PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Variability in the ISO 3382-3 metrics based on repeated acoustic measurements in open-plan offices[1]

Manuj YADAV[2,3,4]; Densil CABRERA[2]; James LOVE[2]; Jungsoo KIM[2], Jonothan HOLMES[2], Hugo CALDWELL[2] and Richard DE DEAR[2]

[2] The University of Sydney, Australia

[3] HörTech gGmbH, Oldenburg, Germany

[4] Carl von Ossietzky University, Oldenburg, Germany

ABSTRACT

This paper investigates variability in the key ISO 3382-3:2012 metrics based on two types of repeated measurements in open-plan offices – Type1, where the same path over workstations was measured from opposite ends, and Type2, where two different measurement paths were measured. Per metric, analyses performed used (i) the *range* of observed values; and (ii) the *observed* values. Results from category (i) analysis: $\overline{\Delta\text{Type1}}$, and 95% confidence intervals were 1.2 m(0.9,1.5) for r_D; 0.8 dB(0.6,1.0) for $D_{2,S}$; 1.2 dB(0.8,1.5) for $L_{p,A,S,4\,m}$; and 1.2 dB(0.7,1.7) for $L_{p,A,B}$. $\overline{\Delta\text{Type2}}$ were between twice and thrice the respective values of $\overline{\Delta\text{Type1}}$. Results from category (ii) analysis: the *reliability*, based on intra-measurement correlation coefficient, was fairly high for all metrics, except for $L_{p,A,S,4\,m}$ for Type2 repeats. The *repeatability limit/coefficient* (r) was 2.5 m for r_D; 1.7 dB for $D_{2,S}$; 2.9 dB for $L_{p,A,S,4\,m}$; and 3.0 dB for $L_{p,B}$, for Type1 repeats. The r values for Type2 repeats were substantially higher except for $D_{2,S}$. Overall, most of the Type1 results seem reasonable considering repeats were conducted in complicated room acoustic environments, while Type2 repeats needs larger sample sizes in future studies. Some recommendations are outlined for ISO 3382-3 vis-à-vis Type1 and Type2 repeats.

Keywords: Office acoustics, Measurement uncertainty

1. INTRODUCTION

ISO 3382-3:2012 (2) is the main international standard for measuring the room acoustics for open-plan offices. The metrics within the standard include those that are based on speech sound pressure level (SPL_{Speech}) – the spatial decay rate of speech per distance doubling (**$D_{2,S}$**), A-weighted SPL of speech at a distance of 4 m (**$L_{p,A,S,4,m}$**); those based on spatial decay of the speech transmission index (STI) – distraction distance (**r_D**), privacy distance (**r_P**), and STI at the nearest workstation; along with the A-weighted background noise level (**$L_{p,A,B}$**). Apart from $L_{p,A,B}$, the salient feature of these metrics is that they characterize aspects of the *spatial decay of speech over workstations* within an office. These metrics are calculated from room acoustic measurements over a more-or-less linear path comprising a set of measurement positions near workstations. ISO 3382-3 requires at least two measurement paths per office. If only one measurement path is possible, then two measurements per path are to be conducted with the source loudspeaker positioned on the opposite ends of the path (2). This study considers the variability of the key ISO 3382-3 metrics – $L_{p,A,B}$, $D_{2,S}$, $L_{p,A,S,4,m}$ and r_D – due to repeated measurements in the same office (i) with two paths that traverse the same set of workstations, but using the source loudspeaker on the opposite ends of the path per measurement, and (ii) with different measurement paths (see section 2.3 for details).

2. METHODS

2.1 Room acoustic measurements

Acoustic measurements were made according to ISO 3382-3 (2) in 27 furnished, unoccupied

[1]This research has previously been published in more detail as (1). Reprint of the copyrighted material for this conference has been cleared by Elsevier under the license number 4596981407529 on May 27, 2019.

offices in 8 buildings with fully operational heating, ventilation, and air conditioning. These offices are either non-contiguous units within a larger floor plate, or single units that cover the entire floor.

The software used to generate, play, record, and analyze the signals was AARAE, which is a MATLAB-hosted GUI-driven environment (3). The measurement signal used was an exponentially swept sinusoid (50 Hz - 20 kHz) of 30 s duration. The signal chain for the measurements included a computer running AARAE, an audio interface (RME Fireface UFX, Haimhausen, Germany), an amplifier (Brüel & Kjær Power Amplifier Type 2716-C, Nærum, Denmark), an omnidirectional loudspeaker (Brüel & Kjær OmniSource Type 4295) with a known sound power output that was calculated using ISO 3745 (4), omnidirectional microphones (Earthworks M30, Milford, USA), and a preamplifier unit (RME OctaMic XTC). For each path, the measurement signal from the source loudspeaker was recorded using between 5-8 microphones simultaneously, where the microphones were located at receiver positions near workstations, as described in ISO 3382-3 (2). The number of microphones used within a given office was always the same – determined by the floor plate of the office measured – whereas the number of microphones used between offices varied. Background noise measurements were made at each microphone position for at least 60 s.

Overall, the measured offices included a reasonably wide variation in several relevant aspects, such as the floor plate geometries (rectangular to curved on one or more sides), absorption profile (carpeted, timber, concrete) and areas (53-719 m^2); the partitions between workstations (cubicle designs, single, double, or no partitions); ceiling heights and shapes (most being flat and measuring 2.7-2.8 m, and 5 vaulted ceilings of heights between 4.1-7.6 m to their apex), and absorption profile (absorptive, hard ceiling, or a combination of the two), density of workstations on the floor (4-24 workstations per 100 m^2), etc. More details are presented in Table 1 in (1)

2.2 Calculation of the ISO 3382-3 metrics

The sinusoidal sweep recordings from microphones per measurement path were used to derive the SPL_{Speech} metrics, i.e., $D_{2,S}$ and $L_{p,A,S,4,m}$. Modulation transfer functions (MTFs) used for calculating STI, and subsequently r_D, used the male octave-band weights and redundancy factors [8]. $L_{p,A,B}$ was calculated as described in ISO 3382-3, from octave-band SPLs that represented spatially averaged values for the microphones along a measurement path (2). In order to avoid potential problems with extrapolation, r_D values were considered acceptable if they were within the range of the measured physical distances extended by 10% on either end (5). Otherwise, in order to use the remaining offices, the background noise was adjusted, with the same gain applied in all bands to all measurement lines in a given office, as described in (5). While not an ISO 3382-3 metric, in order to further characterize the offices, spatially averaged and mid-frequency (average of 500 Hz and 1kHz octave-band values) reverberation times (T_{30}) were calculated from the impulse responses (12^{th} order filters).

2.3 Repeated measurements in offices

Variability within the ISO 3382-3 metrics was considered for two types of repeated measurements:

(i) **Type1**, where two measurements per path were conducted. Here, after conducting one measurement with the loudspeaker on one end of the path, the other measurement was conducted either by placing the loudspeaker on the opposite end, or by swapping the positions of the last microphone and the loudspeaker from the previous measurement. All the paths in the 27 offices were measured using Type1 repeated measurements, with 36 paths measured in total (Table 1).

(ii) **Type2**, where two measurement paths were measured within the same office (i.e., the same acoustic zone). Type2 repeated measurements were only possible in a small subset of all offices measured (7 in total; Table 1): these being the offices for which the allocated access times permitted several repeated measurements. The idea here was to determine the variation in the ISO 3382-3 metrics (2) over two paths that can be considered as likely options to characterize the office overall. These measurement paths included those that were on an adjacent row of workstations relative to the first path, or more-or-less linear paths that were at a small angle to the length of the office, or the main measurement axis.

2.4 Measurement conditions

All the measurements were supervised and conducted by the first author, with assistance from others. Per office, the repeatability conditions included measurements conducted by the same operator over a short time period, using the same method, and the same apparatus. Hence, given the constraints of the definition of an 'office' (see sections 1, 2.1), the guidelines in ISO 5725 1-6 (6), especially regarding repeatability conditions, were followed as closely as possible.

2.5 Statistical analyses

All the analyses were done within the software R [11], using the packages *tidyverse* (data management and plots) [12], *boot* [13], *robustbase* [14], and *robustlmm* [15]. Robust statistical methods were used throughout to reduce the influence of outliers, which included robust linear regression, and mixed-effects models; and using bootstrapping for calculating confidence intervals (CIs). In the following, wherever the use of bootstrapping is indicated, 10^5 bootstrap resamples (each resample the size of the original data, and resampling with replacement) were generated, using the function boot, and the percentile CIs [18] calculated using the function *boot.ci* are reported; both functions are from the package *boot* [13]. Broadly speaking, analyses were performed on the difference between the largest and smallest observed values – the range – per metric, per Type1 and Type2 repeated measurement, which are referred to as ΔType1 and ΔType2, respectively (section 2.5.1); and on the observed values of the metrics (section 2.5.2).

2.5.1 Analysis using ΔType1 and ΔType2

The summary statistics included the mean and bootstrapped CIs (68% and 95%) of ΔType1 and ΔType2, where the mean values were calculated as,

$$\overline{\Delta\text{Type1}} = \frac{1}{36}\sum_{i=1}^{36} \Delta\text{Type1}_i \quad (1)$$

$$\overline{\Delta\text{Type2}} = \frac{1}{7}\sum_{j=1}^{7} \Delta\text{Type2}_j \quad (2)$$

where i and j are the path, and office indices, respectively. Furthermore, for each metric, each ΔType1_i and ΔType2_j values were compared against the means ($\overline{\text{Type1}}_i$, $\overline{\text{Type2}}_j$) of their respective underlying observed values; the latter being indicative of the magnitude of the metric for the i-th, and j-th index, respectively. These comparisons were made using the corresponding effect sizes for robust linear regression, i.e. R^2 values (referred to as $R^2_{\Delta\sim\text{mean}}$ in the following) and its 95% CI; and the significance (using 95% CIs) of the regression slopes, calculated using the *lmrob* function in the *robustbase* package (7). Also, to determine the mutual relationship between the metrics for their respective ΔType1 and ΔType2 values, Kendall's τ correlation coefficients, along with bootstrapped 95% CIs were used.

2.5.2 Analysis using observed values of metrics

An important consideration for repeated measurements is *reliability*, which can be quantified using the intra-measurement correlation coefficient (ICC; generally known as intra-class correlation coefficient) (8) for both Type1 and Type2 repeated measurements, using the following equation:

$$\text{ICC} = \frac{(\sigma_B^2)}{(\sigma_B^2) + (\sigma_W^2)} \quad (3)$$

The ICC (value between 0 and 1, presented alongside its 95% CI in the current study) in equation 3 quantifies *reliability* as the relationship between the magnitude of the measurement error in the observed values per metric (i.e., within-measurement variance, σ_W^2), and the inherent variability in the 'error-free', or 'true' observed values of the repeated measurements per metric (i.e., between-measurement variance, σ_B^2). The idea is that if the reliability is high (i.e., high ICC), then the measurement errors are small in comparison to the true underlying variability between repeated measurements. Conversely, if the reliability is low (i.e., low ICC), measurements are error-prone, and the variability between repeated measurements could be predominantly due to the large underlying measurement errors, rather than due to actual variability between observed values. The measurement reliability was classified according to common criteria (9) as excellent (ICC > 0.75), fair to good (ICC = 0.40–0.75) and poor (ICC ≤ 0.40). In equation 3, σ_B^2 and σ_W^2 were calculated by fitting a robust mixed-effects model per metric, using the *rlmer* function from the *robustlmm* package (10). Each robust mixed-effects model, where the robustness corresponds to minimizing the contribution of outlying values (10), comprised of the fixed-effect of the respective metric (e.g., r_D), and the random-effects due to the repeated measurements in the offices.

To express the *repeatability*, several values could be provided, which are nevertheless dependent on the measurement error, i.e., within-measurement variability (8). These values include σ_W, $\sqrt{k} \times \sigma_W$, where k is the number of repeated measurements made per path/office ((6), part 6) , and the repeatability coefficient (or, limit) r. The repeatability coefficient is calculated as:

$$r = f \times \sqrt{k} \times \sigma_W \quad (4)$$

The term f in equation 4 depends on the probability level to be associated, and on the underlying probability distribution (6). For an assumed normal distribution, and a probability level of 95%, the value of f is 1.96. Hence, for $k = 2$, the repeatability coefficient becomes $r = 2.8\sigma_W$, and the calculated value represents the upper limit within which 95% of the future differences between the Type1 repeated measurements are expected to occur. Similarly, for Type2 measurements, k = 4, and $r = 3.6\sigma_W$ ((6), part 6).

Table 1: Calculated ISO 3382-3 metrics and reverberation times for the all paths, where Type2 repeated measurements are highlighted in bold. The subscripts for the ISO 3382-3 metrics denote the index of the repeated measurement. The symbols '*' and '#' identify 'poor (or insufficient') and 'good' values respectively based on the criteria in Annex A (informative) of ISO 3382-3 (2).

Building (Office.Path)	$T_{30,mid}$ (s)	$L_{p,A,B}$ (dB)	$r_{D(1)}$ (m)	$r_{D(2)}$ (m)	$D_{2,S(1)}$	$D_{2,S(2)}$	$L_{p,A,S,4m(1)}$ (dB)	$L_{p,A,S,4m(2)}$ (dB)
A (1.1)	0.7	40.0	12.9*	13.9*	5.1	4.4*	53.4*	54.0*
A (2.2)	0.7	37.0	16.5*	17.6*	4.8*	2.6*	51.3*	51.4*
B (3.3)	0.4	45.0	9.4	9.7	6.2	5.0	48.3	46.5
B (4.4)	0.4	45.0	9.9	6.9	5.0	5.8	49.1	46.4#
B (5.5)	0.4	45.0	12.2*	8.9	4.3*	5.8	48.8	46.7#
C (6.6)	1.2	41.0	9.2	7.0	3.7*	3.7*	51.7*	50.9*
D (7.7)	0.7	41.0	8.3	9.2	4.8*	5.4	47.5#	48.0#
D (8.8)	0.6	38.0	11.9*	11.0*	5.1	5.5	46.0#	46.2#
E (9.8)	0.5	46.0	7.0	7.7	4.0*	4.8*	53.8*	54.2*
E (10.10)	0.7	40.0	12.1*	9.7	5.0	4.1*	54.4*	51.6*
E (11.11)	0.6	43.0	10.3*	11.6*	4.2*	4.0*	53.2*	54.8*
F (12.12)	0.5	38.0	11.5*	11.0*	5.8	5.2	51.8*	50.3*
F (13.13)	0.5	44.0	8.7	7.9	5.0	5.4	49.0	49.2
F (14.14)	0.5	47.0	9.3	8.5	3.7*	3.0*	53.3*	53.0*
F (15.15)	0.4	54.0	8.7	9.4	3.6*	2.7*	56.0*	55.2*
F (16.16)	0.5	46.0	10.7*	9.1	5.5	3.8*	52.4*	49.0
F (17.17)	0.4	46.0	9.0	9.0	6.2	7.0#	50.7*	51.7*
G (18.18)	0.4	42.0	8.4	8.5	4.6*	4.6*	52.0*	52.4*
G (18.19)	0.4	42.0	8.5	7.6	4.0*	4.5*	52.2*	51.8*
G (18.20)	0.4	42.0	8.7	10.5*	4.7*	5.8	52.3*	54.3*
G (19.21)	0.4	48.0	6.8	6.4	4.8*	3.8*	52.7*	52.8*
G (20.22)	0.4	45.0	8.0	8.0	6.2	5.6	55.1*	53.3*
G (21.23)	0.4	44.0	7.1	6.9	4.9*	5.0	53.7*	53.6*
G (21.24)	0.4	44.0	5.1	7.3	4.3*	4.1*	52.6*	53.9*
G (21.25)	0.4	44.0	7.6	8.2	4.0*	3.1*	52.4*	53.4*
G (22.26)	0.4	44.0	11.8*	10.2*	5.7	5.4	53.8*	53.4*
G (22.27)	0.4	44.0	9.6	10.1*	4.8*	7.1#	51.4*	55.6*
G (23.28)	0.3	40.0	12.2*	11.5*	6.6	6.3	52.7*	52.3*
G (23.29)	0.3	40.0	13.1*	14.2*	5.3	5.8	51.1*	52.4*
G (24.30)	0.4	38.0	10.6*	10.3*	6.6	6.0	52.9*	52.8*
G (24.31)	0.4	38.0	11.4*	12.3*	6.4	5.4	52.9*	52.3*
H (25.32)	0.4	45.0	7.7	8.7	5.4	6.5	50.5*	52.4*
H (26.33)	0.4	40.0	11.5*	12.0*	8.7#	8.3	52.8*	52.3*
H (26.34)	0.4	40.0	8.8	9.7	6.9	6.9	47.3#	47.9#
H (27.35)	0.3	35.0	13.7*	12.0*	8.4#	7.3#	52.4*	48.7
H (27.36)	0.3	35.0	8.9	9.7	7.0#	7.0#	54.1*	53.9*

3. RESULTS AND DISCUSSION

3.1 Acoustic summary of the repeated measurements in offices

The calculated acoustic metrics are presented in Table 1, where each row contains a path with a Type1 repeated measurement, with Type2 repeated measurements in seven offices (18, 21, 22, 23, 24, 26, 27; Table 1). The current research (also published as (1)) presents the largest collection of paths/offices measured using the ISO 3382-3 method (Cabrera et al. (5) reported a subset of the current results), and is the first one that reports on more than one Type2 repeated measurements (Haapakangas et al. (11) reported results from repeated measurements in a single office described in Hongisto et al. (12)). The R^2 values for the linear regressions used to derive r_D and $D_{2,S}$ (2) were fairly high: 0.94±0.05, range: 0.75-1.00 ; and 0.95±0.04, range: 0.85-1.00, respectively. As seen in Table 1, not many paths/offices measured are indicated to have 'good' acoustic conditions as per the calculated values of the metrics, including none for r_D, whereas plenty had 'poor' or 'insufficient' acoustic conditions (2).

3.2 Variability in ΔType1 and ΔType2

Table 2: Results for analyses on ΔType1 and ΔType2 values, as described in section 2.5.1.

Type of repeated measurement	Metric	$\overline{\Delta\text{Type1 or 2}}$, bootstrapped 95%, and 68% CIs	$R^2_{\Delta\sim\text{mean}}$, associated slope, and respective 95% bootstrapped CIs
Type1 (n = 36, except for $L_{p,A,B}$, where n = 13)	r_D (m)	1.17 (0.88,1.51), (1.02,1.33)	<10^{-2} (10^{-5},0.15), 0.15 (-0.35,0.83)
	$D_{2,S}$ (dB)	0.78 (0.59,0.98), (0.68,0.88)	0.11 (10^{-2},0.35), -0.38 (-0.70,0.01)
	$L_{p,A,S,4\,m}$ (dB)	1.16 (0.82,1.53), (0.98,1.34)	0.15 (10^{-2},0.40), -0.40 (-0.89,0.03)
	$L_{p,A,B}$ (dB)	1.20 (0.72,1.73), (0.92,1.44)	0.21 (10^{-3},0.72), -2.13 (-3.97,1.05)
Type2 (n = 7)	r_D (m)	2.57 (1.79,3.42), (2.15,2.99)	0.21 (10^{-3},0.72), 0.83 (-0.12,2.69)
	$D_{2,S}$ (dB)	1.36 (0.98,1.77), (1.16, 1.57)	0.50 (10^{-4},0.95), 1.60 (-0.95,3.48)
	$L_{p,A,S,4\,m}$ (dB)	2.74 (1.33,4.27), (1.97, 3.49)	0.28 (10^{-4},0.71), -0.24 (-0.69,0.29)

3.2.1 ΔType1 results

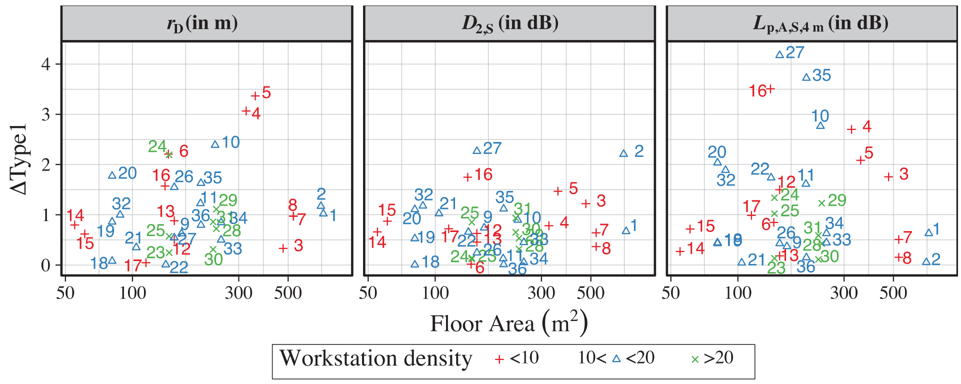

Figure 1: ΔType1 for each path in Table 1, plotted against the office floor area, where the offices are grouped according to their workstation densities.

All the $R^2_{\Delta\sim\text{mean}}$ values in Table 2, except for r_D, indicate some correlation between the ΔType1_i values (i.e., difference per i-th path, per metric, in Table 2), and the corresponding $\overline{\Delta\text{Type1}_i}$ values (i.e., mean per i-th path, per metric in Table 1). However, the 95% CIs of their respective slopes crossed zero, i.e., are non-significant. In other words, overall, ΔType1 values did not vary significantly with the magnitude of the metric. With that in mind, the $\overline{\Delta\text{Type1}}$ values (equation 1), which are based on measurements conducted with the loudspeaker on either end of the *same* path are fairly large, especially for both the SPL_{Speech} metrics ($D_{2,S}$ and $L_{p,A,S,4\,m}$), and $L_{p,A,B}$ (Table 2). However, the large $\overline{\Delta\text{Type1}}$ value for $L_{p,A,B}$ is perhaps less reliable due to the small sample size (n = 13). Hence, the effect of $L_{p,A,B}$ on r_D is not attempted here; moreover, $\overline{\Delta\text{Type1}}$ for r_D is arguably not that large (~1.2 m, Table 2).

Furthermore, while there are no obvious global trends as seen in Figure 1, several paths had a large

ΔType1 value for $L_{\mathrm{p,A,S,4\ m}}$, and to an extent $D_{2,\mathrm{S}}$, which implies that the direction of the measurement along a path is likely to be important in some cases, and needs to be judiciously considered during a measurement. In terms of mutual correlations *between* metrics for their respective $\Delta\mathrm{Type1}_i$ values: Δr_{D} was not significantly correlated to either $\Delta D_{2,\mathrm{S}}$ ($\tau = 0.07; (-0.19, 0.30)$), or to $\Delta L_{\mathrm{p,A,S,4\ m}}$ ($\tau = 0.2; (-0.02, 0.46)$); however, $\Delta D_{2,\mathrm{S}}$ was significantly correlated to $\Delta L_{\mathrm{p,A,S,4\ m}}$ ($\tau = 0.32; (0.03, 0.54)$). The latter, although not surprising given the underlying overlap in the calculation of the SPL_{Speech} metrics, is nonetheless noteworthy, and can be expressed as the following robust linear model:

$$\Delta D_{2,\mathrm{S}(\Delta\mathrm{Type1})} = 0.33 + 0.31 \times \Delta L_{\mathrm{p,A,S,4\ m}(\Delta\mathrm{Type1})} \quad (5)$$

The $R^2 = 0.46$, 95% CI: (0.16,0.79) for equation 5, and the 95% CIs for the intercept and slope are (0.12,0.50) and (0.16,0.47), respectively. Overall, it can be surmised that the SPL_{Speech} based parameters seem more susceptible to a wide variation per Type1 repeats, and any averaging of these values must be done with caution.

3.2.2 ΔType2 results

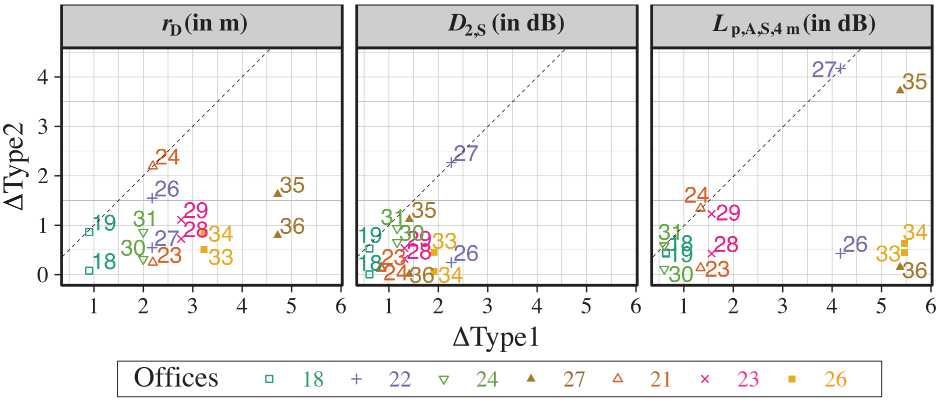

Figure 2: The relationship between the ΔType1 and ΔType2 for the offices where both these types of measurements were performed (see Table 1 and 2 for details about office and path numbers). The points plotted depict the path numbers, with the dashed line depicting the line with slope of 1 and intercept of 0.

Due to the small sample size for ΔType2 ($n = 7$), the corresponding results in Table 2 are at best considered preliminary/exploratory at this stage, and are not discussed here (although they are discussed briefly in (1)). Nevertheless, as surmised in (1) and seen in Figure 2, there are several high values of ΔType1 and ΔType2 in the current dataset. Given the plethora of variations in office topologies and the associated room acoustics variations, ISO 3382-3 perhaps needs to be very specific in what would constitute a repeated measurement that is *not* a Type1 repeat. Furthermore, there might need to be a threshold for tolerance of ΔType2 and perhaps also ΔType1, beyond which measurements cannot be used to characterize an office. Determining just-noticeable differences, especially for r_{D}, might also be beneficial.

As an aside, a comment can be made about the ISO 3382-3 requirement of keeping the loudspeaker and measurement positions at least 2 m away from large reflective surfaces (2). This requirement, which is of course important from a wave theoretical perspective, limits the use of ISO 3382-3 method to mostly medium- to large-sized offices. The current finding of large ΔType2 (and the underlying ΔType1) values provides a practice-based evidence in favour of '2 m rule', which would otherwise be a confounding factor in critically assessing results related to the ISO 3382-3 method in the current, and in future studies.

3.3 Variability in the observed Type1 and Type2 values

Based on the discussion in section 3.2, where it was shown that the findings about Type2 repeated measurements are at best exploratory, and require future studies, the focus in this section will be on Type1 repeated measurements that have larger sample size.

3.3.1 Reliability

ICC values above 0.8 are generally considered between excellent (9) in terms of reliability, which is applicable to all the metrics and both Types 1 and 2, as seen in Table 3, except for $L_{\mathrm{p,A,S,4\ m}}$ for Type2 repeats, which had low ICC, and hence, poor reliability (also non-significant). The latter provides another perspective on the large ΔType2 values, as discussed in section 3.2.2, and reinforces the need to further investigate, and define the scope of Type2 repeated measurements. Also note that the $L_{p,\mathrm{A,B}}$ ICC results are based on thirteen samples, which can be considered a small sample from a statistical perspective. Overall, the high reliability of the ISO 3382-3 metrics implies that the current findings are primarily addressing the true variability between the underlying values, with acceptable measurement errors – providing support for the critical assessment of the ISO 3382-3 measurement method in the current, and similar studies in the future.

Table 3: Some reliability (intra-measurement correlation coefficient: **ICC**) and repeatability metrics (within-measurement standard deviation: $\boldsymbol{\sigma_W}$; and repeatability limit/coefficient: ***r***) for both Type1 and Type2 repeated measurements.

Type1 **(*n* = 36, except for $L_{p,A,B}$, where *n* = 13)**			
Metric	**ICC (95% CI)**	σ_W	***r***
r_D (m)	0.87 (0.77,0.93)	0.90	2.5
$D_{2,\mathrm{S}}$ (dB)	0.85 (0.73,0.92)	0.61	1.7
$L_{\mathrm{p,A,S,4\,m}}$ (dB)	0.82 (0.68,0.90)	1.04	2.9
$L_{p,\mathrm{A,B}}$ (dB)	0.97 (0.91, 0.99)	1.01	3.0
Type2 **(*n* = 7)**			
r_D (m)	0.80 (0.55,0.95)	1.15	4.5
$D_{2,\mathrm{S}}$ (dB)	0.82 (0.56,0.96)	0.44	1.7
$L_{\mathrm{p,A,S,4\,m}}$ (dB)	0.21 (-0.1, 0.71)	1.62	6.3

3.3.2 Repeatability

Table 3, columns three and four, provide the relevant repeatability metrics. In general, the standard deviations reported here do not seem unreasonably large, especially for r_D. The only previous data in this regard was provided by Haapakangas et al. (11), where the variability, expressed as the 68% confidence interval in the ISO 3382-3 metrics, was estimated as ±1 dB for $D_{2,\mathrm{S}}$, ±1.5 dB for $L_{p,\mathrm{A,S,4,m}}$, ±1.5 m for r_D, and ±1 dB $L_{p,\mathrm{A,B}}$. These variability estimates were based on unpublished data about repeated measurements within a single open-plan office (described in Hongisto et al. (12)), which is similar in principle to the current Type2 repeats, although there is likely to be a difference in the respective methods (i.e., current and in Hongisto et. al. (12)), including the number of repeats. Besides including two types of repeated measurements, the current research includes measurements within several offices, which allows providing variability estimates that are generalizable to a wider range of offices. Nevertheless, compared to Haapakangas et al. (11), the standard deviations in the current study ($\boldsymbol{\sigma_W}$ in Table 3) were smaller for $D_{2,\mathrm{S}}$ (for both Type1 and Type2 repeats), for $L_{p,\mathrm{A,S,4,m}}$ (for Type1 repeats; Type2 unreliable), and for r_D (for both Type1 and Type2 repeats); and was almost the same for $L_{p,\mathrm{A,B}}$ (for Type1 repeats).

The repeatability coefficient/limit (*r*) per metric provides a more exhaustive, and moreover, standardized account of the variability within repeated measurements (section 2.5.2). As seen from *r* values in Table 3, for 95% of the instances of Type1 repeats – for measurements by the same observer using the current method, on the same path measured with loudspeakers on either end – the ΔType1 for $L_{\mathrm{p,A,S,4\,m}}$ is unlikely to be more than ~3 dB. Similar assessments can be made for the other metrics for Type1 repeats, whereas the *r* values for Type2 repeats are fairly large, and are not recommended for any use other than informing future research in more offices, and with, moreover, a clearly-defined (perhaps standardized in ISO 3382-3) nature of Type2 repeated measurements.

Overall, the practical implications of the standard deviations, and repeatability limit/coefficients reported here, and whether an appraisal from a theoretical/methodological perspective is required, are open questions.

4. CONCLUSIONS

The current findings represent a substantial step towards quantifying the variability of the ISO 3382-3 metrics in open-plan offices under repeatability conditions. The summary statistics, and repeatability values reported here may be considered reasonable for the most part, especially for r_D; the variability of the SPL_{Speech} metrics may need to be considered further from a practical perspective. In this regard, the excellent reliability values reported here imply that future investigations with repeated measurements are more likely to be studying the 'true' variability between ISO 3382-3 metrics in offices, with acceptable measurement error.

For the Type1 repeats, the results here provide reinforcement for the ISO 3382-3 requirement of at least measuring a path from two opposite direction, if it is the only path being measured. Moreover, we highly recommended that *all* paths be measured from both ends. For exceptionally large ΔType1 values, it would be apt to flag such cases while reporting/using the results, and ISO 3382-3 may benefit from amendments to its reporting format (Figure 3/Table 2 in (2)) to highlight such cases, and perhaps more crucially, include instructions/comments on the relevance and potential remedies to large ΔType1 from a practitioner's perspective. The practical applicability of Type2 results is currently limited due to small sample size in this paper. However, the results here suggest potential updates to the ISO 3382-3 methodology regarding latent issues in Type2, e.g., clarity in its specification, etc.

Ultimately, more work is needed to further strengthen ISO 3382-3 in relation to measurement *repeatability* (from current and future findings based on (6)), other aspects of variability in repeated measurements including *reproducibility* (based on (6)), and perhaps even *uncertainty* at various stages in the methodology (based on (13), similar to (14)).

ACKNOWLEDGMENTS

This study was funded through the Australian Research Council's Discovery Projects scheme (project DP160103978).

REFERENCES

1. Yadav M, Cabrera D, Love J, Kim J, Holmes J, Caldwell H, et al. Reliability and repeatability of ISO 3382-3 metrics based on repeated acoustic measurements in open-plan offices. Appl Acoust. 2019 Jul 1;150:138–46.
2. ISO 3382-3 Acoustics – measurement of room acoustic parameters – Part 3: Open plan offices. International Organization for Standardization, Geneva, Switzerland; 2012.
3. Cabrera D, Jimenez D, Martens WL. Audio and Acoustical Response Analysis Environment (AARAE): a tool to support education and research in acoustics. In: Proceedings of Internoise. Melbourne, Australia; 2014.
4. ISO 3745 Acoustics -- Determination of sound power levels and sound energy levels of noise sources using sound pressure -- Precision methods for anechoic rooms and hemi-anechoic rooms. International Organization for Standardization, Geneva, Switzerland; 2012.
5. Cabrera D, Yadav M, Protheroe D. Critical methodological assessment of the distraction distance used for evaluating room acoustic quality of open-plan offices. Appl Acoust. 2018 Nov;140:132–42.
6. ISO 5725 1-6 Accuracy (trueness and precision) of measurement methods and results. International Organization for Standardization, Geneva, Switzerland; 1994.
7. Maechler M, Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, et al. robustbase: Basic Robust Statistics. R package version 0.93-3. URL http://CRAN.R-project.org/package=robustbase [Internet]. 2018. Available from: http://robustbase.r-forge.r-project.org/
8. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultrasound Obstet Gynecol. 2008 Apr 1;31(4):466–75.
9. Fleiss JL. Design and analysis of clinical experiments. Vol. 73. John Wiley & Sons; 2011.
10. Koller M. robustlmm: An R Package for Robust Estimation of Linear Mixed-Effects Models. J Stat Softw. 2016;75(6):1–24.
11. Haapakangas A, Hongisto V, Eerola M, Kuusisto T. Distraction distance and perceived disturbance by noise—An analysis of 21 open-plan offices. J Acoust Soc Am. 2017 Jan;141(1):127–36.
12. Hongisto V, Haapakangas A, Varjo J, Helenius R, Koskela H. Refurbishment of an open-plan office – Environmental and job satisfaction. J Environ Psychol. 2016 Mar;45:176–91.
13. ISO, IEC, OIML, BIPM. Guide to the Expression of Uncertainty in Measurement. International Organization for Standardization, Geneva, Switzerland; 1995.
14. Dietrich P. Modeling Measurement Uncertainty in Room Acoustics. Acta Polytech [Internet]. 2007 Jan 4 [cited 2018 Dec 31];47(4–5). Available from: https://ojs.cvut.cz/ojs/index.php/ap/article/view/972