Proceedings of the Ninth Triennial Conference of the European Society for the Cognitive Sciences of Music, 17-22 August 2015, Manchester, UK
Ginsborg, J., Lamont, A., Phillips, M., Bramley, S. (Editors)

Web-based testing of congenital amusia with the *Montreal Battery of Evaluation of Amusia*

Jasmin Pfeifer,[*#1] Silke Hamann [*2]

[*] *Amsterdam Center for Language and Communication, University of Amsterdam, The Netherlands*
[#] *Institute for Language and Information, University of Düsseldorf, Germany*
[1]j.pfeifer@uva.nl, [2] silke.hamann@uva.nl

ABSTRACT

In this article we present the results of a web-based testing of 117 German undergraduate students with the Montreal Battery of Evaluation of Amusia (MBEA; Peretz et al. 2003). The MBEA is used to assess congenital amusia, a neuro-developmental disorder present in approximately 4% of the population, according to Kalmus & Fry (1980). Recently, criticism has arisen concerning the usage of the MBEA in relation to the prevalence of congenital amusia in the general population and the statistical evaluation of the results (Henry & McAuley, 2010; 2013, Pfeifer & Hamann 2015).

We compare the results of our web-based study to a group of 111 German students that was tested with a computer-implemented MBEA version under laboratory conditions (Pfeifer & Hamann 2015). We found significant differences between the web-based and the laboratory group based on their sum of correct responses. A Signal Detection Theory analysis of the data, which factors out response bias, however, shows that the discriminatory ability of both groups seems to be fairly similar. The results of the current study are used to critically discuss the validity of a web-based MBEA specifically but also web-based testing more generally as a means of diagnosing congenital amusia.

I. INTRODUCTION

Congenital amusia is an innate perceptual disorder affecting the perception of both music and speech. This disorder is not caused by a hearing deficit or any form of brain lesion (Ayotte, Peretz & Hyde, 2002). As the exact neural underpinnings are still under investigation, no neurological markers can be used to diagnose congenital amusia. Instead, behavioral markers such as pitch perception deficits and a pitch memory deficit are employed. The main tool used to diagnose amusia nowadays is the *Montreal Battery of Evaluation of Amusia* (MBEA; Peretz, Chambod & Hyde, 2003), which was originally developed to confirm acquired amusia in patients with brain lesions.

Peretz *et al.* (2003) used the MBEA to test 160 participants without known neurological problems, who were not selected for musical ability. For each participant, the number of correct responses per MBEA subtest and an average score of the six subtests were calculated. As cut-off scores for congenital amusia, Peretz *et al.* propose 2 standard deviation (SD) below the mean of the 160 participants, thus an average score of correct responses below 21.6, or 76.6%. According to Peretz et al. (2003: 65), the MBEA subtests provide a sensitive measure since less than 20% of their participants obtained perfect scores for each subtest, only 3% had a perfect score for all subtests and less than 2% had average scores that were below 2 SD of the mean (i.e. were diagnosed as amusics). These average scores approximate a normal distribution, though the scores for the individual subtests display a skew to the right.

While the MBEA is mostly conducted in a laboratory, there are two exceptions described in the literature that employ the MBEA in online testing. Stewart and colleagues use two MBEA subtests for pretesting potential congenital amusics via the web (see e.g. Liu et al. 2010, 2013; Williamson & Stewart 2010). Peretz *et al.* (2008) designed a web-based amusia test based largely on the MBEA, which was also employed by Provost (2011). This test is considerably shorter than the MBEA (only 3 subtests with 72 melodies in total) and participants have to spot incongruities in these melodies (off-beat or out-of-key tones) rather than comparing two melodies as in the MBEA. Peretz *et al.* (2008) used the MBEA to validate this web test and found that 19% of people diagnosed as amusic with the MBEA in a laboratory would not have been diagnosed as such with the web-test. This result contradicts the expectations that participants tested online should perform equally well or slightly worse than lab-tested participants due to uncontrolled testing conditions (such as noise, unrestricted amount and length of breaks, etc.). Peretz *et al.* explain their findings with the difference in task between the two tests: for the MBEA (tested in the lab), participants have to compare melodies, which is more demanding than the online test of detecting incongruities, because it requires participants to hold pitch information in their working memory, while the web test does not involve working memory.

A discrepancy between web-based and laboratory results occurs quite often in psychological research, and Krantz & Dalal (2000) comment that this does not demonstrate a lack of validity of web-based experiments, since most variables seem not to be influenced by varying environments. However, for auditory research a stable and quiet environment is crucial, as Krantz & Dalal (2000) point out, and therefore online testing in auditory research in general and a web-based assessment of amusia in particular might be problematic and lead to misdiagnoses.

In the present study, we compare the results of a web-based testing to a testing in the laboratory, where we employ exactly the same test (the full MBEA) in both conditions. Participants were German undergraduate students who had to participate in the experiment to obtain course credits. The data of both testing methods were collected by Pfeifer & Hamann (2015), but only the results of the laboratory testing were analyzed in that study. In the present study, the results of the comparison are discussed together with general advantages and disadvantages of online testing of potential amusics. We propose the use of different cut-off scores for online testing and provide a list of criteria that should be controlled for when testing online.

II. METHOD

A. Participants

131 first year undergraduate students in general linguistics at the Heinrich Heine University Düsseldorf participated in our study for course credit. 14 participants dropped-out of the study nonetheless. A total of 117 participants remained and was analyzed. Of these, 23 reported technical difficulties but these participants finished the study nonetheless and their data were included as the problems were mostly related to the loading of the soundfiles.

The participants were not preselected for the presence or absence of musical disorders such as amusia, or specific levels of musical experience. All participants gave informed written consent to participate in this study and received course credit for their participation. All data were collected in accordance with the declaration of Helsinki.

All participants had normal hearing (as assessed by pure tone audiometry at 250-8000 Hz, where normal hearing was defined as a mean hearing level of 20 dB or less in both ears). 99 of the participants were female and 18 were male. 107 participants were right-handed, 9 left-handed, and 1 was ambidextrous. The age and years of (music) education of the participants can be found in Table 1.

Table 1. Participant details

	Age	Years of education	Years of music education
Mean	22	14.7	6.3
Range	19-36	12-22	0-17

B. Stimuli

The MBEA consists of six subtests, three testing melodic organization (scale, contour and interval subtest), two testing temporal organization (rhythm and meter subtest) and one assessing melodic memory (memory subtest). The musical phrases used in the six subtests were all specifically composed for the MBEA and follow the principles of the Western tonal system. For the metric test, the phrases are polyphonic and have a mean duration of 11 s, for the other five subtests they are monophonic and last 3.8 to 6.4 seconds (mean of 5.1 s). For a more detailed description of the stimuli, see Peretz et al. (2003).

C. Procedure

The participants were informed before the experiment that they should use headphones and take the test in a quiet environment without any distractions. For the first four and the sixth subtest, participants received two examples with feedback before the beginning of each subtest. For the fifth (meter) subtest, participants received four examples, instructing them what a march and a waltz sound like.

At a later point, the participants came to the laboratory for a hearing test and to fill in a questionnaire about their linguistic and musical background. A test administrator was present to answer clarification questions about the questionnaire. At this point, participants could ask questions about the nature of the study.

The procedure of the MBEA is the same for the first four subtests (scale, contour, interval and rhythm): The participants are presented with two practice trials and 31 experimental trials. A trial consists of a target melody and a comparison melody separated by a 2-second silent interval. Each trial is preceded by a warning tone and followed by a 5-second silent interval. 15 trials have comparison melodies that are identical to the target melody and 15 have comparison melodies that are altered in one note (see Peretz et al. 2003 for details). In addition to those 30 trials, each of the first four subtests contains a catch trial (where the pitch of the comparison melody was noticeably different) to ensure that the participants were paying attention and not simply guessing.

For the first four subtests, participants are asked whether the two melodies they hear are the same or different. In the meter subtest the participants have to judge whether the presented melody (a two-phrase sequence in duple or triple meter) is a march or a waltz. In the memory test, participants are also presented with single melodies, half of which already occurred in the previous subtests and they have to indicate for each melody whether they have heard it before during the previous subtests.

III. RESULTS

In the following, we will first provide the sum of correct responses of the web-based group and we will compare the results to that of the group tested under laboratory conditions from Pfeifer & Hamann (2015). We then provide the signal

Table 2. Results of the web-based group compared to mean scores of the laboratory group and percentage of amusics by Pfeifer & Hamann (2015).

Group		**Scale**	**Contour**	**Interval**	**Rhythm**	**Meter**	**Memory**	**Average**
Web-based	**Mean correct responses**	24.97	23.86	23.21	24.87	24.09	26.34	24.56
	SD	3.03	3.38	3.89	3.80	5.29	3.09	3.94
	Perfect score %	0.9	0.9	1.7	5.1	15.4	5.1	0
	Cut-off (2 SD)	18.91	17.1	15.43	17.27	13.51	20.16	16.68
	Cut-off (%)	63.03	57.00	51.43	57.57	45.03	67.20	55.60
	% below cut-off	5.1	5.1	4.3	8.5	6	6	6.7
	% below cut-off (cut-off scores Peretz et al. 2003)	20.5	34.2	29.9	26.5	37.6	14.5	34.6
Laboratory	**Mean correct responses**	24.95	24.68	24.32	25.84	26.07	27.51	25.56
	SD	2.73	3.01	3.29	2.64	3.65	1.77	3.09
	% below cut-off	4.5	7.2	7.2	4.5	6.3	7.2	5.4

Table 3. Distribution analysis for the web-based group per MBEA subtest. Bold indicates $p < 0.001$, italics $p < 0.05$.

Subtest	**Skew**	**SE Skew**	**z Skew**	**Kurtosis**	**SE Kurtosis**	**z Kurtosis**	**Kolmogorov-Smirnov Test**	
							D	**p**
Scale	-0.93	0.22	***-4.18***	0.61	0.44	1.36	0.16	***0.00***
Contour	-0.45	0.22	*-1.99*	-0.43	0.44	-0.96	0.12	***0.00***
Interval	-0.47	0.22	*-2.09*	-0.40	0.44	-0.90	0.09	*0.01*
Rhythm	-0.96	0.22	***-4.31***	0.28	0.44	0.64	0.15	***0.00***
Meter	-0.85	0.22	***-3.80***	0.13	0.44	0.29	0.15	***0.00***
Memory	-1.39	0.22	***-6.22***	1.78	0.44	***4.01***	0.20	***0.00***
Average	-0.94	0.22	***-4.18***	0.62	0.44	1.39	0.28	***0.00***

detection measures *d'* and *c* (criterion location) for the data of the web-based testing.

A. Web-based versus laboratory testing

The mean of correct responses, SD and the percentage of participants below cut-off for both the web-tested group and the group tested in the lab are given in Table 2. The cut-off scores are calculated based on our mean minus 2 SD. For the web-based sample, we additionally calculated the percentage of participants below cut-off employing the cut-off scores established by Peretz *et al.* (2003).

The mean of correct responses for the web-based group is generally lower than for the lab-tested group (though it is almost identical for the scale subtest) and the web-based group shows more variation (SD is larger for every subtest). Based on the mean of correct responses and on the average across all subtests, 6.7% of the web-based participants would be diagnosed as amusics because their mean correct scores fall below the cut-off score of 16.68 (or 55.6%). For the laboratory-tested group, this would be only 5.4% of the participants. If the original cut-off scores from Peretz et al. (2003) were applied to our data, 34.6% of the web-tested participants would be categorized as amusic.

Calculation of skew and kurtosis showed that all subtest scores and the average score for the web-tested group (like that of the lab-based group in Pfeifer & Hamann) exhibit a negative skew, indicating a build up of high scores, see values in Table 3. Especially the memory subtest exhibits a significant (p < 0.001) kurtosis value, indicating that it is not normally distributed. In addition, the Kolmogorov-Smirnov tests yielded significant results as well. The data are therefore not normally distributed and non-parametric statistics are required.

The variances between the web-based and the laboratory groups differed significantly for four of the six subtests (for the results of Levene's Test and Mann-Whitney-U tests, see Table 6 in Pfeifer & Hamann 2015). The contour and interval subtest and the average of all subtests reached significance at $p < 0.05$ and the meter and memory subtests reached significance at $p < 0.01$.

Because of this difference in variances, Pfeifer and Hamann (2015) only analyzed the laboratory group. In the present study, the results of the web-based group are further examined.

B. Analysis of web-based scores with Signal Detection Theory

As was shown by Henry & McAuley (2013) and Pfeifer & Hamann (2015), signal detection theory (Green & Swets 1966; Macmillan & Creelman 2005) seems to be a more appropriate scoring procedure for the MBEA, as it offers a bias-free measure of discriminatory ability. Table 4 gives the means and standard deviations of *d'* and *c* for the web-based group.

An analysis of skew and kurtosis of *d'* showed that the mean scores on the scale, contour, interval and rhythm test are normally distributed, while meter and average mean scores exhibit a significant (p <0.05) negative kurtosis value and the memory subtest exhibits a highly significant skew (p <0.001).

Based on the discriminatory ability (*d'*), cut-off scores were calculated, once with mean minus 1 SD, once with mean minus 2D. The obtained percentage of amusics varies

Table 4. Means and SD of *c* and *d'* for the web-based group, including cut-off scores and z-scores used for normality analysis. Bold indicates *p* < .001, italics *p* < .05.

		Scale	Contour	Interval	Rhythm	Meter	Memory	Average
c	**Mean correct responses**	-0.07	0.17	0.39	0.09	-0.11	-0.27	0.03
	SD	0.52	0.50	0.58	0.55	0.33	0.46	0.54
d'	**Mean correct responses**	2.34	2.00	1.95	2.40	2.27	2.88	2.31
	SD	0.93	0.93	1.04	1.12	1.59	1.03	1.17
	z skew	-0.68	1.58	0.66	-0.56	-0.07	***-2.89***	-0.51
	z kurtosis	-1.13	-0.55	-0.65	-0.81	*-2.28*	-0.16	*-2.54*
% below cut-off (Mean – 1 SD)		17.90	12.80	17.10	15.40	15.40	17.10	16.50
% below cut-off (Mean – 2 SD)		1.70	2.60	1.70	1.70	0.90	2.60	1.40

accordingly (see the last two rows in Table 4). Which cut-off score to use is an arbitrary statistical decision, and is therefore not further discussed in the present paper.

IV. DISCUSSION

In the current paper we analyzed a sample of 117 individuals tested with a web-implemented version of the MBEA (Peretz *et al.* 2003). The data were collected as part of a larger study with a comparison between a laboratory group and this web-based sample. Parts of these data were already compared to a group tested in a laboratory (N = 111), c.f. Pfeifer & Hamann (2015). However, Pfeifer & Hamann focussed their analysis on the laboratory group and discussed in detail the problem of different applied cut-off scores and the possible existence of different amusic subtypes. In the present discussion, we will first focus on the results of the group tested with the MBEA via the web compared to the lab-tested group by Pfeifer & Hamann, then move to implications and limitations of web-based testing with the MBEA and finally discuss general limitations of web-based testing in auditory research.

A. Web-based vs. lab-based testing with the MBEA

Before looking in detail at our results, we have to note that, like in the studies by Wise (2009) and Pfeifer & Hamann (2015), the MBEA cut-off scores proposed by Peretz *et al.* (2003) yielded a very high and improbable percentage of amusics (34.6%) for our web-based group. We therefore used the cut-off scores calculated on the basis of our own data.

For the scoring based on the sum of correct responses, a slightly higher proportion of web-tested participants fell below the cut-off score and thus was diagnosed as amusic (6.7%) than for the group tested in the lab (5.4%). This contrasts with the findings by Peretz *et al.* (2008) who report that 19% of their participants were diagnosed as amusic in the laboratory would have been missed as such by their online test. As explained by Peretz *et al.*, their findings are due to a difference in task: whereas for the on-line test participants had to spot possible incongruities in melodies, for the lab-used MBEA they had to compare two melodies at a time, which is more demanding as it requires the storage of pitch information in the working memory. For the present comparison between lab and web-based testing we used the same test (the full MBEA), hence the differences we found have to be attributed to the testing method.

When looking at the differences in performance in the individual subtests, we found that the scoring based on the sum of correct responses yielded non-normally distributed results that are highly negatively skewed for all subtests. This is in accordance with the findings by Wise (2009) and Henry & McAuley (2010), and led us to use non-parametric statistics for the comparison of the two groups. Mann-Whitney-U tests revealed significant differences between the groups on the contour, interval, meter and memory subtests as well as on the average score, with the performance on the web-based version being worse.

For the further analysis of the scores we then employed the signal detection theory (SDT) measures *d'* and *c* (as suggested by Henry & McAuley 2013). For the lab-based sample tested by Pfeifer & Hamann (2015), the *d'* scores for all subtests and the average score were all distributed normally, indicating that the discriminatory ability of this group was fairly consistent. For the web-based group tested in the present study, only four of the six subtests are distributed normally. The meter subtest exhibits a significant kurtosis value, while the distribution of scores on the memory subtest is highly significantly negatively skewed and platykurtic, i.e. it contains many high scores but also exhibits a long-drawn tail to the left with low scores and an overall flat distribution. For these two subtests, the web-tested group thus shows less discriminatory abilities. Possible explanations for this difference in discriminatory power and also for the statistical difference in correct scores between the web-tested and the lab-tested groups for four of the subtests are discussed below in sections B on MBEA-related issues and section C on web-based testing in general.

B. The MBEA as a web-based test and its limitations

The MBEA was not designed for web-testing, and some of its properties do not seem to make it ideally suited as a web-based test for amusia. In this subsection, we discuss two of these properties, namely length and lack of measures to ensure participants' attention as possible reasons for the low performance of our participants on the last two subtests of the MBEA.

With respect to length, the whole MBEA takes on average 50 minutes to complete under laboratory conditions. While the majority of our participants also completed the web-based version within 50-60 minutes, some took over 90 minutes. This time seems to be too long for a web-based study, as the literature shows. Reips (2002) suggests "a few minutes", Honing & Ladinig (2008) 15 minutes, and Gingras *et al*. (2015) 30 minutes as preferred length for web tests. While Peretz et al. (2008) used a shorter web test loosely based on the MBEA, this yielded misdiagnoses in both directions, as discussed above. Some studies (e.g. Liu et al. 2010, 2013; Williamson and Stewart 2010) only use two subtests of the MBEA for pretesting participants via the web, as mentioned in the introduction.

In order to ensure the participants' attention, the MBEA contains four catch trails (as described in the methods section). All of our participants that finished the web-based test had scored correctly on all four catch trials. However, it is a relatively low number of catch trials and the manipulation in the catch trails stands out so much from the experimental manipulation that anyone paying only the slightest bit of attention should be able to identify them. The catch trials are therefore not enough to ensure a participants' lasting attention, especially in web-testing, where a very quiet and non-distracting environment cannot be assured.

Both factors, length of test and inability to ensure participants' attention, could thus have led to a worse performance of the web-tested group. Especially lack of attention during web testing could contribute to the lower and non-normally distributed discriminatory ability on the last two subtests but especially on the memory subtest as this test relies on how much participants paid attention to and remember from the first four subtests.

C. General limitations of (auditory) web-based testing

There are a number of advantages as well as limitations of web-based testing that are not specific to the MBEA but apply to all, or at least all auditory, web-based studies. Especially the limitations will be discussed here (as possible explanation of our findings). These issues are not new and have been raised before (see for example Mehler 1999; Krantz & Dalal 2000; Reips 2002; Birnbaum 2004, and a discussion on the auditory mailing list: Auditory 2007) but we will discuss them in light of our experiences with the MBEA. Where solutions have been proposed in the literature (e.g. Reips 2002; Birnbaum 2004), they are also outlined.

The obvious **advantages** of web-based testing are that it is relatively easy to gather large heterogeneous data samples as well as to reach specialized populations easily. This can be achieved at much lower costs and at a higher speed than in traditional laboratory settings, while offering a greater external validity and using more automated processes, which makes data analysis faster as well (Reips 2002). Furthermore, web-based studies using highly standardized procedures are easily replicated (Birnbaum 2004) and provide a much more natural listening setting (Honing & Ladinig 2008) that avoids an experimenter bias, i.e. participants do not feel pressured to respond in one way or the other due to the presence of the experimenter.

Though motivation to take part in a study is usually an advantage of web-based testing, our participants had to take part in the study for course credit and therefore were not as intrinsically motivated to participate as other subjects in music-related studies (see e.g. Honing and Ladinig 2008 on their very positive experience with musically interested participants in music-related web testing). However, our laboratory participants also had to take part for course credit, therefore a low motivation cannot explain the difference in performance between the two groups.

The most prominent **disadvantages** of web-based testing in general are the high drop-out rates and multiple submission, both of which are threats to the internal validity of studies (c.f. Reips 2002; Birnbaum 2004).

Generally, a **drop-out rate** of 30-40% has been reported for web-based studies (Reips 2002). In our study, we observed a dropout rate of only 10.7%, which can be attributed to the fact that students had to participate in the study for course credit. In addition, we assured our participants of confidential handling of the data before the study. We also made them aware of the possibility of back-tracing data to participants, and the availability of an explanation of the aim of the experiment after participation, thereby showing that the data were actually analyzed and used in a scientific context. This latter fact greatly interested at least part of the students and many of them not only wanted to be informed about the general outcome of the study but also about their personal results.

Reips (2002) proposes the use of the so-called high hurdle technique against high drop-out rates. With this technique a web-study is designed in such a way that 'obstacles' that test participants' patience are put at the beginning, e.g. the collection of personal data or a screen that takes long to load. With this, it is hoped that all impatient participants or participants that are unwilling to provide personal information are filtered out before the beginning of the actual study. This method not necessarily reduces the drop-out rate but ensures that uninterested participants drop out as early as possible thereby avoiding incomplete datasets. Other measures against a high drop-out rate can be the promise of rewards or a design of the test that is visually appealing or intellectually challenging, as is recommended by Honing & Ladinig (2008) However, as Reips (2002) points out, experiments that are too interesting or engaging might provoke **multiple submissions**. Our MBEA-online version was designed in such a way that it exactly mirrored the visually rather plain instruction screens that were used for the computerized laboratory version of the MBEA implemented with Praat (Boersma & Weenink 2011), again not tempting our participants to perform the test several times. Reips (2002) makes several suggestions for the avoidance or the control of such multiple submissions, such as the collection of personal data for identification, the tracking of IP-addresses, the implementation of a username and password-dependent access. All of these were implemented in our study and we did not have a single multiple submission. The same is observed by Birnbaum (2004) who found only one instance of multiple submission in a dataset of 1000 submissions.

A further disadvantage of web-testing is the **lack of control** pertaining to technical factors (e.g. internet speed or usage of headphones) and environmental factors (like noise or distractions). Both can influence the data considerably (Mehler 1999; Auditory 2007). It has been argued that lack of control is not actually an issue and that web-based studies have a much greater external validity (Reips 2002; Gingras et al. 2015) through their large participant numbers that cancel out the possible noise in the data (McGraw et al 2000). However, Krantz and Dalal (2000) argue that for auditory research, a stable and quiet environment is crucial for the success of the experiment. Auditory (2007) shows that many researchers are in doubt about the use of web-based experiments in auditory research, since it cannot even be controlled whether subjects wear headphones or what the level of background noise is during the experiment.

Concerning internet connection and speed, Reips (2002) advises to pre-load all soundfiles. This takes longer at the beginning of the experiment but ensures then that the experiment can start and run smoothly. To ensure smooth running, experiments should be checked beforehand on different operating systems and with different browsers (Reips 2002). In our web-testing, we followed these recommendations, but nevertheless 23 of our participants encountered technical difficulties. These could mostly be resolved by updating browser versions or installing or updating plugins. However, this factor might have influenced the difference in performance between the two groups in our experiment.

Further issues on lack of control that clearly influenced our results are the environmental factors. We instructed all participants to wear headphones (not to use their computer speakers) and to take the experiment in a quiet room without distractions or interruptions. Whether they followed these instructions could not be checked. Furthermore, they were asked to finish the experiment in one instance and to only take a break when they needed one. As we logged the time of day during which participants took the test and how long they took for every subtest, we could see that data was submitted round-the-clock and that some participants took very long to finish certain blocks, not all did thus follow our instructions. A possible way around this last problem could be the exclusion of participants on the basis of such long times. The long breaks taken by some participants could contribute to the lower and non-normally distributed discriminatory ability especially on the memory subtest.

Lack of control of the environmental factors on the site of the experimenter is thus a crucial point that makes web-based testing unsuitable for the MBEA or for more than just a pre-screening.

A general point of concern with web-based testing not connected to performance is that **of security/privacy concerns**. Via http protocols or Javascript it is possible to track sensitive information about the participant: Which operating system and browser are used, screen resolution, loading times, which link referred them, and even the location can be tracked and logged. Participants need to be informed what data is or even can be collected about them and how it is being stored. However, this information is often not provided, which raises ethical concerns. Indeed, many ethics committees do not approve web-based studies and some journals will not accept web-based studies for publication (Auditory 2007; Honing & Ladinig 2008).

One last concern that is also related to ethics is important to consider when screening for amusia online, be it with the MBEA or any other kind of test, namely that of **diagnosis**. Most participants that will voluntarily seek out a web-based amusia test do this because they suspect that “there is something wrong” with them, i.e. that they have a perceptual deficit and are amusic. These participants naturally take a test like that to get to know their results on that test. However, it is questionable whether and how these participants should be informed of their results. In the case of the present study this was comparatively simple. Participants did not automatically receive their results. Pooled results were presented to all participants. More interested participants could request their personal results, which they were then given including a detailed explanation of what these results meant or did not mean. However, this is not possible with large online samples stemming from the general public. It is questionable whether a relatively simple web-based test should “diagnose” people with a life-long affliction. In a lot of cases when amusics were diagnosed in our laboratory, they were glad when they learned of their amusia because they finally knew “what was going on” with them. However, in a few cases it was almost traumatic for people and these people should not be confronted with a diagnosis like that while sitting alone in front of a computer screen without further explanation.

Finally, this last example also nicely exemplifies another issue of web-based testing: **Self-selection**. While it is argued that a more heterogeneous pool of participants can be reached via internet – and this is certainly true if compared to the normal psychology undergraduate participant pool – the sample one obtains might still be biased. Only people very interested in their own musicality or people doubting their musicality will actively seek out web-based musicality or amusia tests. This also yields a participant pool that does not reflect the normal population but rather two extremes.

While web-based testing thus offers many advantages and is suitable for many kinds of research, we would like to caution that it might not be suitable for the diagnosis of amusia (with the MBEA or another test). Web-based tests can certainly be used as pre-screening tools (as parts of the MBEA are used at the moment) and can be very useful as such. But for the various reasons outlined above, an amusia diagnosis, even if it is no medically recognized diagnosis, should not take place via a web-based test.

Concerning the MBEA, we showed that even though the sum of correct responses differed significantly between our web-tested and laboratory-tested groups, their discriminatory ability was relatively similar. Only the last two subtests showed differences between the two groups, but these can probably be attributed to some properties of the MBEA that make it in its entirety unsuitable for online testing.

ACKNOWLEDGMENT

We would like to thank Henk Schotel for his help with the web-based implementation of the MBEA. This work was supported by the Netherlands Organisation for Scientific Research (NWO) (grant number 322-75-004) to the first author.

REFERENCES

Auditory (2007). *Online listening tests and psychoacoustic experiments with large N*. Discussion on Auditory mailing list, see http://www.auditory.org/mhonarc/2007/threads.html#00531.

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

Boersma, P., & Weenink, D. (2011). Praat: doing phonetics by computer. 5.2.25 ed.

Birnbaum, M. H. (2004). Human research and data collection via the

Gingras, B., Honing, H., Peretz, I., Trainor, L., & Fisher, S. E. (2015). Defining the biological bases of individual differences in musicality. *Philosophical Transactions of the Royal Society B, 370,* 20140092.

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

Hamann, S., Exter, M., Pfeifer, J., & Krause-Burmester, M. (2012). Perceiving Differences in Linguistic and Non-Linguistic Pitch: A Pilot Study With German Congenital Amusics. In E. Cambouropoulos, C. Tsougras, P. Mavromatis & K. Pastiadis (Eds.), *Proceedings of the 12th International Conference on Music Perception and Cognition* (pp. 398-405). Thessaloniki.

Henry, M.J., & Mcauley, J.D. (2010). On the Prevalence of Congenital Amusia. *Music Perception: An Interdisciplinary Journal, 27*, 413-418.

Henry, M.J., & Mcauley, J.D. (2013). Failure to Apply Signal Detection Theory to the Montreal Battery of Evaluation of Amusia May Misdiagnose Amusia. *Music Perception: An Interdisciplinary Journal, 30*, 480-496.

Honing, H., & Ladinig, O. (2008). The potential of the internet for music perception research: A comment on lab-based versus web-based studies. *Empirical Musicology Review, 3*, 4-7.

Kalmus, H., & Fry, D.B. (1980). On tune deafness (dysmelodia): Frequency, development, genetics and musical background. *Annals of Human Genetics, 43*, 369-383.

Krantz, J.H., & Dalal, R. (2000). Validity of Web-Based Psychological Research. In M.H. Birnbaum (Ed.), *Psychological Experiments on the Internet* (pp. 35-60). San Diego: Academic Press.

Liu, F., Jiang, C., Pfordresher, P.Q., Mantell, J.T., Xu, Y., Yang, Y., & Stewart, L. (2013). Individuals with congenital amusia imitate pitches more accurately in singing than in speaking: Implications for music and language processing. *Attention, Perception, & Psychophysics, 75*, 1783-1798.

Liu, F., Patel, A.D., Fourcin, A., & Stewart, L. (2010). Intonation processing in congenital amusia: discrimination, identification and imitation. *Brain, 133*, 1682-1693.

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

McGraw, K., Tew, M. D., & Williams, J. E. (2000). The integrity of web-delivered experiments: Can you trust the data? *Psychological Science, 11*, 502-506.

Mehler, J. (1999). Editorial. *Cognition, 71*, 187-189.

Mullensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PLoS ONE, 9*, e89642.

Peretz, I., Champod, S., & Hyde, K. (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., Paquette, S., & Bouchard, B. (2008). On-line Identification of Congenital Amusia. *Music Perception, 25*, 331-343.

Pfeifer, J., & Hamann, S. (2015). Revising the diagnosis of congenital amusia with the *Montreal Battery of Evaluation of Amusia. Frontiers in Human Neuroscience, 9, 161*.

Provost, M. (2011). *The Prevalence of Congenital Amusia*. Master's thesis, Université de Montréal.

Reips, U.-D. (2000). The Web Experiment Method: Advantages, Disadvantages, and Solutions. In M.H. Birnbaum (Ed.), *Psychological Experiments on the Internet* (pp. 89-117). San Diego: Academic Press.

Reips, U.-D. (2000). Standards for internet-based experimenting. *Experimental Psychology, 49*, 243-256.

Williamson, V.J., & Stewart, L. (2010). Memory for pitch in congenital amusia: Beyond a fine-grained pitch discrimination problem. *Memory, 18*, 657-669.

Wise, K. (2009). *Understanding "tone deafness": A multi-componential analysis of perception, cognition, singing and self-perception in adults reporting musical difficulties*. PhD dissertation, Keele University.