effects using the identity link function. All repeated assessments from all subjects were employed (some participants had more than two assessments). All analyses were conducted in *R*[40].

3. RESULTS

3.1. J-TMT

3.1.1. Item-level analysis—Figure 1 shows the facility and discrimination of all 130 items (tone-pairs) in the original test plus 11 hyper-items for HC and SZ combined. Only the hyper-items corresponding to (i) PD (5 levels), (ii) average base frequency of a tone-pair (3 levels) and (iii) order of tones within a pair (3 levels) are shown. The last characteristic (iv) order of items in a PD level (with 26 levels) is not shown for clarity of the figure.

For the three lowest PDs, many items meet criteria for facility -- those falling between the horizontal dashed lines on the left panel of Figure 1; they are indicated in solid black as they

also meet the criteria for discrimination. Within the three lowest PDs, pairs that do not meet criteria for facility are primarily those in which the two tones are identical – they have facilities above 0.75, i.e., they are easy to correctly identify as same tones, and discrimination values below the cutoff of 0.2, i.e., they cannot differentiate between subjects with high and low abilities. Within the two largest PD levels almost all tone pairs have facility above 0.75, indicating that they are easy to correctly identify as same or different. However, within these "easy" PD levels almost all items have discrimination above 0.2, although the pairs of different tones have on average larger discrimination than the pairs of identical tones. It is interesting to notice that as the PD level increases, pairs of identical tones tend to have higher discrimination values, i.e., they better differentiate between low and high ability individuals. Note, that pairs of identical tones are necessary in the test and cannot be eliminated.

Item facility and discrimination values were also computed separately for each diagnostic group (results for individual items not shown and available upon request). As expected, the facilities of all pairs (except those with identical tones) were higher for HC than for SZ. However, the discrimination values of all tone-pairs were quite similar for the SZ and HC groups, indicating that J-TMT "works" well for both diagnostic groups. Figure 2 shows results for the five PD hyper-items for SZ and HC separately.

3.1.2. Effect of item characteristics—The $\chi^2(1)$ test statistics for independence between correct answer and diagnosis are computed for all items separately and higher values of χ^2 indicate a greater difference between SZ and HC with respect to correct answers. The panels on Figure 3 show the dependence of the χ^2 test-statistics on the characteristics of the items. The bottom left panel shows that there is a variation in the distributions of χ^2 test statistics corresponding to tone pairs at different PD levels. Tone-pairs at 5% and at 10% PD have higher means and medians of the test statistics than those from more difficult (2.5%) or easier (20% and 50%) PD levels. The 2-way interaction term diagnosis-by-characteristic PD from the GLMM analysis is highly significant ($\chi^2(4)=53.9$, $p<0.001$), consistent with the visual impression from Figure 3. Thus, the PD characteristic was considered when shortening the test.

The bottom middle panel on Figure 3 suggests that tone pairs where the higher tone is presented first and those where the lower tone is presented first do not differ with respect to how well the items differentiate between SZ and HC. Note, that the pairs where the tones are the same have much lower χ^2 test statistics, consistent with the observed very high facilities and very low discrimination of those items. Although items in which the tones are the same discriminate between the diagnostic groups less well, as noted above, their inclusion in the task is necessary for introducing variation and reducing guessing. Therefore pairs with same tones cannot be eliminated and were not included in the GLMM analysis. The interaction of diagnosis-by-characteristic tone order in a pair (with 2 levels only, no "same" level) in the GLMM is not significant, formally confirming the lack of effect of order of tones in a pair.

The bottom right panel of Figure 3 shows the dependence of the χ^2 values for the individual items as a function of average base frequency of the tones in a pair. The χ^2 values for tone pairs with medium level of the average base frequency (1000Hz) appears higher than those

for low (500Hz) and for high (2000Hz) frequencies. However, the interaction term of average base frequency-by-diagnosis in the GLMM is not statistically significant.

Finally, the panel on the top of Figure 3 shows the dependence of the χ^2-statistics on the order of an item within a PD level. Each of the 26 levels of this item characteristic contains 5 items – one from each PD level. We were expecting that items appearing earlier in the sequence might have different ability to differentiate between SZ and HC than items later in the sequence. No such systematic relationship is apparent from the plot. The interaction term diagnosis-by-characteristic order of the item in the PD level is not statistically significant in the GLMM analysis. Keeping in mind that items 1 to 13 within each PD level are unique, while the items 14 to 26 are a replication in the same order, we can notice that the first and second 13 items exhibit similar ability to differentiate between SZ and HC.

When the above analyses regarding each item characteristics were repeated controlling for all the other item-characteristics, the results did not change qualitatively (those results are available from the authors upon request). There was no evidence that any combination of two item characteristics jointly affects the ability of the items to differentiate between HC and SZ, as indicated by the p-values for the 3-way interaction terms between diagnosis, item characteristic 1 and item characteristic 2 for all combinations of characteristics 1 and 2 (all p-values >0.15).

3.1.3. Shortening the test—For shortening the J-TMT two approaches are considered. The first approach is based on the fact that the second 13 items within each PD level are a repetition of the first 13 items and have a similar ability to differentiate between groups as the first 13. The second approach is suggested by the differential effect of PD on between group differences.

3.1.3.1. Reducing the number of tone-pairs: Before reducing the number of tone pairs within a difficulty level, we investigated the similarity between performance of the first 13 and second 13 items and whether it differed by diagnostic group. First, we estimated the intra-class correlation coefficients (ICCs) between the total percent correct over pairs 1–13 and over pairs 14–26. The overall ICC is 0.92, suggesting a high level of similarity between performance on the first and second sets of tone pairs. The ICC for HC is higher than the ICC for SZ, but after controlling for average percent correct, which is necessary because the percent values close to the ceiling of 100% have less variance than percent values in the middle 50%, these ICCs did not differ (p=0.14).

Second, we estimated the area under the curve (AUC) of the receiver operating characteristics (ROC) for the total correct obtained using (a) all 26 items per PD level, (b) only items 1 to 13 and (c) only items 14 to 26. The results are shown on Figure 4, top row. The differences between SZ and HC using only the first 13 of the 26 tone-pairs in the original J-TMT are similar to the differences based on all 26 pairs and also to the second 13 of the tone-pairs. This suggests little benefit from using all 26 pairs compared to using only 13. Thus, one shortened version is to simply use all 5 PDs with only the first 13 pairs in each PD.

3.1.3.2. Reducing the number of PDs: To further shorten the test, using only the first 13 items from each PD level, summary scores (percent correct) were obtained for all possible subsets of the 5 PDs. We compared these subsets with respect to how well the summary scores differentiated between SZ and HC. Based on AUC, the best subset is {5%, 10%, 50%} with AUC=0.7939, followed closely by {5%, 10%, 20%, 50%} with AUC=0.7936 and {2.5%, 5%, 10%, 50%} with AUC=0.7880.

3.1.4. Summary scores—We also consider alternative summary scores to the commonly employed "percent correct" computed from the J-TMT. An alternative sometimes used in similar tests is the smallest PD at which at least 75% of the answers are correct, which we denote as PD75. PD75 from the first 13, the second 13, and from all 26 items show good differentiation between patients and controls, similar to the "percent correct" summary score (Figure 4, bottom row panels).

3.1.5. Test-Retest Reliability—Test-retest reliability for percent correct from the original J-TMT is high for SZ and HC and does not worsen meaningfully for either shortened version (i.e., only the first 13 pairs for each PD, or only the first 13 pairs for a test with PDs of 5%, 10%, and 50% (see Table 2). Test-retest reliability for D75 is somewhat lower. Thus, the recommendation is to utilize percent correct and employ only half of the tones at each PD, which is half the length of the original test. For an even shorter version, half of the tone-pairs and only 3 PDs (5%, 10%, 50%) can be used, reducing the length further to only 30% of the original (from ~15 to ~5 min.)

3.2. Auditory Emotion Recognition (JL-AER)

3.2.1. Item-level analysis—Figure 5 shows facility and discrimination for all items in the original test and 14 hyper-items for SZ and HC combined. The hyper-items correspond to item characteristics: (i) emotion (6 levels); (ii) speakers (4 levels); (iii) sentence form (2 levels); and (iv) emotion intensity (2 levels). Thirty-four of the original 88 items met criteria for both facility and discrimination -- they are indicated in black on Figure 5 and also are shown in Table 3. The neutral items have the highest facilities, indicating that the lack of emotion is easy to identify. The lowest facilities are for items with "happy" or "disgust" emotion and for items that are of "weak" intensity. It is interesting to notice on Figure 5 (as well as on Figure 1), that the discrimination values of the hyper-items are quite high, even though they consists of items that all have much lower discriminations. This can be explained by the fact that when items in a test are trying to measure the same latent construct, the items are expected to be positively correlated between each other. In such case, larger subsets of the items will tend to measure the latent construct more precisely then smaller subsets of items, and of course, typically more precisely than single items. Therefore, the correlations between a hyper-item (the average of items in a subset of several of the total items in the test) and the average of the remaining items would tend to be higher than the correlation between one item and the average of the remaining items. This correlation on average tends to increase as the number of items in the hyper-item increases from 1 to half of the total number of items.

Facility and discrimination values estimated separately for individual items for SZ and HC (results for individual items not shown and available upon request) showed similar patterns and the expected lower facility across all items for SZ compared to HC. Unexpectedly, the discrimination values for HC were lower than for SZ. This is in contrast to the J-TMT test, which has similar discrimination values in the two diagnostic groups, and suggests that the JL-AER captures a construct that is less well delineated in HC than it is in SZ patients. Figure 6 shows results for the emotion hyper-items for each diagnostic group separately. With the exception of "happy" the facilities were higher for HC than for SZ. Note that even though the facility of the hyper-item "neutral" is high for both SZ and HC, the difference between the facilities of "neutral" for the two diagnostic groups is the largest among all emotion hyper-items. The discrimination values (except for angry, for which the discrimination values are similar) are higher for SZ than HC, indicating that the emotion hyper-items (except angry) differentiate better between SZ low and high ability subjects, than they differentiate between low and high ability HCs. This is in contrast to J-TMT, which has similar discrimination performance for SZ and HC. Also in comparison to J-TMT, the discrimination values of JL-AER are lower.

3.2.2. Effect of item characteristics—From Figure 6 we see that all of the hyper-items meet the criteria for facility and discrimination. In contrast with the individual items, the discrimination values of the hyper-items in the combined sample of HC and SZ were all quite high, indicating that the hyper-items can all differentiate between persons with low vs. high ability on the JL-AER test. At an item level, we tested for the independence between correct answer to each individual item and diagnosis via χ^2 tests for independence. The four panels on Figure 7 show how the χ^2 test-statistics depend on each of the four item characteristics (emotion, speaker, form, and intensity). As mentioned earlier, eight utterances did not have an emotion (i.e., emotion was neutral) and these items were not characterized by intensity of emotion. Note on the left most panel of Figure 7 that the test statistics corresponding to the neutral utterances have the highest values, i.e., the null hypothesis of independence between diagnosis and correct answer is rejected most often. This result is consistent with the observation on the left panel of Figure 6 showing the largest difference between the facilities of HC and SZ for the neutral hyper-item.

In the formal tests for dependence of the ability of items to differentiate between SZ and HC of an item characteristic, none of the interaction terms diagnosis-by-characteristic in the GLMMs were significant (after removing the neutral category from the *emotion* characteristic and the missing intensity category associated with the neutral emotion from the *intensity* characteristic). Clearly, using only neutral items in the AER test is not an option, as different emotions are required. The single item characteristic analyses were repeated using models that controlled for all other item characteristics; the results were qualitatively the same, indicating that no item characteristic is associated with differential response. In addition, no 3-way interaction was statistically significant when testing for joint effect of two characteristics on the ability of an item for differentiate between HC and SZ (all p-value >0.18). Those results suggest that since none of the item characteristics has an effect on the ability of the items to differentiate between SZ and HC, an alternative strategy to shorten the test should be considered.

3.2.3. Shortening the test—Because no item characteristics showed an effect on ability of the items to distinguish between HC vs. SZ, the test was shortened based on the item-level facility and discrimination values. The 34 items that met the criteria for facility and discrimination across groups were chosen for the shortened battery (*short34*). Those criteria produced a well-balanced test that included items from each of the 14 categories. The selected items are shown in Table 3 and include 6 happy, 7 sad, 4 anger, 8 fear, 3 disgust, and 6 neutral items. Twenty of these are statements and 14 are questions. Fourteen are weak and 14 are strong and 6 have no intensity (are neutral). A previously proposed abbreviated 32-item version (*short*32) was developed based on pitch properties of items[2, 16, 23] and contains 23 of the same items as the *short*34 (see Table 3). The full version, *short32* and *short34* have comparable AUCs, with AUC_{Full}=0.7751, $AUC_{short34}$=0.7769 and $AUC_{short32}$=0.7645 (see Figure 8).

3.2.4. Test-Retest Reliability—As seen in Table 2, the test-retest reliabilities of HC and SZ are high for both the original test and *short34.* The test-retest reliability of the *short32* is somewhat lower, but still acceptable. We conclude that both *short34* and *short32* versions can be used in clinical trials. In neither version is the test-retest reliability compromised by the omission of more than half of the original 88 items, thus allowing a reduction to 40% of the original duration (from ~45 min to ~20 min long).

4. DISCUSSION

This paper reports on the psychometric properties of behavioral auditory tests that have been used in schizophrenia research to assess sensory function and prosody. To understand the measured phenomenon and possibly develop shortened versions of the tasks suitable for clinical research and practice, we performed an item-level analysis and also studied the effects of various item characteristics on the ability of the items to differentiate between SZ patients and HC. Our aim is consistent with the goals of such initiatives as CNTRICS and Social Cognition Psychometric Evaluation (SCOPE)[8, 11]. We recommend tasks that can have meaningful applications in clinical trials. In addition, the detailed characterization of properties of these auditory tests can contribute to development of further measures of auditory functioning.

The J-TMT assesses pitch perception at five different PDs between two tones, and although it is relatively short to begin with (~15 min to administer), when administered as part of battery of tests, a shorter version would be desirable. The original test shows high sensitivity and specificity as seen in AUC values, which is consistent with large effect sizes reported previously[2, 22]. The results presented here show that facility (i.e., probability of an item to be answered correctly) is within acceptable limits, particularly for items at the more difficult PDs. However, pairs with “same” tones were generally above the facility cut-off, i.e., they are easy to answer correctly. Discrimination (i.e., correlation between the response to an item and the average response of all other items, a measure of how well the item can differentiate between individuals with low and vs. high ability on the construct assessed by the test) is good for most items, except for the “same” items, indicating that pairs of equal tones do not discriminate well between subjects with good and bad pitch perception. “Same” tone-pairs, however, clearly need to be included in the test. As expected, the average

facilities of the items are lower in patients than in controls. However, the average discrimination values are similar between SZ and HC, indicating that the J-TMT works well in both SZ and HC and would also be appropriate for assessing basic tone perception even among healthy subjects. The greatest differences between SZ and HC were in correctly answering items with PD of 5% and 10%.

An obvious way to shorten the J-TMT is by removing repeated tone-pairs within each PD level (i.e., half of all items). This successfully produced a similar AUC curve as the original test. Because PD showed a differential ability to distinguish groups, PD level was explored to even further shorten the test. Utilizing only half the tone-pairs and three of the five PDs produces a test without compromising its sensitivity and specificity. The original and shortened versions have similar test-retest reliability, all above 0.78 for patients and controls. Thus, all three versions can be used in treatment studies, with the shortest version less than half the time as the original. Several papers report on an adaptive up-down transfer staircase method to determine matching thresholds[33, 41, 42], akin to our PD75 scores, which may also be considered in developing tests to assess pitch perception.

The JL-AER, used to assess emotional prosody, has high AUC for differentiating patients from controls, like the J-TMT. This test takes ~45 min to administer and would particularly benefit from shortening. Thirty-four of the original 88 items showed facility and discrimination that met suggested guidelines. As expected, healthy controls performed on average better on all emotions, with the exception of happy, where both diagnostic groups had similar and low rates of correct answers. However, the discrimination values were generally higher for patients than for controls, which suggests that although JL-AER is a good test for assessing emotional prosody in SZ and can differentiate well between SZ and HC, it might not be optimal for differentiating healthy subjects with respect to low and high levels of auditory emotion recognition.

We note that both patients and controls performed best on utterances with “neutral” emotion, and these items also showed the best ability to differentiate between patients and controls. As reported in other studies, the difference between SZ and HC in the probability to answer correctly did not depend on the emotion[2, 22] (excluding neutral items). The lack of dependence of differentiation between SZ and HC on other item characteristics (i.e., speaker and form of the sentence) has not previously been explored and is first documented here. Low, but not high, intensity has previously been shown to differentiate between SZ and HC for the emotion of angry and this emotion appears to depend more on intensity than pitch cues for recognition [2, 25]. However, as seen in the present study, while weak intensity has lower facility than strong intensity for both groups, weak vs. strong intensity does not have a differential effect between groups when all emotions are considered.

Because none of the item characteristics affected the ability of utterances to discriminate between diagnostic groups, the strategy employed in shortening the J-TMT could not be used. Instead we focused on the subset of items 34 items individually meeting the facility and discrimination criteria (*Short34*). The *Short*34 has good AUC and test-retest reliability>0.7 for both SZ and HC. Interestingly, a 32-item short version (*Short*32), which was previously developed based on items having pitch and intensity characteristics near the

mean levels for specific emotions[2, 16, 23], contains 23 of the same items as the *Short*34. The test-retest reliability of *Short*34 is slightly higher than *Short*32, particularly for HC, but the test-retest reliability of *Short*32 is also acceptable. This provides validation for the previously shortened version in addition to developing a new 34-item version.

The most serious limitation of the present study is that the properties of the proposed short versions were assessed from the same sample used to develop those instruments. Our group has begun collection of data from new participants, which will allow proper independent assessment of their properties. Potential limitations are also the higher prevalence of females in the HC group. The possible effect of subjects' gender on correctly answering the items in the two tests was evaluated based on GLMM for probability for correct answer that included interactions of gender with item characteristics and diagnosis. These analyses provided no evidence for effect of gender on the ability of items to differentiate between patients and controls. Finally, the SZ group includes both inpatients and outpatients. While there is some indication for poorer pitch perception in inpatients, a replication sample of outpatients from the University of Pennsylvania showed similar performance on the J-TMT as the mixed sample of inpatients and outpatients from the present site (Nathan Kline Institute)[2]. Further multi-site studies should be undertaken.

In conclusion, the J-TMT and JL-AER may be useful in clinical trials of social cognitive remediation and perceptual treatments, both of which are receiving great attention[35, 43, 44] due to the serious consequences of these disturbances. Here we present shorter version of these tests that require less than 50% of the time necessary for administering the original tests, without compromising their psychometric properties. Our work further increases those auditory tasks' utility in clinical trials and clinical practice. The large differences between SZ and HC reported here also highlight the profound nature of auditory deficits and need for remediation.

Acknowledgments

Funding: Supported by NIH grant R01 MH084848.

Abbreviations

AER	Auditory Emotion Recognition
CNTRICS	Cognitive Neuroscience Treatment Research to Improve Cognition in Schizophrenia
HC	Healthy Control
JL-AER	AER task by Juslin and Laukka (2001)[1]
J-TMT	TMT task by Gold et al. (2012) [2]
PD	Pitch Differential
SZ	Schizophrenia patients
TMT	Tone Matching Task

References

1. Juslin PN, Laukka P. Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. Emotion. Dec; 2001 1(4):381–412. [PubMed: 12901399]
2. Gold R, Butler P, Revheim N, et al. Auditory emotion recognition impairments in schizophrenia: relationship to acoustic features and cognition. Am J Psychiatry. Apr; 2012 169(4):424–432. [PubMed: 22362394]
3. Green MF, Leitman DI. Social cognition in schizophrenia. Schizophr Bull. Jul; 2008 34(4):670–672. [PubMed: 18495642]
4. Couture SM, Penn DL, Roberts DL. The functional significance of social cognition in schizophrenia: a review. Schizophr Bull. Oct; 2006 32(Suppl 1):S44–63. [PubMed: 16916889]
5. Fett AK, Viechtbauer W, Dominguez MD, Penn DL, van Os J, Krabbendam L. The relationship between neurocognition and social cognition with functional outcomes in schizophrenia: a meta-analysis. Neurosci Biobehav Rev. Jan; 2011 35(3):573–588. [PubMed: 20620163]
6. Mancuso F, Horan WP, Kern RS, Green MF. Social cognition in psychosis: multidimensional structure, clinical correlates, and relationship with functional outcome. Schizophr Res. Feb; 2010 125(2–3):143–151. [PubMed: 21112743]
7. Addington J, Girard TA, Christensen BK, Addington D. Social cognition mediates illness-related and cognitive influences on social function in patients with schizophrenia-spectrum disorders. J Psychiatry Neurosci. Jan; 2010 35(1):49–54. [PubMed: 20040246]
8. Carter CS, Barch DM, Buchanan RW, Bullmore E, Krystal JH, Cohen J, Geyer M, Green M, Nuechterlein KH, Robbins T, Silverstein S, Smith EE, Strauss M, Wykes T, Heinssen R. Identifying cognitive mechanisms targeted for treatment development in schizophrenia: an overview of the first meeting of the Cognitive Neuroscience Treatment Research to Improve Cognition in Schizophrenia Initiative. Biol Psychiatry. Jul 1; 2008 64(1):4–10. [PubMed: 18466880]
9. Carter CS, Barch DM, Gur R, Pinkham A, Ochsner K. CNTRICS Final Task Selection: Social Cognitive and Affective Neuroscience-Based Measures. Schizophr Bull. Nov 14.2008
10. Butler PD, Silverstein SM, Dakin SC. Visual perception and its impairment in schizophrenia. Biol Psychiatry. Jul 1; 2008 64(1):40–47. [PubMed: 18549875]
11. Pinkham AE, Penn DL, Green MF, Buck B, Healey K, Harvey PD. The Social Cognition Psychometric Evaluation Study: Results of the Expert Survey and RAND Panel. Schizophr Bull. May 31.2013
12. Silverstein SM, Keane BP, Barch DM, Carter CS, Gold JM, Kovacs I, MacDonald A 3rd, Ragland JD, Strauss ME. Optimization and validation of a visual integration test for schizophrenia research. Schizophr Bull. Jan; 2011 38(1):125–134. [PubMed: 22021658]
13. Hoekert M, Kahn RS, Pijnenborg M, Aleman A. Impaired recognition and expression of emotional prosody in schizophrenia: review and meta-analysis. Schizophenia Research. 2007; 96:135–145.
14. Edwards J, Pattison PE, Jackson HJ, Wales RJ. Facial affect and affective prosody recognition in first-episode schizophrenia. Schizophr Res. Mar 30; 2001 48(2–3):235–253. [PubMed: 11295377]
15. Kerr SL, Neale JM. Emotion perception in schizophrenia: specific deficit or further evidence of generalized poor performance? J Abnorm Psychol. 1993; 102(2):312–318. [PubMed: 8315144]
16. Kantrowitz JT, Hoptman MJ, Leitman DI, Silipo G, Javitt DC. The 5% difference: early sensory processing predicts sarcasm perception in schizophrenia and schizo-affective disorder. Psychol Med. Apr 24.2013 :1–12.
17. Orbelo DM, Grim MA, Talbott RE, Ross ED. Impaired comprehension of affective prosody in elderly subjects is not predicted by age-related hearing loss or age-related cognitive decline. J Geriatr Psychiatry Neurol. Mar; 2005 18(1):25–32. [PubMed: 15681625]
18. Strous RD, Cowan N, Ritter W, Javitt DC. Auditory sensory (“echoic”) memory dysfunction in schizophrenia. Am J Psychiatry. Oct; 1995 152(10):1517–1519. [PubMed: 7573594]
19. Javitt DC, Strous RD, Grochowski S, Ritter W, Cowan N. Impaired precision, but normal retention, of auditory sensory (“echoic”) memory information in schizophrenia. J Abnorm Psychol. May; 1997 106(2):315–324. [PubMed: 9131851]

20. Holcomb HH, Ritzl EK, Medoff DR, Nevitt J, Gordon B, Tamminga CA. Tone discrimination performance in schizophrenic patients and normal volunteers: impact of stimulus presentation levels and frequency differences. Psychiatry Research. 1995; 57:75–82. [PubMed: 7568562]
21. Wexler BE, Stevens AA, Bowers AA, Sernyak MJ, Goldman-Rakic PS. Word and tone working memory deficits in schizophrenia. Arch Gen Psychiatry. Dec; 1998 55(12):1093–1096. [PubMed: 9862552]
22. Jahshan C, Wynn JK, Green MF. Relationship between auditory processing and affective prosody in schizophrenia. Schizophr Res. Feb; 2013 143(2–3):348–353. [PubMed: 23276478]
23. Kantrowitz JT, Leitman DI, Lehrfeld JM, Laukka P, Juslin PN, Butler PD, Silipo G, Javitt DC. Reduction in tonal discriminations predicts receptive emotion processing deficits in schizophrenia and schizoaffective disorder. Schizophr Bull. Jan; 2013 39(1):86–93. [PubMed: 21725063]
24. Leitman DI, Foxe JJ, Butler PD, Saperstein A, Revheim N, Javitt DC. Sensory contributions to impaired prosodic processing in schizophrenia. Biol Psychiatry. Jul 1; 2005 58(1):56–61. [PubMed: 15992523]
25. Leitman DI, Laukka P, Juslin PN, Saccente E, Butler P, Javitt DC. Getting the cue: sensory contributions to auditory emotion recognition impairments in schizophrenia. Schizophr Bull. May; 2010 36(3):545–556. [PubMed: 18791077]
26. Javitt DC. When doors of perception close: bottom-up models of disrupted cognition in schizophrenia. Annu Rev Clin Psychol. 2009; 5:249–275. [PubMed: 19327031]
27. Leitman DI, Hoptman MJ, Foxe JJ, et al. The neural substrates of impaired prosodic detection in schizophrenia and its sensorial antecedents. Am J Psychiatry. 2007; 164(3):474–482. [PubMed: 17329473]
28. Leitman DI, Wolf DH, Laukka P, Ragland JD, Valdez JN, Turetsky BI, Gur RE, Gur RC. Not pitch perfect: sensory contributions to affective communication impairment in schizophrenia. Biol Psychiatry. Oct 1; 2011 70(7):611–618. [PubMed: 21762876]
29. Leitman DI, Wolf DH, Ragland JD, Laukka P, Loughead J, Valdez JN, Javitt DC, Turetsky BI, Gur RC. “It’s Not What You Say, But How You Say it”: A Reciprocal Temporo-frontal Network for Affective Prosody. Front Hum Neurosci. 2010; 4:19. [PubMed: 20204074]
30. Shaw RJ, Dong M, Lim KO, Faustman WO, Pouget ER, Alpert M. The relationship between affect expression and affect recognition in schizophrenia. Schizophr Res. Jun 22; 1999 37(3):245–250. [PubMed: 10403196]
31. Kucharska-Pietura K, David AS, Masiak M, Phillips ML. Perception of facial and vocal affect by people with schizophrenia in early and late stages of illness. Br J Psychiatry. Dec.2005 187:523–528. [PubMed: 16319404]
32. Bozikas VP, Kosmidis MH, Anezoulaki D, Giannakou M, Andreou C, Karavatos A. Impaired perception of affective prosody in schizophrenia. J Neuropsychiatry Clin Neurosci. Winter;2006 18(1):81–85. [PubMed: 16525074]
33. Leitman DI, Ziwich R, Pasternak R, Javitt DC. Theory of Mind (ToM) and counterfactuality deficits in schizophrenia: misperception or misinterpretation? Psychol Med. Aug; 2006 36(8): 1075–1083. [PubMed: 16700967]
34. Fisher M, Holland C, Subramaniam K, Vinogradov S. Neuroplasticity-based cognitive training in schizophrenia: an interim report on the effects 6 months later. Schizophr Bull. Jul; 36(4):869–879. [PubMed: 19269924]
35. Hooker CI, Bruce L, Fisher M, Verosky SC, Miyakawa A, D’Esposito M, Vinogradov S. The influence of combined cognitive plus social-cognitive training on amygdala response during face emotion recognition in schizophrenia. Psychiatry Res. Aug 30; 2013 213(2):99–107. [PubMed: 23746615]
36. First, MB.; Spitzer, RL.; Gibbon, M.; Williams, JBW. Structured Clinical Interview for DSM-IV Axis I Disorders- Patient Edition. New York: New York State Psychiatric Institute; 1997.
37. Gregory, RJ. Psychological Testing: History, Principles, and Applications. 6. Boston: Allyn & Bacon; 2011.
38. McCullogh, CE.; Searle, SR.; Neuhaus, JM. Generalized, Linear and Mixed Models. New Jersey: John Wiley & Sons; 2008.

39. Zhou, X-H.; Obuchowski, NA.; McClish, DK. Statistical Methods in Diagnostic Medicine. New York, NY: Wiley and Sons; 2002.

40. RDCT. R: A language and environment for statistical computing. R Foundation for Statistical Computing;

41. Rabinowicz EF, Silipo G, Goldman R, Javitt DC. Auditory sensory dysfunction in schizophrenia: imprecision or distractibility? Arch Gen Psychiatry. 2000; 57(12):1149–1155. [PubMed: 11115328]

42. Leitman DI, Sehatpour P, Higgins BA, Foxe JJ, Silipo G, Javitt DC. Sensory deficits and distributed hierarchical dysfunction in schizophrenia. Am J Psychiatry. Jul; 2010 167(7):818–827. [PubMed: 20478875]

43. Horan WP, Kern RS, Tripp C, Hellemann G, Wynn JK, Bell M, Marder SR, Green MF. Efficacy and specificity of social cognitive skills training for outpatients with psychotic disorders. J Psychiatr Res. Aug; 45(8):1113–1122. [PubMed: 21377168]

44. Lindenmayer JP, McGurk SR, Khan A, et al. Improving social cognition in schizophrenia: a pilot intervention combining computerized social cognition training with cognitive remediation. Schizophr Bull. May; 2012 39(3):507–517. [PubMed: 23125396]

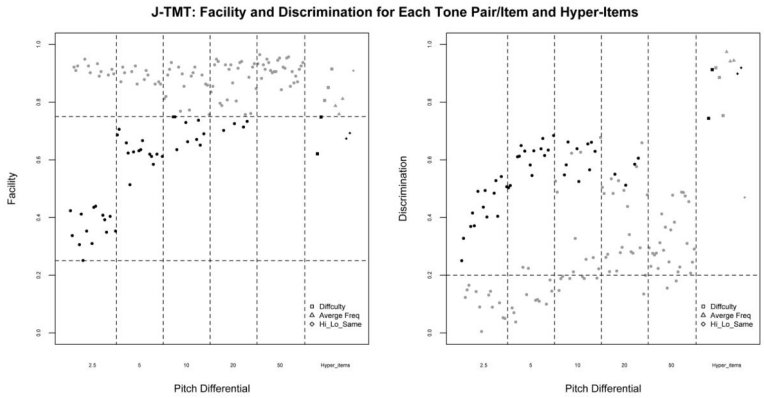

Figure 1. J-TMT: Facility and Discrimination for Each Tone Pair/Item and Hyper-Items

Facility and discrimination of tone-pairs in the original J-TMT for patients and controls combined. The tone-pairs are ordered horizontally from the first pair within the smallest PD to the last pair within the largest PD. The “hyper-items” at the far right are in the order: (i) PD (5 squares left to right: 2.5 to 50%); (ii) average frequency (3 triangles left to right: low, medium, high); (iii) order of tones within a pair (3 diamonds left to right: higher first, lower first, same). The horizontal lines mark the cut-off guidelines from CTT. Black symbols indicate items that meet guidelines for both facility and discrimination.

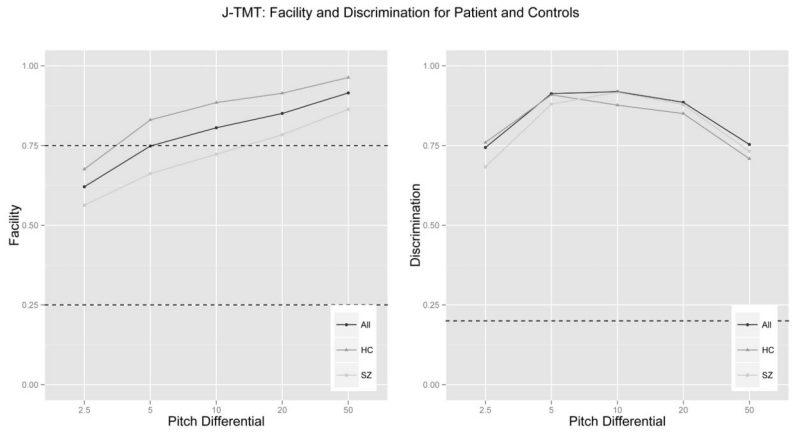

Figure 2.

J-TMT: Facility and Discrimination for Patient and Controls

J-TMT: Facility and discrimination of the hyper-items corresponding to 5 PDs for patients, controls, and both groups together.

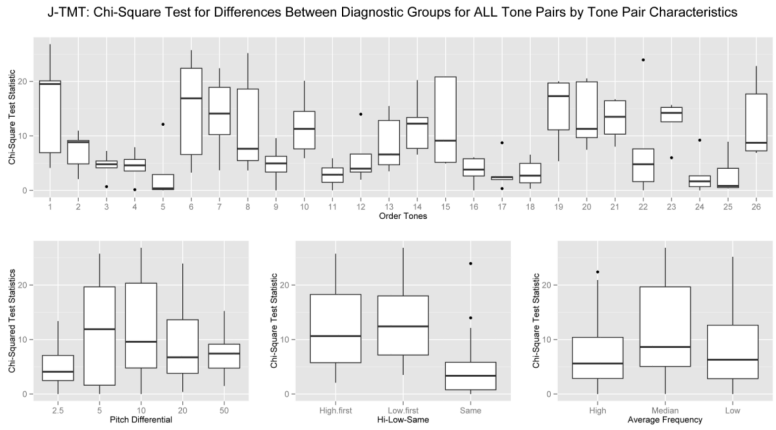

Figure 3.

J-TMT: Chi-Square Test for Differences Between Diagnostic Groups for ALL Tone Pairs by Tone Pair Characteristics

J-TMT: Box-plots of the 130 χ^2 test statistics testing for independence between correct response and diagnosis by item characteristics: by PD (bottom left panel), by order of tones within a pair (bottom middle panel), by average pitch of the two tones in a pair (bottom right panel) and by sequence number of a tone-pair within each PD (top panel). Each box represents the interquartile range, the horizontal line within the box represents the median, and the points beyond the whiskers represent the outliers for the set of $\chi^2(1)$ test statistics for all items at the given level of the characteristic. Note, since identical tones are necessary for the J-TMT, only the tone-pairs from categories ‘lower-first’ and ‘higher-first’ were included in the test assessing the effect of the characteristic on the ability of items to differentiate between SZ and HC. Note, that the critical α=0.05 value for a test for independence is $\chi^2(1)$=3.84.

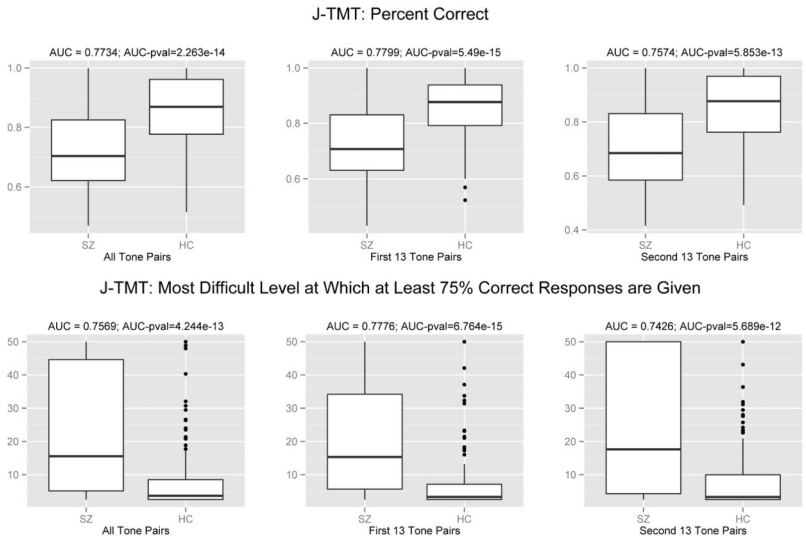

Figure 4.

J-TMT: How well the summary scores differentiate between SZ and HC. Percent correct (top panel) and PD75 (bottom panel) are computed using all items, the first 13 and the second 13 items and are shown by diagnostic group. The AUC corresponding to these summary scores are given together with a p-value from t-tests comparing the diagnostic groups.

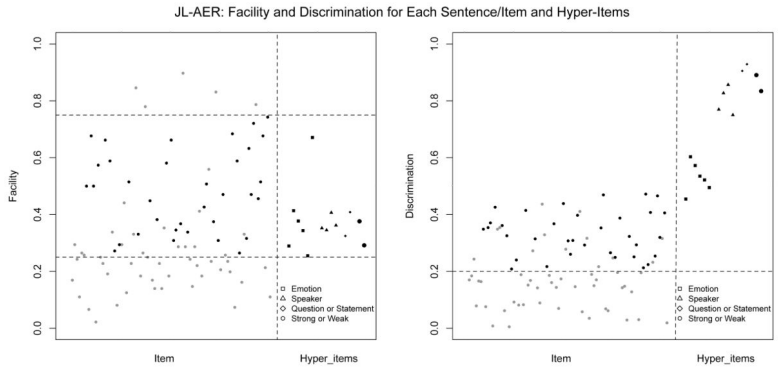

Figure 5.

JL-AER: Facility and Discrimination for Each Sentence/Item and Hyper-Items

Facility and discrimination of the utterances in the original JL-AER test for patients and controls combined. The “hyper-items” at far right are in the order: (i) emotions (6 triangles left to right: happy, sad, angry, fear, disgust, neutral); (ii) speaker (4 squares left to right); (iii) question or statement (2 diamonds left to right); (iv) strong or weak (2 circles left to right). Black symbols indicate items that meet guidelines for facility and discrimination.

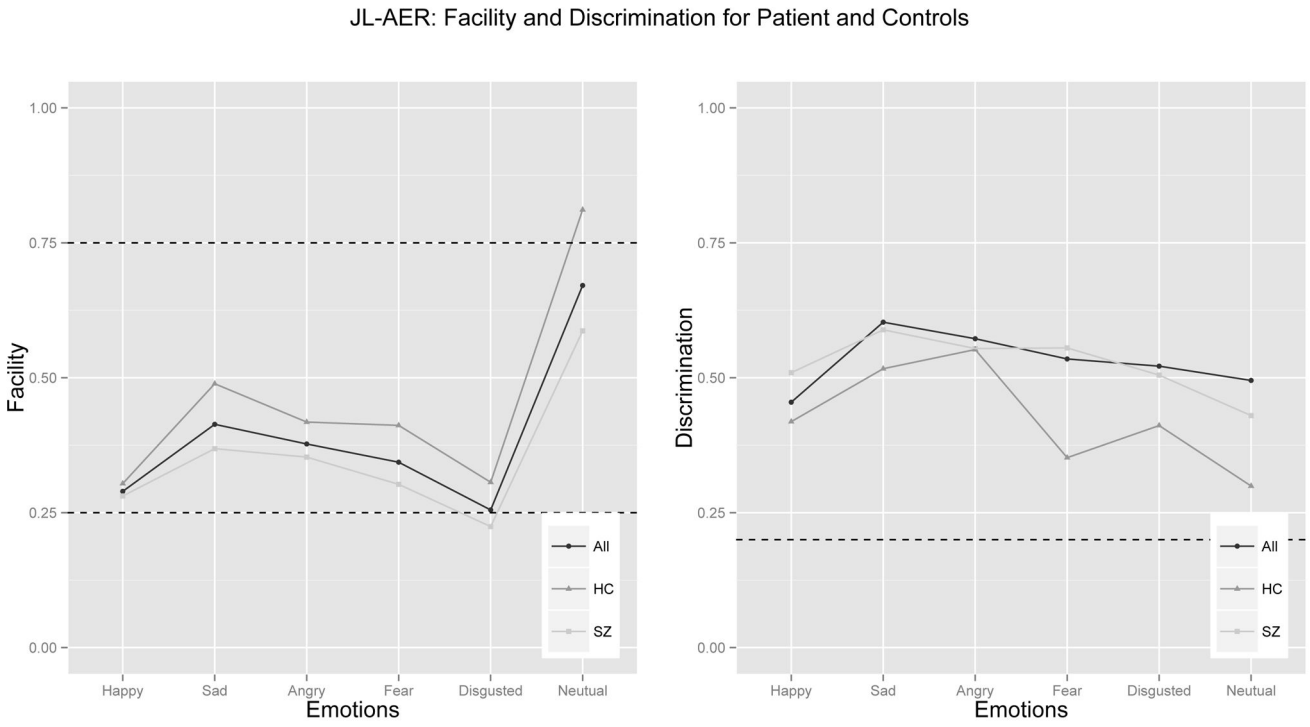

Figure 6.

JL-AER: Facility and Discrimination for Patient and Controls

JL-AER: Facility and discrimination of hyper-items corresponding to the six levels of the emotion characteristic for patients, controls, and both groups together.

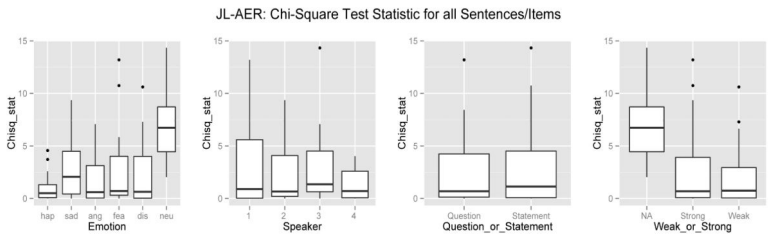

Figure 7.

JL-AER: Chi-Square Test Statistic for all Sentences/Items

JL-AER: Box-plots of the $\chi^2(1)$ test statistics comparing SZ vs. HC with respect to response to individual items by: emotion, speaker, question/statement, and intensity.

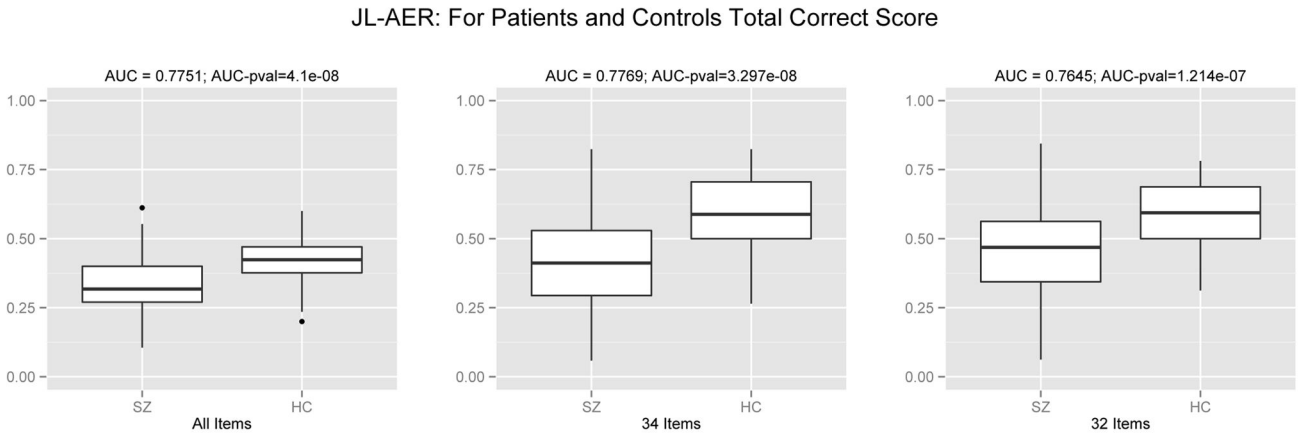

Figure 8.

JL-AER: For Patients and Controls Total Correct Score

JL-AER: How well the summary scores differentiate between SZ and HC. Percent correct for the original, *Short*34 and *Short*32.

Table 1

The table provides clinical and demographic characteristics of the sample used for each auditory test.

	J-TMT		JL-AER	
	SZ, n=124	**HC, n=131**	**SZ, n=85**	**HC, n=51**
Demographics				
Age, mean ± SD	38.6±10.7	34.6±11.5*	39.9±10.3	37.5±10.8
Gender (M/F)	105/19	86/45**	70/15	36/15
SES parents ± SD	38.6±14.7	43.9±12.9*	37.1±13.9	42.3±14.0
SES participants ± SD	26.4±11.1	41.9±11.7**	25.9±11.0	43.1±10.0**
IQ mean ± SD	95.4±10.1	106.2±11.1**	95.3±10.3	105.4±10.4**
Highest grade achieved ± SD	11.9±2.3	14.8±2.2**	11.9±2.3	14.7±1.9**
SZ Clinical Characteristics				
CPZ Equivalents ± SD	875.5±708.6		851.3±694.7	
Atypical/typical/combination/none	86/12/24/2		56/9/18/2	
PANSS Total ± SD	74.3±13.2		74.7±14.0	
PANSS Positive ± SD	18.8±5.6		19.1±5.9	
PANSS Negative ± SD	18.8±4.6		18.5±4.4	
PANSS General Psychopathology ± SD	36.8±7.3		37.1±7.7	
Illness duration (years) ± SD	16.4±9.4		17.1±9.4	
Schizophrenia/Schizoaffective	99/25		66/19	

* $p<0.01$;

** $p<0.001$

Note: Socioeconomic status was measured with the 4-factor Hollingshead Scale. IQ was assessed with the Quick Test. M, male; F, female; PANSS, Positive and Negative Syndrome Scale.

Table 2

Test-retest reliability of the original summary measures and proposed shorter versions, by diagnosis.

Measure	SZ	HC
J-TMT SZ: n@T1[1]=124, n@T2[2]=52; HC: n@T1=131, n@T2=32		
% Correct: 5 PD[3], 26 items per PD (Original)	0.830	0.821
% Correct: 5 PDs, 13 items per PD	0.786	0.806
% Correct: 3 PDs (5, 10, 50), 13 items per PD	0.785	0.833
PD75[4], 26 items	0.744	0.701
PD75, 13 items	0.668	0.628
JL-AER SZ: n@T1=85, n@T2=35; HC: n@T1=51, n@T2=18		
% Correct, 86 items (Original)	0.804	0.716
% Correct, 32 items (*Short*32)	0.681	0.716
% Correct, 34 items (*Short*34)	0.706	0.855

[1] n@T1: number of subjects with observations at Time 1

[2] n@T2: number of subjects with observations at Time 2

[3] PD = Pitch Differential

[4] PD75 = the most difficult level at which subject had at least 75% correct responses

Compr Psychiatry. Author manuscript; available in PMC 2015 December 25.

Table 3

The subset of JL-AER items selected for the proposed 34 items short version (*Short34*). In bold are the items that belong also to the existing abbreviated version of 32 items (*Short32*).

Items	Emotion	Form	Weak/Strong	Speaker	Facility	Discriminat	$\chi^2(1)$	p-value
emo07	Disgust	statement	Strong	C	0.500	0.349	4.518	0.034
emo09	**Neutral**	**question**		**C**	**0.676**	**0.354**	**5.160**	**0.023**
emo10	**Sad**	**statement**	**Weak**	**B**	**0.500**	**0.371**	**4.518**	**0.034**
emo12	**Sad**	**statement**	**Strong**	**C**	**0.574**	**0.426**	**1.355**	**0.244**
emo15	**Happy**	**statement**	**Weak**	**C**	**0.662**	**0.361**	**0.009**	**0.925**
emo17	**Neutral**	**statement**		**D**	**0.588**	**0.326**	**3.918**	**0.048**
emo19	**Fear**	**statement**	**Weak**	**B**	**0.272**	**0.208**	**0.418**	**0.518**
emo21	Sad	statement	Weak	A	0.294	0.240	3.060	0.080
emo25	**Fear**	**question**	**Strong**	**A**	**0.515**	**0.415**	**13.196**	**0.000**
emo29	Fear	question	Strong	C	0.331	0.315	4.484	0.034
emo34	**Anger**	**statement**	**Weak**	**D**	**0.449**	**0.217**	**0.050**	**0.824**
emo37	**Anger**	**question**	**Weak**	**D**	**0.382**	**0.368**	**1.196**	**0.274**
emo41	**Happy**	**question**	**Weak**	**C**	**0.581**	**0.439**	**1.065**	**0.302**
emo43	**Neutral**	**question**		**B**	**0.662**	**0.308**	**4.634**	**0.031**
emo44	**Happy**	**question**	**Strong**	**B**	**0.309**	**0.260**	**2.067**	**0.151**
emo45	**Happy**	**statement**	**Strong**	**C**	**0.346**	**0.309**	**0.626**	**0.429**
emo47	Sad	statement	Strong	D	0.368	0.398	0.076	0.783
emo50	**Happy**	**statement**	**Weak**	**B**	**0.338**	**0.293**	**0.710**	**0.400**
emo57	Fear	question	Weak	A	0.426	0.353	5.844	0.016
emo58	Disgust	statement	Weak	B	0.507	0.469	7.298	0.007
emo61	**Anger**	**statement**	**Strong**	**C**	**0.375**	**0.266**	**0.755**	**0.385**
emo63	Sad	question	Weak	D	0.309	0.249	2.067	0.151
emo65	**Fear**	**statement**	**Weak**	**A**	**0.471**	**0.388**	**0.787**	**0.375**
emo69	Neutral	question		A	0.684	0.323	8.436	0.004
emo71	**Fear**	**question**	**Strong**	**D**	**0.588**	**0.251**	**2.623**	**0.105**
emo72	Disgust	statement	Weak	C	0.265	0.293	0.645	0.422

Items	Emotion	Form	Weak/Strong	Speaker	Facility	Discriminat	$\chi^2(1)$	p-value
emo75	Fear	question	Strong	B	0.316	0.213	0.274	0.600
emo76	**Happy**	**question**	**Strong**	**C**	**0.632**	**0.472**	**3.719**	**0.054**
emo77	**Anger**	**question**	**Strong**	**C**	**0.471**	**0.224**	**7.083**	**0.008**
emo78	**Sad**	**statement**	**Strong**	**B**	**0.721**	**0.408**	**9.359**	**0.002**
emo80	**Sad**	**statement**	**Weak**	**C**	**0.456**	**0.254**	**6.648**	**0.010**
emo81	**Fear**	**statement**	**Strong**	**A**	**0.515**	**0.466**	**10.747**	**0.001**
emo82	**Neutral**	**statement**		**C**	**0.676**	**0.320**	**14.335**	**0.000**
emo84	Neutral	statement		A	0.743	0.406	9.544	0.002