Behavior Research Methods (2018) 50:57–83
https://doi.org/10.3758/s13428-017-1005-4

How honest are the signals? A protocol for validating wearable sensors

Varol Onur Kayhan[1,2] · Zheng (Chris) Chen[1] · Kimberly A. French[3] · Tammy D. Allen[4] · Kristen Salomon[4] · Alison Watkins[1]

Published online: 12 January 2018

Abstract
There is growing interest among organizational researchers in tapping into alternative sources of data beyond self-reports to provide a new avenue for measuring behavioral constructs. Use of alternative data sources such as wearable sensors is necessary for developing theory and enhancing organizational practice. Although wearable sensors are now commercially available, the veracity of the data they capture is largely unknown and mostly based on manufacturers' claims. The goal of this research is to test the validity and reliability of data captured by one such wearable badge (by Humanyze) in the context of structured meetings where all individuals wear a badge for the duration of the encounter. We developed a series of studies, each targeting a specific sensor of this badge that is relevant for structured meetings, and we make specific recommendations for badge data usage based on our validation results. We have incorporated the insights from our studies on a website that researchers can use to conduct validation tests for their badges, upload their data, and assess the validity of the data. We discuss this website in the corresponding studies.

Keywords Wearable sensors · Unobtrusive measures · Machine learning

Introduction

There is growing interest among organizational researchers in tapping into alternative sources of data beyond self-reports to provide a new avenue for measuring behavioral constructs. Use of alternative data sources is necessary for developing theory and enhancing organizational practice. One such alternative data source is wearable sensors that collect data about organizational members' day-to-day activities and interactions. Although wearable sensors are now commercially available, the veracity of the data they capture is largely unknown and mostly based on manufacturers' claims.

The goal of the present research is to test the validity and reliability of data captured by a particular wearable sensor, the Sociometric badge, an unobtrusive device originally developed by the MIT Media Laboratory and later commercialized by Sociometric Solutions (now Humanyze). The badge uses four sensors – a microphone, an infrared sensor, a Bluetooth detector, and an accelerometer – to capture data about the wearer's voice, face-to-face interactions, proximity to other wearers, and body motion. Despite the possibilities offered by this new technology, ambiguities exist regarding the validity of the data captured. The evidence supporting their use is based primarily on studies conducted by the research team involved in their development. To address this issue, Chaffin, et al. (2017) conducted a series of studies to validate the data captured for *unstructured meetings*, in which participants move in and out of conversations and can interact with anyone, including those who do not wear badges. Our study

✉ Varol Onur Kayhan

Zheng (Chris) Chen
zhengchen@mail.usf.edu

Kimberly A. French
KFrench0429@gmail.com

Tammy D. Allen
tallen@usf.edu

Kristen Salomon
ksalomon@usf.edu

Alison Watkins
awatkins@usfsp.edu

[1] Kate Tiedemann College of Business, University of South Florida St. Petersburg, St. Petersburg, FL, USA

[2] St. Petersburg, USA

[3] School of Psychology, Georgia Institute of Technology, Atlanta, GA, USA

[4] College of Arts and Sciences, University of South Florida, Tampa, FL, USA

builds on their work by collecting data and investigating the metrics for *structured meetings*, in which all individuals present wear a badge for the duration of the encounter (Sociometric Solutions, 2015). Table 1 presents and defines the metrics generated by the four types of sensors in these badges and differentiates the metrics tested by Chaffin et al. (2017) and by the present study.

We have developed a series of validation studies, each targeting a specific badge sensor and metrics relevant for structured meetings recommended by Sociometric Solution's manual (Sociometric Solutions, 2015). Based on the validation results, we make specific recommendations for data usage. We also incorporate the insights obtained from our studies on a website that allows researchers to conduct validation tests with their own badges, upload their data, and assess the validity of the data captured by the badges (http://www.badgevalidation.com). We discuss this website as we present the corresponding studies.

Our research extends existing knowledge concerning these badges, and wearable sensors in general, in four key ways. First, we focus on structured meetings and extend the work of Chaffin et al. (2017). While Chaffin et al. (2017) focus on unstructured meetings and predominantly test the Bluetooth and infrared sensors that are more relevant to unstructured interactions, we conduct an in-depth analysis of both the accelerometer and the microphone sensors (Study 2 and Study 3) due to their ability to measure finer-grained details relevant to structured interactions. Our motivation to focus on structured meetings is that the amount of time spent in meetings by organizational members is increasing every year. Reports indicate that managers spend 35–50 % of their time in meetings (Dockweiler, 2014). Considering that these meetings may involve anything from developing new products or services to setting organizational strategy, data collected in these meetings can not only advance our understanding of theories that govern interactions among organizational members but also shed light on grand theories. For example, volume, vocal variability, or body movement – which have been linked to dominance, power, and status (Hall, Coats, & LeBeau, 2005) – can be measured using the microphone and accelerometer sensors, and thus provide insights into the corresponding theories. Similarly, the same two sensors can capture vocal dynamics and mirroring, which have been shown to explain negotiation outcomes (Curhan & Pentland, 2007).

Second, we develop a data validation protocol for the microphone and accelerometer sensors that can be used by other researchers to validate their own badges (Study 2 and Study 3). While each research study that uses these badges may have its own measurement protocol, our study provides a guideline on how to check the validity of the badges in a systematic fashion before following a measurement protocol. This is important because it can not only unveil differences in badges (due to the manufacturing or the calibration processes), but also inform future research on how to ensure validity of other types of wearable sensors – besides the badges examined in this study – that can capture sound, movement, or any other type of stimuli.

Third, when we uncover data validity issues, we provide alternative strategies—such as use of machine learning—that yield more accurate data. Fourth, we extend previous work by testing basic assumptions (time synchronization and gendered vocal patterns) that have not previously been examined but have critical implications for the use of the data captured by the badges (Study 1). The sections that follow present in detail our three studies that were aimed at developing a validation protocol using 10 Sociometric badges.

The lessons learned from our research extend beyond the specific wearable badge we test. The current speed of technology leads to the development of new or updated wearables on a regular basis. While these wearables provide new avenues for advancing organizational theories, it is incumbent upon researchers to understand their capabilities and to determine the validity of the data they capture and generate. Our research provides insights about how to do this in a systematic and controlled fashion using a specific type of wearable. However, this may guide future research on the collection of data using any type of wearable, irrespective of its brand.

Background

We developed a validation protocol for the Sociometric badge, an unobtrusive device originally developed by the MIT Media Laboratory and later commercialized by Sociometric Solutions (now Humanyze). The badge is worn around the neck on a lanyard and is the approximate size and shape of an ID tag (Fig. 1). It uses four sensors (a microphone, an infrared sensor, a Bluetooth detector, and an accelerometer) to capture data about the wearer's voice, face-to-face interactions, proximity to other wearers, and body motion. Even though these data can be considered "raw," they are generated using the badge firmware and do not necessarily reflect the true values for the external stimuli observed in the data collection setting (Sociometric Solutions, 2014). For example, a stable tone at 170 Hz is captured as 307 Hz by the badge due to the ways in which its proprietary firmware processes this tone and records it in its internal memory. The Sociometric Solutions software – which is required for exporting the data from the badges – also generates new data by processing the raw data collected by the badge. For example, the accelerometer sensor collects activity data (raw data based on the wearer's physical activity), and these are then processed by the Sociometrics Solutions software to derive binary data about *walking* – to indicate whether the badge wearer was walking (1) or not (0) at each given second in time. We therefore use the term "raw data" in this paper to refer to data that

Table 1. Differences between Chaffin et al. (2017) and the current study

Sensor	Metric	Description	Unstructured meetings	Structured meetings
Bluetooth	Proximity	Number of times badges' Bluetooth sensors detect each other	O	N/A[1]
Infrared	Face-to-face interaction	Number of times badges' infrared sensors detect each other	O	N/A[1]
Microphone	Frequency	Frequency of sound captured by badges	O	X
	Amplitude	Amplitude of sound captured by badges	O	N/A[2]
	Volume	Volume of sound captured by badges		X
	Pitch	Pitch of sound captured by badges		X
	Speech	Speech detection, overlap, turn-taking between speakers	O	X
Accelerometer	Left–right posture	The degree of tilting between left and right		X
	Front–back posture	The degree of tilting between front and back		X
	Activity	The level of physical activity		X
	Mirroring	The level of similarity of body movement between two badges		X

Note. O: Tested in Chaffin et al.; X: Tested in current study. The current study tested all metrics recommend by Sociometric Solutions (2015)

[1] Bluetooth and infrared are not particularly relevant to structured meetings, because all participations are present in one meeting and engaged in one conversation

[2] Not recommended for structured meetings by Sociometic Solutions (2015)

are only processed by the badge firmware and "derived metrics" to refer to data that have been further processed by the Sociometric Solutions software. The file exported from the badge includes both raw data and derived metrics. We note that the badge does not audio record verbatim speech.

The Sociometric badges have been used to predict organizationally relevant outcomes such as job attitudes and performance (Olguin-Olguin & Pentland, 2010b), job satisfaction (Olguin-Olguin, Waber, et al., 2009), workspace design (Orbach, Demko, Doyle, Waber, & Pentland, 2015), personal and group interaction satisfaction (Waber, Olguin-Olguin, Kim, & Pentland, 2008), network cohesion (Wu, Waber, Aral, Brynjolfsson, & Pentland, 2008), creativity (Tripathi & Burleson, 2012), personality traits (Olguin-Olguin, Gloor, & Pentland, 2009), group performance (Olguin-Olguin, Gloor, et al. 2009; Olguin-Olguin & Pentland, 2010a), and group collaboration (Kim, Chang, Holland, & Pentland, 2008). A brief summary of these studies showing the types of sensors and metrics used in prior work is presented in Table 2.

Fig. 1 Sociometric badge

In this paper, we present three studies (see Table 3). In Study 1, we test the *synchronicity* assumption, which posits that the badges' internal clocks are in sync with each other as well as with the real-world clock. While this assumption is taken for granted in earlier work, including Chaffin et al. (2017), we show that there is a lack of synchronicity between badges, which creates issues for metrics derived by the Sociometric Solutions software. In Study 2, we focus on the microphone sensor and test the validity of both its raw data (volume, frequency, and pitch) and the derived metrics (speaking, silence, overlap, and turn-taking). In Study 3, we focus on the accelerometer sensor and test posture and movement (raw data) as well as mirroring (derived metric). We note that all studies concerning the microphone (Study 1 and 2) were conducted in an isolated quiet office to minimize background noise.

Study 1: Synchronicity assumptions

In Study 1, we tested two assumptions about badge functioning by examining (a) the degree to which badge clocks are synced with the real-world clock (and thus with one another) and (b) the degree to which a badge's microphone and accelerometer are synced (i.e., capture the same event at the same time). The badges we used in Study 1 (and the subsequent

Table 2. Summary of prior work

Study	Sensor	Metric	Dependent variable
Kim et al., 2008	Accelerometer Microphone	Activity Speaking	Group collaboration
Waber et al., 2008	Infrared Bluetooth Accelerometer Microphone	Face-to-face interactions Proximity Activity Speaking	Personal and group interaction satisfaction, productivity
Wu et al., 2008	Infrared Bluetooth Microphone	Face-to-face interactions Proximity Tonal variation	Network cohesion
Olguin-Olguin, Gloor, et al., 2009	Infrared Bluetooth Accelerometer Microphone	Face-to-face interactions Proximity Activity Speaking	Personality traits, group performance
Olguin-Olguin, Waber, et al., 2009	Infrared Bluetooth Accelerometer Microphone	Face-to-face interactions Proximity Activity Speaking	Job satisfaction, group interaction
Olguin-Olguin & Pentland, 2010a	Accelerometer Microphone	Activity Speaking	Group performance
Olguin-Olguin & Pentland, 2010b	Infrared Bluetooth	Face-to-face interactions Proximity	Job attitudes and performance
Tripathi & Burleson, 2012	Infrared Bluetooth	Face-to-face interactions Proximity	Creativity
Orbach et al., 2015	Infrared Bluetooth	Face-to-face interactions Proximity	Workspace design

studies) had firmware version 3.1.2669. We changed the firmware's setting for collecting data from the default setting of 0.5 s to 0.1 s, as recommended for structured meetings (Sociometric Solutions, 2015).

Study 1a: Synchronicity between badges

It is important to synchronize the badges with the real-world clock to ensure that the internal clocks do not drift. Such

Table 3. Summary of current work

Study	Sensor tested	Metrics used	Data type	Study characteristics	Treatment
Study 1a	Microphone	Frequency	Raw	10 badges, 1 session	30-s stable tone
Study 1b	Microphone and accelerometer	Frequency and activity	Raw	10 badges, 1 session	30-s stable tone
Study 2a	Microphone	Volume	Raw	10 badges, 3 sessions	Three volume settings (ambient, normal, normal × 3)
Study 2b	Microphone	Frequency	Raw	10 badges, 1 session	Five types of frequencies (stable tone, ambient, sweep tone, male speech, and female speech)
Study 2c	Microphone	Pitch	Raw	10 badges, 3 sessions	Machine-generated tones and scripted speaking
Study 2d	Microphone	Speaking, silent, overlapping, turn-taking	Derived	2 badges, 1 session (conducted twice using different pairs)	Simulated conversation between a male and a female speaker (machine-generated scripted speaking)
Study 3a	Accelerometer	Posture			
		i. front–back	Raw	10 badges, 3 sessions	Three types of controlled movement (stationary, front, back)
		ii. left–right	Raw	10 badges, 3 sessions	Three types of controlled movement (stationary, left, right)
Study 3b	Accelerometer	Activity	Raw	10 badges, 1 session	Three controlled activity levels (stationary, moderate, high)
Study 3c	Accelerometer	Mirroring	Derived	10 badges, 3 sessions	Three controlled activity levels (stationary, moderate, high)

drifting could result in invalid combinations and comparisons of data across badges regarding phenomena as they occur in real time. The Sociometric Solutions (2014) user manual suggests synchronizing the internal clocks by installing the Sociometrics Solutions software on a computer and connecting the badges to this computer through a USB port. In this study, we tested whether the internal clocks were in sync with the real-world clock and with one another.

Experimental procedure After connecting all the badges to our computer as recommended, we turned the badges on. Then, at 11:07:00 a.m., we played a stable tone (170 Hz) for 30 s. After turning the badges off, we exported the data by seconds. The 30-s stable tone generated 30 rows of consistent frequency values in the exported data file. For each badge, we checked whether the values started at the marked time (11:07:00 a.m.).

Results The badges failed to capture the stable tone at the marked time or with the same timestamp (Table 7 of Appendix 1). The badges were nearly 3 min ahead of the real-world clock. Between-badge comparison showed that the internal clocks were anywhere between 1 and 35 s apart from each other.

A repetition of this experiment on another day (after synching the badges with the computer) generated similar results (Table 8 of Appendix 1). The badges were nearly 3.5 min ahead of the real-world clock, and their internal clocks were anywhere between 0 and 12 s apart from each other.

Discussion and recommendations The results show that the badges were not in sync with the real-world clock, nor with each other. Further, the time difference between the same two badges was not consistent across experimental sessions. For example, the 4-s delay observed between the first two badges (Badge 1 and 2) in the first experiment (see Table 7 of Appendix 1) was not observed in the second experiment (see Table 8 of Appendix 1).

The lack of synchronicity between badges is problematic for a variety of reasons. If data captured by badges do not have the same timestamp for the same event, any analysis that relies on data from multiple badges will lead to inaccurate results because the same event will be interpreted as multiple events. For example, consider the first two badges in the first experiment: Badge 2 registers the stable tone 4 s later than Badge 1. If the delay is not accounted for, the data will show that an external stimulus (such as a stable tone) occurred at two different time points. As we discuss later, this is particularly problematic for metrics that the Sociometric Solutions software derives from raw data, because the software observes the same event happening at different points in time (e.g., ten different times, as shown in Table 7 of Appendix 1). Therefore, derived metrics that rely on the synchronicity of events (such as turn-taking or mirroring of body movements) will be largely inaccurate.

To resolve this issue, we played a 30-s stable tone at the beginning of each data collection using the badges. The stable tone is a consistent sound at 170 Hz, the average frequency of the human voice (Titze, 1994). The 30-s stable tone generated 30 rows of consistent frequency data (when analyzed by seconds) in the frequency section of the exported data file. We used these rows as a marker to determine the extent of drift across badges, align badge timelines, and achieve synchronicity (in all subsequent studies discussed in this paper). It is important to note that this does not sync the badges' internal clocks. Rather, it enables to align the data after the data are exported from the badges and helps eliminate the drift during data analysis. We also recommend that this approach be used each time the badges are turned on and off since the drift in internal clocks is not consistent between any pair of badges.

However, this approach does not alleviate the problems of derived metrics that rely on data captured by multiple badges, because Sociometric Solutions software cannot account for the drift, and thus computes the derived metrics based on data captured in different timelines. We make separate recommendations to address the derived metrics later in this paper.

We also note that when we turned off the badges immediately after an experimental treatment, the export process truncated the data, causing us to lose valuable data. We speculate that this is also due to the lack of synchronicity between the badges' internal clocks and the real-world clock. To resolve this issue, we left the badges on for a minimum of 3 min at the end of each data collection.

Due to the importance of synchronicity and its implications for the data collected by the badges, we have developed a website (http://www.badgevalidation.com) where researchers can download the 170-Hz stable tone sound file, upload their data, and assess the level of synchronicity between their badges. (Figure 4 of Appendix 1 shows an example analysis of synchronicity on this website.)

Study 1b: Microphone and accelerometer test

Next, we tested whether the data captured by a badge's microphone and accelerometer were synchronized – that is, whether the microphone and accelerometer sensors captured the same event with the same timestamp when an event simultaneously triggered both sensors.

Experimental procedure We employed two experimental conditions. For the first condition, we kept the badges stationary for 30 s in silence (i.e., ambient noise in a quiet office), and for the second, one of the researchers wore all badges and walked for 30 s while the stable tone played. We anticipated that if the microphone and accelerometer worked in sync, then the timestamp of the stable tone's frequency data (raw data)

would coincide with the timestamp of the activity data (raw data capturing the level of physical activity) in the exported file. We captured data for only one session because it was not possible to precisely replicate the walking condition.

Results To evaluate the microphone and accelerometer synchronization, we visually inspected the alignment of the raw data from each sensor and also examined the covariation of the frequency values (raw data) and walking values (derived metric).

Visual inspection We examined the frequency and the activity data in the exported file. Per the Sociometric Solutions (2014) user manual, the activity data captured by the accelerometer are continuous in nature and indicate how physically active a person is while wearing a badge. Values closer to 0 indicate no activity, while values higher than 0.002 indicate activity—the higher the value, the greater the activity. We opted for a visual inspection rather than a statistical analysis (e.g., ANOVA) because statistical analyses can obscure a time lag in the data. To this end, we superimposed the frequency data on top of the activity data and plotted the activity values for the 60-s duration (separately for each badge). The chart (Figure 5 of Appendix 2) shows that the microphones and accelerometers were indeed in sync: there was a jump in the activity values at the end of the 30-s mark, which is the beginning of the walking condition.

Frequency and walking covariation The Sociometric Solutions software uses a badge wearer's activity data to determine whether that person is walking or not at a specific point in time. Because this is a derived metric, it is binary in nature: it is 1 if the person is walking and 0 if the person is stationary. The algorithm derives the walking metric for each second of data. This similarly provided us with an opportunity to check whether the badges indicated walking while the stable tone played. To this end, we performed binary coding on the frequency values: we coded all ambient frequencies as 0 and all stable tone frequencies as 1. Therefore, we had 30 s of 0s (for ambient noise while the badges were stationary), and 30 s of 1s (for a stable tone while the badges were moving). We conducted a correlation analysis between the frequency and derived walking values for each badge to statistically examine their covariation. A correlation coefficient of 1.00 would indicate perfect covariation between the frequency and derived walking metrics. The correlation values were between $r = 0.59$ and $r = 0.69$. A Chi-square test further showed that the walking metric could only account for 52–65% of the stable frequency (i.e., objective walking; Chi-square values across badges ranged from 21 to 29 with all $p < 0.001$). We visually inspected the data to determine why this relationship was weaker than expected. Our inspection revealed that there was an average delay of 11.5 s between the beginning of the stable tone and the first indication of walking. This indicates that while the activity values (raw data) showed that the microphone and accelerometer worked in sync, the walking values (derived metric) did not support this.

Discussion and recommendations The results show that the badges' microphone and accelerometer sensors work in sync. This means that they rely on the same internal clock to capture data. Therefore, an event that triggers both the microphone and accelerometer simultaneously has the same timestamp across the data captured by these two sensors. This is important for data integrity and analyses that rely on second-level data.

However, it is worth noting that the synchronicity between the sensors was observed for the raw activity data, but not for the walking metric derived by the Sociometric Solutions software. This is likely due to the settings of the proprietary algorithm used by the Sociometric Solutions software to derive this metric. We recommend that researchers rely on raw activity data rather than the binary-coded walking metric in situations where synchronicity of events is examined while badge wearers are walking.

Study 2: Microphone

Study 2a: Volume

In this study, our goal was to test whether the badges captured changes in speech volume (raw data) accurately. We created three experimental conditions: no speech (ambient noise), Speech 1, and Speech 2. The volume setting of the Speech 2 condition was three times the volume setting of the Speech 1 condition.

Experimental procedure We created machine-generated speech using a freeware text-to-speech program (http://text-to-speech.imtranslator.net/). The speech, hereafter referred to as Speech 1, was 20 s long and was generated using a female voice that read a specific passage of text. We then used TechSmith's Camtasia program to generate a second copy of the same speech (hereafter referred to as Speech 2) at three times the original volume. Also using Camtasia, we combined these two speeches in a single 60-s recording: the first 20 s consisted of Speech 1, the second 20 s consisted of silence (i.e., ambient noise), and the last 20 s consisted of Speech 2. For the data collection, we placed all 10 badges approximately 3 feet away from a speaker in a quiet room. Data were collected for three sessions, with the badges turned off between sessions.

Results We conducted an ANOVA to test the effects of session, badge, and condition (the latter comprising three

categories: ambient, Speech 1, and Speech 2) on volume. We anticipated significant differences for the conditions, but not for sessions or badges. The results showed that the independent variables explained 72 % of the variance in volume, and pairwise comparisons between conditions were significant ($p < 0.001$). Thus, the badges could detect differences in volume across conditions. However, session and badge were also significant (at $p = 0.010$ and $p < 0.001$, respectively), suggesting that changes in volume were captured differently across badges and sessions (Table 9 of Appendix 3).

In an effort to shed light on differences across badges and sessions, we examined the volume data captured by all badges in each session (see Fig. 6 of Appendix 3). Although the badges captured the changes in volume correctly, the volume levels differed for each badge. For example, one badge registered a maximum volume of 0.044, while another badge registered a maximum volume of 0.027 in the same session. We standardized the volume data by badge. An ANOVA using these standardized values suggested that while condition was significant ($p < .001$) and badge was not ($p = 1.00$), there were still session-level differences ($p = 0.012$). We further standardized the data by session. The ANOVA results suggested that condition was significant ($p < 0.001$). However, neither badge ($p = 1.00$) nor session ($p = 1.00$) was significant.

Discussion and recommendations The results show that the badges capture changes in volume accurately. However, the badges register different values within and between sessions, even for the same experimental condition. Two possible explanations for this are: (1) the experimental sessions were tainted by background noise, making it difficult to replicate the results across sessions, and (2) the microphones do not have the same level of sensitivity for volume and therefore register different values. Notably, the variance explained by the volume condition was greater than the variance explained by badge or session. Moving forward, we recommend that volume data be standardized by badge to examine the relative changes in volume.

Study 2b: Frequency

Next, we tested whether the microphone could distinguish different sound frequencies (raw data). We replicated and extended Chaffin et al.'s (2017) microphone experiment by examining whether the badge microphone can distinguish different sounds in the environment. In addition to the three conditions that Chaffin et al. (2017) tested (stable tone, sweep tone, and ambient noise), we added male and female speech to the experimental design.

Experimental procedure We created a recording that included five 30-s segments: (1) a stable tone (a consistent 170-Hz tone), (2) ambient noise (i.e., silence), (3) a sweep tone (a frequency ranging from 20 Hz to 2,000 Hz), (4) male speech, and (5) female speech. The male and female speech were created using the text-to-speech program, which read the same passage of text aloud. The total duration of the recording was 150 s. We placed the ten badges approximately 3 f. away from a pair of speakers in a quiet room. We played the recording from beginning to end for three sessions, with the badges turned off between sessions.

Results We conducted an ANOVA to check whether the badges distinguished between the sound categories based on audio frequency. We entered badge, session, and sound (the latter comprising five categories: stable, sweep, ambient, male, and female) and their interactions as independent variables. We anticipated significant sound differences, but no badge or session differences. Such a pattern would suggest that the microphones can distinguish sound conditions and do not differ based on session or badge.

The results suggested that session, badge, and sound were significant ($p < .009$) and that all variables explained 71.8 % of the variance in frequency (Table 10 of Appendix 4). All pairwise comparisons between the sound categories were significant at $p < .001$, indicating that the badges could detect and distinguish each sound category.

Similar to the results in Study 2a, significant badge and session parameters were found. It appears that the same sound was captured with different frequency values by different badges. To account for these differences, we standardized all frequency data within each badge, allowing us to eliminate absolute differences in frequency across badges. We then reran the ANOVA with the standardized frequencies. As expected, the badge variable became nonsignificant in the ANOVA test, but the session variable was still significant ($p = .01$), indicating that the badges captured inconsistent data across sessions. Again, we further standardized the data by session. In this case, both the badge and the session variables were nonsignificant ($p = .99$ and $p = .21$, respectively), while the sound parameter was significant ($p < .001$).

Discussion and recommendations The results indicate that the badges can accurately detect changes in frequency. The frequency conditions explained most of the variance in frequency values registered by the badges. However, although the relative changes in frequency are equivalent, badge microphones might not have the same set-point level of frequency across badges and sessions. Consequently, absolute comparisons of frequency levels across badges or sessions will contain errors. This issue is similar to the volume results of Study 2a. Researchers can standardize data within each badge and session to identify and compare changes in frequency.

Study 2c: Pitch

In this study, our goal was to test whether the badges capture changes in pitch (raw data).

Experimental procedure We used five sounds characterized by different pitch values: (1) a stable tone (a consistent 170 Hz tone), (2) ambient noise, (3) a sweep tone (frequencies ranging from 20 to 2,000 Hz), (4) male speech, and (5) female speech. The male and female speeches were machine generated, as discussed earlier. We created a 210-s recording in which the following conditions were played for 30 s each: stable tone, ambient noise, sweep tone, stable tone, male speech, stable tone, female speech. We captured data for three sessions, with the badges turned off between sessions.

Results We conducted an ANOVA using the badges' pitch data as the dependent variable. The independent variables were sound (comprising five categories: stable, ambient, sweep, male, female), session, and badge. The results (Table 11 of Appendix 5) suggest that the badges captured the variation in pitch ($p < 0.001$). There were no significant differences between sessions ($p = 0.15$) or badges ($p = 0.50$). Most of the pairwise sound comparisons were significantly different ($p < 0.001$), with three exceptions: comparisons between ambient and stable ($p = 0.07$), female and sweeping ($p = 0.47$), and male and stable ($p = 0.51$) sounds were nonsignificant.

Discussion and recommendations The results show that the badges can distinguish changes in pitch, and that there are no differences in pitch across badges and sessions. Thus, both changes in pitch and the absolute values of pitch are comparable across badges and sessions. However, some of the nonsignificant pairwise comparisons raise concerns regarding the accuracy of the pitch values captured by the badges. Two factors may explain the nonsignificant findings. First, the pitch values of the experimental conditions may in reality be similar, and thus nonsignificant comparisons are an indication of pitch data validity. This is difficult to test, because pitch values registered on an oscilloscope are different than the values captured by the badges. Even though the pitch values are considered raw data, they are captured using Sociometric Solutions' proprietary firmware, which may involve nonlinear transformations of the true pitch or data other than the true pitch (Sociometric Solutions, 2014).

The second, and perhaps more plausible, explanation is that the missing values in the data impede pairwise comparisons. An examination of the data reveals that the badges captured pitch values sporadically: 69 % of the data were missing in Session 1, 70 % of the data were missing in Session 2, and 72 % were missing in Session 3. The missing data were not due to experimental error or device malfunction, as the badges captured other types of data, such as frequency, for the same time frames. This could explain why the ambient noise and stable tone were not statistically significantly different from each other.

We recommend that researchers use caution when using the pitch data provided in the exported file. The extent of missing values can hamper any attempt to use pitch as a construct in research studies. Even though several imputing techniques (such as using average values in lieu of missing values) can handle missing values, researchers must exercise caution since these techniques can lead to spurious findings.

Study 2d: Speech

In this study, our goal was to examine whether the badges could identify speech conditions (speaking, listening, silence, and overlapping speech) as well as conversation characteristics (number of speaking segments, turn-taking, pausing, successful interruption, and unsuccessful interruption). Note that all speech conditions and conversation characteristics are derived metrics that the Sociometric Solutions (2015) manual recommends be used by researchers for structured meetings.

Experimental procedure We placed two laptops with external speakers approximately 3 f. apart in a quiet room; the speakers faced each other and were elevated 10 in. above the desk the laptops were on. We placed a badge approximately 3 in. directly beneath each speaker, as if the speakers were wearing the badges. We created two recordings, one using the text-to-speech female voice and the other using the text-to-speech male voice, as outlined earlier. Both recordings were based on the same text, and each lasted 4 min 45 s. Each recording included speaking segments (15 s each) and silent segments (5 s each). Because the badges identify any pause longer than 0.5 s as the beginning of a new speaking segment (Sociometric Solutions, 2014), the speaking segments were continuous, with no pauses lasting longer than 0.5 s. We played the female and male recordings from the separate laptops, simulating a back-and-forth conversation between the speakers. The speech conditions (speaking, listening, silence, and overlapping speech) included in the recordings generated a conversation that could be characterized by other metrics such as the number of speaking segments, the number of turns taken, the number of self-turns, the number of successful interruptions, and the number of unsuccessful interruptions. We followed the definitions in the Sociometric Solutions manual (2014) shown in Table 12 of Appendix 6 while generating this simulated conversation. For example, we used 3 s of overlapping speech for a successful interruption (the maximum is 5 s) and 7 s for an unsuccessful interruption (the minimum is 5 s). The speaking segments of each speaker along the timeline are presented in Fig. 2.

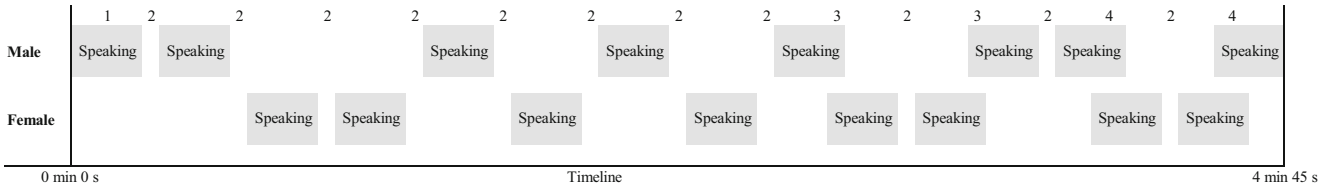


Fig. 2 Study 2d. Speaking segments of the male and female speakers along the timeline. 1 = speaking segment (each is 15 s long); 2 = silent segment (each is 5 s long); 3 = overlapping speech (3 s long); 4 = overlapping speech (7 s long). Note that "1" appears only once in order to streamline the figure presentation

The data were collected in a single session. We initially used all badges in the same session (five in front of the male speaker and five in front of the female speaker). However, use of more than one pair created problems during speech assignment. Specifically, the Sociometrics Solutions algorithm had difficulty assigning speech to the data, flagging all speech as overlap among all badges. We therefore opted to use a single pair of randomly selected badges from our pool of ten badges. To ensure validity, we repeated the experiment with a second pair of badges (randomly selected) from the pool.

Results *Pair 1.* Each speech condition (speaking, silence, overlapping, and listening) is a continuous metric derived by the Sociometric Solutions (Sociometric Solutions, 2014) software. Values are provided for each second, representing the proportion of each second scored as relevant for that speech condition. For example, 0.6 for listening in a given second indicates that the badge scored 6/10 of that second as time during which the individual was listening. We conducted a correlation analysis between the actual and captured values. The correlation analysis is presented in Table 13 of Appendix 6. For speaking, the correlation between actual female speaking and female speaking captured by the female badge was $r = 0.48$ ($p < 0.001$), whereas the correlation between actual male speaking and male speaking captured by the male badge was $r = 0.41$ ($p < 0.001$). For overlap, the correlation was low ($r = 0.25$, $p < 0.001$) for each badge. Both badges accurately captured silence ($r = 0.86$, $p < 0.001$ for each badge). For listening, the male badge had a correlation value of $r = 0.48$ ($p < 0.001$), while the female badge had a value of $r = 0.41$ ($p < 0.001$).

Next, we looked at the conversation characteristics derived from the speech conditions by the Sociometric Solutions software: number of turns taken, number of self-turns, number of speaking segments, number of successful interruptions, and number of unsuccessful interruptions. The definitions of these – per the Sociometric Solutions (2014) manual – are provided in Table 12 of Appendix 6. The results (presented in Table 4) show that the conversation characteristics captured by the badges were overestimated by a large margin. For example, the female badge reported the total number of turns taken as 54 and the male badge reported these as 67, even though the actual values were both 6.

Pair 2. The results were similar to those for pair 1 (see Table 14 of Appendix 6). For speaking, the correlation between actual and captured female speaking (by the female badge) was $r = 0.52$ ($p < 0.001$); whereas the correlation between actual and captured male speaking (by the male badge) was $r = 0.22$ ($p < 0.001$). The correlation for overlap was low ($r = 0.21$, $p < 0.001$ reported by the male badge; $r = 0.23$, $p < 0.001$ reported by the female badge). However, silence was captured accurately by both badges ($r = 0.90$, $p < 0.001$ reported by the male badge; $r = 0.79$, $p < 0.001$ reported by the female badge). For listening, the male badge captured its own listening with a correlation value of $r = 0.41$ ($p < 0.001$), while the female badge captured its own listening with $r = 0.12$ ($p < 0.001$). Turn-taking values were overestimated by a large margin (see Table 15 of Appendix 6). For example, the female badge captured the total number of turns taken as 62 and the male badge captured these as 79, even though the actual values were both 6.

Machine learning for speech detection To improve speech detection and assignment, we applied machine learning on the data captured by the badges using SAS Enterprise Miner 13.1. Machine learning requires a *training* data set to identify patterns and then uses a second *validation* data set to check whether the same patterns can be successfully detected in unseen data. Our entire data set consisted of 285 data points since we exported the badge data with a resolution of one second (and our experimental procedure lasted 285 s). We split the data using simple random partitioning such that 40

Table 4. Comparison of actual and captured values for speech experiment (Pair 1)

	Actual values		Captured values	
	Female	Male	Female	Male
No. of turns taken by this badge	6	6	54	67
No. of self-turns	3	2	84	44
No. of speaking segments	8	8	92	54
No. of successful interruptions	1	1	14	9
No. of unsuccessful interruptions	1	1	53	45

% of the data were used for training and the remaining 60 % for validation.

We used the frequency, volume, and pitch data captured by a badge as input variables. Our first output variable was female's self-speaking: whether the female badge captured female speaking (labeled as 1) or not (labeled as 0). Due to the binary nature of this classification, not speaking meant that there was silence, overlap, or male speech. The model is evaluated according to accuracy, which is the percentage of data points that are correctly classified into speaking or not speaking. The accuracy of the model is also compared to a baseline accuracy.

Baseline accuracy. We calculated the baseline accuracy using the naïve rule, which posits that each data point should be classified as a member of the majority category (Shmueli, Patel, & Bruce, 2010). This is because the majority category maximizes the accuracy by minimizing the misclassification error. Suppose we are trying to predict club membership with a data set consisting of 100 individuals. If only 40 of these individuals are members of the club, the naïve rule suggests that all individuals be classified as nonmembers since nonmembers constitute the majority category. Therefore, the baseline accuracy in this example is 60 %, because 60 % of the individuals will be correctly classified if all individuals are classified as nonmembers. Obviously, a model's accuracy should be higher than the baseline accuracy, otherwise the model has no utility. Applying this rule to our study provided a baseline accuracy of 65 %. This is because the actual speaking segments of one of the speakers without any overlap constituted 100 s of the 285-s timeline (35 %). For instance, the female speaker's speaking segments without overlaps in Fig. 2 were 100 s (or 35 %) of the entire timeline shown in the figure. Therefore, the majority category was "not speaking" (i.e., 185 s as the combination of listening, silence, and overlap), which accounted for 65 % of the timeline according to this badge. With all data points classified as "not speaking," the naïve rule produces a baseline accuracy of 65 %.

With this baseline accuracy in hand, we conducted the first analysis by setting the output variable to female speaking (1=Yes; 0=No) captured by the female badge, and the input variables to the frequency, volume, and pitch data captured by the female badge. We built neural network, decision tree, nearest neighbor, and logistic regression models. The neural network model had a multilayer perceptron architecture. It had three neurons in the hidden layer. We used the activation and combination functions set by the NEURAL procedure of SAS. The training technique was set as "default" so that SAS could try all available techniques – such as back propagation, quick propagation, quasi-Newton, etc. – and determine the best technique to use on the data. The maximum number of iterations to converge on the solution was set to 50. The decision tree model used the p-value of the F-test associated with the variance of each node as the splitting criterion. Input variables could be used more than once. The maximum branch and depth of the tree were set to two and five, respectively. The nearest neighbor model used the reduced dimensionality tree (RD-Tree) method to find nearest neighbors. The number of nearest neighbors was set to 16, which was the default value. And last, the logistic regression model used default settings with the link function set to logit. The model used only the main effects and excluded interaction and polynomial terms. The optimization technique was set as "default" so that SAS could determine the best technique depending on the number of parameters used in the model.

The decision tree model had the highest accuracy with 97 % (see Table 16 of Appendix 6 for the classification table). This was a significant improvement over the baseline (65 % accuracy) and the other algorithms (64–69 %).

Next, we examined the male speech captured by the female badge (considered listening by this badge). We used the same set of models (with the same settings) and the same three input variables (i.e., frequency, volume, and pitch data captured by the female badge). In this case, our output variable was listening (labeled as 1) or not (labeled as 0). Not listening meant that there was silence or the female was speaking. The decision tree model achieved 97 % accuracy, which was greater than the baseline accuracy (65 %) and that of the other algorithms (92–94 %). See Table 17 of Appendix 6 for the classification table.

We repeated these analyses for the data captured by the male badge. The logistic regression model identified male speaking captured by the male badge (i.e., self-speaking) with 91% accuracy (see Table 18 of Appendix 6 for the classification table). This is an improvement over the baseline accuracy (65 %) and that of the other models (87–90 %). For listening, the decision tree model was most accurate (87 %), with a baseline of 65 % (see Table 19 of Appendix 6 for the classification table). The other algorithms achieved accuracies ranging from 60–73 %.[1]

We used machine learning for overlapping speech as well. In this case, the output variable was overlapping speech (labeled as 1) or not (labeled as 0). However, this posed a challenge, because the duration of overlapping speech in the experimental manipulation was short (a total of 20 s out of 285 s). Therefore, the naïve rule suggested a baseline accuracy of 93 % which made it difficult to build better models. For this reason, we used the *oversampling* technique that is commonly used in machine learning to overcome these types of skewed data (Linoff & Berry, 2010). To this end, we built a training data set (based on data from the female badge) that had the following composition: 10 s of overlapping speech, 10 s of silence, 10 s of male speech, and 10 s of female speech (all

[1] The accuracy values reported in this study will likely differ from those in other studies because individual participants (and combinations of participants) may have different frequency, volume, and pitch values.

data were randomly selected from the original data set). This ensured that overlapping speech constituted 25 % of the data. Then we built models using the decision tree, neural network, nearest neighbor, and logistic regression algorithms. We validated the models with a validation data set consisting of the unused data (with the original distribution). The decision tree model was the most accurate (94 %) and was better than both the baseline (93 %) and the other models (ranging from 91 % to 93 %). The models built on data collected by the male badge generated even better results. The decision tree algorithm achieved 96 % accuracy, while the others ranged from 93 % to 95 %.

Last, we combined the predictions made by the algorithms to generate the conversation characteristics as turn-taking, interruptions, and so forth. To this end, we rank ordered the algorithms based on their accuracy and ran the entire data set through them separately. We used the overlapping speech model first (to label overlapping speech), then we used the female speech model (captured by the female badge) to label female speech, and finally, we used the male speech model (captured by the male badge) to label male speech. If the same data point was labeled by more than one model, we used the label generated by the most accurate model for that data point. After completing the labeling, we used the definitions provided in Table 12 of Appendix 6 to determine the conversation characteristics (turn taking, interruptions, etc.; see Table 5). The conversation characteristics produced by this approach were more realistic and closer to the objective data. Even though the values were still higher than the objective values, the relative differences between the speech characteristics of the female and male speakers were consistent.

Discussion and recommendations This study shows that the Sociometric Solutions software that derives (from raw data) who is speaking and when is not accurate: the correlations between the captured and the actual speaking, listening, and overlap are relatively low; the exception is the captured and actual silence values. This could be in part due to the synchronicity between the badges: the badges' internal clocks are rarely in sync, causing them to report the same event at different times along the timeline. Even a few seconds of difference between a pair of badges results in an incorrect assignment of events between the badges. Although events can be manually synchronized after the data are exported, Sociometric Solutions software uses the asynchronous data to derive these metrics. Therefore, the algorithm encounters the same event multiple times along the timeline: a badge may capture the beginning of a speaking segment at 12:00:00 p.m. while another badge captures it at 12:00:05 p.m. Thus, the algorithm will interpret the same speaking segment as an interruption when it compares the data captured by these two badges. This leads the algorithm to make incorrect assignments for speech, overlap, and interruption. Thus, the inaccuracy of speech assignment is a major concern for research studies that aim to analyze conversations between study participants.

Alternatively, machine learning can be used for the detection of speaking and listening. However, researchers must use caution here: the frequency, volume, and pitch signature of each voice are different. To use machine learning, researchers must create a *training* data set that includes speaking segments for each participant in order to learn the participants' unique patterns of frequency, volume, and pitch. Even though this training data set can be generated in various ways, we recommend that researchers build it based on a structured conversation between the badge wearers before actual data collection. For example, if data will be collected from two badge wearers, a minimum of 3 min of (preferably neutral) conversation can be prompted among the wearers. We recommend impromptu speech rather than the reading of predetermined text, to capture participants' natural frequency, volume, and pitch. During the conversation, each wearer should talk for 1 min without interruption. This will ensure that the wearers and ambient noise have equal representations in the data set (each constituting 33 % of the conversation). Note that researchers must also know the beginning and ending times for each speaking segment, rather than rely on the badge's internal timestamp, and must label these in the exported data set (which should have a resolution of 1 s). Only then can machine learning algorithms use the data as training data for predicting speech within actual data.

Table 5. Comparison of actual and predicted values for speech experiment (machine learning)

	Actual values		Predicted by machine learning	
	Female	Male	Female	Male
No. of turns taken by this badge	6	6	13	14
No. of self-turns	3	2	8	8
No. of speaking segments	8	8	12	11
No. of successful interruptions	1	1	3	1
No. of unsuccessful interruptions	1	1	1	2

Study 3: Accelerometer

Study 3a: Posture

The goal of this study was to examine whether the badges captured changes in posture (raw data). According to the Sociometric Solutions (2014) manual, posture is "the absolute angular velocity for every badge at every timestamp" (p. 3). Posture is measured along two separate dimensions: front–back and left–right. The badges register front–back tilting

using values between 0 and 90. Ninety indicates that a badge is in a vertically upright position, while 0 indicates that the badge is tilted forward or backward from the upright position by 90°. The badges register left–right tilting using values between -90 and 90, where -90 indicates tilting to the right by 90° and 90 indicates tilting to the left by 90°.

Experimental procedure We developed two separate experimental procedures, one for front–back tilting and the other for left–right tilting. For front–back tilting, we used our ten badges to create a 5 × 2 block: we taped five badges side by side (with all the badges facing the front), then stacked the two blocks so that all badges faced the same direction. Then we secured this block (with the badges facing the front) onto the seat back of the driver's seat in an automobile. The seat was a power seat that enabled us to tilt its back in two directions (front and back) with the push of a button. The arrangement of the badges and their position on the seat back is shown in Fig. 3a.

We set the initial position of the seat back all the way to the front. Then we tilted the seat all the way to the back and then all the way to the front again by pressing the seat's automatic button. This allowed us to control the range as well as the speed of tilting– thus ensuring repeatability. We collected data for a total of three sessions, with the badges turned off between sessions. Each session consisted of a full front–back cycle with three categories of objective movement: a stationary period in which the seat did not move, a back period in which the seat was tilted toward the back, and a front period in which the seat was tilted toward the front.

We repeated the same experimental procedure outlined for front–back testing to conduct left–right testing by changing the orientation of the badges. First, we taped the ten badges to create a 2 × 5 block: two badges were taped side by side in a 5-stack block. Then we secured this block onto the seat back with the badges facing left. Therefore, tilting the seat front and back was captured as left and right tilting by the badges (see Fig. 3b for the arrangement and position of the badges on the seat back). We collected data for three sessions.

Results *Front–back.* We conducted an ANOVA to detect the change in front–back tilting values using session, badge, and posture category (objectively determined as either stationary, front, or back) as independent variables. One badge did not collect any data and one badge stopped collecting data midway; therefore, we excluded these two badges from the analysis. The results showed that the badges captured objective posture accurately ($p < .001$; Table 20 of Appendix 7). Pairwise comparison of values between the stationary and front categories ($p < .001$) and between the stationary and back categories ($p < .001$) were significant. However, pairwise comparison of posture between the front and back categories was not significant ($p = 0.29$). This is because the range of tilting toward the front (increasing from approximately 20 to 80) is the same as the range for tilting toward the back (decreasing from approximately 80 to 20), and both conditions had an equal duration and rate of change. Therefore, the means for the front and the back categories were the same.

The results also showed that session was nonsignificant ($p = 0.95$) but badge was significant ($p = 0.02$; Table 20 of Appendix 7). In order to shed light on the significance of the badge variable in the ANOVA, we examined the front–back posture values of the eight badges across the three sessions (see Fig. 7 of Appendix 7). As can be seen in the figure, the badge front–back values changed consistently across conditions, but the level or "set" points differed by badge. For example, one badge registered a value of 84 when stationary (consistently in all three sessions), but another consistently registered a value of 74 when stationary in all three sessions. After standardizing the data by badge, the ANOVA showed

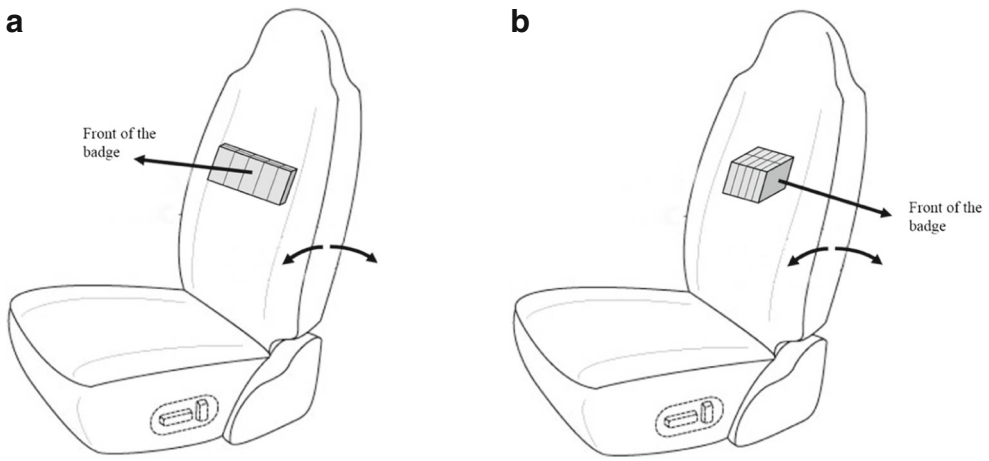


Fig. 3. **a** Study 3a. Experimental setup for front–back movement. **b** Study 3a. Experimental setup for left–right movement

that posture was significant ($p < 0.001$). However, neither badge ($p = 1.00$), nor session ($p = 0.95$) was significant.

Left–right. Similar to the previous analysis, we conducted an ANOVA on the data collected for the left–right tilting values using session, badge, and posture (as three objective categories: stationary, left, and right) as independent variables. Three badges did not collect any data and were excluded from the analysis. The results (Table 21 of Appendix 7) showed that posture was significant ($p < 0.001$), indicating that the change in left–right posture was captured accurately. Pairwise comparisons of posture values between the stationary and left categories ($p < 0.001$) as well as between the stationary and right categories ($p < 0.001$) were significant. However, pairwise comparison of posture values between the left and right categories was not significant ($p = 0.60$), because the badges captured similar data: when badges were tilted toward the left, the posture values consistently increased from approximately 20 to 88, whereas when they were tilted toward the right, they consistently decreased from approximately 88 to 20 at the same rate and duration. Thus, the means were not different from one another. The results also showed that session was nonsignificant ($p = 0.39$), suggesting that changes in posture were captured consistently across all sessions. However, badge was significant ($p < 0.001$), suggesting that the badges captured different values.

In order to identify the reason behind the significance of the badge variable, we plotted the data for each badge across the three sessions (see Fig. 8 of Appendix 7). The findings are similar to those for the front–back analysis: all badges consistently captured changes across the stationary, left, and right conditions, although the badges showed different levels of tilting. For example, one badge registered a value of 87 when it was stationary (consistently in all three sessions), but the same value was captured as 77 by another badge (also consistently in all three sessions). After the data were standardized by badge, the ANOVA showed that posture was significant ($p < 0.001$). However, neither badge ($p = 1.00$) nor session ($p = 0.39$) was significant.

Discussion and recommendations Our analyses show that the accelerometer inside each badge that captures the front–back and left–right tilting that are used to infer posture is reliable, as the measurements were replicated across multiple sessions. However, each badge registered a slightly different value for the same position. This might be due to two reasons. First, it is possible that the badges registered different values because of their relative positions with respect to the other badges in each experiment. For example, we arranged the badges in a 2 × 5 block during the left–right experiment (two badges were taped side by side in a 5-stack block), so the badge in front of the block was nearly 5 in. away from the badge in the back. The two badges could have registered different values for the same tilting movement due to this short distance between the badges. Second, it is possible that each accelerometer has a different calibration and therefore captures the same position with a different value.

Regardless of the reason, these inaccurate values pose a concern for between-badge comparability. One of the ways to address this is to standardize the values separately for front–back and left–right posture and then focus on the changes in posture rather than the magnitudes of the values.

Study 3b: Movement

According to the Sociometric Solutions (2014) manual, the raw data that the accelerometers capture for "activity" show the physical activity level of a badge wearer. Based on the activity values captured by a badge, one can determine the wearer's level of activity using three categories: low activity (or stationary), moderate activity, and high activity (such as walking). The goal of this study was to examine whether the badges could distinguish between activity levels.

Experimental procedure We created three experimental conditions: (1) no activity, (2) moderate activity, and (3) high activity. In the no activity condition, all badges were stationary. To simulate moderate activity, we taped the ten badges together to create a 2 × 5 block (two badges were taped side by side in a 5-stack block), and then we secured this block onto the seat back of a massage chair. We used the vibration of the seat back (when the massage chair was turned on) as a proxy for moderate movement. We simulated the high activity condition by having one researcher wear all the badges (in a 2 × 5 block arrangement) and walk on a treadmill with a speed setting of 2 mph. The experimental conditions were captured in a single session due to the difficulty of precisely replicating the walking condition. One of the badges stopped collecting data midway and was therefore excluded from the analyses.

Results We conducted an ANOVA in which the dependent variable was the activity values captured by the badges and the independent variables were badge and movement (as three objective categories: stationary, moderate, high). The results (Table 22 of Appendix 8) showed that movement was significant ($p < 0.001$). The pairwise comparisons between all movement categories were significant as well ($p < 0.001$). The badge variable was nonsignificant ($p = 0.45$), indicating that there were no badge-level differences.

Discussion and recommendations This study showed that badges can accurately distinguish between stationary, moderate, and high activity levels. The lack of between-badge differences shows that absolute activity values can be used to make comparisons across badges. Such comparisons can help determine the different levels of physical activity among badge wearers.

Study 3c: Mirroring

The Sociometric Solutions (2014) manual defines mirroring as a similarity between two individuals' body movement activity. This is a pairwise derived metric that the Sociometric Solutions software calculates separately each second for each pair of badges, based on all permutations. The metric is a continuous value between 0 and 1: values closer to 0 indicate that there is no mirroring, while values closer to 1 indicate that there is high mirroring. The goal of this study was to test whether the mirroring metric derived by the Sociometric Solutions software was accurate.

Experimental procedure We used the data obtained from Study 3b to test the accuracy of the mirroring metric. Recall that in Study 3b, the badges were taped together in a 2 × 5 block arrangement and exposed to three movement conditions: stationary (no movement), moderate movement (using the vibrations generated in a massage chair), and high movement (through walking with the badges on a treadmill). Because the badges were taped to one another in a solid block arrangement, we expected to observe high mirroring between the badges. As with Study 3b, we excluded one of the badges from the analysis since it stopped collecting data midway through the experiment.

Results We analyzed the descriptive statistics of the mirroring data captured for nine badges (72 pairwise comparisons). The mean mirroring value was 0.10, with a standard deviation of 0.09 (see Appendix 9). The maximum mirroring value observed between two badges was 0.84. The minimum value observed was 0. The histogram of values (Fig. 9 of Appendix 9) shows that most values were closer to 0 than to 1.

Another way to operationalize activity mirroring would be through the correlation between two badges' activity values. A high correlation would indicate high agreement, or mirroring. The correlation values, presented in Table 24 of Appendix 9, show that there was a high pairwise correlation between the activity data of the badges (the lowest correlation between badges was 0.92, $p < 0.001$). We repeated the analysis on the data collected for the two other sessions, and the results were the same: the lowest correlation in the second session was 0.93 ($p < 0.001$), and the lowest correlation in the third session was 0.92 ($p < 0.001$). These results suggest that pairwise correlation of activity data is a better indicator of mirroring than the default mirroring metric in the exported file.

Discussion and recommendations The descriptive statistics for the mirroring data show that the badges do not accurately capture mirroring, even when they are subjected to the same movement. This raises a concern about the utility of the mirroring metric in the exported file. A plausible reason for this inaccuracy is the lack of synchronicity between the badges. As discussed in Study 1, the badges' internal clocks are rarely in sync with each other. The badge-derived metrics (e.g., mirroring, activity, turn-taking) are computed from the raw data by the software program upon processing. Thus, the derived metrics are computed *before* researchers have the chance to manually sync badge timelines using a syncing signal (e.g., a stable tone). Consequently, derived metrics that use data from two badges (such as mirroring) will be inaccurate because such metrics are computed on misaligned timestamps – for example, when one badge captures a movement at time T and another captures the same movement at T+ 15 s. This lack of synchronicity might have caused the Sociometric Solutions software to generate low mirroring values even though there was mirroring.

Instead of using the mirroring values produced by the Sociometric Solutions software, we suggest that researchers use a proxy measure, namely, pairwise correlation of the activity data for each badge. Because the badges capture activity data, the correlation value between two badges' activity data may show the extent to which these two badges mirror each other.

General discussion and recommendations

As Chaffin et al. (2017) note, their work on the validation of Sociometric badges "only scratched the surface of what needs to be done to realize the full potential of these devices for all forms of behavioral research" (p. 29). The Sociometric badge opens new doors for social science. Like any other wearable sensor, it helps quantify behaviors captured in real time, providing an exciting new avenue for data collection beyond that of retrospective self-reports. However, our studies add to the concerns raised by Chaffin et al. (2017): the Sociometric badges are not tools that can be taken out of the box and immediately put to use. Overall, our studies show that the raw sensor data are generally accurate, although badges may differ in terms of sensitivity, which limits the ability to generate accurate off-the-shelf cross-badge comparisons. Moreover, the algorithms for computing higher-level (derived) values (e.g., turn-taking, mirroring) may not be accurate, especially when such algorithms rely on data from multiple badges. To resolve these issues, we outline procedures for synchronizing data and checking the assumption that the badges are working appropriately. These procedures are critical for further inferences using the badge data. In the section that follows, we present our recommendations for researchers. A summary of our studies and their implications can be found in Table 6.

Table 6. Summary of findings

Study	Sensor tested	Metrics used	Badge-level differences	Session-level differences	Findings and recommendations
Study 1a	Microphone	Frequency	NA	NA	Internal clocks are not synced with the real-world clock or one another. **Recommendation:** Play a stable tone to manually synchronize data during data analysis. Leave the badges on for three minutes following data collection.
Study 1b	Microphone and accelerometer	Frequency and activity	NA	NA	Microphone and accelerometer work in sync. An event that triggers both sensors simultaneously will be captured accurately by the badges.
Study 2a	Microphone	Volume	Yes	Yes	Absolute comparisons of volume within and between badges cannot be conducted. **Recommendation**: Standardize data for each badge and session in order to compare *changes* in volume.
Study 2b	Microphone	Frequency	Yes	Yes	Absolute comparisons of frequency within and between badges cannot be conducted. **Recommendation**: Standardize data for each badge and session in order to compare *changes* in frequency.
Study 2c	Microphone	Pitch	No	No	The badges may not distinguish the pitch of sound sources. There is a substantial amount of missing data in pitch values. **Recommendation:** Pitch values should not be used.
Study 2d	Microphone	Speaking, silence, overlapping, turn-taking	NA	NA	Sociometric Solutions software overestimates speaking, listening, and overlapping speech. **Recommendation:** Use machine learning to predict speaking, listening, overlapping speech. Then derive speech metrics (turn-taking, etc.) from the predicted values.
Study 3a	Accelerometer	Posture			
		i. front–back	No	Yes	Absolute comparisons of front–back movement between badges cannot be conducted. **Recommendation**: Standardize data for each badge in order to compare *changes* in front–back movement.
		ii. left–right	No	Yes	Absolute comparisons of left–right movement between badges cannot be conducted. **Recommendation**: Standardize data for each badge in order to compare *changes* in left–right movement.
Study 3b	Accelerometer	Activity	No	NA	The badges can distinguish different levels of activity. **Recommendation**: Data can be used as is.
Study 3c	Accelerometer	Mirroring	NA	NA	Sociometric Solutions software's mirroring metric underrepresents the degree of mirroring. **Recommendation:** Use correlations between raw activity data to assess mirroring. Higher correlations indicate greater mirroring.

Recommendations

Before data collection, researchers need to conduct multiple pretests to become familiar with the badges, and specifically the data they capture. We do not recommend the use of badges on actual participants in a lab or field setting without understanding how the data can be interpreted or used.

Synchronicity issues One of the first issues that researchers need to address is synchronicity. In our experience, none of the badges were in sync with one another or with the real-world clock. Therefore, we always played a stable tone (170 Hz for 30 s) to mark the beginning and ending times of our data collection. We also played the same stable tone to mark important points in the timeline – such as the introduction of new stimuli – during data collection. The frequency and the length of the stable tone need not be the same as those used in this study. As long as researchers can identify a steady stream of frequency values for a

specific amount of time, they can use any type of stable tone for any amount of time. However, we do not recommend less than 10 s of stable tone since this makes it difficult to identify the stream of values in the raw data.

Microphone, raw data As part of this research, we tested three types of raw data captured by the badge microphone: frequency, volume, and pitch. First, researchers should keep in mind that the values of these data are different than the actual values observed on an oscilloscope or another device. Therefore, we recommend that researchers focus on differences (or changes) in these values rather than on the absolute values. Second, of the three metrics tested, pitch is the least useful due to the amount of missing (or uncaptured) values. If this problem persists with the newer versions of the badge firmware (or newer badge models), we do not recommend using badges for this metric. Third, researchers should be cautious about the reliability of volume: even though our studies show that badges can accurately distinguish between different levels of volume, there are still within- and between-badge differences. Therefore, we recommend standardization of values when a research question calls for between-badge comparisons.

Microphone, derived metrics

In addition to the raw metrics captured by the microphone, we also tested derived metrics – such as speaking, silence, overlap, and turn-taking – that the Sociometric Solutions software generates from the raw data. In our experience, these metrics were inaccurate and, in the case of turn-taking, grossly overestimated. Therefore, we do not recommend that researchers rely on the speech-related metrics for a data set. Instead, researchers can use machine learning to identify the speaking and listening segments for each badge (based on the data captured by that badge) and then generate their own values for metrics like turn-taking, speaking, and so forth.

Movement, raw data We also tested the validity of the accelerometer raw metrics: posture and activity. Although the accelerometers accurately distinguished differences in posture and activity, the within- and between-badge differences are cause for concern about the reliability of the values. Therefore, we recommend that researchers standardize the posture and activity values within each badge when a research question requires cross-badge comparisons.

Movement, derived metrics Finally, we tested mirroring, which is a metric derived by the Sociometric Solutions software, to check whether the mirroring values were accurate. In our experience, the mirroring metric was not usable. Even though all badges made the same movements in our experimental conditions, the pairwise values indicated no mirroring between any badge pair. Researchers can instead use the correlation of movement across badges as a proxy for mirroring.

Conclusion

In a series of studies, we expanded previous efforts to examine the validity of Sociometric badge data. Researchers should exercise caution in using the metrics captured by the badges. We encourage researchers to use their own algorithms or machine learning to derive metrics from the values the badges capture, and we provide guidance on how to do so. The procedures we describe for optimizing the use of these badges may seem daunting. However, the ability to extract fine-scale speech patterns in structured settings in real time sets the stage for research that can advance our understanding of interpersonal interactions within the workplace. For example, badge data can be used to capture interpersonal dynamics in supervisor–employee and mentor–protégé dyads, which to date have been primarily captured through survey data. With automatization of the standardizing and machine-learning procedures, analyzing data from the badges will take less time than coding video recordings. Further, these badges can be used in real-world settings where video recording is not feasible. We hope that the present study provides another step toward better understanding the capabilities of these devices and that it will encourage further research.

We encourage researchers who intend to use the Sociometric badge in future studies to visit our website (http://www.badgevalidation.com) for more information about how to perform certain tests and ensure the validity of the badges. The website supports all experiments related to the microphone tests conducted in Studies 1a, 2a, 2b, and 2c. Through this website, researchers can download our prerecorded sound files, collect data using badges, and upload their data files to check the validity of their data. The website does not support Study 1d since this study relies on data that were already validated in studies through 1a to 1c. The website does not support any accelerometer tests, as we are not able to account for the movement manipulations that might be employed by other researchers.

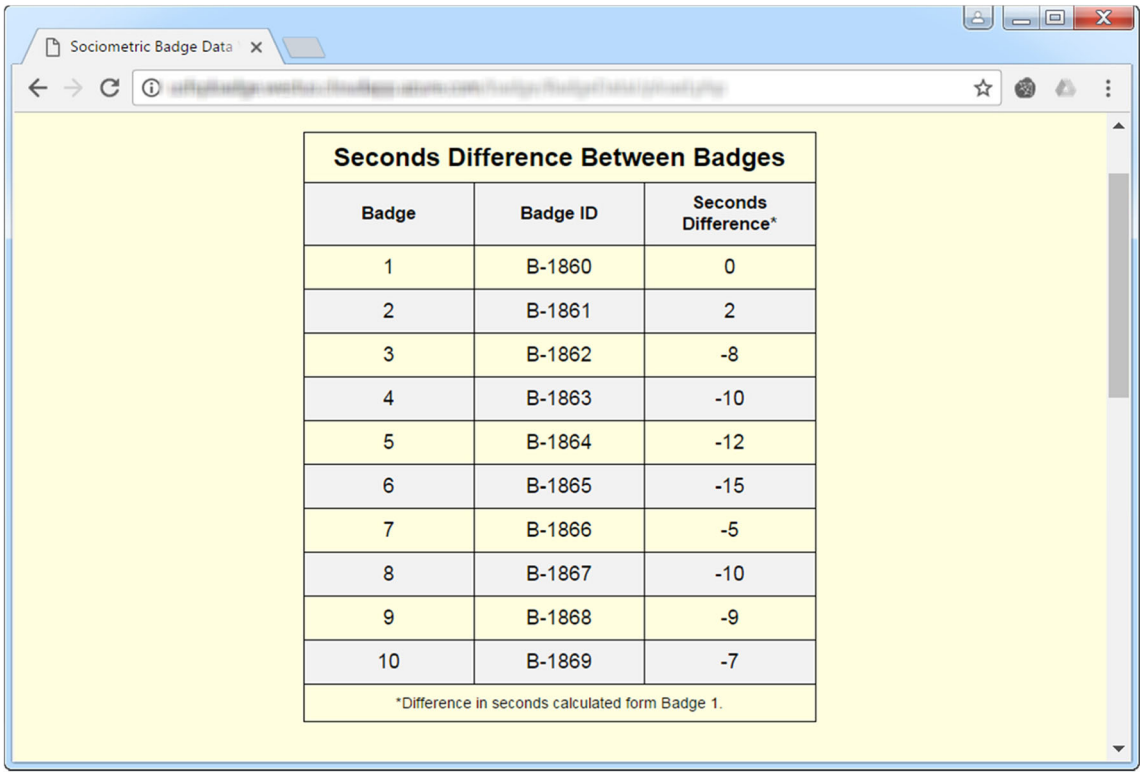
Appendix 1

Table 7. Synchronicity of badges' internal clocks

Badge	Beginning of stable tone in the real world (hh:mm:ss)	Beginning of stable tone observed in the data (hh:mm:ss)	Difference (mm:ss)
Badge1	11:07:00 a.m.	11:10:13 a.m.	03:13
Badge2	11:07:00 a.m.	11:10:17 a.m.	03:17
Badge3	11:07:00 a.m.	11:09:55 a.m.	02:55
Badge4	11:07:00 a.m.	11:09:52 a.m.	02:52
Badge5	11:07:00 a.m.	11:09:47 a.m.	02:47
Badge6	11:07:00 a.m.	11:09:42 a.m.	02:42
Badge7	11:07:00 a.m.	11:10:01 a.m.	03:01
Badge8	11:07:00 a.m.	11:09:51 a.m.	02:51
Badge9	11:07:00 a.m.	11:09:50 a.m.	02:50
Badge10	11:07:00 a.m.	11:09:57 a.m.	02:57

Table 8. Synchronicity of badges' internal clocks (repeated experiment)

Badge	Beginning of stable tone in the real world (hh:mm:ss)	Beginning of stable tone observed in the data (hh:mm:ss)	Difference (mm:ss)
Badge1	12:41:00 p.m.	12:44:35 p.m.	03:35
Badge2	12:41:00 p.m.	12:44:35 p.m.	03:35
Badge3	12:41:00 p.m.	12:44:28 p.m.	03:28
Badge4	12:41:00 p.m.	12:44:27 p.m.	03:27
Badge5	12:41:00 p.m.	12:44:26 p.m.	03:26
Badge6	12:41:00 p.m.	12:44:23 p.m.	03:23
Badge7	12:41:00 p.m.	12:44:31 p.m.	03:31
Badge8	12:41:00 p.m.	(did not collect data)	NA
Badge9	12:41:00 p.m.	12:44:26 p.m.	03:26
Badge10	12:41:00 p.m.	12:44:28 p.m.	03:28



Seconds Difference Between Badges

Badge	Badge ID	Seconds Difference*
1	B-1860	0
2	B-1861	2
3	B-1862	-8
4	B-1863	-10
5	B-1864	-12
6	B-1865	-15
7	B-1866	-5
8	B-1867	-10
9	B-1868	-9
10	B-1869	-7

*Difference in seconds calculated form Badge 1.

Fig. 4. Example synchronicity analysis available at http://www.badgevalidation.com.

Appendix 2

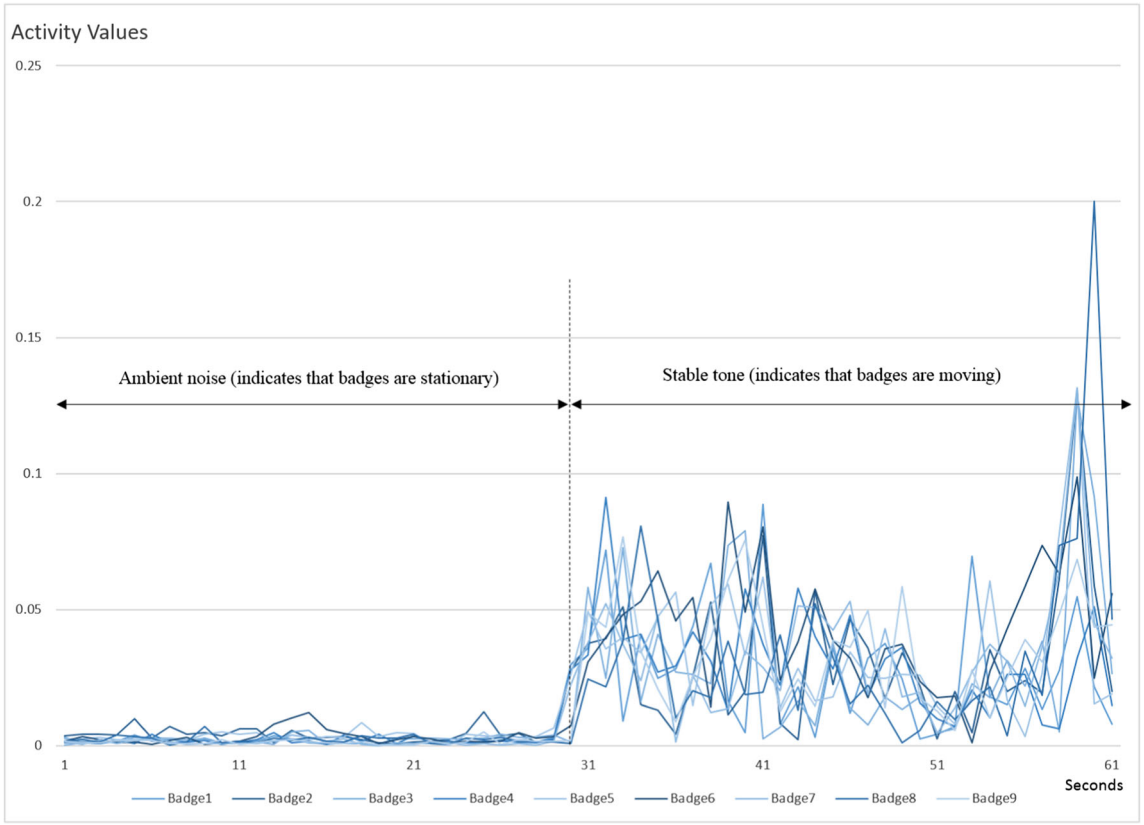


Fig. 5. Study 1b. Activity data captured by each badge during ambient noise and stable tone.

Appendix 3

Table 9. ANOVA test (dependent variable: volume)

Variable	*df*	*F*	Significance	Partial Eta Squared
Session	2	4.66	.010	.005
Badge	9	47.59	<.001	.200
Condition	2	2,046.79	<.001	.705
Session × Badge	18	1.19	.265	.012
Session × Condition	4	0.87	.484	.002
Badge × Condition	18	2.16	.003	.022
Session × Badge × Condition	36	0.29	1.000	.006
Error	1,710			
Total	1,800			

Note. Adjusted *R* squared = 0.715.

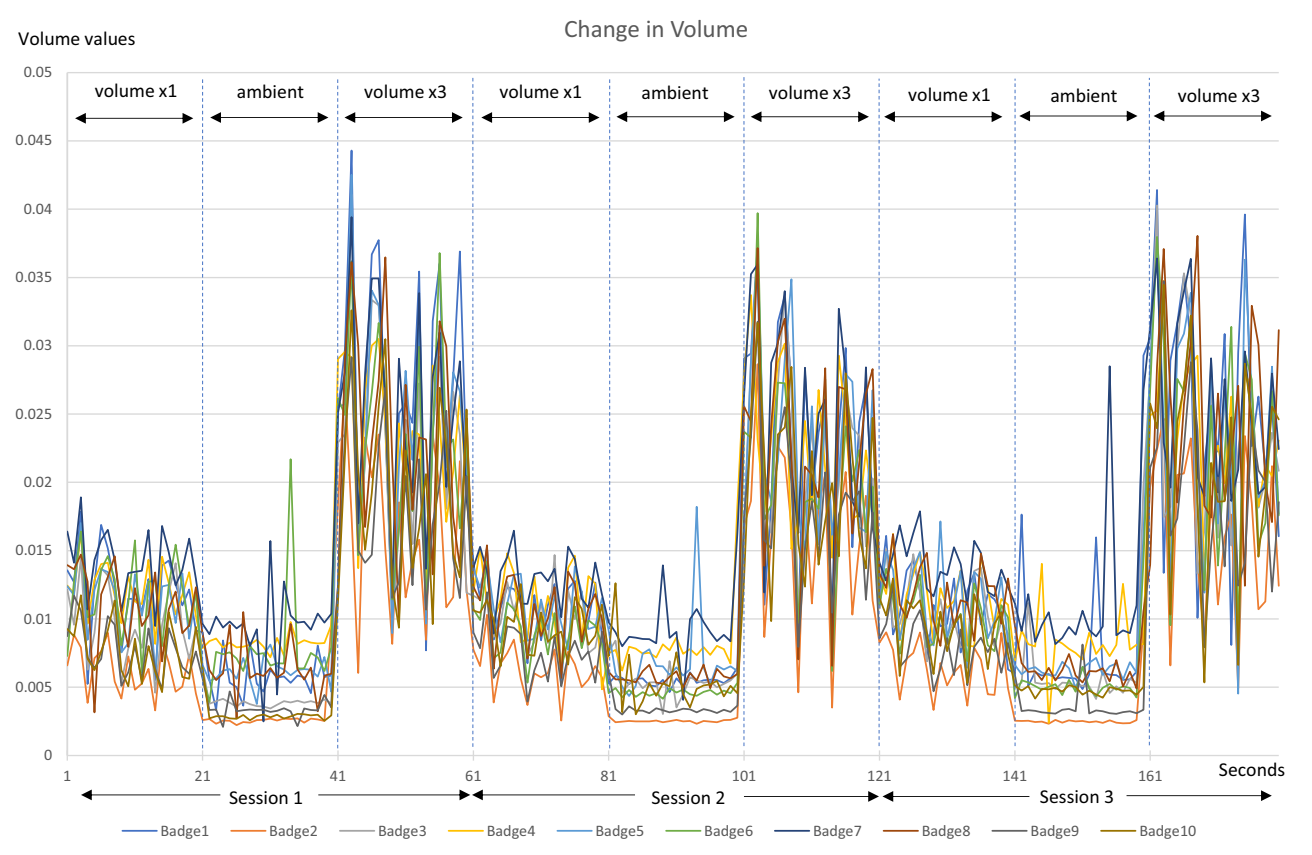


Fig. 6. Study 2a.Volume data captured by each badge in each session.

Appendix 4

Table 10. ANOVA test (dependent variable: frequency)

Variable	*df*	*F*	Significance	Partial Eta Squared
Session	2	4.71	.009	.002
Badge	9	3.52	<.001	.005
Sound	4	4,019.53	<.001	.722
Session × Badge	18	0.46	.974	.001
Session × Sound	8	7.96	<.001	.010
Badge × Sound	36	2.30	<.001	.013
Session × Badge × Sound	72	0.48	1.000	.006
Error	6,178			
Total	6,328			

Note. Adjusted *R* squared = 0.718.

Appendix 5

Table 11. ANOVA test (dependent variable: pitch)

Variable	*df*	*F*	Significance	Partial Eta Squared
Session	2	1.89	.152	0.002
Badge	9	0.93	.498	0.005
Sound	4	79.51	<.001	0.155
Session × Badge	18	0.91	.572	0.009
Session × Sound	8	2.05	.038	0.009
Badge × Sound	36	1.12	.293	0.023
Session × Badge × Sound	64	1.12	.239	0.040
Error	1,732			
Total	1,874			

Note. Adjusted *R* squared = 0.170.

Appendix 6

Table 12. Definitions provided by sociometric solutions (2014) manual (p.38)

Turn taking concepts	Definitions
Speaking Segment	Any continuous, uninterrupted length of speech made by a single person.
Turns	Turns are speaking segments that occur after and within 10 seconds of another speaking segment. By default a speech segment must be made within 10 seconds after the previous one ended in order to be considered a turn. Note that the two speech segments need not be from two different people to count as a turn—a person can pause and then start speaking again. This would count as two speech segments, and one "self-turn."
Self-Turns	A speaker starts speaking, pauses for greater than 0.5 seconds (but less than 10 seconds), and then resumes speaking.
Successful Interruptions	Person A is talking. Person B starts talking over A. If Person A talks for less than 5 out of the next 10 seconds, then Person B successfully interrupted Person A.
Unsuccessful Interruptions	Person A is talking. Person B starts talking over A. If Person A talks for more than 5 out of the next 10 seconds, then Person B [un]successfully interrupted Person A.
Pause	A pause is a period of time within which there is no speaking. All pauses are between .5s and 10s. (Anything less than .5s gets marked as continuous speech, anything greater than 10s gets marked as the end of the turn-taking exchange.)

Table 13. Correlation matrix for speaking experiment (pair 1)

			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	Female voice	Speaking	1															
2		Actual speaking	.48**	1														
3		Overlap	-.40**	.18**	1													
4		Actual overlap	-.10	-.20**	.25**	1												
5		Listening	-.23**	-.33**	-.40**	-.09	1											
6		Actual listening	-.29**	-.54**	.19**	-.20**	.41**	1										
7		Silence	-.18**	-.34**	-.53**	-.13*	-.20**	-.34**	1									
8		Actual silence	-.16**	-.40**	-.56**	-.15*	-.03	-.40**	.86**	1								
9	Male voice	Speaking									1							
10		Actual speaking									.41**	1						
11		Overlap									-.40**	.19**	1					
12		Actual overlap									-.09	-.20**	.25**	1				
13		Listening									-.23**	-.29**	-.40**	-0.10	1			
14		Actual listening									-.33**	-.54**	.18**	-.20**	.48**	1		
15		Silence									-.20**	-.34**	-.53**	-.13*	-.18**	-.34**	1	
16		Actual silence									-.03	-.40**	-.56**	-.15*	-.16**	-.40**	.86**	1

* $p < 0.05$. ** $p < 0.01$.

Table 14. Correlation matrix for speaking experiment (Pair 2)

			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	Female voice	Speaking	1															
2		Actual speaking	.52**	1														
3		Overlap	-.40**	.12**	1													
4		Actual overlap	-.14*	-.20**	.23**	1												
5		Listening	-.20**	-.32**	-.38**	-.05	1											
6		Actual listening	-.22**	-.54**	.31**	-.20**	.12**	1										
7		Silence	-.16**	-.33**	-.63**	-.13*	-.11	-.29**	1									
8		Actual silence	-.25**	-.40**	-.62**	-.15*	.27**	-.40**	.79**	1								
9	Male voice	Speaking									1							
10		Actual speaking									.22**	1						
11		Overlap									-.39**	.35**	1					
12		Actual overlap									-.02	-.20**	.21**	1				
13		Listening									-.20**	-.30**	-.41**	-.14*	1			
14		Actual listening									-.23**	-.54**	.15**	-.20**	.41**	1		
15		Silence									-.11	-.36**	-.64**	-.14*	-.15*	-.36**	1	
16		Actual silence									.02	-.40**	-.71**	-.15*	-.04	-.40**	.90**	1

* $p < 0.05$. ** $p < 0.01$.

Table 15. Comparison of actual and captured values for speech experiment (Pair 2)

	Actual values		Captured values	
	Female	Male	Female	Male
No. of turns taken by the badge	6	6	62	79
No. of self-turns	3	2	77	51
No. of speaking segments	8	8	92	63
No. of successful interruptions	1	1	8	19
No. of unsuccessful interruptions	1	1	71	43

Table 16. Classification table of the decision tree model

Female Badge	Predicted 1 (Female - speaking)	Predicted 0 (Female - not speaking)
Actual 1 (Female - speaking)	61	-
Actual 0 (Female - not speaking)	5	107

Table 17. Classification table of the decision tree model

Female Badge	Predicted 1 (Female - listening)	Predicted 0 (Female - not listening)
Actual 1 (Female - listening)	56	5
Actual 0 (Female - not listening)	1	111

Table 18. Classification table of the logistic regression model

Male Badge	Predicted 1 (Male - speaking)	Predicted 0 (Male - not speaking)
Actual 1 (Male - speaking)	54	7
Actual 0 (Male - not speaking)	9	103

Table 19. Classification table of the decision tree model

Male Badge	Predicted 1 (Male - listening)	Predicted 0 (Male - not listening)
Actual 1 (Male - listening)	53	8
Actual 0 (Male - not listening)	15	97

Appendix 7

Table 20. Results of ANOVA (dependent variable: front–back values)

Variable	*df*	*F*	Significance	Partial Eta Squared
Session	2	0.05	.95	.000
Badge	7	2.31	.02	.005
Posture	2	53.76	<.001	.035
Session × Badge	14	0.01	1.00	.000
Session × Posture	4	0.16	.96	.000
Badge × Posture	14	0.02	1.00	.000
Session × Badge × Posture	28	0.03	1.00	.000
Error	3,000			
Total	3,072			

Note. Adjusted R squared = 0.021.

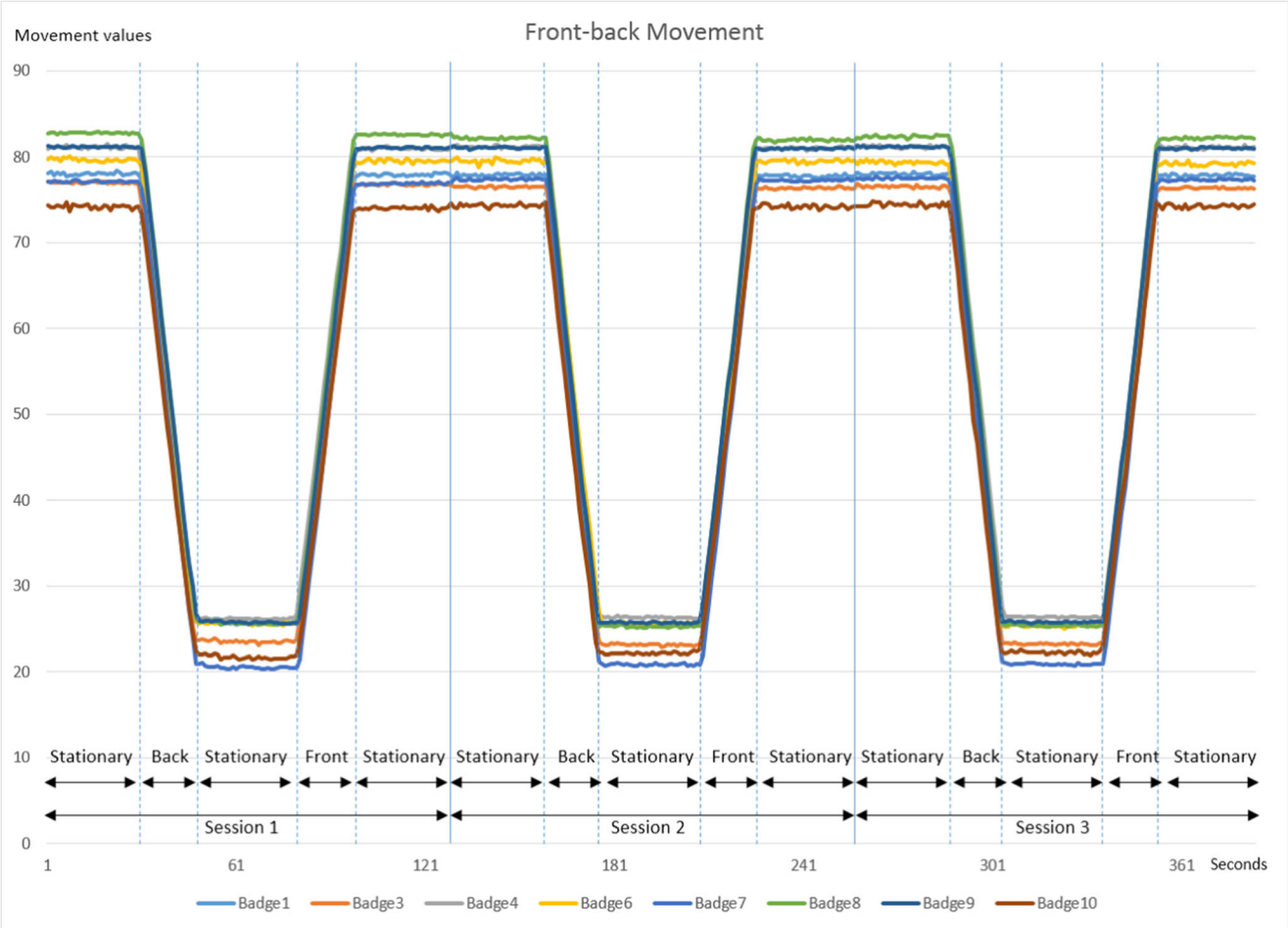


Fig. 7. Study 3a. Stationary, front, back conditions in Posture experiment.

Table 21. Results of ANOVA (dependent variable: left–right values)

Source	*df*	*F*	Significance	Partial Eta Squared
Session	2	0.95	.39	.001
Badge	6	5.26	<.001	.012
Posture	2	47.95	<.001	.036
Session × Badge	12	0.05	1.00	.000
Session × Posture	4	0.23	.92	.000
Badge × Posture	12	0.01	1.00	.000
Session × Badge × Posture	24	0.04	1.00	.000
Error	2,597			
Total	2,660			

Note. Adjusted *R* squared = 0.032.

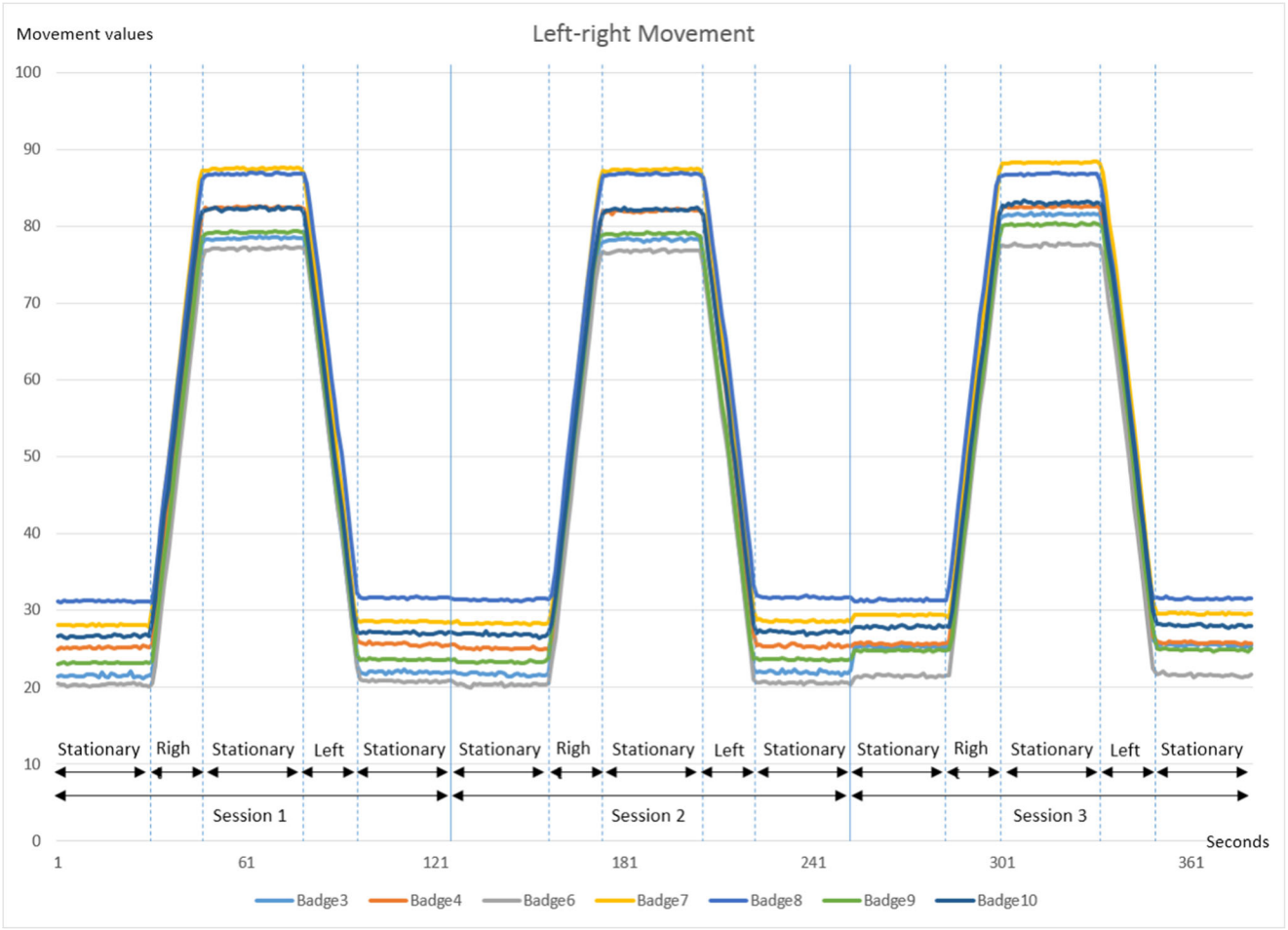


Fig. 8. Study 3a. Stationary, left, right conditions in Posture experiment.

Appendix 8

Table 22. Results of ANOVA (dependent variable: activity values)

Source	*df*	*F*	Significance	Partial Eta Squared
Badge	8	0.98	.45	.005
Movement	2	9,709.73	<.001	.923
Badge × Movement	16	2.07	.008	.020
Error	1,624			
Total	1,651			

Note. Adjusted *R* squared = 0.922.

Appendix 9

Table 23. Descriptive statistics for activity mirroring values

	Session 1
Minimum	0.00
Maximum	0.84
Mean	0.10
Standard Deviation	0.09

Table 24. Correlation of activity data between badges

	Badge1	Badge2	Badge4	Badge5	Badge6	Badge7	Badge8	Badge9	Badge10
Badge1	1								
Badge2	.97	1							
Badge4	.96	.95	1						
Badge5	.95	.94	.95	1					
Badge6	.96	.95	.94	.96	1				
Badge7	.95	.94	.92	.94	.92	1			
Badge8	.95	.95	.94	.95	.95	.93	1		
Badge9	.97	.96	.96	.95	.95	.93	.95	1	
Badge10	.96	.97	.94	.95	.94	.94	.97	.95	1

Note. All values are significant at $p < .001$.

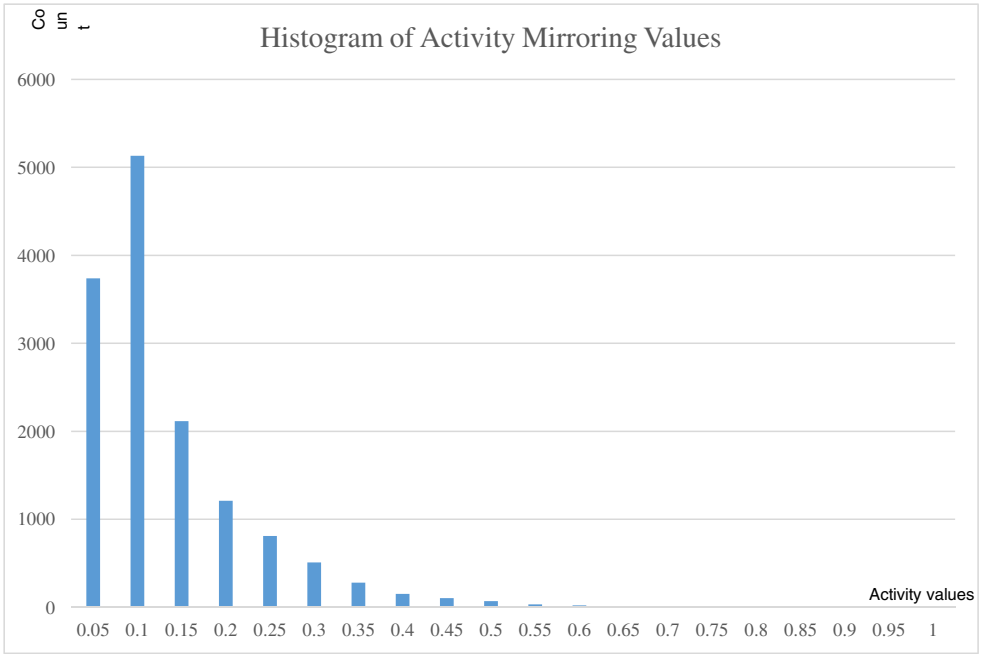


Fig. 9. Study 3c. Histogram of the activity mirroring values.

References

Chaffin, D., Heidl, R., Hollenbeck, J. R., Howe, M., Yu, A., Voorhees, C., & Calantone, R. (2017). The promise and perils of wearable sensors in organizational research. *Organizational Research Methods, 20*, 3-31. doi:https://doi.org/10.1177/1094428115617004

Curhan, J. R., & Pentland, A. (2007). Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology, 92*(3), 802-811. doi: https://doi.org/10.1037/0021-9010.92.3.802

Dockweiler, S. (2014). How much time do we spend in meetings? (Hint: It's Scary). Retrieved from https://www.themuse.com/advice/how-much-time-do-we-spend-in-meetings-hint-its-scary

Hall, J. A., Coats, E. J., & LeBeau, L. S. (2005). Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological Bulletin, 131*(6), 898-924. doi:https://doi.org/10.1037/0033-2909.131.6.898

Kim, T., Chang, A., Holland, L., & Pentland, A. S. (2008). Meeting mediator: Enhancing group collaborationusing sociometric feedback. In *Proceedings of the 2008 ACM conference on Computer supported co-operative work*, San Diego, CA, pp.457-466.

Linoff, G. S., & Berry, M. J. (2010). *Data mining techniques: For marketing, sales, and customer relationship management* (3rd ed.). Indianapolis, IN: John Wiley & Sons.

Olguin-Olguin, D., Gloor, P. A., & Pentland, A. S. (2009). Capturing individual and group behavior with wearable sensors. In

Proceedings of the 2009 AAAI Spring Symposium on Human Behavior Modeling, Palo Alto, CA, pp.68–74.

Olguin-Olguin, D., & Pentland, A. (2010a). *Assessing group performance from collective behavior*. Paper presented at the ACM Conference on Computer Supported Cooperative Work, Workshop on Collective Intelligence In Organizations, Savannah, GA.

Olguin-Olguin, D., & Pentland, A. (2010b). Sensor-based organisational design and engineering. *International Journal of Organisational Design and Engineering, 1*(1-2), 69-97. doi:https://doi.org/10.1504/IJODE.2010.035187

Olguin-Olguin, D., Waber, B. N., Kim, T., Mohan, A., Ara, K., & Pentland, A. (2009). Sensible organizations: Technology and methodology for automatically measuring organizational behavior. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39*(1), 43-55. doi:https://doi.org/10.1109/TSMCB.2008.2006638

Orbach, M., Demko, M., Doyle, J., Waber, B. N., & Pentland, A. (2015). Sensing informal networks in organizations. *American Behavioral Scientist, 59*(4), 508-524. doi:https://doi.org/10.1177/0002764214556810

Shmueli, G., Patel, N., & Bruce, P. (2010). *Data mining for business intelligence: Concepts, techniques, and applications in Microsoft Office Excel with XLMiner* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Sociometric Solutions. (2014). Sociometric badge 03-02. Preliminary user guide. Revision 1.21.

Sociometric Solutions. (2015). Data collection and export recommendations for researchers.

Titze, I. R. (1994). *Principles of voice production* (1st ed.). Englewood Cliffs, N.J: Prentice Hall.

Tripathi, P., & Burleson, W. (2012). Predicting creativity in the wild: Experience sample and sociometric modeling of teams. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, Seattle, WA, pp.1203-1212.

Waber, B. N., Olguin-Olguin, D., Kim, T., & Pentland, A. S. (2008). *Understanding Organizational Behavior with Wearable Sensing Technology*. Available at SSRN: https://ssrn.com/abstract=1263992

Wu, L., Waber, B. N., Aral, S., Brynjolfsson, E., & Pentland, A. S. (2008). Mining face-to-face interaction networks using sociometric badges: Predicting productivity in an IT configuration task. In *Proceedings of the 29th International Conference on Information Systems*, Paris, France, pp.1–19.