Failure to Apply Signal Detection Theory to the Montreal Battery of Evaluation of Amusia May Misdiagnose Amusia

Molly J. Henry
Michigan State University & Bowling Green State University

J. Devin McAuley
Michigan State University

This article considers a signal detection theory (SDT) approach to evaluation of performance on the Montreal Battery of Evaluation of Amusia (MBEA). One hundred fifty-five individuals completed the original binary response version of the MBEA ($n = 62$) or a confidence rating version (MBEA-C; $n = 93$). Confidence ratings afforded construction of empirical receiver operator characteristic (ROC) curves and derivation of bias-free performance measures against which we compared the standard performance metric, proportion correct (PC), and an alternative signal detection metric, d'. Across the board, PC was tainted by response bias and underestimated performance as indexed by A_z, a nonparametric ROC-based performance measure. Signal detection analyses further revealed that some individuals performing worse than the standard PC-based cutoff for amusia diagnosis showed large response biases. Given that PC is contaminated by response bias, this suggests the possibility that categorizing individuals as having amusia or not, using a PC-based cutoff, may inadvertently misclassify some individuals with normal perceptual sensitivity as amusic simply because they have large response biases. In line with this possibility, a comparison of amusia classification using d'- and PC-based cutoffs showed potential misclassification of 33% of the examined cases.

Received: April 29, 2011, accepted September 30, 2012.

Key words: Montreal Battery of Evaluation of Amusia (MBEA), signal detection theory (SDT), amusia, ROC curves, music perception

Congenital amusia, or *tone-deafness*, is a lifelong impairment in musical ability that has been primarily linked to a pitch processing deficit (Hyde & Peretz, 2004) that is unrelated to normal hearing acuity, general neurological functioning, or exposure to music (Ayotte, Peretz, & Hyde, 2002). The assessment tool that has been most widely used over the past decade to diagnose congenital and acquired forms of amusia (Ayotte et al., 2002; Cuddy, Balkwill, Peretz, & Holden, 2005; Hyde & Peretz, 2004; Peretz et al., 2008) is the Montreal Battery of Evaluation of Amusia (MBEA; Peretz, Champod, & Hyde, 2003); for an online version of the test, see http://www.brams.umontreal.ca/amusia-general/. The MBEA consists of six subtests, which assess melodic organization (Scale, Contour, and Interval), temporal organization (Rhythm and Meter), and musical memory (Memory). For four of the subtests (Scale, Contour, Interval, and Rhythm), listeners are presented with pairs of melodies and asked to judge whether the two melodies are the *same* or *different*, while for the remaining two subtests (Meter and Memory), listeners are presented with a single melody on each trial. For the Meter subtest, listeners judge whether the presented melody is a *march* or *waltz*, while for the Memory subtest, listeners judge whether the melody is one that they've heard before on the previous subtests (an *old* melody) or is previously unheard (a *new* melody).

The tests of melodic organization differ in the type of melodic change that is introduced on *different* trials. For the Scale subtest, the *different* melody contains one note that violates the key of the first melody, while keeping the overall melodic contour intact. For the Contour subtest, *different* melodies contain one note that violates the contour of the first melody in each pair without disrupting the key. For the Interval subtest, the altered note changes the pitch interval while preserving the melodic contour and key. For the Rhythm subtest, rather than a melodic change, *different* melodies are created by altering the onset time of one note so that the preceding and following inter-note onset intervals are changed. For all same-different subtests, half of the melody pairs are the *same* and half are *different*. For the Meter subtest, half of the melodies are *marches* (identified by a repeating strong-weak subjective accent pattern) and half are *waltzes* (identified by a repeating strong-weak-weak subjective accent pattern). For the final Memory subtest, half of the melodies are *old* and half are *new*.

Music Perception, volume 30, issue 5, pp. 480–496, issn 0730-7829, electronic issn 1533-8312. DOI: 10.1525/mp.2013.30.5.480

To assess performance on the MBEA, proportion correct (PC) on each of the subtests is typically averaged to produce a composite PC score; an equivalent method is simply to sum the number of correct responses on each of the subtests (as in, e.g., McDonald & Stewart, 2008; Tillmann, Schulze, & Foxton, 2009; Williamson, McDonald, Deutsch, Griffiths, & Stewart, 2010). The diagnostic criteria for amusia using the MBEA vary (Ayotte et al., 2002; Douglas & Bilkey, 2007; Loui, Alsop, & Schlaug, 2009), but one commonly used method is to determine whether an individual's composite PC or raw score falls more than two standard deviations (SDs) below the mean score of a normative sample (Peretz et al., 2003).

The use of PC (or equivalently a raw score) as a performance metric for the MBEA has been shown by Peretz and colleagues (2003) to demonstrate a number of favorable psychometric properties including approximate normality, test-rest reliability, and convergent validity. However, PC may not be the best measure to assess MBEA performance because PC, by itself, does not permit a distinction between listeners' ability to hear differences between the melodies (i.e., perceptual sensitivity) and any general tendency to make one response or the other (i.e., response bias). Moreover, PC decreases with increasing response bias, independent from sensitivity. Thus, as we will show below, using a PC-based criterion on the MBEA to categorize an individual as amusic or not (typical of many previous studies) has the potential to misclassify non-amusic individuals as amusic simply because they have a large response bias.

To address this issue, the present article takes a signal detection theory (SDT) approach to measuring MBEA performance. To permit a comprehensive SDT analysis, we asked listeners to provide confidence ratings rather than binary judgments. We will refer to the confidence-rating version of the test as the MBEA-C to distinguish it from the MBEA. The potential benefit of using confidence ratings and associated signal detection measures to evaluate MBEA performance is that it affords derivation of nonparametric sensitivity measures against which we can compare PC and d' (an alternative signal detection performance metric) to assess any potential biases in these measures. The remainder of the introduction provides a short tutorial on SDT and its application to the MBEA and MBEA-C, followed by an overview of the present study. MATLAB code for calculation of the SDT measures of MBEA performance is included in the Appendix, and can be downloaded from http://psychology.msu.edu/TAPlab/publications.htm.

A Signal Detection Theory (SDT) Perspective on the MBEA

SDT (Green & Swets, 1966) is a principled psychophysical approach to measuring performance that has been applied to a wide range of experimental tasks and domains (Durlach & Braida, 1969; Ratcliff, McKoon, & Tindall, 1994; Rousseau, Rogeaux, & O'Mahoney, 1999; Snodgrass & Corwin, 1988; Yonelinas, 1999). In general, the SDT approach assumes a decision model that provides a distinction between an individual's ability to discriminate between stimulus classes (i.e., sensitivity) and their tendency to make one response or the other (i.e., response bias). For the subtests of the MBEA, sensitivity refers to listeners' ability to discriminate between *same* versus *different* melodies, *march* versus *waltz* meters, and *new* versus *old* melodies. High sensitivity refers to good ability to discriminate between stimulus classes and low sensitivity refers to poor ability to discriminate between stimulus classes. Bias, in contrast, refers to listeners' general tendency to make one response or the other (e.g., a general tendency to respond "same" or "different").

Sensitivity and response bias make independent contributions to MBEA performance, but they cannot be separated when PC is used as the dependent variable. Taking as an example the same-different subtests, it is critical to understand that poor performance (as indexed by PC) can be caused by responding "same" to *different* trials, responding "different" to *same* trials, or a combination of the two. Making errors of both types equally often decreases sensitivity; that is, the participant is less able to discriminate between the two categories of melodies. On the other hand, responding "same" to *different* trials and not vice versa indicates the presence of a response bias. SDT allows separation of these two contributions to performance.

Within SDT, the specific decision model varies with different task characteristics (MacMillan & Creelman, 2005), but for all of the subtests of the MBEA the decision model can be conceptualized as follows. The brain's response to a stimulus is assumed to be imperfect (i.e., noisy); as a result, stimuli comprising two classes form two normal distributions that vary along an information dimension that is used to make a decision about which stimulus class was presented (see Figure 1). For the subtests of the MBEA, there are two stimulus classes associated with *same* versus *different*,[1] *march*

[1] Although the Scale, Contour, Interval, and Rhythm subtests require a same versus different response, the task does not constitute a same-different task in strict signal detection terms. The reason is that the intact

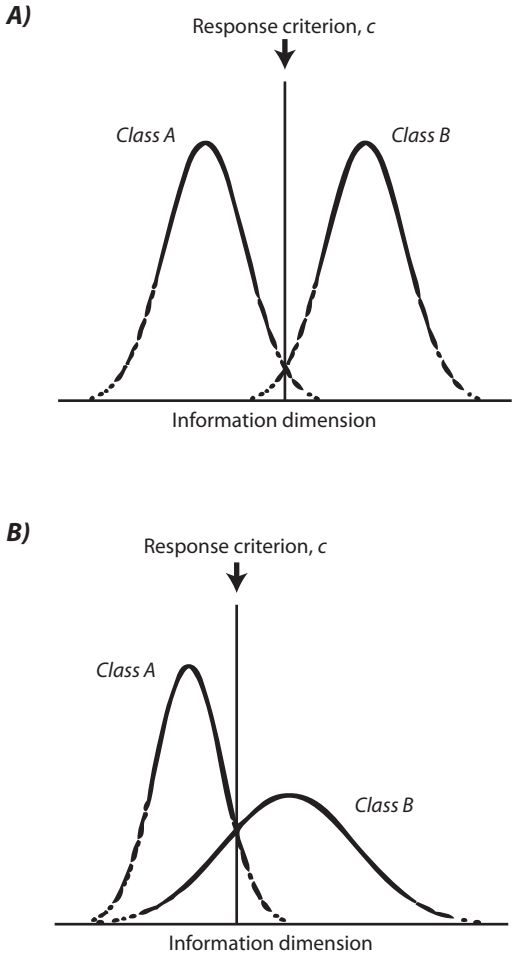

FIGURE 1. A schematic of the SDT model. Stimuli from each of two stimulus classes (Class A and Class B) are represented internally along an information dimension; the location of a response criterion, *c*, along the information dimension determines the decision about which stimulus class was presented. In the figure, stimuli that fall to the left of the criterion are associated with a Class A response and stimuli falling to the right of the criterion are associated with a Class B response. Since the distributions overlap, some Class A stimuli will be associated with a Class B response, and vice versa. In Panel A, the distributions underlying stimulus Classes A and B satisfy the equal-variance assumption of SDT. In Panel B, however, the equal-variance assumption is violated; the Class B variance is larger than the Class A variance.

version of each melody is only presented as the first melody in the pair. A true same-different task requires that both stimulus classes (i.e., same, different) must be presented with equal likelihood in both positions; that is, as the first and second melody in each pair. The current task is more precisely a 'reminder' task, because instances from only one stimulus distribution (i.e., same) are presented in the first position of each melody pair, followed with equal likelihood by a stimulus from either distribution (i.e., same, different).

versus *waltz*, and *new* versus *old*. When a stimulus from a particular class is presented, individuals are assumed to compare the sample to the location of a decision criterion in order to decide whether the stimulus is from Class A (e.g., *same*, *waltz*, or *new*) or Class B (e.g., *different*, *march*, or *old*). If the value sampled from the respective distribution is above (i.e., to the right of) the decision criterion, the response is Class B (e.g., *different*, *march*, *old*), whereas if the value is below (i.e., to the left of) the criterion, the response is Class A (e.g., *same*, *waltz*, *new*). When a participant correctly responds Class B to a Class B stimulus, this is considered a hit, and when a participant correctly responds Class A to a Class A stimulus, this is considered a correct rejection.

Note that because the two distributions overlap, participants will sometimes make errors. Some of the time, the sampled value for a stimulus from Class A will fall above the criterion, and the participant will incorrectly respond Class B—a false alarm. In hypothesis-testing terms, this is equivalent to making a Type I error. Conversely, some of the time, the sampled value for a stimulus from Class B will fall below the criterion, and the participant will incorrectly respond Class A—a miss. This is equivalent to making a Type II error. For the same-different subtests, for example, a false alarm would amount to responding "different" when the melodies are the same, and a miss would amount to responding "same" when the melodies are different.

When the criterion moves to the left, both the hit rate (HR) and the false alarm rate (FAR) increase and in the limit approach 1.0, whereas when the criterion moves to the right, both the hit rate and the false alarm rate decrease and in the limit approach 0. Movement of the criterion corresponds to changes in response bias. For the same-different subtests, movement of the criterion to the left corresponds to a bias to say "different" (a liberal response bias) and movement of the criterion to the right corresponds to a bias to say "same" (a conservative response bias). There is no response bias when the criterion is exactly halfway between the two distributions (as shown in Figure 1A); that is, the participant is equally likely to respond "same" or "different."

The distance between the means of the two distributions provides a measure of an observer's ability to discriminate the two stimulus classes—independent of the placement of the criterion. As the two distributions move closer together, it is more difficult to tell the difference between the two stimulus classes and sensitivity is lower, whereas when they are farther apart, it is easier to tell the difference between the two stimulus classes and sensitivity is higher. Critically, criterion location is independent from the distance between the distributions.

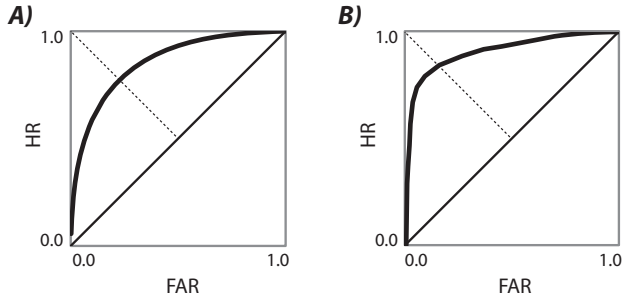

FIGURE 2. Normal-model ROC curves. Hit rate (HR) is plotted as a function of false alarm rate (FAR). Panel A shows a theoretical ROC curve for a case when the equal-variance assumption is satisfied. The ROC curve is symmetrical around the minor diagonal, and d' is the same for every (FAR, HR) coordinate pair. Panel B shows a theoretical ROC curve for a case when the equal-variance assumption has been violated. The ROC curve is asymmetrical around the minor diagonal; thus every (FAR, HR) pair corresponds not only to a different value of c but also to a different d' value. Here, d' no longer constitutes an unbiased measure of performance.

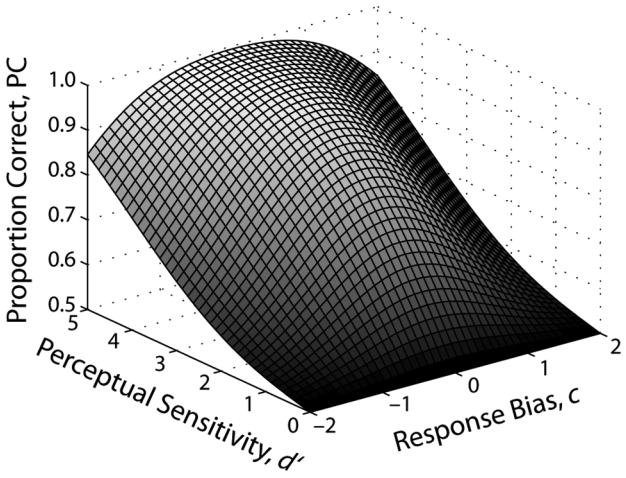

FIGURE 3. Three-dimensional plot of PC as a function of sensitivity (d') and the response criterion, c. Except for the bias-free case ($c = 0$), PC underestimates performance, and the degree of underestimation increases as the value of c becomes either more positive or more negative.

An important feature of the basic signal detection model is that the information axis is typically represented in standardized (z-score) units. This allows for both the placement of the criterion and the separation of the two distributions to be measured in units of standard deviation, which are comparable across tasks and conditions.

Perceptual sensitivity (d') and criterion location (c) can be calculated for each participant and subtest using the proportion of hits (HR) and the proportion of false alarms (FAR). Sensitivity, d', is determined by z(HR) – z(FAR), and the criterion location, c, is determined by -0.5^{*}[z(HR) + z(FAR)]. Since both d' and c are measured in standard deviation units (i.e., z-scores), an important assumption of the basic SDT approach is that the two distributions have equal variance (as in Figure 1A). Otherwise (see Figure 1B), estimates of d' and c would depend on whether they were calculated with respect to the standard deviation of Stimulus Class A or the standard deviation of Stimulus Class B. Critically, d' and c only represent independent performance measures if the equal-variance assumption is satisfied. When this assumption is violated, d' is no longer independent of response bias, and instead varies with c.

The independence of sensitivity and bias under the equal-variance assumption can be understood by representing HR and FAR on what is referred to as an implied receiver operator characteristic (ROC; or relative operating characteristic) curve. Examples are shown in Figure 2. On this graph, FAR is plotted on the x-axis and HR is plotted on the y-axis. If the equal-variance assumption holds, then the ROC curve (shown in Figure 2A) is symmetrical around the minor diagonal, and thus traces an iso-sensitivity curve, meaning that all points on the curve have the same d' value, but different criterion values, c.

Notably, when sensitivity, d', is held constant and the criterion, c, is allowed to vary, then PC also varies in a systematic manner. Figure 3 shows a three-dimensional plot of PC as a function of d' and c. PC can be calculated from d' and c according to the following formula:

$$PC = \frac{\Phi\left(\frac{d'}{2} - c\right) + \left[1 - \Phi\left(\frac{-d'}{2} - c\right)\right]}{2} \qquad (1)$$

where Φ returns the corresponding value of the cumulative normal distribution function. When $c = 0$, PC is maximal. With increasingly positive or negative values of c, corresponding to an increasingly conservative or liberal response criterion, values of PC decrease. This means that for any non-zero response criterion, c, PC underestimates the perceptual sensitivity of the participant. Moreover, the degree of bias is not a linear function of d'. Instead, the degree of bias due to shifts of the response criterion increases up to a maximum at $d' = 2.75$ and decreases thereafter. Notably, the maximum amount of bias estimated here is 28% in terms of PC. For a d' of 2.75 – quite high sensitivity – an individual with zero response bias would achieve PC = .92. However, an individual with *equal sensitivity* (the same value of d'), but a response criterion, $c = \pm 2.0$, would

achieve only PC = .63. Thus, if a participant responds using a relatively extreme response criterion, *c*, then PC becomes a misleading measure of performance. Moreover, in the context of the MBEA, if a PC-based cut-off is used to determine whether or not an individual is amusic, then an individual with a relatively extreme response criterion and normal perceptual sensitivity has the potential to be misclassified.

From a SDT perspective, the problem of amusia classification is further compounded when the equal-variance assumption does not hold. In this case, the implied ROC curve is asymmetrical around the minor diagonal (Figure 2B) and every (FAR, HR) point on the curve no longer corresponds to the same d' value. This means that d' would also be a biased measure of performance. In this case, nonparametric measures of sensitivity must be derived from empirical ROC curves;[2] we describe this procedure below (see Appendix for MATLAB code that can be used to calculate these measures). One way to test the equal-variance assumption is to generate an empirical ROC curve, which involves explicitly manipulating criterion location by having participants make confidence ratings. Requiring a confidence rating response effectively moves the criterion along the information dimension from left (when the listener responds "very sure same") to right (when the listener responds "very sure different"). To construct an empirical ROC curve, both the proportions of "same" ("waltz," "new") responses and the proportions of "different" ("march," "old") responses must be considered for each confidence rating category. Summing proportions of each response type as the response criterion moves from left to right along the information dimension yields cumulative response proportions for each response category; plotting cumulative proportions of "different" responses as a function of cumulative proportions of "same" responses yields an empirical ROC curve. When cumulative response proportions are z-transformed, the plotted zROC approximates a straight line. When the equal-variance assumption is satisfied, the slope of the zROC line is 1; this is because the zROC slope is a ratio of the standard deviations underlying the two stimulus classes. A zROC slope different from 1 indicates a violation of the equal-variance assumption.

[2] We note here that the measures we describe are nonparametric in the sense that they do not rely on assumptions about the form of the distributions underlying the two stimulus classes. This is different from using nonparametric statistics on a non-normally distributed dependent measure, which does not eliminate problems with PC as a measure of performance. Our main point is that PC as a measure of performance is tainted by response bias. The nonparametric measures we introduce are designed to correct for response bias and (if present) violation of the equal-variance assumption of SDT.

Overview

Participants in the present study either completed the original version of the MBEA with its binary response format or the MBEA-C involving a 6-point confidence rating scale. Aside from the response mode, the two tests were identical. The main motivation for obtaining confidence ratings with the MBEA-C is that it permitted the construction of empirical ROC curves and a test of whether the equal-variance signal detection assumption holds for the MBEA. Failure to satisfy this assumption would mean that both PC and d' would be biased performance measures and a better choice would be nonparametric ROC-based statistics. However, satisfaction of this assumption allowed us to use the nonparametric measures as a benchmark against which to test PC and d', thereby confirming that d', but not PC, is an unbiased measure of MBEA performance.

We were also interested in comparing the response criterion, *c*, for individuals whose MBEA performance fell more than 2 *SD* below mean PC on the MBEA for amusia diagnosis to individuals who were considered to have normal musical abilities. One important possibility regarding poor MBEA performance associated with amusia is that, since PC cannot separate contributions of sensitivity and response bias, amusia classification may be based in part on response bias in addition to overall poor sensitivity.

Finally, we were interested in comparing empirical ROC curves for amusics versus non-amusics. The potential benefits of such a comparison are highlighted by studies showing differences in the shape of ROC curves for individuals demonstrating memory impairments due to amnesia (Aly, Knight, & Yonelinas, 2010), hippocampal damage (Vann et al., 2009), or normal aging (Parks, DeCarli, Jacoby, & Yonelinas, 2010) relative to control participants. Thus, in considering why individuals do very poorly on the MBEA, the ROC-based analyses of the MBEA-C may highlight novel performance differences between these subgroups that help clarify the nature of amusia.

Method

PARTICIPANTS

One hundred fifty-five individuals with self-reported normal hearing from the Bowling Green State University and Michigan State University communities completed either the MBEA-C or the original version of the

TABLE 1. *Demographic Information for the Participants Completing Both the Original Version of the MBEA with Binary Eesponses and the MBEA-C Involving Confidence Rating Responses.*

	MBEA-C	MBEA
n (female)	93 (68)	62 (39)
Age (years ± *SD*)	20.6 (3.3)	20.4 (3.3)
Education (years ± *SD*)	15.2 (2.4)	14.5 (1.8)
Music Training (years ± *SD*)	4.2 (4.0)	3.9 (4.3)

Note: Aside from the response mode, the two versions were otherwise identical. Sample size (*n*) is shown with the number of female participants in parentheses, while mean age, education, and music training (all given in years) are shown with standard deviation (*SD*) in parentheses.

MBEA in exchange for course credit or a cash payment of $10/hr. Ninety-three individuals (68 female) completed the MBEA-C (age, $M = 20.6$ yrs, $SD = 3.3$ yrs; education, $M = 15.2$ yrs, $SD = 2.4$ yrs; formal music training, $M = 4.2$ yrs, SD = 4.0 yrs). Sixty-two individuals (39 female) completed the original version of the MBEA (age, $M = 20.4$ yrs, SD = 3.3 yrs; education, $M = 14.5$ yrs, $SD = 1.8$ yrs; formal music training, $M = 3.9$ yrs, $SD = 4.3$ yrs). Demographic information for all participants is summarized in Table 1. The two participant groups did not differ in terms of age ($p = .73$) or years of music training ($p = .66$), but were marginally different with respect to years of education ($p = .06$), with the sample completing the MBEA-C having on average slightly more education (MBEA: $M = 14.5$ years, $SD = 1.8$; MBEA-C: $M = 15.2$ years, $SD = 2.4$).

STIMULI, EQUIPMENT, AND PROCEDURE

The original MBEA and MBEA-C consist of thirty novel melodies plus fifteen additional melodies that serve as new melodies on the Memory subtest (Peretz et al., 2003). Each same-different subtest also includes one catch trial where the difference between melodies is very obvious; catch trials were included to ensure that participants were attending to the task. Mean melody duration is 5.1 s for the melodies presented in the Scale, Contour, Interval, Rhythm, and Memory subtests and 11 s for the melodies presented in the Meter subtest. Melodies comprising the Meter subtest also included an accompaniment that emphasized the repeating strong-weak-strong-weak or strong-weak-weak-strong-weak-weak accent patterns associated with marches and waltzes, respectively. For the Scale, Contour, Interval, and Rhythm subtests, participants heard a pair of melodies on each trial and then rated how confident they were that the two melodies in each pair were the same or different on a scale ranging from "1" ("sure same") to "6" ("sure different"); participants completing the original version of the MBEA, simply responded "same" or "different." For all same-different subtests, half of the melody pairs were the *same* and half were *different*. For the Scale subtest, *different* melodies contain one note that violates the key of the intact version of the melody, while keeping the overall melodic contour intact. For the Contour subtest, *different* melodies contain one note that violates the contour of the intact melody in each pair without disrupting the key. For the Interval subtest, the altered note changes the pitch interval while preserving the melodic contour and key. For the Rhythm subtest, rather than a melodic change, *different* melodies are created by shifting the temporal location of one note so that the preceding and following inter-note onset intervals are altered.

For the Meter subtest, a single melody was presented on each trial and participants were asked to rate the extent to which they were confident that the melody was a march or waltz on a scale ranging from "1" ("sure march") to "6" ("sure waltz"); participants completing the original version of the MBEA simply responded "march" or "waltz." Prior to completing the Meter subtest, participants were told that marches sound like groups of two, with an alternating strong-weak-strong-weak accent pattern, and waltzes sound like groups of three, with a strong-weak-weak-strong-weak-weak accent pattern. They then heard examples of a march and a waltz and completed four training trials with feedback. Half of the melodies on the Meter subtest were marches and half were waltzes.

For the final Memory subtest, 15 *old* and 15 *new* melodies were presented; *old* melodies had been heard previously in the earlier subtests, while *new* melodies had not been heard previously, but had similar characteristics to the *old* melodies. Participants completing the MBEA-C rated the familiarity of the melody on a scale ranging from "1" ("sure new") to "6" ("sure old"); participants completing the original version of the MBEA simply responded "new" or "old."

Both the MBEA-C and original version of the MBEA were adapted to be administered using E-Prime software (Psychology Software Tools, Inc.) running in a Microsoft Windows environment on a Dell Optiplex computer. All stimuli were presented at a comfortable volume (∼70 dB) over Sennheiser HD280 headphones. Subtests were presented in the following order: Scale, Contour, Interval, Rhythm, Meter, and Memory. This order of subtest presentation is the same as in Peretz et al. (2003). Following administration of the MBEA-C or original version of the MBEA, listeners filled out several surveys that included questions about

participant age, gender, music training, and level of education. Overall, the battery and additional surveys took between 60 and 90 min to administer.

DATA ANALYSIS

First, catch trials were removed from the analysis; all participants performed 100% correct on catch trials. To permit a comparison between the MBEA-C and original version of the MBEA, confidence ratings for the MBEA-C were collapsed into binary response categories; ratings of "1," "2," and "3" were coded as "same," "march," or "new" responses (depending on the subtest), while ratings of "4," "5," and "6" were coded as "different," "waltz," or "old." Then, binary response proportions were used to calculate PC, d', and c for participants completing the MBEA-C and original version of the MBEA. PC was taken as the proportion of correct responses out of 30 trials on each subtest; d' and c were calculated according to the following standard formulas:

$$d' = \mathrm{z(HR)} - \mathrm{z(FAR)} \tag{2}$$

$$c = -1/2[\mathrm{z(HR)} + \mathrm{z(FAR)}] \tag{3}$$

Hits were defined as the proportions of "different," "march," or "old" responses given for *different*, *march*, or *old* trials, respectively, and false alarms were defined as the proportions of "different," "march," or "old" responses given for *same*, *waltz*, or *new* trials, respectively. For all subtests, values of d' equal to 0 correspond to chance performance, and larger values of d' correspond to increased perceptual sensitivity and better performance. For all but the Meter subtest, negative values of c can be interpreted as a liberal response strategy (i.e., a tendency to respond "different" or "old") whereas positive values of c can be interpreted as a conservative response strategy (i.e., a tendency to respond "same" or "new"). Values of $c = 0$ indicate no response bias. We note here that the value of c for the Meter subtest does not meaningfully align with either a liberal or conservative response strategy, but rather reflects a tendency to respond "march" or "waltz," respectively.

Confidence ratings for participants completing the MBEA-C were then used to construct empirical ROC curves for each subtest. When an individual responds using the extremes on the rating scale (e.g., "very sure same" and "very sure different") without making use of the middle of the scale (i.e., they fail to make use of the full rating scale), it is not possible to construct empirical ROC curves for that individual. This is especially likely for participants who perform extremely well on the MBEA. Thus, for the current study, we aggregated data over small numbers of participants ($n = 4$) and iteratively constructed empirical ROC curves based on average data in the following manner. First, we randomly selected data for four participants. We chose four because this was the number of amusic participants in our MBEA-C sample (see below), and we wanted to match sample sizes when comparing amusic and non-amusic individuals. Then, we calculated proportions of "different"/"march"/"old" responses and "same"/"waltz"/"new" responses for each rating category averaged over the four randomly selected participants and used these values to construct ROC curves by plotting cumulative response proportions as described in the Introduction. Next, we z-transformed the cumulative response proportions to create zROCs, correcting for proportions of 0 and 1 using $1/2N$ and $1 - 1/2N$, respectively, where $N = 15$ trials (Macmillan & Creelman, 2005).

Several dependent measures were then derived from the empirical ROCs. First, the slope, s, of the zROCs for each subtest is given by:

$$s = d'_2/d'_1 \tag{4}$$

where, theoretically, d'_1 corresponds to the horizontal distance from the zROC to the major diagonal at the point where $z(HR) = 0$, and d'_2 corresponds to the vertical distance from the zROC to the major diagonal where $z(FAR) = 0$. Practically, s is estimated from the best-fit regression line through the zROC data points. The values d'_1 and d'_2 depend on the standard deviations of the distributions underlying the stimulus classes for a given subtest. The slope of the zROC, s, gives the ratio of the standard deviations of the distributions underlying the two stimulus classes; thus, when the standard deviations of the two distributions are equal, $s = 1$, and d' is an accurate measure of perceptual sensitivity. When, $s \neq 1$, a more appropriate measure of perceptual sensitivity is d_a, which corrects for violation of the equal-variance assumption:

$$d_a = \left(2/\left(1 + s^2\right)\right)^{1/2} * [\mathrm{z(HR)} - s * \mathrm{z(FAR)}] \tag{5}$$

The measure d_a is given in units of root-mean-square standard deviation for the two stimulus classes. We also calculated A_z, which corresponds to the area under the normal-model ROC and increases from .5 at zero sensitivity (i.e., chance) to 1.0 for perfect performance. A_z is a nonparametric performance measure, corrected for non-unit slope zROCs and response bias, and can be compared to PC:

TABLE 2. *The Dependent Measures PC, d′, and c (with SD in Parenthesis) for each Subtest of the MBEA-C and MBEA.*

		Subtest						
		Scale	Contour	Interval	Rhythm	Meter	Memory	Overall
MBEA-C (n = 93)	PC	0.84	0.81	0.8	0.84	0.84	0.9	0.84
		(0.09)	(0.12)	(0.12)	(0.11)	(0.16)	(0.09)	(0.08)
	d'	2.93	2.68	2.54	2.93	2.34	2.74	2.69
		(0.71)	(0.90)	(0.94)	(0.78)	(1.18)	(0.76)	(0.63)
	c	−0.04	0.07	0.21	0.06	−0.18	0.02	0.02
		(0.45)	(0.38)	(0.38)	(0.39)	(0.26)	(0.28)	(0.22)
MBEA ($n = 62$)	PC	0.83	0.79	0.77	0.79	0.76	0.88	0.80
		(0.10)	(0.13)	(0.13)	(0.12)	(0.21)	(0.10)	(0.10)
	d'	2.89	2.53	2.37	2.56	1.74	2.54	2.44
		(0.73)	(0.95)	(0.98)	(0.94)	(1.52)	(0.81)	(0.72)
	c	−0.05	0.06	0.21	0.04	−0.11	0.08	0.04
		(0.49)	(0.50)	(0.50)	(0.50)	(0.25)	(0.33)	(0.30)

$$A_z = \Phi(d_a/\sqrt{2}) \quad (6)$$

On each of 10,000 iterations, we estimated s, d_a, and A_z as described above. These estimates formed a sampling distribution from which we estimated a test statistic for each dependent variable (Ernst, 2004).

Results

COMPARISON OF THE MBEA-C TO THE MBEA

We first compared performance on the MBEA for the two response modes. Table 2 summarizes performance on the MBEA-C and the original version of the MBEA for PC, d', and c for each subtest and combined across subtests. Overall performance on the MBEA-C was slightly better than performance on the MBEA for both PC (.84 $\pm$ 0.01 versus .80 $\pm$ .01, $t(153) = -2.52$, $p < .05$, Cohen's $d = 2.43$) and d' (2.00 $\pm$ 0.09 versus 2.24 $\pm$ 0.07, $t(153) = -2.30$, $p < .05$, Cohen's $d = 2.23$). In contrast, no differences were observed in response criterion, c, for the two versions of the test (MBEA-C: $c = 0.02 \pm 0.02$; MBEA: $c = 0.04 \pm 0.04$, $t(153) = 0.44$, $p = .66$, Cohen's $d = 0.45$). To explore the slight performance advantage on the MBEA-C further, we compared performance on the MBEA-C and the MBEA for each subtest by conducting separate families of Bonferroni-corrected independent-samples t-tests (per-comparison $\alpha = .017$). For both PC and d', performance on the Rhythm and Meter subtests was significantly better on the MBEA-C than on the MBEA (PC, Rhythm: $t(153) = -2.52$, $p = .01$, Cohen's $d = 2.63$; PC, Meter: $t(153) = -2.91$, $p = 0.004$, Cohen's $d = 2.55$; d', Rhythm: $t(153) = -2.63$, $p = .009$, Cohen's $d = 2.57$; d', Meter: $t(153) = 2.74$, $p = .007$, Cohen's $d = 2.63$), but did not differ for any of the other subtests. This suggests that the slight performance advantage observed for the MBEA-C compared to the original MBEA was driven by the Rhythm and Meter subtests only. Exploratory analyses of these subtest differences revealed that performance on the Rhythm subtest correlated significantly with years of education (PC: $r(141) = .18$, $p = .03$, d': $r(141) = .18$, $p = .03$), and when this variable was included as a covariate, the difference between the MBEA and the MBEA-C on the Rhythm subtest did not reach statistical significance with a corrected α-level (PC: $p = .04$, d': $p = .03$). Including years of education as a covariate did not influence results for the Meter subtest.

ROC ANALYSIS OF THE MBEA-C

Next, we used confidence ratings from the MBEA-C to construct empirical ROC curves for each subtest. Figures 4 and 5 show normal-model ROCs and zROCs, respectively, based on data aggregated over all participants. Table 3 reports the signal detection measures s, d_a, and A_z separately for each subtest calculated according to the iterative method described above. The slopes of the zROCs, s, provide an assessment of whether the equal-variance signal detection model holds for each subtest. The s = 1 case corresponds to the situation when the distributions for Class A and Class B have equal variance. Critically, when $s \neq 1$, d' varies with the criterion, c, and thus represents a biased measure. For zROCs with non-unit slope ($s \neq 1$), the measures d_a and A_z provide unbiased alternatives to d' and PC, respectively.

Monte Carlo permutation tests were conducted to test for differences between the estimated signal detection measures and their expected values under an equal-variance assumption. On each iteration, estimated

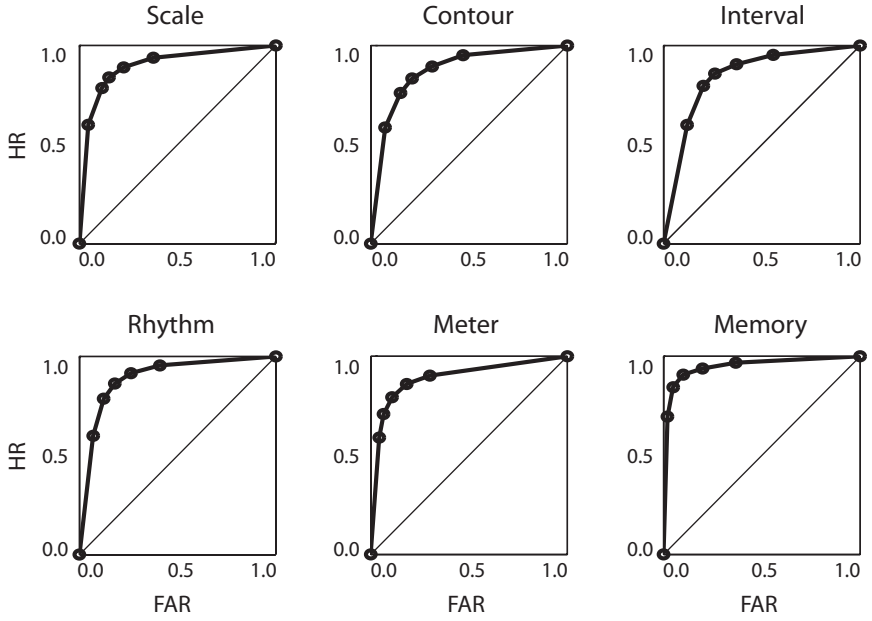

FIGURE 4. Empirical normal-model ROC curves for the MBEA-C. HR is plotted as a function of FAR separately for each of the six subtests.

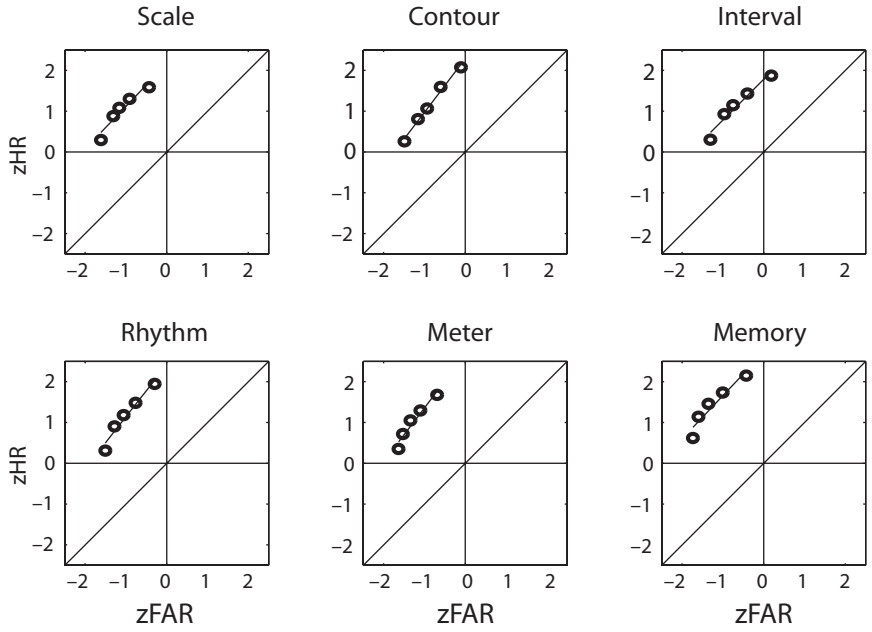

FIGURE 5. Z-transformed ROC curves (zROCS) for the MBEA-C. zHR is plotted against zFAR separately for each of the six subtests.

slopes were compared against 1, estimates of d_a were compared against average d' for each 4-participant sample, and estimates of A_z were compared against average PC. Differences between observed and expected values on each iteration were used to form a sampling distribution, with mean equal to zero under the null hypothesis, that is, that observed and expected values did not differ. If the null hypothesis value (i.e., zero) fell more than 2.33 SD away from the true mean of the sampling distribution (corresponding to $p < .01$), the observed and expected values were considered to be significantly different. Slopes did not differ significantly

TABLE 3. *ROC-based Dependent Measures s, d_a, and A_z for Each Subtest of the MBEA-C.*

	Subtest					
	Scale	Contour	Interval	Rhythm	Meter	Memory
s	0.95	1.01	1.01	1.10	0.88	0.73
d_a	2.00	1.78	1.64	1.97	1.96	2.44
A_z	0.92	0.89	0.87	0.91	0.90	0.95

from their expected values of 1 for any subtest (Scale: $z = -0.16$, $p = .44$; Contour: $z = 0.05$, $p = .48$; Interval: $z = 0.02$, $p = .49$; Rhythm: $z = 0.23$, $p = .41$; Meter: $z = -0.41$, $p = .34$; Memory: $z = -1.07$, $p = .14$), indicating satisfaction of the equal-variance assumption. Consistent with this, d_a values were not significantly different from d' for any subtest (Scale: $z = -1.50$, $p = .07$; Contour: $z = -1.48$, $p = .07$; Interval: $z = -1.71$, $p = .04$; Rhythm: $z = -1.66$, $p = .06$; Meter: $z = -1.52$, $p = .06$; Memory: $z = -1.45$, $p = .07$), indicating that d' is a minimally biased measure of performance for the MBEA. However, PC values significantly underestimated performance for all of the same-different subtests (Scale: $z = 3.52$, $p < .001$; Contour: $z = 3.92$, $p < .001$; Interval: $z = 3.21$, $p < .001$; Rhythm: $z = 3.27$, $p < .001$); this difference did not reach statistical significance for the Meter ($z = 2.11$, $p = .02$) or Memory subtest ($z = 2.01$, $p = .02$).

In sum, inspection of empirical ROC curves revealed that PC is generally a suboptimal measure of sensitivity for the MBEA. For all of the same-different subtests, PC underestimated sensitivity relative to A_z, which provides an analogue to PC that is corrected for response bias. The equal-variance assumption of SDT was also found to be satisfied for all subtests, meaning that d' is not tainted by (i.e., is independent of) response bias, thus indicating that d' is preferable to PC as a measure of MBEA performance. This was confirmed by comparisons of d' to the nonparametric measure d_a, which indicated that d' did not differ significantly from d_a for any subtest.

SIGNAL DETECTION EVALUATION OF NONAMUSIC AND AMUSIC LISTENERS

Next, we compared nonamusic and amusic samples of listeners using signal detection measures derived from confidence rating responses on the MBEA-C. To classify listeners as amusic, we adopted the commonly used convention of considering amusic individuals to be those with composite PC scores more than 2 *SD* below the mean of our full sample (Peretz et al., 2003); an equivalent method is simply to sum the number of correct responses on each of the subtests. Although the above results argue against the use of PC as the primary dependent variable to summarize MBEA performance, we used PC to form the amusic group so that we could evaluate the results of this practice using alternative signal detection measures. For our sample of 93 participants, four met the criterion for amusia diagnosis, consistent with the number expected when using a 2 *SD* below the mean cutoff in a slightly negatively skewed distribution (Henry & McAuley, 2010). Nonamusic ($M = 4.32$, $SD = 3.99$) participants were found to have significantly more music training than amusic listeners ($M = 0.50$, $SD = 1.00$) based on a single sample *t*-test comparing nonamusic music training data against the mean of the amusic sample, $t(87) = 3.88$, $p < .001$.[3]

Results for all of the dependent measures are summarized in Table 4 separately for amusic and nonamusic samples. To compare the scores for amusic to nonamusic participants, we conducted single-sample *t*-tests for the nonamusic data against the test value equal to the mean of the amusic group. As expected based on the use of PC to diagnose amusia, average values of d' were lower for the amusics than for the non amusics for each of the subtests [Scale: $t(88) = 17.86$, Cohen's $d = 1.89$; Contour: $t(88) = 15.30$, Cohen's $d = 1.62$; Interval: $t(88) = 17.47$, Cohen's $d = 1.85$; Rhythm: $t(88) = 20.76$, Cohen's $d = 2.20$; Meter: $t(88) = 14.16$, Cohen's $d = 1.50$; Memory: $t(88) = 20.39$, Cohen's $d = 2.16$; all *p*s < .001]. More interesting was a consideration of differences in the criterion score, *c*. For all of the same-different subtests, amusics were more conservative than nonamusics [Scale: $t(88) = 13.17$, Cohen's $d = 1.40$; Contour: $t(88) = 6.51$, Cohen's $d = 0.69$; Interval: $t(88) = 10.67$; Cohen's $d = 1.13$; Rhythm: $t(88) = 4.05$,

[3] Although the nonamusic group tended to have more music training than the amusic group, we point out that the observed group difference results from a significant correlation across the entire sample between years of musical training and performance, as indexed by both PC, $r(90) = .44$, $p < .001$, and d', $r(90) = .45$, $p < .001$. Thus, the relationship between music training and performance on the MBEA is more continuous and it is not simply the case that nonamusic individuals are musically trained, while amusic individuals are not.

TABLE 4. *Dependent Measures Derived From Binary Response Proportions (i.e., PC, d′, and c shown with SD) and From Empirical ROC Curves for Nonamusics and Amusics.*

		Subtest					
		Scale	Contour	Interval	Rhythm	Meter	Memory
Nonamusics ($n = 89$)	PC	0.85 (0.08)	0.82 (0.11)	0.81 (0.11)	0.85 (0.10)	0.85 (0.15)	0.92 (0.08)
	d'	2.30 (0.68)	2.06 (0.84)	1.96 (0.80)	2.31 (0.76)	2.40 (1.17)	2.81 (0.71)
	c	−0.06 (0.40)	0.06 (0.36)	0.18 (0.38)	0.06 (0.38)	−0.19 (0.26)	0.03 (0.27)
	s	0.93	1.03	1.01	1.08	0.87	0.73
	da	2.06	1.84	1.71	2.04	2.01	2.52
	Az	0.92	0.90	0.88	0.92	0.91	0.96
Amusics ($n = 4$)	PC	0.67 (0.11)	0.59 (0.11)	0.58 (0.17)	0.60 (0.03)	0.62 (0.16)	0.73 (0.06)
	d'	1.01 (0.70)	0.70 (0.83)	0.49 (0.94)	0.64 (0.29)	0.65 (0.92)	1.28 (0.31)
	c	0.50 (0.96)	0.31 (0.78)	0.61 (0.37)	0.22 (0.69)	−0.02 (0.19)	−0.23 (0.37)
	s	1.00	0.62	0.79	1.02	0.90	0.66
	da	0.88	0.59	0.20	0.59	0.52	1.10
	Az	0.73	0.66	0.56	0.66	0.64	0.78

Cohen's $d = 0.43$; all ps < .001]. That is, relative to nonamusics, amusic participants responded "same" more often than "different." Mean c values for amusic and nonamusic participants are reported in Table 4. Response biases also differed between amusic and non-amusic groups for the Meter, $t(88) = 6.15$, $p < .001$, Cohen's d = 0.65, and Memory, $t(88) = 9.00$, $p < .001$, Cohen's $d = 0.95$, subtests. For the Memory subtest, amusics tended to use a more liberal response criterion than nonamusics, responding "old" more often than "new." For the Meter subtest, nonamusics had slightly more liberal response criteria than amusics. However, this difference is difficult to interpret since negative values of c for this subtest only indicate a slight tendency to respond "march" more often. The reliable difference in estimates of c for amusics and nonamusics across all of the MBEA subtests raises the possibility that differences in response bias could impact amusia classification (a possibility we consider more directly below).[4]

We also compared ROC-based measures for the amusic and nonamusic groups for each subtest using the same permutation test described above. Figures 6 and 7 show normal-model ROCs and zROCs, respectively, for both groups. Comparison of s values for nonamusics relative to amusics revealed that slopes did not differ significantly between groups for any subtest [Scale: $z = -0.17$, $p = .43$; Contour: $z = 1.50$, $p = .07$; Interval: $z = .24$; Rhythm: $z = 0.18$, p .43; Meter: $z = -0.14$, $p = .44$; Memory: $z = 0.22$, $p = .41$]. Not surprisingly, d_a and A_z were significantly different between amusic and nonamusic groups for all same-different subtests [d_a – Scale: $z = 3.22$, $p < .001$; Contour: $z = 3.21$, $p < .001$; Interval: $z = 4.13$, $p < .001$; Rhythm: $z = 3.29$, $p < .001$; A_z – Scale: $z = 2.62$, $p = .004$; Contour: $z = 3.54$, $p < .001$; Interval: $z = 6.06$, $p < .001$; Rhythm: $z = 2.49$, $p = .006$]. Moreover, d_a values differed significantly between groups for the Memory subtests, $z = 3.30$, $p < .001$. However, d_a values for the Meter subtest, $z = 1.99$, $p = .02$, and A_z values for both the Meter, $z = 1.12$, $p = .13$, and Memory, $z = 1.80$, $p = .04$, subtests did not differ between amusic and nonamusic listeners.

In sum, a comparison of amusics with and nonamusic (normal) listeners revealed differences in response bias, with amusics responding more conservatively on the same-different subtests and more liberally on the Memory test; that is, individuals classified with amusia using the standard PC cut-off, tended to respond "same" more often than "different" and "old" more often than "new." Amusic and nonamusic samples did not differ in terms

[4] To validate our method of testing the nonamusic data against the mean of the amusic group using a single-sample t-test, we supplemented the analysis with permutation tests. On each of 10,000 iterations, mean d' and c were calculated for random samples of four nonamusic participants, and a sampling distribution was formed from the difference scores (amusic vs. non-amusic), from which a test statistic was calculated. The results using this method were identical and larger in magnitude (d': min $z = 18.99$; c: min $z = 36.06$; all ps < .0001).

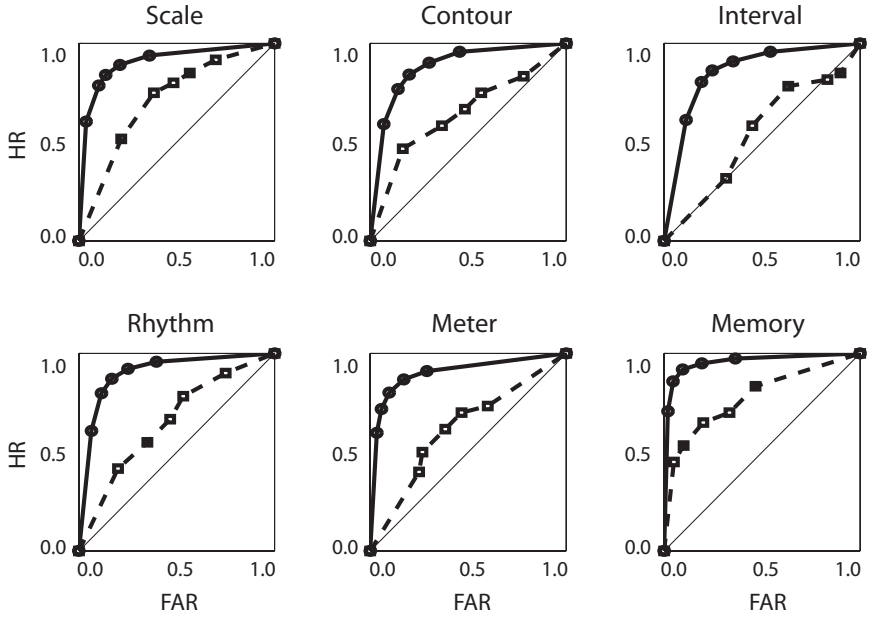

FIGURE 6. Empirical normal-model ROC curves for the MBEA-C shown separately for nonamusics (solid lines) and individuals classified as amusic based on a *2-SD* cutoff (dashed lines).

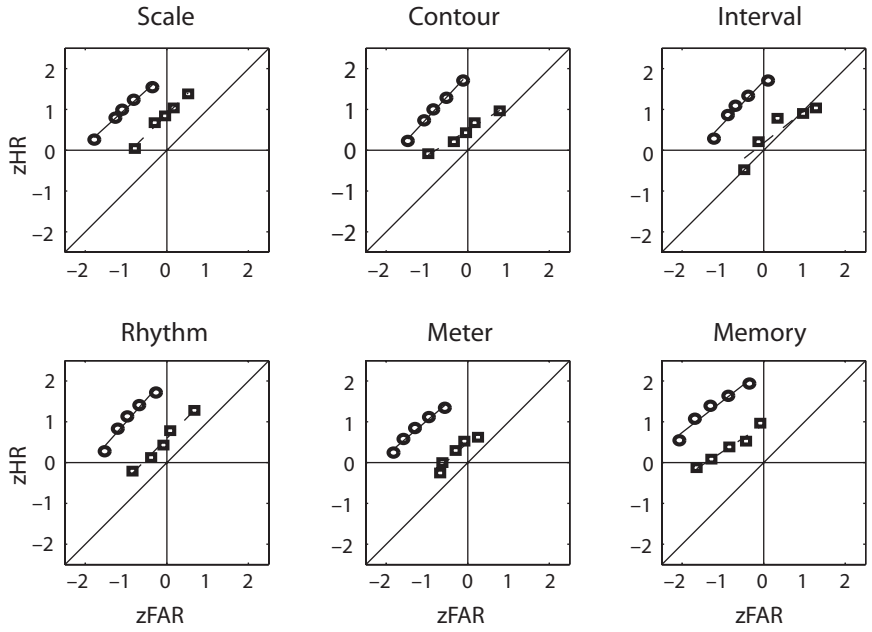

FIGURE 7. Z-transformed ROC curves (zROCs) for the MBEA-C shown separately for nonamusics (solid lines) and individuals classified as amusic based on a *2-SD* cutoff (dashed lines).

of zROC slope, indicating that d' as a dependent measure, unlike PC, is not tainted by response bias, and should therefore be preferred to PC as a measure of MBEA performance.

COMPARISON OF PC AND D′ FOR DIAGNOSIS OF AMUSIA

As a final means of comparing PC and d' as measures for setting a diagnostic criterion, we pooled all participants from the MBEA and MBEA-C and formed amusic

groups using 2-*SD* cutoffs based on both dependent measures separately. Based on PC (cutoff: PC = 65%),[5] 6 of 155 individuals were categorized as amusic. Based on d' (cutoff: $d' = 1.23$), 5 of 155 individuals were categorized as amusic. Critically, only four of those were also present in the amusic group formed on the basis of a PC cutoff. Indeed, two of six individuals (33%) who would have been diagnosed as amusic based on PC were not present in the d'-based group, and closer inspection of the data for these two individuals revealed that they had large positive response criteria ($c = 0.51$ and 1.08 averaged over the pitch subtests), but normal perceptual sensitivity ($d' = 1.65$ and 1.72, respectively, for the pitch subtests).[6] Indeed, based on Equation 1, if these individuals were unbiased in their responding, their pitch-subtest PC values would have been .80 and .81, respectively, well within the normal performance range. Thus, this analysis demonstrates the susceptibility of PC to shifts in the response criterion, *c*. For our sample, it led to a 33% (2/6) misclassification rate.

Discussion

Research on normal and disordered music perception has garnered increased attention in the past decade, alongside interest in the relationship between music and language (Patel, 2008). In this regard, research on congenital amusia has taken center stage. In the present study, we evaluated current methods for diagnosing amusia and more generally assessing music perception ability using the Montreal Battery of Evaluation of Amusia. Specifically, we conducted a comprehensive signal detection analysis using the original binary-response version of the MBEA and a confidence-rating version, the MBEA-C. The use of confidence ratings afforded a more comprehensive signal detection analysis of the MBEA than is possible with the traditional binary-response format.

An overall comparison of the MBEA-C with the MBEA revealed slightly better performance on the MBEA-C than on the MBEA. However, the difference between the two versions of the test was driven by the two temporal organization subtests (i.e., Rhythm, Meter). For the Rhythm subtest, this difference was no longer significant when we took into account differences in education for the samples completing the MBEA and the MBEA-C, however, the difference for the Meter subtest remained. One potential explanation for better performance on the temporal organization subtests is related to the report of Peretz, Brattico, Järvenpää, and Tervaniemi (2009) regarding a set of amusics who failed a mistuned pitch detection test when required to make a binary response, but showed some sensitivity to the target pitch when allowed to provide a graded (confidence-rating) response. It is possible that confidence-rating responses allowed for slightly better performance, because sensitivity can be masked by a binary-response requirement. However, given that this previous result was in the pitch domain, it is unclear why in the present study a performance benefit with confidence ratings would be only observed for the temporal organization subtests. It is possible that this finding for the temporal organization subtests is simply spurious and would fail to replicate in another study. More work is needed to assess this.

Three primary questions of interest for the SDT analyses were: 1) whether PC (the standard performance index for the MBEA) is biased and thus perhaps not the best measure for assessing and comparing individual performance, 2) whether amusics and nonamusics differ in their response bias, as measured by *c*, and 3) whether d' (an alternative signal detection performance index) provides a useful alternative to PC.

With respect to the first question, we found that PC is indeed a biased measure of MBEA performance. We used confidence-rating responses to calculate A_z, a bias-free nonparametric performance measure that is directly comparable to PC. Comparison of PC with A_z using a permutation-based approach revealed that PC consistently underestimates performance for all of the same-different subtests. Our SDT analyses revealed that this was due to lower PC values associated with shifts in the location of the response criterion, *c*, rather than due to a true decrease in sensitivity. The theoretical relationship between the response criterion, *c*, and PC is shown in Figure 3; for any fixed sensitivity, non-zero response bias causes decreases in PC that are not attributable to decreased sensitivity. Because of this dependency, we suggest the use of d' as an alternative measure of MBEA performance. Toward this end, we showed that the equal-variance assumption of SDT holds for all of the MBEA subtests, which means that d' does not

[5] We note that our PC-based cutoff was somewhat lower than cutoffs from other normative samples using the MBEA. For example, the cutoff for amusic diagnosis emerging from the initial study of Peretz et al. (2003) was 77%, whereas ours was 65%, which is a much lower and necessarily more conservative criterion. We do not currently have an explanation for the overall difference in performance between our sample and the original Peretz norms. However, we note that several other studies have reported cutoff scores in between these values (Cuddy et al., 2005: 72%, Peretz et al., 2008: 74% for young adults, 70% for older adults), suggesting that some variability in cutoff scores across samples should be expected.

[6] One additional individual was classified as amusic using only a d'-based cutoff, but was considered nonamusic based on PC. We note that this individual fell on the border of diagnosis with PC as well (composite PC = 66%, relative to a 65% cutoff).

suffer from the same contamination from response bias that PC does. One advantage of d' is that it can be calculated from binary response proportions obtained using the standard binary-response version of the MBEA.

With respect to the second question, we found that participants classified as amusic using the PC-based criterion were much more conservative in their responding (they had reliably more extreme values of *c*) than nonamusic participants. Several other studies are worth noting in this regard. First, Peretz et al. (2009) found that the relatively poor performance of amusics relative to nonamusics on a task that required detection of an out-of-key note was in part due to frequent 'congruous' responses to incongruous melodies containing a mistuned note. Second, Williamson et al. (2010) reported larger proportions of misses than false alarms on a pitch memory task requiring same-different responses (i.e., more "same" than "different" responses) for amusics relative to nonamusics. Finally, Omigie and Stewart (2011) found differences between response biases for amusic versus nonamusic participants. However, response bias in this study was defined somewhat differently than the standard SDT definition in order to assess implicit versus explicit learning (Kunimoto, Miller, & Pashler, 2001; Tunney & Shanks, 2003). As far as we are aware, however, the current study is the first to explicitly report systematic differences in response bias between amusic and nonamusic participants using estimates of the response criterion, *c*, on the MBEA.

There are at least two possible interpretations of the observed response bias differences between amusic and nonamusic individuals. On the one hand, it could be the case that individuals with a musical impairment are more strongly biased than nonamusic individuals. On this account, shifts in the response criterion are a symptom. However, on the other hand, because PC is tainted by response bias and a PC-based criterion is typically used to create samples of amusic and nonamusic participants in many research studies on congenital amusia, simply having a large response bias could *cause* an individual with normal sensitivity to be diagnosed as amusic. Indeed, these possibilities are not mutually exclusive. We have shown that 33% of our diagnosed amusics showed perceptual sensitivity in the normal range (based on d') but had large response biases that led to amusia diagnosis when using a PC-based cutoff. These individuals constitute a clear case of the latter situation, where increased response bias causes an amusia diagnosis. However, the rest our amusic sample showed high response bias concomitant with low sensitivity, suggesting that for these individuals, extreme response bias may be a symptom of a true musical disorder. Using PC as a dependent measure does not allow separating these two possibilities. Thus, the solution offered in this article is to abandon the use of a PC-based criterion for amusia classification in favor of an unbiased performance metric. In this regard, with respect to the third primary question, we've shown that for the MBEA, d' is an unbiased alternative.

For cases where having a conservative response criterion is a symptom of amusia rather than contributing to potential misdiagnosis, one obvious question is why these amusic individuals would demonstrate such a conservative response strategy. One possibility is motivational in origin. Previous work on regulatory focus theory has shown that when an individual performs a diagnostic test with a preconceived notion about how she should perform, one potential consequence is adoption of a more conservative response criterion (Crowe & Higgins, 1997). In the domain of the MBEA, previous work has shown that nonmusicians (who were critically told that they were likely to perform poorly relative to musicians on the task) were more conservative than musicians (who were told they were likely to perform well; McAuley, Henry, & Tuft, 2011). On this basis, our suggestion is that individuals who expect to perform poorly on a task that they know will be diagnostic (i.e., the MBEA) may adopt a more conservative response criterion in an effort to hedge their losses. A test of this hypothesis is currently underway.

It is interesting here to consider a possible connection between (1) potential misclassification of a musical impairment based on extreme response biases and (2) perception-action mismatch observed for some amusic individuals. Consider, for example, that although some amusics perform very poorly when asked to identify the direction of a pitch change (by responding 'lower' or 'higher'), they are capable of correctly singing the direction of the change (Loui, Guenther, Mathys, & Schlaug, 2008). One possibility that deserves some consideration based on the present study is that the quantification of perceptual deficits may be based, at least in part, on response biases, whereas production tasks by nature require responses that are bias-free. This suggestion is supported by a recent study indicating that the degree of a perception-action mismatch in amusics depends on the nature of the perception task (Williamson, Liu, Peryer, Grierson, & Stewart, 2012). Perception thresholds were found to be much higher than production thresholds when amusics were asked to identify the direction of a pitch change (as in Loui et al., 2008); however, when

perception abilities were evaluated using an AXB task intended to reduce response bias, no such dissociation was found. In general, future work should carefully consider the role of response bias in conclusions regarding the nature of amusia.

Finally, it is worth noting that although in the current study we were able to examine only a small number of diagnosed amusics ($n = 4$ on the MBEA-C), the conclusions we have drawn regarding the biases inherent in PC as a measure of performance are based on a signal detection analysis of our full sample of 93 participants. Thus, our primary conclusion (PC is a biased measure of performance on the MBEA and should be avoided) could not be an artifact of the small number of amusic participants. Nonetheless, an important future goal of amusia research should be to better explicate the nature of response biases in amusics using larger sample sizes.

In sum, we prescribe use of the signal detection measure, d', in evaluation of MBEA performance and diagnosis of amusia. The current study shows that PC is a biased performance measure, due to its variation with response criterion location. Moreover, amusic individuals (diagnosed based on PC) were more biased than nonamusic listeners. However, we argue that this is, in some cases, not a hallmark of amusia per se, but instead can be a consequence of using an inappropriate dependent measure that has the potential to misdiagnose amusia when individuals use relatively extreme response criteria but have sensitivity in the normal range. The use of a bias-free performance metric for amusia diagnosis is necessary to avoid this problem. The SDT analyses we performed based on confidence rating data revealed satisfaction of the equal-variances assumption of SDT for all subtests of the MBEA, thus supporting the use of d' as a preferred performance metric and diagnostic criterion. From this perspective, the use of bias-free criteria tightens the criteria for amusic classification by restricting diagnoses to subgroups of "poor performers" who show true sensitivity deficits.

Author Note

We thank Bryan Gruschow and Elizabeth Wieland for assistance with data collection and organization. We are also grateful to Barbara Tillmann for comments that greatly improved this manuscript and to Björn Herrmann for fruitful discussion related to permutation-based analyses.

Correspondence concerning this article should be addressed to Molly J. Henry, Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstrasse 1a, Leipzig, Germany. Email: henry@cbs.mpg.de or to J. Devin McAuley, Department of Psychology, Michigan State University, East Lansing, MI 48824. E-mail: dmcauley@msu.edu

References

Aly, M., Knight, R. T., & Yonelinas, A. P. (2010). Faces are special but not too special: Spared face recognition in amnesia is based on familiarity. *Neuropsychologia, 48*, 3941-3948.

Ayotte, J., Peretz, I., & Hyde, K. (2002). Congenital amusia: A group study of adults afflicted with a music-specific disorder. *Brain, 125*, 238-251.

Crowe, E., & Higgins, E. T. (1997). Regulatory focus and strategic inclinations: Promotion and prevention in decision-making. *Organizational Behavior and Human Decision Processes, 69*, 117-132.

Cuddy, L. L., Balkwill, L. L., Peretz, I., & Holden, R. R. (2005). Musical difficulties are rare: A study of "tone deafness" among university students. *Annals of the New York Academy of Sciences, 1060*, 311-324.

Douglas, K. M., & Bilkey, D. K. (2007). Amusia is associated with deficits in spatial processing. *Nature Neuroscience, 10*, 915-921.

Durlach, N. I., & Braida, L. D. (1969). Intensity perception I: Preliminary theory of intensity resolution. *Journal of the Acoustical Society of America, 46*, 372-383.

Ernst, M. D. (2004). Permutation methods: A basis for exact inference. *Statistical Science, 19*, 676-685.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Henry, M. J., & McAuley, J. D. (2010). On the prevalence of congenital amusia. *Music Perception, 27*, 413-418.

Hyde, K. L., & Peretz, I. (2004). Brains that are out of tune but in time. *Psychological Science, 15*, 356-360.

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition, 10*, 294-340.

Loui, P., Alsop, D., & Schlaug, G. (2009). Tone deafness: A new disconnection syndrome? *The Journal of Neuroscience, 29*, 10215-10220.

Loui, P., Guenther, F., Mathys, C., & Schlaug, G. (2008). Action–perception mismatch in tone-deafness. *Current Biology, 18*, 331-332.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, NJ: Lawrence Erlbaum Associates.

McAuley, J. D., Henry, M. J., & Tuft, S. (2011). Musician advantages in music perception: An issue of motivation, not just ability. *Music Perception, 28*, 505-518.

McDonald, C., & Stewart, L. (2008). Uses and functions of music in congenital amusia. *Music Perception, 25*, 345-355.

Omigie, D., & Stewart, L. (2011). Preserved statistical learning of tonal and linguistic material in congenital amusia. *Frontiers in Psychology, 2*, 1-11.

Parks, C. M., DeCarli, C., Jacoby, L. L., & Yonelinas, A. P. (2010). Aging effects on recollection and familiarity: The role of white matter hyperintensities. *Aging, Neuropsychology, Cognition, 17*, 422-438.

Patel, A. D. (2008). *Music, language, and the brain*. New York: Oxford University Press.

Peretz, I., Brattico, E., Järvenpää, M., & Tervaniemi, M. (2009). The amusic brain: In tune, out of key, and unaware. *Brain, 132*, 1277-1286.

Peretz, I., Champod, A. S., & Hyde, K. L. (2003). Varieties of musical disorders: The Montreal Battery of Evaluation of Amusia. *Annals of the New York Academy of Sciences, 999*, 58-75.

Peretz, I., Gosselin, N., Tillman, B., Cuddy, L. L., Gagnon, B., Trimmer, C. G. et al. (2008). On-line identification of congenital amusia. *Music Perception, 25*, 331-343.

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Experimental generality of data from recognition memory receiver-operating characteristic functions and implications for the glocal memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 763-785.

Rousseau, B., Rogeaux, M., & O'Mahoney, M. (1999). Mustard discrimination by same-different and triangle tests: Aspects of irritation, memory, and t criteria. *Food Quality and Preference, 10*, 173-184.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117*, 34-50.

Tillmann, B., Schulze, K., & Foxton, J. (2009). Congenital amusia: A short-term memory deficit for non-verbal, but not verbal sounds. *Brain and Cognition, 71*, 259-264.

Tunney, R. J., & Shanks, D. R. (2003). Subjective measures of awareness and implicit cognition. *Memory and Cognition, 31*, 1060-1071.

Vann, S. D., Tsilvilis, D., Denby, C. E., Quamme, J. R., Yonelinas, A. P., Aggleton, J. P. et al. (2009). Impaired recollection but spared familiarity in patients with extended hippocampal system damage revealed by 3 convergent methods. *Proceedings of the National Academy of Sciences, 106*, 5442-5447.

Williamson, V. J., Liu, F., Peryer, G., Grierson, M., & Stewart, L. (2012). Perception and action de-coupling in congenital amusia: Sensitivity to task. *Neuropsychologia, 50*, 172-180.

Williamson, V. J., McDonald, C., Deutsch, D., Griffiths, T. D., & Stewart, L. (2010). Faster decline of pitch memory over time in congenital amusia. *Advances in Cognitive Psychology, 6*, 15-22.

Yonelinas, A. P. (1999). The contribution of recollection and familiarity to recognition and source-memory judgments: A formal dual-process model and an analysis of receiver operator characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1415-1434.

Appendix

```
function [Az da] = roc_analysis(yesratings,noratings)
% ------------------------------------------------ %
% Molly J. Henry & J. Devin McAuley
% henry@cbs.mpg.de, mcauley@msu.edu
%
% This function calculates ROC curves and nonparametric ROC-based dependent
% measures s (zROC slope), da, and Az, from confidence-rating data for a
% yes-no signal detection design.
%
% See Henry & McAuley, ''Failure to apply signal detection theory to the
% Montreal Battery of Evaluation of Amusia may misdiagnose amusia'' for a
% full description of the dependent measures.
%
% Inputs:
% yesratings: A vector containing raw confidence rating counts for yes
%  trials
% noratings: A vector containing raw confidence rating counts for no trials
% totalyes: The total number of presented trials from the yes category
% totalno: The total number of presented trials from the no category
%
% Note: Vectors of ratings must contain a minimum of 4 elements.
%
% Outputs:
% Az: A nonparametric, bias-free analogue to proportion correct (PC).
% da: A nonparametric, bias-free analogue to d'.
%
% ------------------------------------------------ %

% Check inputs
if length(yesratings) <= 3 || length(noratings) <= 3; disp('Confidence ratings must be more than
three values!'); return; end;
yesratings = yesratings(:); noratings = noratings(:);

% Calculate cumulative response proportions
pyes = yesratings ./ sum(yesratings);
pno = noratings ./ sum(noratings);
cpyes = pyes; cpno = pno;
for ii = 1:length(pyes)-1
 cpyes(ii+1,1) = cpyes(ii+1,1) + cpyes(ii,1);
 cpno(ii+1,1) = cpno(ii+1,1) + cpno(ii,1);
end
% Calculate zROCs and slopes
cpyes = cpyes(2:end-1,1); cpno = cpno(2:end-1,1);
 % First, correct for 0s and 1s
 cpyes(find(cpyes == 0)) = 1 / 2*(sum(yesratings));
 cpyes(find(cpyes == 1)) = 1 - (1 / 2*(sum(yesratings)));
 cpno(find(cpno == 0)) = 1 / 2*(sum(noratings));
 cpno(find(cpno == 1)) = 1 - (1 / 2*(sum(noratings)));
zyes = norminv(cpyes); zno = norminv(cpno);
C = polyfit(zno,zyes,1);
s = C(1,1); int = C(1,2);

% Calculate dependent measures
da = ((2/(1 + s^2))^0.5)*int;
Az = normcdf((da/sqrt(2)));

end
```