Published in final edited form as:

Hear Res. 2011 December ; 282(1-2): 265–271. doi:10.1016/j.heares.2011.07.004.

Reliability of Distortion-Product Otoacoustic Emissions in the Common Marmoset (*Callithrix jacchus*)

Michelle D. Valero and Rama Ratnam

University of Texas at San Antonio, TX 78249, Department of Biology

Abstract

This study examines the test-retest reliability of distortion-product otoacoustic emissions (DPOAEs) in ketamine-anesthetized common marmosets (*Callithrix jacchus*). DPOAE gain functions were measured at 16 f_2 -frequencies between 3 – 24 kHz. Test-retest reliability was assessed at the following time intervals: 1) Interleaved, in which two gain functions were obtained at each frequency before advancing to the next frequency, 2) Immediate, wherein one gain function was collected at all f_2 -frequencies and the retest was immediately performed without removing the probe tip, 3) Short-term, in which the retest followed a 10-min period with the probe removed, and 4) Long-term, wherein the retest was performed at least one week after the initial test. Reliability was assessed using four correlation coefficients used in the literature. Test-retest reliability was best in the interleaved interval and worst in the short-term interval. In general, reliability was best when primary-tone levels were high. Correlation coefficients decreased at frequencies above 12-kHz in the short-term and long-term intervals, but the decrease was more substantial in females than in males in the long-term interval. At frequencies below 12 kHz, same-day measurements (2, 3) were less repeatable, regardless of whether the probe was removed, which may be due to time under anesthesia. These results have implications for DPOAE studies where repeated measures are required and when treatment or group differences are small.

Keywords

Distortion-product otoacoustic emissions; DPOAE; repeatability; reliability; primate; common marmoset; normal hearing; sex difference

1. Introduction

In an earlier study of distortion-product otoacoustic emissions (DPOAEs) in the common marmoset (*Callithrix jacchus*), we noted that the mean DPOAE levels of six female marmosets were larger than those measured in six males by up to 12 dB (Valero *et al.*, 2008). Observations of sex differences in DPOAEs are often reported, but typically the difference is much smaller than those observed in marmosets (see McFadden *et al.*, 2009 for a review). This prompted the question: is this large sex difference real, and more generally, how reproducible are DPOAE measurements in marmosets? To answer this question, we assessed variability in marmoset DPOAE levels so as to distinguish normal variability (due to experimental measurement error) from a meaningful difference in DPOAE level (L_{DP}).

Test/retest reliability of DPOAEs has been examined thoroughly in humans to assess the clinical usefulness of this measure (for recent studies, see Sockalingam *et al.*, 2007; Wagner *et al.*, 2008; Stuart *et al.*, 2009; Keppler *et al.*, 2010). However, reliability of DPOAEs has been examined less thoroughly in animals, and in only a few animal models (Harding *et al.*, 2002; Rasmussen *et al.*, 2005; Parrazini *et al.*, 2006; Park *et al.*, 1995). In many studies, test/retest reliability was assessed using correlation coefficients, including both *interclass*

correlation coefficients (*i.e.*, Pearson's product moment) and several *intraclass* correlation coefficients (for a review of intraclass correlation coefficients, see McGraw & Wong, 1996). These methods were used here to assess reproducibility of marmoset DPOAEs and are discussed in the current report.

The general findings of this paper are in agreement with previous reports of reliability in humans. DPOAEs are more stable at mid-range frequencies and high primary-tone levels (e.g., Franklin *et al.*, 1992; Roede *et al.*, 1993; Zhao & Stephens, 1999; Wagner *et al.*, 2008; Stuart *et al.*, 2009; Keppler *et al.*, 2010), and DPOAEs are more repeatable when the probe has not been removed and replaced (e.g., Zhao & Stephens, 1999; Wagner *et al.*, 2008). Further, a sex difference was observed in the current group of marmosets, but the difference was smaller than previously reported, and more in line with human studies (see McFadden *et al.*, 2009). A key finding was that DPOAEs were less variable in male marmosets over long-term intervals when compared to females. This may explain the discrepancy in the sex differences between our current and previous studies (Valero *et al.*, 2008).

2. Materials and Methods

2.1 Subjects

We measured DPOAEs in the right ear of 12 adult common marmosets (6 males, 6 females) housed at the Southwest National Primate Research Center (SNPRC) at the Texas Biomedical Research Institute (TBRI; formerly the Southwest Foundation for Biomedical Research) in San Antonio, TX. All subjects were housed in a single room in individual cages with the exception of one female who was housed with her sibling in a separate building. DPOAEs were measured in a sound attenuating chamber (MAC-3, Industrial Acoustics Company, Inc.). Subjects were anesthetized with intramuscular injections of ketamine hydrochloride (Ketaset, $\mu = 25.9$ mg/kg) and butorphanol tartrate ($\mu = 1.40$ mg/kg). Middle ear health was confirmed otoscopically and any cerumen in the external auditory canal was removed. Subjects were positioned prone upon a non-electric heat source during measurements, head position was secured, and rectal temperature was monitored. An ER10B+ (Etymotic Research, Inc.) probe was fitted tightly in the ear canal with a rubber immittance tip and secured with an adjacent stand. These procedures were approved by the TBRI Institutional Animal Care and Use Committee (IACUC; 1044-CJ-4).

2.2 Procedures and Equipment

Acoustic stimuli were delivered by an analog input/output device (PXI-4461, National Instruments, Inc.) housed in a National Instruments chassis (PXI-1042Q). Primary tones were delivered separately through two high-frequency speakers (EC-1, Tucker-Davis Technologies, Inc.), which were powered by an electrostatic speaker driver (ED-1, Tucker-Davis Technologies, Inc.). The EC-1 speakers were coupled to the subject via two separate speaker ports in the ER10B+ (Etymotic Research, Inc.). The microphone output of the ER10B+ was amplified by 20 dB by a low-noise preamplifier (Etymotic Research, Inc.) and digitized by the PXI-4461 at a 50-kHz sampling rate with 24-bit resolution.

The DPOAE procedures used here were a modified CUBDIS procedure (Allen, 1990) and have been described previously (Valero *et al.*, 2008). Briefly, we verified that we had achieved a tight seal in the ear canal by presenting a single click and measuring the frequency response of the subject's ear canal. A smooth curve suggested a tight fit in the ear canal. The primary tones were calibrated at each frequency step to generate each tone at 65 dB SPL in the ear canal, and the voltage output was adjusted to change the levels of the primary tones from $L_1 = 38 - 74$ dB SPL ($L_1 - L_2 = 10$). The recorded primary tone levels had to be within 2 dB of the target primary-tone level or calibration was repeated. We

selected the frequencies of the primary tones and the DPOAE so that each 1024-point segment contained an integral number of cycles. The f_2/f_1 ratio was as close to the target value (1.21) as this sampling constraint allowed. In order to remove possible artifacts, the first two segments and the strongest 15% of the other segments (taken from a histogram of RMS values) were excluded from the final average. The remaining segments (about 165) were averaged in the time domain with a sliding window of 2048 points and 50% overlap. A power spectrum was computed from the averaged waveform. The power in the frequency bin containing the DPOAE and the primary tones were converted to SPL. The power in six frequency bins surrounding the DPOAE frequency (but separated from the DPOAE frequency by at least 3 bins) was averaged and converted to SPL to provide an estimate of the noise floor.

2.3 Testing schedule

Primary tones ranging from $f_2 = 3 - 24$ kHz were presented in 1/5-octave steps for a total of 16 frequencies ($f_2/f_1 = 1.21$; Valero *et al.*, 2008). At each frequency, the level of the primary tones were increased from $L_1 = 38$ to 74 dB SPL ($L_1 - L_2 = 10$) in 6-dB increments to produce a gain function (L_{DP} vs. primary-tone level).

For the *interleaved* time-interval, two gain functions (test/retest) were measured at each frequency before proceeding to the next frequency. In the *immediate* time-interval the retest immediately followed the initial test without repositioning of the probe tip. In the *short-term* interval the probe tip was removed for 10 minutes and then replaced just prior to the retest. The test and retest in the *long-term* time interval were performed on different days, but were always the initial test of the day. The tests in this interval were selected to minimize time between tests, which ranged from 0.3 - 4.3 months (mean = 1.8 ± 1.4 mo).

2.4 Data Analysis

DPOAEs met our inclusion criteria only if the L_{DP} in both the test and retest exceeded -20 dB SPL and were greater than 6 dB above the measured noise floor for a given trial (e.g., Dreisbach *et al.*, 2006). We found this to be a stricter criterion than when L_{DP} had to exceed -20 dB and one SD above the mean noise floor (e.g., Cacace *et al.*, 1996).

Four correlation coefficients were calculated. The Pearson-product moment correlation, r (an *interclass* correlation coefficient) was calculated from

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}, \quad (1)$$

where x and y correspond to test and retest values. Three *intraclass* correlation coefficients; ICC(A,1), ICC(C,1), and ICC(C,k); the last of which is equivalent to Cronbach's α (e.g., McGraw & Wong, 1996), were calculated from

$$ICC(A, 1) = \frac{\sigma_r^2 - \sigma_{rc}^2 / (k - 1)}{\sigma_r^2 + \sigma_c^2 + (\sigma_{rc}^2 + \sigma_e^2)} = \frac{MS_R - MS_E}{MS_R + (k - 1)MS_E + \frac{k(MS_C - MS_E)}{n}}, \quad (2)$$

$$ICC(C, 1) = \frac{\sigma_r^2 - \sigma_{rc}^2 / (k - 1)}{\sigma_r^2 + (\sigma_{rc}^2 + \sigma_e^2)} = \frac{MS_R - MS_E}{MS_R + (k - 1)MS_E}, \quad (3)$$

$$ICC(C, k) \equiv \alpha = \frac{\sigma_r^2}{\sigma_r^2 + (\sigma_{rc}^2 + \sigma_e^2)/k} = \frac{MS_R - MS_E}{MS_R}, \quad (4)$$

where σ_r^2 is the row variance, σ_{rc}^2 represent the interaction variance, σ_c^2 is column variance, σ_e^2 is the variance due to error, MS is the mean square of each source of variance (MS subscripts follow those explained above for variance), n is the number of values in each column, and k is the number of columns (trials).

Pearson's r , Cronbach's α , ICC(A,1) and ICC(C,1) were examined in each time interval as a function of primary-tone frequency and level. When examined as a function of frequency, data collected with $L_1 = 50$ to 74 dB SPL were pooled for each subject, so that each subject had a maximum of 5 values in each column (excepting those that did not met our inclusion criteria). When examined as a function of primary-tone level, data were pooled across each frequency, so that each subject had a maximum of 16 data points per column (one point per frequency). These ICCs were used to assess the degree of absolute agreement (A,1) and consistency (C,1 and k) between test and retest (McGraw & Wong, 1996). These were single score ICCs calculated from the two-way, repeated measures ANOVA table which was constructed from pooled data.

The first measurement of each test day was averaged for each subject ($n = 3 - 5$ trials per subject, mean = 3.9, mode = 4) and then grouped according to sex to assess any differences in the mean group L_{DP} . The ICC(C,1) reported for males vs. females was calculated using the same test and retest measurements that were used for the group long-term analyses.

3. Results

Mean DPOAE audiograms (DPgrams; L_{DP} vs. f_2 -frequency) measured with $L_1 = 50$ and 74 dB are shown for the interleaved (Fig. 1A), immediate (Fig. 1B), short-term (Fig. 1C), and long-term (Fig. 1D) intervals. DPgrams showed maxima at approximately 9 and 18 kHz and a notch between 12 – 13.7 kHz. The error bars represent one SD around the mean. Although the DPgrams reported here were measured at 1/5-octave intervals, they are in agreement with DPgrams measured in 1/2-octave steps from an earlier study (Valero *et al.*, 2008).

The mean L_{DP} of test and retest measurements were most similar in the interleaved time-intervals (Fig. 1A), which had the shortest intertrial interval (about 30 sec): the difference between mean L_{DP} s (ΔL_{DP}) at each frequency was 1.7 dB when $L_1 = 50$ dB and 0.6 dB when $L_1 = 74$ dB. In the immediate interval (Fig. 1B), ΔL_{DP} was larger than in the interleaved interval: 0.15 – 5.3 dB when $L_1 = 50$ dB SPL and 0.15 – 3.4 dB when $L_1 = 74$ dB SPL. ΔL_{DP} was largest in the short-term interval (Fig. 1C): 3.6 – 13.3 dB when $L_1 = 50$ dB SPL and 0.6 – 9.6 dB when $L_1 = 74$ dB SPL. When $L_1 = 74$ dB SPL, the ΔL_{DP} in the long-term interval (Fig. 1D) was 2.1 dB, and it ranged from 0.12 – 4.9 dB when $L_1 = 74$ dB.

We assessed the relationship between ΔL_{DP} and the number of subjects represented by each point. This was because DPOAE measurements were more repeatable in some subjects than others, and not all subjects met the inclusion criteria at each frequency, especially when $L_1 = 50$ dB SPL. Each point in Figure 1 represents the mean L_{DP} of up to 12 subjects (up to 10 in the short-term interval), if L_{DP} for both the test and retest met our inclusion criteria. The correlation was negative (i.e., ΔL_{DP} tended to be larger when fewer subjects were represented by the mean L_{DP} s) in the interleaved, short-term, and long-term intervals, but positive in the immediate interval (ΔL_{DP} was smaller when fewer subjects were represented by the mean L_{DP} s). However, in all intervals the correlation was weak ($r^2 < 0.2$). At the

lowest intensities ($L_1 = 38$ and 44 dB SPL), where variability is typically higher, few DPOAEs met our strict inclusion criteria (18.2% and 45.8%, respectively), so these data were excluded from analysis.

DPOAEs were most stable in the interleaved interval at each frequency (Fig. 2A), as indicated by Pearson's r , ICC(A,1), and ICC(C,1), and Cronbach's α (ICC(C,k)). Each correlation coefficient exceeded 0.99 at every test-frequency in this interval (Fig. 2A). In fact, ICC(C,k) indicated good reliability in each interval (it exceeded 0.70) at each frequency, except for the lowest three frequencies in the short-term interval (Fig. 2C) and the two highest frequencies in the long-term interval (Fig. 2D). Similar results were obtained using Pearson's r and ICC(C,1), although the values were lower than ICC(C,k) values. However, the coefficient ICC(A,1) was more sensitive to disagreement between test and retest values, as it takes into consideration column variance. ICC(A,1) indicated that measurements in the immediate interval were not highly reproducible at 3 kHz, and in the short-term interval ICC(A,1) was less than 0.7 at frequencies below 6 kHz and above 20.8 kHz.

Similar results were obtained when data were examined as a function of primary-tone level (Fig. 3). In the interleaved interval, the four reliability coefficients were greater than 0.97 at all levels. ICC(C,k) exceeded 0.7 in every time interval and primary-tone level. The other three correlation coefficients were reduced compared to ICC(C,k). Pearson's r and ICC(C,1) indicated good reliability in the immediate interval at all levels and in the long term interval when $L_1 = 56$ dB SPL. In the short-term interval, these coefficients exceeded 0.7 only at the highest two levels. However, ICC(A,1) was much lower than the other correlation coefficients in the short-term interval, never exceeding 0.7.

In the short-term interval, ICC(C,1) was greater than ICC(A,1) at every frequency (Fig. 2C) and level (Fig. 3C), which indicates that systematic error contributed to the variability of DPOAE measurements. This was also seen in the immediate interval (Fig. 2B), primarily at $f_2 = 4.5$ kHz and $f_2 = 20.8$ kHz.

Intraclass correlation coefficients are appropriate for variables within the same class (our definition of class follows that of Fisher, 1938: variables must share both variance and metric). Thus, we first confirmed equal variance in each time interval using a student's t -test (under the null hypothesis of equal variance; $p > 0.05$). As a function of frequency, variance was unequal in the immediate time-interval when $f_2 = 3.9$ kHz and 13.8 kHz ($0.02 < p < 0.05$), when $f_2 = 24$ kHz ($0.001 < p < 0.002$), and in the short-term time-interval when $f_2 = 18.2$ kHz ($0.02 < p < 0.05$). Unequal variance was observed in the immediate interval when $L_1 = 50$ dB SPL ($0.02 < p < 0.05$), in the short-term interval when $L_1 = 56$ dB and 74 dB SPL ($0.002 < p < 0.005$), and the long-term interval when $L_1 = 62$ dB ($0.002 < p < 0.005$) and 68 dB SPL ($p < 0.001$).

Mean DPOAE levels were larger in females than males in 77.5 % of the stimulus combinations (e.g., $f_2 = 2978$ kHz, $L_1 = 50$ dB SPL). When DPOAE levels were averaged across $L_1 = 50$ dB SPL, the mean female L_{DP} s were 0.2 – 4.5 dB larger than in males (Fig. 4A), with the exception of $f_2 = 7.9$ and 9.0 kHz, where the mean L_{DP} of males was 0.6 and 1.1 dB larger than in females. When averaged across frequencies, mean female L_{DP} s exceeded those of the males at each primary-tone level by 0.7 – 1.8 dB SPL (Fig. 4B).

Long-term repeatability also was examined according to sex of the subject. The reliability coefficient ICC(C,1) was greater in males than females at all frequencies except 9.1 kHz, where reliability was similar between males and females (Fig. 4C). ICC(C,1) remained above 0.85 from $f_2 = 2.9$ – 20.8 kHz in male subjects and decreased to 0.63 at $f_2 = 24$ kHz. For female subjects, ICC(C,1) values exceeded 0.85 at frequencies > 12 kHz (excepting 4.5

kHz where $ICC(C,1) = 0.81$). Above 12 kHz $ICC(C,1)$ ranged from 0.75 (at $f_2 = 15.8$ kHz) to 0.33 (at $f_2 = 24$ kHz). As primary-tone levels increased (Fig. 4D) there was not a steady increase in male $ICC(C,1)$ values, but the reliability coefficients increased greatly from $L_1 = 50 - 74$ dB in female subjects.

4. Discussion

4.1 Statistics

A proper interpretation of the reliability coefficients reported here requires that we first discuss the relationship between the statistical methods used in the literature and in this paper.

4.2 Pearson's r vs. $ICC(C,1)$

Pearson's r , an interclass correlation coefficient, is an appropriate tool for assessing the relationship between variables in different measurement classes. For variables in the same class, use of an intraclass correlation coefficient (ICC) may serve as a better means of assessing homogeneity (McGraw & Wang, 1996). In fact, Pearson's r is a linearity index, representing a linear transformation of one data set to another ($y = ax + b$), whereas $ICC(C,1)$ is an additivity index, representing the transformation of one data set to another by the addition of some constant ($y = x + b$) (McGraw & Wong, 1996). As such, $ICC(C,1)$ is sensitive to unequal variance and is reduced relative to Pearson's r in such circumstances (McGraw & Wong, 1996). In this study, the $ICC(C,1)$ values were virtually identical to Pearson's r in most cases, excepting those where we observed unequal variance (Figs. 2B,C; Figs. 3B-D).

4.3 $ICC(A,1)$ vs. $ICC(C,1)$

$ICC(C,1)$ and $ICC(A,1)$ are single score ICCs constructed using a two-way mixed-effect model (with fixed column (trial) effects and random row (subject) effects). $ICC(C,1)$ measures the consistency between two sets of data and detects random, but not systematic, error (it excludes column variance); $ICC(A,1)$, does include column variance and is thus more sensitive to the effects of trial than $ICC(C,1)$. A comparison of these two ICCs provides a simple means of assessing whether systematic error is present in the data set and how well the data is (McGraw & Wong, 1996). For example, a comparison of $ICC(C,1)$ and $ICC(A,1)$ in the short-term (Figs. 2C, 3C) and immediate interval (Figs. 2B, 3B) indicates that there is systematic error present at some frequencies, but no systematic error is seen in the long-term interval (Figs. 2D, 3D). The use of $ICC(C,1)$ alone could overestimate reliability, and would not allow for the identification of systematic error.

When the effect of trial is relevant, as it would be when comparing DPOAE levels before and after exposure to a treatment, $ICC(A,1)$ may be beneficial. A comparison of $ICC(C,1)$ and $ICC(A,1)$ in control and experimental subjects (before and after treatment) could indicate whether L_{DP} systematically changed. When the effect of trial is irrelevant, the $ICC(C,1)$ could be sufficient.

4.4 $ICC(C,k)$

$ICC(C,k)$, or Cronbach's α , values were greater than all other reliability coefficient values at every primary-tone level and frequency in each test condition. In some cases, the difference in reliability coefficient values was inconsequential, such as in the interleaved interval where all values exceeded 0.99. However, Pearson's r , $ICC(A,1)$, and $ICC(C,1)$ were each less than 0.7 at some frequencies (Fig. 2) and levels (Fig. 3) where the $ICC(C,k)$ values exceeded 0.85. It may be that $ICC(C,k)$ overestimates reliability when applied to single score data, and

it may be a more accurate measure of reliability when applied to averaged data (when systematic error is not of interest) (e.g., McGraw & Wong, 1996).

4.5 Reliability of DPOAEs in non-human primates

There is little repeatability data available in non-human primates for comparison with this study (to our knowledge). Lasky *et al.* (1995) reported data under conditions similar to our immediate interval. The mean absolute difference in L_{DP} ($|\Delta L_{DP}|$) was less than 1.0 dB when $L_1 = L_2 = 60$ and 65 dB SPL and $f_2 = 0.54 - 20$ kHz, and this difference increased as primary-tone levels decreased. In another study (Park *et al.*, 1995), DPOAEs were measured repeatedly in a single rhesus monkey at higher primary-tone levels ($L_1 = L_2 = 75$ dB) and a lower frequency range (1 – 8 kHz), changed by up to 10 dB at some frequencies over long-term intervals. The results from both studies were similar to what we observed in marmosets, although $|\Delta L_{DP}|$ was slightly higher in marmosets (3.9 dB) when $L_1 = 68$ dB and $f_2 < 20.8$ kHz.

4.6 Immediate test-retest repeatability

Test-retest reliability was best when the probe-tip was not removed between test and retest (Figs. 2A,B; 3A,B). In marmosets, Pearson's r was 0.86 when $L_1 = 68$ dB SPL. This was similar to r values reported in humans tested under similar conditions (mean $r = 0.85$; Zhao & Stephens, 1999). In another human study, ICC(C,k) (Cronbach's α) exceeded 0.95, regardless of primary-tone level or frequency (Wagner *et al.*, 2008). Keppler *et al.* (2010) reported ICC(C,k) values exceeding 0.91 when $L_1/L_2 = 65/55$ dB between 0.8 – 8 kHz. ICC(C,k) values in marmosets were lower than what has been reported in humans: when pooled across levels, ICC(C,k) values in marmosets exceeded 0.92 at frequencies > 3 kHz. When pooled across frequencies, ICC(C,k) increased from 0.77 – 0.86 as primary-tone levels increased.

The reliability reported here was similar to that observed in humans at moderate to high primary-tone levels, but correlation coefficients were lower as the primary-tone levels were decreased. This could be attributed to sample size; 80 ears were examined by Wagner *et al.* (2008) and 56 by Keppler *et al.* (2010). In the current study, 12 ears were examined. Zhao & Stephens (1999) also examined 12 ears in humans and reported reliability similar to the marmoset.

It is likely, however, that something other than a small sample size contributed to the difference in reliability, such as a systematic decrease in L_{DP} between test and retest. For instance, the mean L_{DP} s were, on average, 3.1 and 1.4 dB weaker in the retest than in the initial test when $L_1 = 50$ and 74 dB SPL, respectively (Fig. 1B). In fact, a comparison of ICC(C,1) and ICC(A,1) indicated that systematic error was present (Figs. 2B, 3B). A systematic decrease in L_{DP} could be attributed to an increase in middle ear or ear canal pressure, which could impede forward and reverse transmission, but these effects would likely be small. Garner, Neely, and Gorga (2008) reported that changes in middle ear transmission only accounted for a small amount of the variability in human DPOAEs, and at only a few frequencies. Another contributing factor could be the use of injectable anesthesia, which would not maintain a constant depth of anesthesia throughout the entirety of the procedure.

4.7 Short-term test-retest reliability

Reliability was poorer in the short-term interval than in the immediate interval in our study and in each of the human studies which examined both intervals (Zhao & Stephens, 1999; Beattie *et al.*, 2003; Keppler *et al.*, 2010). Here we provide a comparison of reliability data at $L_1 = 68$ dB in marmosets and $L_1 = L_2 = 70$ dB in humans. In marmosets, Pearson's r was

0.72. Pearson's r values were in general better in humans, ranging from 0.74 – 0.95 (Zhao & Stephens, 1999; Ng & McPherson, 2005). Using ICC(C,k) as a measure of reliability, Keppler *et al.* (2010) reported that DPOAEs were highly reliable in conditions similar to our short-term interval: ICC(C,k) exceeded 0.93 when $L_1/L_2 = 65/55$ and was 0.84 in marmosets ($L_1 = 68$ dB SPL).

It should be noted that the reliability was worst in this interval at the lower primary-tone levels ($L_1 = 50$ and 56 dB SPL). Mean DPOAE levels decreased by up to 13.3 dB in the retest when compared to the initial test (Fig. 1C), and ICC(A,1) indicated that systematic error was present at most frequencies (Fig. 2C) and all levels (Fig. 3C). The systematic error was not likely due to removing and replacing the probe, as it was observed in the immediate interval (Fig. 2B) and not in the long-term interval (Fig. 2D). It is possible that effects of time under anesthesia could have led to a systematic reduction in L_{DP} . It should be noted that there was one subject in this interval whose L_{DP} s mostly increased in the retest when compared to the test.

4.8 Long-term test-retest reliability

Long-term reliability is important when repeated measurements are made over a period of time, such as when examining prolonged noise exposure, or the effects of long-term treatment with potentially ototoxic drugs. Long-term repeatability has been extensively examined in humans, and good reliability is generally reported. Here as well, we observed good reliability in marmoset DPOAEs at the frequencies typically assessed in humans. However, the large number of statistical measures used in the literature does not allow for an at-a-glance comparison, as detailed below.

In adult humans, 98.4% and 87.5% of DPOAEs measured at low (2 - 8 kHz) and high (10 – 16 kHz) frequencies, respectively, fell within 10-dB of each other ($L_1/L_2 = 70/55$ dB SPL; Dreisbach *et al.*, 2006). In marmosets, all of the DPOAEs measured with $L_1 = 74$ dB SPL varied by less than 10-dB at low frequencies (3 – 8 kHz) and 95% changed by less than 10-dB at high frequencies (9 – 16 kHz), respectively. At frequencies below 8 kHz, the mean absolute ΔL_{DP} between test and retest measurements was 1.7 dB in humans ($L_1/L_2 = 70/64$ dB; Roede *et al.*, 1993) and 2.1 dB in marmosets ($L_1 = 74$ dB). Similar results were obtained with moderate primary-tone levels in both humans and marmosets.

Pearson's r indicated good repeatability in human DPgrams measured $L_1 = L_2 = 70$ dB, ranging from 0.74 – 0.9 (Zhao & Stephens, 1999; Ng & McPherson, 2005). In marmosets, Pearson's r was 0.83 when $L_1 = 68$ dB. ICC(C,k) indicated good repeatability in the human studies as well, exceeding 0.74 (Wagner *et al.*, 2008), 0.85 (Franklin *et al.*, 1992), and 0.92 (Keppler *et al.*, 2010) at moderate to high primary-tone levels. In marmosets, ICC(C,k) exceeded 0.85 when $L_1 = 56$ dB SPL.

Long-term repeatability of marmoset DPOAEs was also comparable to that reported in children. For instance, Sockalingam *et al.* (2007) reported ICC(C,1) values ranging from 0.64 – 0.89 when $L_1/L_2 = 65/55$ dB SPL and $f_2 = 2.5 - 10$ kHz. In this study, ICC(C,1) values had nearly identical range (0.64 – 0.9) under similar stimulus conditions ($f_2 = 10.8$ kHz, $L_1 = 62$ dB SPL).

Long-term repeatability was similar to that reported in Wistar rats as well (Rasmussen *et al.*, 2005). This was assessed using the standard error of test/retest differences, which was calculated from

$$SEM\Delta = SD \sqrt{1 - r}, \quad (5)$$

where SD was that of the mean ΔL_{DP} and r was Pearson's. At $f_2 = 4, 8,$ and 16 kHz, SEM Δ ranged from about $1.5 - 3$ dB when $L_1 = 50$ dB in Wistar rats, and in marmosets SEM Δ ranged from $0.73 - 2.67$ dB at similar frequencies and levels.

4.9 Sex Differences

We previously reported large sex differences in DPOAEs measured from a different set of marmosets (Valero *et al.*, 2008), a historically uncommon finding in DPOAE studies. In the current group of subjects, mean DPOAE levels were larger in females than males at most frequencies and levels. However, the difference in mean L_{DP} was smaller in this group than the group in our earlier study. The data reported in the previous study were obtained with equilevel primary tones, but DPOAEs obtained with $L_1 - L_2 = 10$ were available for analysis, and the sex differences remained large (effect sizes were 0.57 or larger at all frequencies; data not shown).

This prompted us to examine sex differences in DPOAE repeatability. The correlation coefficients (ICC(C,1)) were lower in females than males at most frequencies, suggesting that the repeatability of DPOAE measurements was better in the male subjects than female subjects at these frequencies. Further, correlation coefficients were reduced in females at the frequencies at which the largest sex differences were observed in the previous study.

It is unclear why the female DPOAEs would be less reliable over time, but it is tempting to suggest that there may be effects of sex hormones on DPOAE levels, as they were not less reliable in same day measurements (data not shown). There is evidence that endogenous estrogen and progesterone can affect the efficacy of the efferent system and TEOAE levels (Al-Mana *et al.*, 2010), and exposure to exogenous estrogen and progesterone has been shown to affect DPOAE levels (e.g., Guimaraes *et al.*, 2006; Price *et al.*, 2009) and/or SOAE prevalence (McFadden *et al.*, 1998). However, there is currently no evidence that DPOAEs vary systematically over the course of the human menstrual cycle (Yellin & Stillman, 1999), but it would be interesting to examine this in marmosets, as blood serum concentrations of sex hormones are much higher in the marmoset (Harlow *et al.*, 1983), ovulation is suppressed in subordinate females (see Saltzman *et al.*, 2009 for a review), and the marmoset estrous cycle differs from the human menstrual cycle in several ways (e.g., Silvestri & Fraser, 2007). We were unable to tease out any effects of estrous, as we did not control for time of cycle in the study.

5. Conclusions

The reliability of DPOAE measurements in the common marmoset is similar to that reported in humans in most intervals, but in the same-day measurements there was a systematic decrease in the mean group L_{DP} that may have been due to some effect of the increased time under anesthesia. A follow up study using isoflurane anesthesia, where dosage can be kept constant throughout the procedure, may tell us whether the change in depth of anesthesia is the source of this systematic change.

Acknowledgments

This work was supported by a research grant (R03DC009050) from the National Institute on Deafness and other Communication Disorders (NIDCD). The authors are grateful for the data acquisition software provided by Edward Pasanen, from Dennis McFadden's lab, which was written while supported by research grant DC00153 from the NIDCD. Figures were made on IgorPro software. We are thankful to the reviewers, the Specialized Neuroscience Research Program (SNRP) at UTSA, and Pamela Tijerina for the helpful suggestions made on earlier versions of this manuscript. We also thank Donna Layne-Colon and the staff at TBRI, who take great care of our marmoset colony.

Abbreviations

DPOAEs	Distortion-product otoacoustic emissions
TEOAEs	Transient-evoked otoacoustic emissions
L_{DP}	DPOAE level
f₁	lower-frequency primary-tone
f₂	higher-frequency primary-tone
L₁	Level of lower-frequency primary-tone
L₂	Level of higher-frequency primary-tone
OHC	Outer hair cell
SEM	Standard error of measurement
SEMA	Standard error of the difference in measurements
ICC	Intraclass correlation coefficient

References

- Al-Mana D, Ceranic B, Djahanbakhch O, Luxan LM. Alteration of auditory function during the ovarian cycle. *Hear Res.* 2010; 268:114–122. [PubMed: 20685243]
- Allen JB. User Manual for the CUBDIS Distortion Product Measurement System. AT&T Bell Labs. 1990
- Beattie RC, Ireland A. Effects of sample size on noise floor and distortion product otoacoustic emissions. *Scand. Audiol.* 2000; 29:93–102. [PubMed: 10888346]
- Beattie RC, Kensworthy OT, Luna C. Immediate and short-term reliability of distortion-product otoacoustic emissions. *Int. J. Audiol.* 2003; 42:348–354. [PubMed: 14570243]
- Cacace AT, McClelland WA, Weiner J, McFarland DJ. Individual differences and the reliability of 2f1-f2 distortion-product otoacoustic emissions: effects of time-of-day, stimulus variables, and gender. *J. Speech Hear. Res.* 1996; 39:1138–1148. [PubMed: 8959599]
- Crocker, L.; Algina, J. Introduction to classical and modern test theory. New York: Holt, Rinehart, and Winston; 1986.
- Demorest ME, Walden BE. Psychometric principles in the selection, interpretation, and evaluation of communication self-assessment inventories. *J. Speech Hear. Disord.* 1984; 49:226–240. [PubMed: 6748618]
- Dreisbach LE, Long KM, Lees SE. Repeatability of high-frequency distortion-product otoacoustic emissions in normal-hearing adults. *Ear Hear.* 2006; 27:466–479. [PubMed: 16957498]
- Fisher, RA. Statistical Methods for Research Workers. 7th ed. Edinburgh: Oliver and Boyd; 1938.
- Franklin DJ, Marcy J, McCoy MS, Martin GK, Lonsbury-Martin BL. Test/retest reliability of distortion-product and transiently evoked otoacoustic emissions. *Ear. Hear.* 1992; 12:417–429. [PubMed: 1487104]
- Frisina RD. Age-related hearing loss; ear and brain mechanisms. *Ann. N.Y. Acad. Sci.* 2009; 1170:708–717. [PubMed: 19686217]
- Garner CA, Neely ST, Gorga MP. Sources of variability in distortion product otoacoustic emissions. *J. Acoust. Soc. Am.* 124:1054–1067. [PubMed: 18681596]
- Gaskill SA, Brown AM. The behavior of the acoustic distortion product, 2f₁ - f₂, from the human ear and its relations to auditory sensitivity. *J. Acoust. Soc. Am.* 1990; 88:821–839. [PubMed: 2212308]
- Gorga MP, Neely ST, Ohlrich B, Hoover B, Redner J, Peters J. From laboratory to clinic: a large scale study of distortion-product otoacoustic emissions in ears with normal hearing and hearing loss. *Ear. Hear.* 1997; 18:440–455. [PubMed: 9416447]

- Guimaraes P, Frisina ST, Mapes F, Tadros SF, Frisina DR, Frisina RD. Progesterin negatively affects hearing in aged women. *Proc. Natl. Acad. Sci. U.S.A.* 2006; 103:14246–14249. [PubMed: 16959884]
- Hall AJ, Lutman ME. Methods for early identification of noise-induced hearing loss. *Audiol.* 1999; 38:277–280.
- Hallenbeck H, Dancer J. Distortion-product otoacoustic emissions in ears with normal hearing sensitivity: test-retest variability. *Percept. & Mot. Skills.* 2003; 97:990–992.
- Harding GW, Bohne BA, Ahmad M. DPOAE level shifts and ABR threshold shifts compared to detailed analysis of histopathological damage from noise. *Hear. Res.* 2002; 174:158–171. [PubMed: 12433407]
- Harlow CR, Gems S, Hearn JP, Hodges JK. The relationship between plasma progesterone and the timing of ovulation and early embryonic development in the marmoset monkey. *J. Zool.* 1983; 201:273–282.
- Horner KC, Lenoir M, Bock GR. Distortion product otoacoustic emissions in hearing impaired mutant mice. *J. Acoust. Soc. Am.* 1985; 78:1603–1611. [PubMed: 4067076]
- Johnson TA, Neely ST, Garner CA, Gorga MP. Influence of primary-level and primary-frequency ratios on human distortion product otoacoustic emissions. *J. Acoust. Soc. Am.* 2006; 119:418–428. [PubMed: 16454296]
- Kemp DT. Evidence of a mechanical nonlinearity and frequency selective wave amplification in the cochlea. *Arch. Otolaryngol.* 1979; 224:37–45.
- Kemp, DT. Exploring cochlear status with otoacoustic emissions. The potential for new applications. In: Robinette, MS.; Glatke, TJ., editors. *Otoacoustic Emissions: Clinical Applications*. 2nd ed.. New York: Thieme; 2002. p. 1-47.
- Keppeler H, Dhooge I, Maes L, D'haenens W, Bocksteal A, Philips B, Swinnen F, Vinck B. Transient-evoked and distortion product otoacoustic emissions: a short-term test-retest reliability study. *Int. J. Audiol.* 2010; 49:99–109. [PubMed: 20151884]
- Kimberley BP. Distortion-product emission features and the prediction of pure tone thresholds. *Laryngoscope.* 1995; 105:349–353. [PubMed: 7715375]
- Kummer P, Janssen T, Arnold W. The level and growth behavior of the 2f1 – f2 distortion product otoacoustic emission and its relationship to auditory sensitivity in normal hearing and cochlear hearing loss. *J. Acoust. Soc. Am.* 1998; 103:3431–3444. [PubMed: 9637030]
- Kummer P, Janssen T, Hulin P, Arnold W. Optimal L(1)-L(2) primary tone level separation remains independent of test frequency in humans. *Hear. Res.* 2000; 146:47–56. [PubMed: 10913883]
- Lasky RE, Snodgrass E, Hecox K. Distortion-product otoacoustic emission input/output functions as a function of frequency in human adults. *J. Am. Acad. Audiol.* 1994; 5:183–194. [PubMed: 8075414]
- Lasky RE, Snodgrass EB, Laughlin NK, Hecox KE. Distortion product otoacoustic emissions in *Macaca mulatta* and humans. *Hear. Res.* 1995; 89:35–51. [PubMed: 8600131]
- Lonsbury-Martin BL, Cutler WM, Martin G. Evidence for the influence of aging on distortion product otoacoustic emissions in humans. *J. Acoust. Soc. Am.* 1990; 89:1749–1759. [PubMed: 2045583]
- McFadden DM, Pasanen EG, Raper J, Lange HS, Wallen K. Sex differences in otoacoustic emissions measured in rhesus monkeys (*Macaca mulatta*). *Horm. & Beh.* 2006; 50:274–284.
- McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol. Methods.* 1996; 1:30–46.
- Mills DM, Feeney MP, Drake EJ, Folsom RC, Sheppard L, Seixas NS. Developing standards for distortion product otoacoustic emission measurements. *J. Acoust. Soc. Am.* 2007; 122:2203–2214. [PubMed: 17902856]
- Moulin A, Collet L, Duclaux R. Contralateral auditory stimulation alters acoustic distortion products in humans. *Hear. Res.* 1993; 65:193–210. [PubMed: 8458751]
- Ng IH, McPherson B. Test-Retest reliability of distortion-product otoacoustic emissions in the 1 – 7 kHz range. *Audiol. Med.* 2005; 3:108–115.
- Norton SJ. Emerging role of evoked otoacoustic emissions in neonatal hearing screening. *Am. J. Otol.* 1994; 15:4–12. [PubMed: 8109628]

- Parazzini M, Wilson HK, Bell S, Tognola G, Ravazzani P, Lutman ME. The generation mechanisms and repeatability of 2f1 - f2 distortion-product otoacoustic emission: study on normally hearing subjects. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2006; 1:2122–2125. [PubMed: 17946498]
- Park JY, Clark WW, Coticchia JM, Esselman GH, Fredrickson JM. Distortion product otoacoustic emissions in rhesus (*Macaca mulatta*) monkey ears: normative findings. *Hear. Res.* 1995; 86:147–162. [PubMed: 8567411]
- Price K, Zhu X, Guimaraes PF, Vasilyeva ON, Frisina RD. Hormone replacement therapy diminishes hearing in peri-menopausal mice. *Hear. Res.* 2009; 252:29–36. [PubMed: 19269311]
- Rasmussen AN, Osterhammel PA, Lund SP, Kristiansen GB, Andersen S. A system for measuring distortion product otoacoustic emissions at ultrasonic frequencies in rodents. *Int. J. Audiol.* 2005; 44:237–243. [PubMed: 16011052]
- Roede J, Harris FP, Probst R, Xu L. Repeatability of distortion product otoacoustic emissions in normally hearing humans. *Audiology.* 1993; 32:273–281. [PubMed: 8216026]
- Saltzman W, Digby LJ, Abbott DH. Reproductive skew in female common marmosets: what can proximate mechanisms tell us about ultimate causes? *Proc. Biol. Sci.* 2009; 276:389–399. [PubMed: 18945663]
- Shera C. Mechanisms of mammalian otoacoustic emission and their implications for the clinical utility of otoacoustic emissions. *Ear. Hear.* 2004; 25:86–97. [PubMed: 15064654]
- Shrout PE, Fleiss JL. Intraclass correlation coefficients: Uses in assessing reliability. *Psychol. Bull.* 1979; 86:420–428. [PubMed: 18839484]
- Silvestri A, Fraser HM. Oestrogen and progesterone receptors in the marmoset endometrium: changes during the ovulatory cycle, early pregnancy and after inhibition of vascular endothelial growth factor, GnRH or ovariectomy. *Reproduction.* 2007; 134:341–353. [PubMed: 17660243]
- Sockalingam R, Kei J, Choi DH. Test-retest reliability of distortion-product otoacoustic emissions in children with normal hearing: A preliminary study. *Int. J. Audiol.* 2007; 46:351–354. [PubMed: 17680466]
- Stuart A, Passmore AL, Culbertson DS, Jones SM. Test-retest reliability of low-level evoked distortion-product otoacoustic emissions. *J. Speech Lang. Hear. Res.* 2009; 52:671–681. [PubMed: 18952856]
- Torre PT, Fowler CG. Age-related changes in auditory function of rhesus monkeys (*Macaca mulatta*). *Hear. Res.* 2000; 142:131–140. [PubMed: 10748335]
- Valero MD, Pasanen EG, McFadden D, Ratnam R. Distortion-product otoacoustic emissions in the common marmoset (*Callithrix jacchus*): parameter optimization. *Hear. Res.* 2008; 243:57–68. [PubMed: 18586424]
- Wagner W, Heppelmann G, Vonthein R, Zenner HP. Test-retest repeatability of distortion product otoacoustic emissions. *Ear. Hear.* 2008; 29:378–391. [PubMed: 18382378]
- Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J. Strength Cond. Res.* 2005; 19:231–240. [PubMed: 15705040]
- Yellin MW, Stillman RD. Otoacoustic emissions in normally-cycling females. *J. Am. Acad. Audiol.* 1999; 10:400–408. [PubMed: 10949944]
- Zhao F, Stephens D. Test-retest variability of distortion-product otoacoustic emissions in human ears with normal hearing. *Scand. Audiol.* 1999; 28:171–178. [PubMed: 10489866]
- Zurek PP, Clark WW, Kim DO. The behavior of acoustic distortion products in the ear canals of chinchillas with normal or damaged ears. *J. Acoust. Soc. Am.* 1982; 72:774–780. [PubMed: 7130536]

Figure 1.

Mean DPgrams showing L_{DP} as a function of f_2 frequency. Each plot is averaged across all subjects within each time-interval. Error bars indicate one SD. DPOAEs were most reproducible in the interleaved interval and in the long-term interval at low frequencies. They were least reproducible in the short-term interval.

Figure 2.

Reliability coefficients as a function of f_2 frequency. DPOAE data were pooled across $L_1 = 50 - 74$ dB SPL in the A) Interleaved, B) Immediate, C) Short-term, and D) Long-term intervals. Low frequencies were less reproducible in same-day measurements, and high-frequency measurements were less reproducible when the probe was replaced.

Figure 3.

Reliability coefficients as a function of L_1 primary-tone level. DPOAE data were pooled across all frequencies in the A) Interleaved, B) Immediate, C) Short-term, and D) Long-term intervals. Reproducibility increased with increasing primary-tone levels.

Figure 4. Mean DPgrams (A) and gain functions (B) are shown for male and female subjects. Error bars represent one SD. ICC(C,1) are shown for males and females as a function of C) f_2 -frequency and D) L_1 primary-tone level.