Article

Organizational Research Methods
1-29

DOI: 10.1177/1094428115617004
orm.sagepub.com

The Promise and Perils of Wearable Sensors in Organizational Research

Daniel Chaffin[1], Ralph Heidl[2], John R. Hollenbeck[3], Michael Howe[4], Andrew Yu[3], Clay Voorhees[3], and Roger Calantone[3]

Abstract

Rapid advances in mobile computing technology have the potential to revolutionize organizational research by facilitating new methods of data collection. The emergence of wearable electronic sensors in particular harbors the promise of making the large-scale collection of high-resolution data related to human interactions and social behavior economically viable. Popular press and practitioner-oriented research outlets have begun to tout the game-changing potential of wearable sensors for both researchers and practitioners. We systematically examine the utility of current wearable sensor technology for capturing behavioral constructs at the individual and team levels. In the process, we provide a model for performing validation work in this new domain of measurement. Our findings highlight the need for organizational researchers to take an active role in the development of wearable sensor systems to ensure that the measures derived from these devices and sensors allow us to leverage and extend the extant knowledge base. We also offer a caution regarding the potential sources of error arising from wearable sensors in behavioral research.

Keywords

wearable sensors, group research methodology, big data

The advent of "big data" collection and computing is revolutionizing the world, and thus it should not be surprising that this development would eventually touch the lives of organizational researchers (George, Haas, & Pentland, 2014; Kozlowski, Chao, Chang, & Fernandez, in press). Advances in mobile computing and sensor technology in particular have created the opportunity to transcend the limits of traditional data collection instruments. These new technological developments have

[1]University of Nebraska Kearney, Kearney, NE, USA
[2]University of Oregon, Eugene, OR, USA
[3]Michigan State University, East Lansing, MI, USA
[4]Iowa State University, Ames, IA, USA

Corresponding Author:
Daniel Chaffin, University of Nebraska Kearney, 1917 W 24th Street, West Center 405C, Kearney, NE 68849, USA.
Email: chaffintd@unk.edu

the potential to significantly advance research on individuals, teams, and multi-team systems. However, the scholarly community has yet to address the questions related to the integration of these new measurement methods into the extant knowledge base.

Group research has generally focused on constructs related to individual behavior within groups (e.g., boundary spanning), group process (e.g., leadership emergence), and group structure (e.g., interaction patterns) (Greenberg & Baron, 1995; Marks, Zaccaro, & Mathieu, 2000; Mathieu, Heffner, Goodwin, Salas, & Cannon-Bowers, 2000; Sparrowe, Liden, Wayne, & Kraimer, 2001). Traditionally, constructs in this domain have been gathered using retrospective self-reports obtained from team members. The problems associated with retrospective self reports such as to measure these attributes have been well documented, such as social desirability bias, halo, and leniency effects (Donaldson & Grant-Vallone, 2002; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003; Spector & Brannick, 2010). Some researchers have attempted to overcome these limitations through video coding of interpersonal interactions. However, this approach tends to be resource intensive and is usually restricted to short-term laboratory contexts. Recent technological advances in mobile computing create the possibility of collecting high-resolution data related to social interactions in unrestricted space over extended time periods.

Wearable sensors (WSs) are mobile devices containing electronic components that record the environmental context of the device-bearing person. For example, mobile devices fitted with microphones and Bluetooth modules can generate data streams describing ambient sound and proximity to other devices. These low-order data streams have the potential to then provide the foundation for higher-order measures of individual behavior and social interactions. The benefits of WS technology have been a prominent topic in the practitioner and popular press outlets (e.g., Silverman, 2013). Scholarly interest in this technology has also been significant, as evidenced by the Organizational Behavior Division of the Academy of Management, which bestowed the 2013 Outstanding Practitioner-Oriented Publication Award to the author of an article based on WS-generated data (Pentland, 2012).

Relative to the substantial scholarly interest in leveraging WS technology, the related body of research is somewhat limited, and there are many questions as to how to best employ WS-derived data for measuring established behavioral and social constructs (Kim, McFee, Olguin, Waber, & Pentland, 2012; Olguin & Pentland, 2010; Pentland, 2012). We begin the process of trying to integrate this new measurement capacity into the extant knowledge base by conducting four distinct studies. These studies are aimed at (a) establishing construct validation protocols for WSs for different types of research studies and (b) providing evidence from the application of such protocols in WS deployment conditions that range from short-term laboratory experiments, with strict control over environmental conditions, to long-term field studies, with little control over time or space. That is, the sequence of studies considers WS-generated data streams ranging in duration from minutes to weeks and systematically evaluates the utility of this novel data gathering method for the measurement of interaction patterns in contexts where, in most cases, we have known true scores or the best available alternative.

Our results provide initial evidence of both the promise and the perils of using WSs in behavioral research. We also present research protocols that show how organizational researchers can take an active role in the continued advancement of wearable sensor systems. The involvement of active researchers in this area is sorely needed to ensure that the development of measures integrates with the extant knowledge base. Like any measurement method, there are limits to what can be accomplished with WSs. However, when appropriately deployed, configured, and analyzed, WSs can capture variables such as boundary spanning, leadership emergence, and group structure without the need for problematic retrospective self-reports, reports from others, or direct experimenter observation.

Wearable Sensors

As more WSs are offered on the market, researchers are faced with choices related to both the sensor composition and configuration of these platforms. The WSs used in all of our studies were primarily developed to measure interaction patterns among individuals and have been used in several previous studies (e.g., Kim et al., 2012; Olguin et al., 2009; Olguin & Pentland, 2010), including the aforementioned paper (Pentland, 2012) that was recognized by the Academy of Management in 2013. This WS is produced by Sociometric Solutions and is a white "badge" about the same size as a deck of playing cards worn around the neck of participants on a lanyard. Specifically, this WS uses a Bluetooth sensor to measure physical proximity, an infrared detector to measure face-to-face positioning, an accelerometer for measuring body movement and posture, and microphones to measure verbal activity (Olguin et al., 2009). After undergoing a series of computations, the raw data from these sensors are used to create measures of lower level behavioral dimensions, such as body movement, co-location, and verbal activity. These basic measures can then be used to create more abstract constructs, such as network centrality, social dominance, cohesion, and so on.

We note that this is not the only WS in existence, and while the decisions related to sensor composition and configuration are critical, the selection of the appropriate WS should be informed by the focal research question. A systematic comparison of different WS options currently available is outside the scope of this paper. It is important to highlight that WS technology is rapidly developing and platforms for delivering sensors are changing quickly. Still, regardless of whatever platform one uses, many of the component sensors that are used in these devices (Bluetooth, microphones, infrared, and accelerometers) are commodities that remain relatively constant across platforms. As such, the primary focus of our research is on the individual component sensors rather than on a specific platform. That is, even though Google Glass has been discontinued, all of the components that went into that specific platform live on. Other new platforms will eventually be developed and replace or complement existing ones. Just as alternative devices for delivering music have come and gone over the years (phonographs, eight-track players, cassettes, mp3, etc.), there will always be music. Thus, measurement work in this area of the organizational sciences should be directed at how to employ the core *component sensors* as opposed to any one specific device.

Prior to getting into a detailed description of each of the four studies in this article, we should note that a comprehensive construct validation effort of WSs is complicated for several reasons. First, there are different sources of error, and the sources of these errors are not all equal when it comes to construct validity concerns. For example, even though the WSs we study all come from the same manufacturer, there are still variations in the sensitivity of each WS. This creates both "within-" and "between-" WS error variance even when the WSs are exposed to the exact same environmental stimuli.

Within-WS variability can be attributed simply to the unreliability of any one of the component sensors (e.g., microphone or Bluetooth). Whereas this particular source of error variance may be regrettable in a perfect-world sense, it does not do serious damage to WS-based measures. A WS assesses the surrounding environment several times each minute and generates an extremely large number of assessments when worn for any length of time. This is analogous to a test with thousands of items, and because random errors tend to cancel out over time, even small correlations between each item and the test as a whole would generate a highly reliable assessment (Nunnally & Bernstein, 1994).

However, there is also some degree of between-WS variability because component WSs may have different mean levels of detection when exposed to the same environment. For example, the microphone in one WS might simply be more sensitive than the microphone in a different WS—a phenomenon that would be familiar to anyone who ever worked in a contemporary sound studio—but perhaps not an organizational researcher who has not worked closely with microphones. Unlike within-WS variance, which is a relatively benign problem for validity, between-WS

Table 1. Summary of Studies.

	Channel(s) Tested	Study Characteristics	Treatment	Treatment Source	Measures	Study Setting
Study 1	Bluetooth and infrared	3 Sessions 24 WSs 19 conditions	Distance and barriers	Fixed board	Raw	Lab
	Microphone	3 sessions 24 WSs 3 conditions	Tone/background noise	Speaker	Raw	Lab
Study 2	Microphone	18 Sessions 24 WSs	Scripted conversations	Human participants	Processed	Lab
Study 3	Bluetooth and microphone	24 Sessions 20 WSs	Boundary spanners, emerging team structure	Experimental condition Human participants	Raw and processed	Lab
Study 4	Bluetooth and infrared	20 Sessions 13 WSs	Field setting with perceived interactions	Human participants	Raw	Field
	Microphone	10 Sessions 13 WSs	Unscripted conversations	Human participants	Processed	Field

Note: WSs = wearable sensors.

variability causes bias that accumulates over time, if not detected, and can threaten validity. This is a serious problem because it means that some WSs will always over-detect relative to other WSs, and the effect of this error compounds over time (rather than cancelling out), leading to systematic errors in measurement. For example, an individual wearing a WS with a more sensitive Bluetooth sensor will always report a more central position than is warranted in the proximity network relative to other individuals.

Second, even if one ignores variability within and between WSs, the fact that measurements take place at a raw level (i.e., the sensors themselves), the basic level (e.g., co-location and verbal activity), and at a higher level (e.g., dominance, mirroring, etc.) further complicates matters. The relationships between the raw data and higher level measures are not straightforward due to the complexity (and proprietary nature) of the post-processing algorithms. In particular, the proprietary nature of the post-processing algorithms makes it difficult to explain why different results at higher levels emerge from the exact same data taken at raw levels.

Third, as we noted previously, WS platforms and individual components present a moving target. Thus, any construct validation effort has to be appreciated as a single snapshot in time. Still, one snapshot in time is useful for serving as a benchmark for the future, where additional snapshots can be strung together to create the evolving trajectory of this technology. The information provided at one point in time can also focus future development efforts. This is especially the case with the initial development of WSs because this development has been dominated by engineers, who lack expertise in psychometrics and may not have intimate familiarity with the nature of history of central constructs in the organizational sciences.

With these caveats in mind, Table 1 provides an overview of the four studies conducted as part of this effort. In general, the studies move from short-term, highly controlled, and small space contexts to long-term, totally uncontrolled, and unrestricted space contexts. Taken as a whole, the four studies presented here provide initial evidence regarding the construct validity of raw, basic, and higher level measures derived from WSs across a range of contexts where they might be deployed in programs of research involving individuals and groups.

Study 1

The purpose of Study 1 was to evaluate the raw data recorded via Bluetooth, infrared, and microphone sensors as well as the ability to use these sensors to derive measures of co-location and verbal activity. In a field study context, the WSs may be prone to error due to jostling of the WS, between-subject differences, environmental noise, and other sources of contamination. As a baseline, the focus of Study 1 was on the ability to detect true score variance in a context where these aspects could be well controlled (i.e., a lab setting).

Evidence Regarding Co-Location

Co-Location and Bluetooth. We performed a range of tests to evaluate the ability of WSs to detect co-location via Bluetooth and infrared. First, we discuss the method and results for Bluetooth. In this test, we placed 12 WSs each on two corkboards and placed these boards on easels at varying distances (i.e., 13 distances ranging from 1 meter to 40 meters) and with varying barriers between them. Because the detection interval of Bluetooth sensors is approximately 30 seconds, the boards were left in place for 3 minutes at each experimental condition to allow for repeated detections. Each condition was repeated three times, and the WSs were powered down and then powered back up in between each session to evaluate any variability that may result from this process. This approach allowed us to simultaneously compare multiple WSs in multiple sessions and partition both within- and between-WS variance. This design also allowed us to compare the relative magnitude of within- and between-WS variance to the amount of variance arising from varying the experimental conditions, that is, the environmental stimulus.

The Bluetooth sensor generates a categorical measure of detection (on/off) comprised of a time/date stamp, sender WS number, receiver WS number, and a radio signal strength indicator (RSSI) that varies based on the intensity of signal between WSs ranging from –65 to –95 with the larger value (–65) representing a higher strength when compared to the lower value (–95). We only considered detections across the corkboards at distances varying between 1 and 40 meters. We counted the number of detections per WS and divided by 36 (3 minutes and 12 potential detected WSs) to arrive at an average detection count per minute for each WS. The box and whisker plot in Figure 1 presents these values. The number of observations for this plot is 936, which is the result of 24 WSs $\times$ 3 sessions $\times$ 13 experimental conditions (i.e., distances). Bluetooth technology is expected to detect to a distance of 10 meters, so in a perfect world, each WS would provide 2 detections per minute at each of the distances up to 10 meters, with no detections at distances greater than 10 meters (Hallberg, Nilsson, & Synnes, 2003). This 2.0 detection count may be unattainable in practice because each Bluetooth module independently cycles every 30 seconds; therefore, it is unlikely to reach this technological maximum. Still, the WS should detect at least 1.0 per minute per co-located WS, or else one would draw the false inference that the two WSs were not co-located when in fact they were.

It should also be noted that the WS platform allows for researchers to adjust the sensitivity of the Bluetooth sensor, and how one sets this parameter will affect results. Thus, this setting becomes an important further consideration in the decision-making process for researchers employing WSs. For the first run of this test, we used the maximum sensitivity and relied on Bluetooth RSSI as a threshold in follow-up analyses, as previously recommended (Olguin et al., 2009). In subsequent studies reported in this article that focus on broader behavioral and social constructs, such as boundary spanning, leadership emergence, and social structure, we will see that decisions regarding how to set parameters and cutoffs will have a major impact on inferences. Thus, our results here foreshadow those results.

As we noted earlier, the performance for manufactured component sensors can vary from the technical specifications, so it should not come as a surprise that there is variability both within and

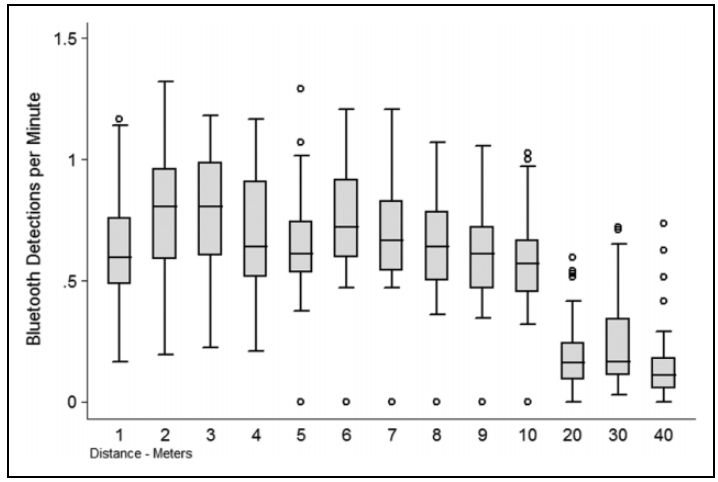

Figure 1. $N = 936$ All Bluetooth detection count per minute by distance. *Note.* The plots are a typical box and whisker such that the mean is indicated by the horizontal line, the 25th and 75th percentiles are indicated by the lower and upper shaded box, the whiskers and dots represent more extreme values in detection count.

between Bluetooth sensors. However, taken as a whole, the results provided mixed support for the potential for the Bluetooth sensor to serve as a pure measure of co-location. First, the count of detections at 20 meters is far less than that at 10 meters, which is somewhat comforting. However, the fact that WSs detect at distances well beyond 10 meters is concerning because detections at long distances increase the risk of *over*-detection. Forty meters is far beyond the distance one would expect for meaningful interactions between two people, and yet, in some instances, the WSs would still report co-location of individuals separated by this distance.

In addition, the count of detections observed for any one distance varies a great deal. The lack of uniform detection counts at the same distance means that there is a potential that some WSs may over- or under-detect when compared to the true score. As we note earlier, the degree to which this error variance is within versus between WSs is important in terms of whether or not the error would either cancel out or compound over time. Our subsequent analyses examine this issue more specifically. As an aside, while also investigated, we noted very little variance attributable to WS by distance interactions.

Although informative, Figure 1 is limited because it does not consider RSSI. According to previous calibration efforts, an RSSI of –80 represents an appropriate threshold for face-to-face interactions (Olguin et al., 2009). Prior research has shown that a distance of 1 to 4 meters is an appropriate estimate for personal and social space (Hall, 1990). Thus, in an ideal case, filtering detections based on RSSI would result in a high detection count at 1 to 4 meters with a sharp drop-off in detection count at distances greater than 4 meters.

To test this threshold, we reduced the data set to include only detections of –80 RSSI or greater. Again, for RSSI signal strength, the larger value such as –70 is a stronger signal than an RSSI of –90. Figure 2 illustrates the results. In Figure 2, the detection count does appear to decline at a sharper rate as distances increase when compared to Figure 1. In addition, the detection count is significantly reduced at longer distances. However, there remains variance in the number of detections as well as detections at distances beyond 4 meters. Collectively, these results suggest that *how* you set the threshold can affect the performance of Bluetooth as a measure of co-location, at least in this specific context. Nevertheless, there remains substantial error variance in detection count at each specified

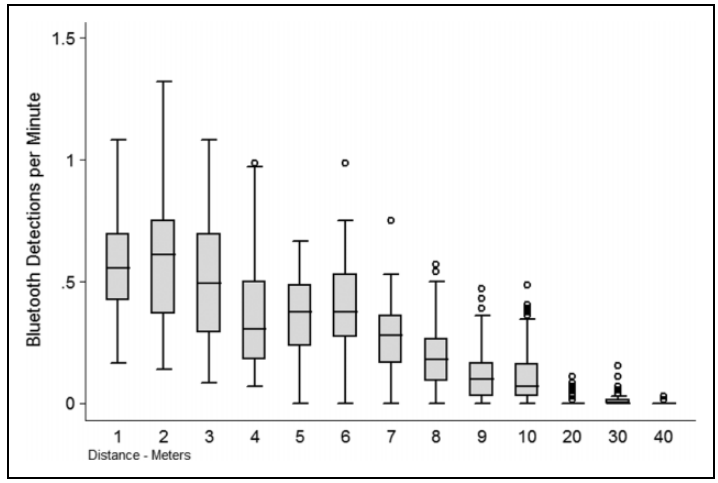

Figure 2. $N = 936$ Bluetooth detection count per minute RSSI > -80.

distance. While not shown, other detection thresholds were also investigated, and none consistently provided the desired drop-off as the distance between WSs exceeded 4 meters.

Figures 1 and 2 illustrate the role of distance in assessing co-location, but the analyses reported there do not account for the presence of obstacles, such as office walls or clothing (e.g., jacket) when it comes to detecting co-location via Bluetooth. That is, two individuals may be within 4 meters of each other, but if they are separated by a wall, they would be precluded from interaction, voiding the notion of co-location. Thus, we tested the degree to which various obstacles affected the inferences regarding co-location that might be derived from Bluetooth.

Figure 3 represents the Bluetooth detection count per minute using an RSSI threshold of –80 with physical barriers that might be encountered in a typical work environment such as a body, coat, wall, and window. For comparison purposes, the distance between the boards were approximately 1.5 meters for each of the barriers. The sample size for this plot is 432, which is the result of 24 WSs $\times$ 3 sessions $\times$ 6 experimental conditions. In an ideal case, this figure would indicate a normal detection count for the coat and a count of zero for the other barriers. This would indicate that the WS can maintain its detection capabilities when clothing is worn over the WS but that the WS does not detect co-location across barriers that would normally impede face-to-face communication.

The results illustrated in Figure 3 indicate that the detection count remains high for the coat, but it also remains high for many other barriers. These results suggest that the WSs are robust to measuring co-location when clothing is worn over the WS; however, the WSs may over-report co-location in the presence of other barriers that effectively preclude interaction between adjacent parties, such as a cubicle wall. Therefore, it is essential that researchers who deploy WSs be aware of the precise physical layout of contexts where the WSs are to be employed.

Although the figures discussed previously illustrate the ability of Bluetooth to detect co-location when there is a known signal, these analyses do not discriminate between errors that can be traced to within- versus between-WS variability. As emphasized earlier, if a large portion of the variance in detection counts is due to differences within WSs, then much of this error will cancel out over long deployments where, in essence, there may be thousands of ''items'' and individuals' scores will converge on their true scores. In contrast, if the variability is between WSs, then the error will compound over long deployments and substantially bias the results for specific individuals.

Table 2 provides results from our analysis that partitions the error variance into these two sources. These data are the same as those represented in Figures 2 and 3, thus representing only detections of

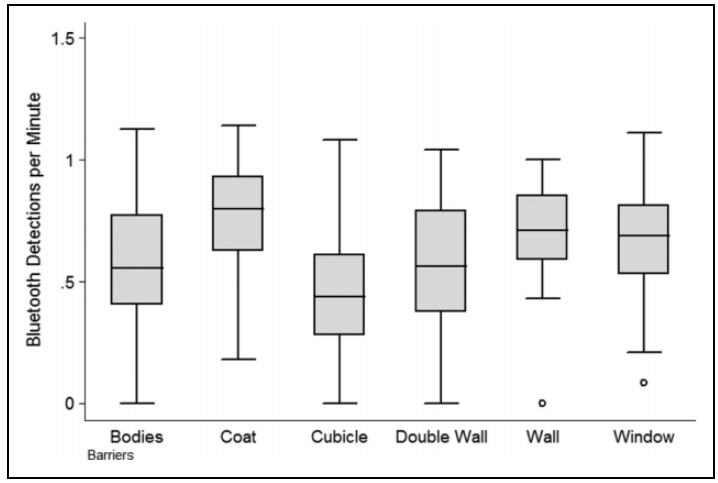

Figure 3. $N = 432$ Bluetooth detection count per minute RSSI > -80.

Table 2. Results of Ordinary Least Squares (OLS) Regression Predicting Bluetooth Detection Count.

Bluetooth Detection Count per Minute	Model 1	Model 2	Model 3	Model 4
Session fixed effects	Included	Included	Included	Included
Experimental condition fixed effects	Not included	Included	Included	Included
WS fixed effects	Not included	Not included	Included	Included
Session × WS Fixed Effects	Not included	Not included	Not Included	Included
R^2	.005	.595	.678	.692
ΔR^2		.591**	.082**	.014

Note: $N = 1{,}368$. Radio signal strength indicator (RSSI) > –80. WS = wearable sensor.
**$p < .01$.

RSSI –80 or greater. The sample size is 24 WSs × 3 sessions × 19 conditions resulting in 1,368 observations (the conditions in this analysis include both distances and barriers). In this analysis, the WS detection count is the dependent variable, and a series of regression models are used to partition variance. In the first regression, we include the session fixed effects to measure the amount of variance due to the repeated sessions. The results of this Model 1 suggest that only 0.5% of the variance is due to session, which is what one would expect since the sessions were identical replications of each other. Model 2 includes both session and distance fixed effects. The results indicate that conditions matter—explaining an incremental 59% of the variance in Bluetooth detection count. This indicates that variance in detection counts is reflective of the experienced conditions.

In Model 3 from Table 2, we include the WS fixed effect that captures the amount of between WS variance. These fixed effects predict an incremental 8.2% of the variance in Bluetooth detection count, which is all attributable to between-WS variance. In Model 4, we include a measure for within-WS variance by including an interaction between the session and WS fixed effect dummy codes. The inclusion of these fixed effects assesses the potential for within-WS variance beyond session, experimental condition, and WS fixed effects. The inclusion of these effects explains an incremental 1.4% of the variance in detection count. Finally, in a separate analysis, we included a WS experimental condition interaction, which explained 0.7% of the variance in detection count.

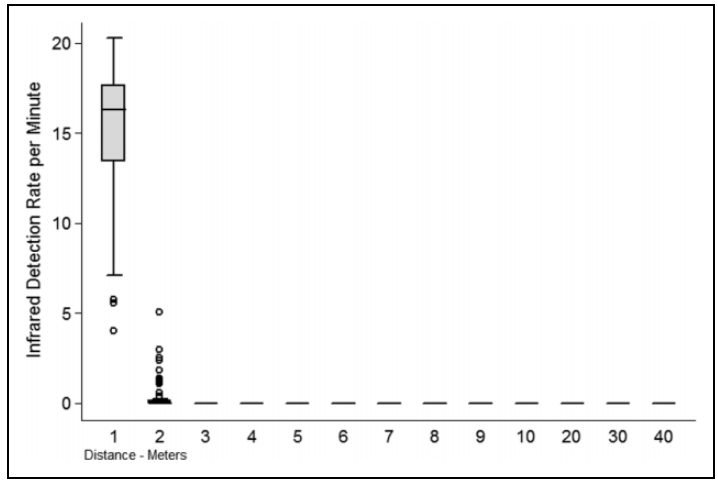

Figure 4. $N = 936$ Infrared detections per minute by distance.

Therefore, while there is some meaningful between-WS variance, it does not appear that there is significant within-WS variance in different sessions or experimental conditions.

In summary, almost 60% of the variance is due to variations in the experimental conditions, which is encouraging; however, 8% of the variance was systematic bias attributed to specific WSs that would not average out but instead compound over time. This means that this variance would be misattributed to *individuals* when in reality, it should be attributed to the specific *WS* worn by the individuals.

Co-Location and Infrared. In addition to Bluetooth, co-location can also be derived from the infrared sensor. Compared to Bluetooth, infrared detection requires more strict conditions in order to indicate proximity. With the WSs studied as part of this research, infrared detection should occur within 1.5 meters of separation, provided the faces of the WSs are within 15 degrees of being parallel with one another based on the technical constraints of the technology (Olguin et al., 2009). Following the same protocol that we developed and discussed previously with Bluetooth, we examined the infrared detection count for WSs summed by WS, session, and distance to create 936 observations (i.e., 24 WSs $\times$ 3 sessions $\times$ 13 distances). Ideally, the detection count should be high at 1 meter and diminish to 0 at greater distances. Also, the detection count should not vary by WS.

The results of this analysis are documented in the box plot in Figure 4, where we plot the detection count per minute at each of the distances. This figure shows that the count of detection is approximately 16 detections per minute at 1 meter and that this detection count significantly decreases, with only a few outlier detections at 2 meters and no detections at greater distances. Most people would consider 3 to 4 meters a reasonable distance for face-to-face social interactions, and therefore one might conclude that infrared sensors under-detect co-location. This figure also shows that there is a small amount of variance in the detection counts at 1 meter that is attributable to between-WS variability at this range. Again, a person with a more sensitive infrared sensor would be attributed more co-location inferences than a person with a less sensitive WS, and this difference, which should really be attributed to the WS, would not cancel out over long-term deployments.

As was the case for Bluetooth, we also examined the impact of common barriers on infrared detection using the same protocol. As one would expect, infrared detections are very sensitive to any physical break in line of sight. Therefore, we would expect that detection counts would be zero across all barriers, even though ideally a coat would not preclude an inference of co-location.

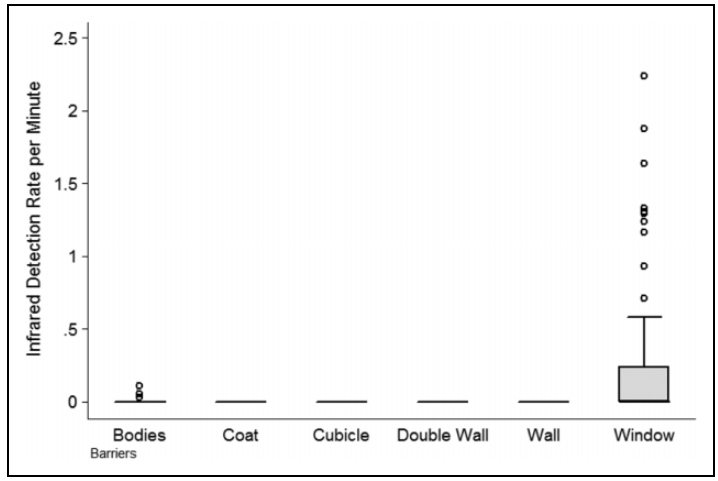

Figure 5. $N = 432$ Infrared detections per minute by barrier.

Table 3. Results of Ordinary Least Squares (OLS) Regression Predicting Infrared Detection Count.

Infrared Detection Count per Minute	Model 1	Model 2	Model 3	Model 4
Session fixed effects	Included	Included	Included	Included
Experimental condition fixed effects	Not included	Included	Included	Included
WS fixed effects	Not included	Not included	Included	Included
Session × WS Fixed Effects	Not included	Not included	Not included	Included
R^2	.000	.944	.947	.948
ΔR^2		.944**	.003**	.001

Note: $N = 1{,}368$. WS = wearable sensor.
**$p < .01$.

Figure 5 illustrates the actual detection counts for each of the barrier experimental conditions, showing that there are essentially no detections through any of the opaque barriers. However, there is a detection count greater than zero through a window, and there was a high level of between-WS variability in terms of how the window affected reports of co-location.

While helpful in understanding the potential of WSs to capture physical proximity via infrared, these figures do not precisely measure the amount of variance resulting from the two different sources of error. In Table 3, we use ordinary least squares (OLS) regression across all the infrared detection counts across both distance and barrier experimental conditions. The number of observations for this study is 1,368 based on 24 WS × 3 sessions × 19 conditions. Using infrared detection counts as the dependent variable in Model 1, we include the session fixed effects, which should explain no variance since each of the sessions were exact replications. These findings are largely consistent with the results shown in Table 3. In Model 2, we include the experimental condition, which explains 94.4% of the variance in detection count. This is a strong indicator that detection count is reflecting the actual conditions. In Model 3, we include the fixed effects for the WSs. Model 3 documents that only 0.3% of the variance in detection count is attributable to between WS variability. Finally, in Model 4, we include the interaction of session and WS fixed effects. The inclusion of these interaction variables explains only 0.1% of the variance in infrared detection count. In sum, these results suggest that nearly all of the variance in detection count for infrared sensors is a result of differences in the true signal rather than between- or within-WS variability. Despite the lack of

between- and within-WS variance, the inability for infrared to detect co-location at 3 to 4 meters severely limits the utility of the sensor for assessing co-location, and this is likely to exacerbate in real-world contexts where this sensor is attached to a platform worn loosely round the neck and chest.

Because the dependent variables in all these tests were frequency counts, they technically require a different model specification. Thus, we replicated all these analysis using a Poisson model approximation and found that the results converged across methods. We report ordinary least square results here because they provide a means for partitioning variance across alternative sources that is commonly understood (i.e., R^2). Results based on Poisson regression are available from the authors.

Summary. Regardless of the analytical method, the results associated with assessing co-location via Bluetooth and infrared suggests that the majority of the variance in both sensors can be explained by differences in proximity. Moreover, we find that there is between-WS variance in Bluetooth detection counts that will not average out and will compound over longer deployments. We also find little evidence of within-WS variance, suggesting that the WSs appear relatively stable in detection counts for infrared and Bluetooth across deployment sessions. We also show that the threshold value that is chosen will have a strong impact on the results of Bluetooth, which again foreshadows results we will discuss later in this article when we attempt to measure higher level behavioral and social constructs. Finally, the inability of infrared to detect at 3 to 4 meters severely limits the sensors utility in measuring co-location.

Evidence Regarding Verbal Activity

We performed a range of tests to assess the ability of the microphone to detect verbal activity. In the first test, we placed 24 WSs in a stack (because microphone ports are on top of the unit) with all of the microphones pointed directly at a stereo speaker 1.5 meters away. To conduct this test, we used two tones; the first tone covered frequencies from 20 Hz to 2,000 Hz (which we refer to as a "sweep" tone) over a period of 30 seconds, and the second tone was stable at 170 Hz for 10 seconds (this is the average frequency of human voice; Titze, 1994). To further put this in perspective, the standard 88 key piano has a range of 32 Hz to 4,186 Hz, with "middle C" on the piano registering at 261 Hz.

Each experimental session took place in a quiet, isolated office and included 10 seconds of ambient noise, 30 seconds of the sweep tone, 10 seconds of ambient noise, 10 seconds of the stable tone, and concluded with 10 seconds of ambient noise. This procedure was repeated across 3 sessions, turning the WSs off and back on between sessions. It should be noted that there are multiple data streams that come from the microphones. First, there is a measure of raw volume. Second, the microphone provides both amplitude and frequency levels as well as variability within four frequency bands (Olguin et al., 2009). Our analysis covers overall volume and the filtered data within the first frequency band intended to correspond to human vocalization, including frequencies between 85 to 222 Hz (Olguin et al., 2009). Volume and amplitude both refer to the overall loudness of sound (size of the sound wave). Frequency refers to the pitch of the sound (concentration of sound waves).

In this analysis, we focus only on the first frequency band. All microphone measures are automatically aggregated every 8 milliseconds within the device, which in turn outputs an average of these values once per second. We coded each second of the data using a dummy code for each session, each WS, and the experimental condition (including background noise, sweep tone, and stable tone). Because there were multiple seconds for each experimental condition, these values were then collapsed by averaging per experimental condition. Therefore, the number of observations for this analysis was 24 WS $\times$ 3 experimental conditions $\times$ 3 sessions for a total of 216.

Table 4. Results of Ordinary Least Squares (OLS) Regression Predicting Volume.

Volume	Model 1	Model 2	Model 3	Model 4
Sweep tone		.012** (.001)	.012** (.001)	.012** (.001)
Stable tone		.013** (.001)	.013** (.000)	.013** (.000)
Constant	.019** (.001)	.010** (.001)	.004** (.000)	.004** (.001)
Session fixed effects	Included	Included	Included	Included
WS fixed effects	Not included	Not included	Included	Included
Session × Microphone Fixed Effects	Not included	Not included	Not included	Included
R^2	.003	.672	.961	.973
ΔR^2		.670**	.233**	.012

Note: $N = 216$. Standard errors in parentheses. Background noise is base condition for this regression analysis with the tones being dummy codes for the other respective conditions. WS = wearable sensor.
**$p < .01$.

Table 5. Results of Ordinary Least Squares (OLS) Regression Analysis for Predicting Amplitude.

Amplitude	Model 1	Model 2	Model 3	Model 4
Sweep tone		.004** (.000)	.004** (.000)	.004** (.000)
Stable tone		.000** (.000)	.000** (.000)	.001** (.000)
Constant	.005** (.000)	.003** (.000)	.003** (.000)	.003** (.000)
Session fixed effects	Included	Included	Included	Included
WS fixed effects	Not included	Not included	Included	Included
Session × WS fixed effects	Not included	Not included	Not included	Included
R^2	.003	.947	.968	.977
ΔR^2		.944**	.038**	.009

Note: $N = 216$. Standard errors in parentheses. Background noise is base condition for this regression analysis with the tones being dummy codes for the other respective conditions. WS = wearable sensor.
**$p < .01$.

In Tables 4 and 5, we conduct a regression analysis for raw volume and filtered amplitude, respectively. In these analyses, the dependent variables are the resultant measures of volume and amplitude, and the independent variables are dummy codes that capture the tones, sessions, and WSs. In an ideal case, all of the variance in our dependent variables should be attributed to the tones.

As shown in Table 4, Model 1 includes the session fixed effects, which account for 0.3% of the variance in volume. We attribute this to subtle differences in ambient background noise between sessions. Even though the room and surrounding area seemed quiet when conducting tests, it was not a completely soundproof room. In Model 2, we include the 2 experimental condition dummy codes with background noise as the base condition. The coefficients for both tones are positive and significant and taken together explained 67% of the variance in volume.

In Model 3, we include the effect for WS, and this accounts for 23.3% of the variance in volume. Thus, whereas the means of volume for each of the tones is statistically different from background noise, there remains a high level of variance between WSs. This is troubling because bias between WSs may cause a researcher to inappropriately assign speaking to some individuals due to WS differences rather than true variance in volume. One potential solution to this would be to employ the first derivative of this measure. However, this may create more problems than it solves because this creates a measure of "within-person" speaking, which is rarely the index of interest for most behavioral researchers. Most behavioral research is interested in "between-person" differences in

speaking (that is, who speaks the most and the least), not "within-person" differences in speaking (that is, is one specific person speaking more or less than he or she usually does).

In Model 4, we include an interaction between the session and WS fixed effects in order to measure the consistency with which the WSs detect their environment. The inclusion of these interaction variables explains an incremental 1.2% of the variance in volume; therefore, it appears that any given WS is relatively consistent in detecting sound.

In Table 5, we repeat this analysis employing speaking band amplitude as the dependent variable. In Model 1, the session fixed effects account for 0.3% of variance in amplitude. In Model 2, as expected, the coefficient for both the sweep and constant tone are positive and statistically significant. The inclusion of these dummy code measures of tone explains 94.4% of the variance in amplitude. This suggests that nearly all of the variance in amplitude is a result of environmental sound. In Model 3, we include WS fixed effects, which explains an incremental 3.8% of the variance in amplitude. This is significantly less than the between-WS effects observed when predicting volume. Finally, in Model 4, we include the interaction of the session and WS fixed effects, which accounts for only 0.9% of the variance in amplitude, suggesting that overall, the WSs are relatively reliable in detecting sound across deployment sessions.

Summary. These tests seem to suggest that simple amplitude measures of select frequency bands may be less susceptible to microphone-induced bias. However, there are two caveats to this conclusion. First, the tests reported here were very short in duration, and the potential for longer tests to provide more valid data cannot be ruled out. Second, the tones tested here were not actual human voices. Thus, with these two caveats in mind, we turn to Study 2, which addresses both of these limitations.

Study 2

In Study 2, we focused specifically on the validity of the microphone and the WS proprietary analytics to properly assign speaking. Unlike all the tests performed as part of the first study, Study 2 involved human participants interacting with each other in a laboratory setting. Because it involves human participants who may vary in both speech tone and volume, this study is less tightly controlled than Study 1, but in terms of foreshadowing, it is more tightly controlled than Study 3. Study 2 essentially simulates how one might use WSs as part of a laboratory study to simply measure verbal activity (as an alternative to employing coded videotapes of interactions). Study 3, on the other hand, simulates how one might use WSs as part of a study assessing behavioral constructs, such as boundary spanning and emergent leadership.

In Study 2, four undergraduate participants, sitting 1.5 meters apart, wore *two* WSs each and read a structured script according to the outline in Table 6. During this period in our study, the WSs we were studying had a flaw in the firmware that led to a random failure in the microphone that manifested itself in a flat-line audio reading across the entire session. This problem with flat-lining was later resolved with a firmware update, but following the manufacturers recommendations at the time, the research participants wore two WSs in order to minimize data loss.

The script included a variety of speakers, speaking times, and conversation structures. In order to maintain conformity with the prescribed script, each session was supervised by a research assistant who kept participants on track. The purpose of this study is to determine the potential of the WS to identify the speaker and assess speaking time accurately. This setting provides an opportunity to assess the effectiveness of speech detection using the WSs in conjunction with their supporting speech detection algorithms in a context where the true score was known.

The data were collected across 8 sessions with different subjects in each session, and, as mentioned previously, each of the 4 participants was wearing 2 WSs, which resulted in a total of 64 WS samples and 16 sessions. Among these sessions, 8 WSs failed, resulting in 6 failed team

Table 6. Seconds Spoken by Each Participant.

	Speaker A	Speaker B	Speaker C	Speaker D
Minute 1	30	10	10	10
Minute 2	10	30	10	10
Minute 3	10	10	30	10
Minute 4	10	10	10	30

sessions. Therefore, the number of observations for this study is 160, which results from 10 sessions $\times$ 4 roles $\times$ 4 minutes. The actual WS indicated speaking seconds are reported in Table 7.

It is important to note that there are an array of speech detection algorithms available to researchers to analyze vocal activity patterns (Proakis, 1999). In addition, the provided proprietary analytical software supports a number of settings to identify the speaker and assess speaking time. Thus, this is another area where researchers would need to make important decisions on which algorithm to employ based on their research question and contextual setting. In order to make this determination for our study, we employed every possible setting combination available at the time for a total of eight analyses. Using these outputs, we evaluated each according to their accuracy, that is, the degree to which the output represented the true score value. In order to conserve length, we include results for just two of the algorithms here, namely, the manufacturer recommended algorithm and an optimized algorithm that was the most accurate setting combination in this context.

In Table 7, we show the correlations among these different predictors and the actual speaking time. This table shows that the correlation among speech as measured by the two algorithms is $r = .47$, $p < .01$. The results in Table 7 suggest that the correlation between the manufacturer recommended algorithm and actual speaking is $r = .15$, $p = ns$, whereas the correlation between the optimized algorithm and actual speaking is $r = .36$ $p < .01$. The divergent performance of these algorithms with respect to the detection of actual speaking in this experiment illustrates that for both theoretical and conceptual reasons, it is critical to establish a proper match between WS, speech detection logic, research question, and environment. Moreover, the low correlations for the algorithms suggest limitations in the current technology for accurately measuring speaking time.

Table 7. Correlations, Means, and Standard Deviations of Speaking Measures.

	Measure	Mean	*SD*	1	2
1.	Speaking time	15.00	8.69	—	
2.	Manufacturer recommended algorithm speaking time	23.71	15.81	0.15	—
3.	Optimized algorithm speaking time	15.31	9.61	0.36**	0.47**

Note: $N = 160$.
**$p < .01$.

Summary. Taken as a whole, the results from Study 1 and Study 2 indicate that the WSs we examined performed better when assessing co-location relative to verbal activity. However, it also needs to be kept in mind that these studies were both short in duration, and longer time intervals would generate many more measurement opportunities. Thus, even though signal detection is low, across thousands of detections, this could still generate reliable measures. This is analogous to a situation where a very long test, like the Scholastic Aptitude Test (SAT), can generate reliable estimates of aptitude despite very low ($r = .05$) item-total correlations. With this in mind, we performed two additional

studies that lengthened the duration for detection and where we targeted broader behavioral and social constructs.

Study 3

Studies 3a and 3b attempt to evaluate the ability of WSs to assess boundary spanning behavior and leadership emergence, respectively. Like Study 2, in Study 3a, we have known "true scores" in a relatively controlled setting where we knew precisely who was and who was not a boundary spanner. In Study 3b, although we do not have true scores for leadership emergence, we had the best known alternative—aggregated ratings from other team members regarding the focal individual, where we show high levels of agreement between raters. Although boundary spanning and leadership emergence are hardly the only concerns researchers have with respect to team dynamics, it is a reasonable first step for assessing the extent WSs can be utilized to measure constructs of interest to organizational researchers and hence show relevance for the extant knowledge base.

Evidence Regarding Boundary Spanning

A boundary spanner refers to an individual who coordinates work-related activities *between* established formal boundaries (Davison & Hollenbeck, 2012; Marrone, 2010). Traditional measures of boundary spanning behavior often rely on self-reports that are subject to well-known biases, especially when individuals are asked to respond to boundary spanning activities over an extended period of time (e.g., week, months, quarters) (Podsakoff & Organ, 1986). Thus, it was imperative in this context to have a known "true score" for validating WS-based measures of boundary spanning behavior.

Study 3a was conducted in a laboratory context where the individuals wearing WSs were working in three independent teams of four to five members each separated in isolated rooms of a larger research suite. While these teams worked, a team of observers ($n = 3$ to 6 who also wore WSs) moved back and forth between the rooms, sometimes aggregating themselves in a separate room of the suite to compare notes. This was all conducted as part of a class, and the observers were, with few exceptions, not members of the research team but merely advisors to the students in this class. In this context, the observers are known boundary spanners moving between the boundaries of the teams, whereas the team members themselves rarely, if ever, crossed boundaries during the course of the session. We gathered data across 13 separate sessions, and in the end, we had data from 235 individuals, of which 31% were known boundary spanners and 69% were known to not be boundary spanners.

Across all sessions, the WS reported a total 190,607 total Bluetooth detections, and we analyzed these data to see if one could discriminate boundary spanners from non–boundary spanners using Bluetooth detections set at alternative levels. That is, boundary spanning status (coded 0 or 1) should show a significant positive correlation with between-team detections and a statistically significant negative correlation with within-team connections because the non–boundary spanners were together at all times. Based on the results from Study 1, we also varied the detection setting to see if this had any impact on the resulting number of detections or the correlations between boundary spanning status and within- and between-team detections.

Table 8 shows the descriptive statistics and correlations between the total number of within-team ties and between-team ties at varying Bluetooth RSSI cutoff values ranging from –70 to –90. For RSSI signal strength, the larger value of RSSI, such as –70, would be a stronger signal relative to –90. Within-team ties were calculated as the sum of all Bluetooth detections between a WS worn by an individual and all other WSs that were assigned to his or her team, including boundary spanners that were also assigned to their own team. Between-team ties were calculated as the sum of all

Table 8. Descriptive Statistics and Correlations of Study Variables.

Variables	Mean	SD	Bspan	WTT
Boundary spanner	0.31	0.46	—	
RSSI > –90				
Within-team	259.88	139.36	–0.63*	—
Between-team	176.85	119.28	0.43*	–0.07
RSSI > -85				
Within-team	171.28	102.96	–0.60*	—
Between-team	82.14	69.73	0.42*	–0.06
RSSI > –80				
Within-team	79.40	59.46	–0.47*	—
Between-team	26.74	29.91	0.32*	0.06
RSSI > –75				
Within-team	32.94	31.44	–0.32*	—
Between-team	8.50	11.53	0.25*	0.18*
RSSI > –70				
Within-team	9.03	11.11	–0.21*	—
Between-team	2.06	3.80	0.16*	0.32*

Note: $N = 235$. RSSI = radio signal strength indicator; Bspan = boundary spanner; WTT = within-team ties; BTT = between Team Ties.
*$p < .05$.

Bluetooth detections from other WSs that were worn by individuals *not assigned* on the same team. As noted in Study 1, the Bluetooth sensor generates a categorical measure of detection (on/off), a time/date stamp, sender WS number, receiver WS number, and an RSSI ranging from –65 to –95 that generally varies with the distance between WSs.

The first and second columns of Table 8 show that a significant number of detections are removed at more stringent thresholds. That is, we see much higher means and standard deviations at the more liberal –90 RSSI threshold and much lower means and standard deviations at the more conservative –70 RSSI threshold. Thus, as one would expect, the threshold setting has a major impact on the mean and standard deviation of detections.

More critically, the correlations reported in the third and fourth columns of Table 8 show that the threshold setting also has an impact on the validity of the WS for detecting boundary spanners. The correlation between known boundary spanning status and between team detections was highest ($r = .43$, $p < .05$) at more liberal settings and then decreased steadily as setting became more stringent ($r = .16$, $p < .05$). It should be noted that the manufacturer recommended setting provided by the manufacturer (RSSI > –80) did not result in the highest validity in this context ($r = .32$, $p < .05$), and instead, a researcher in this context would generate a more valid measure by using a more lenient threshold when it came to predicting the boundary spanning behavior of individuals.

Table 8 shows that this same pattern of results was reinforced when we examined within-team ties. Non–boundary spanners should show more within-team ties relative to boundary spanners because they never left the room that they occupied with their other team members. Indeed, the correlation between boundary spanning status and within-team detections was negative and statistically significant. This correlation was strongly influenced by the set threshold, however, and again, the highest validity ($r = –.63$, $p < .05$) was found with the most liberal threshold, and this decreased steadily as the threshold became more stringent, bottoming out at –.21 ($p < .05$). It was also the case that the manufacturer recommended setting did not produce the highest validity in this context ($r = –.47$, $p < .05$).

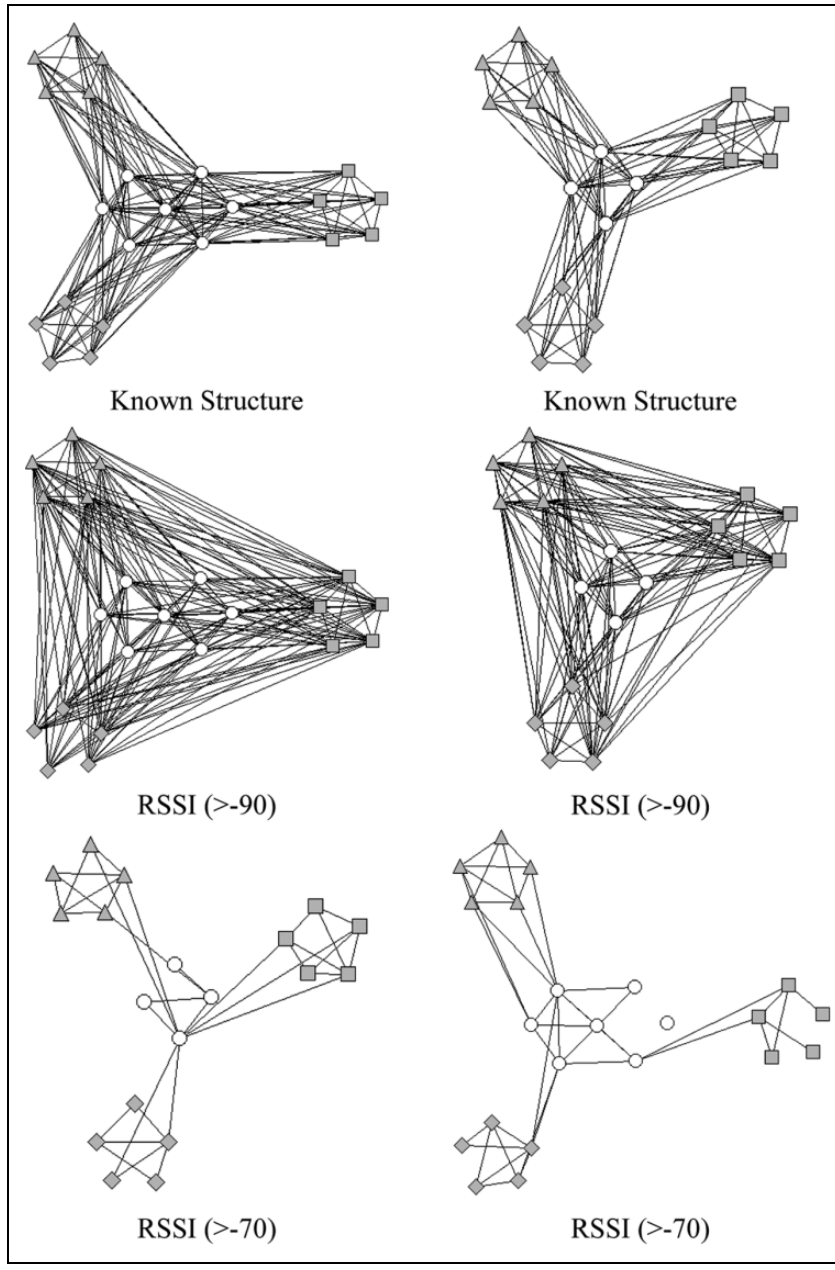

Figure 6. Bluetooth detection networks at different RSSI. *Notes:* RSSI = Radio Signal Strength Indicator [-70, -80, -90]; ○ = boundary spanner.

Going beyond simply the count of ties, it is also instructive to see how well the social structure of these teams is revealed by the WSs at various thresholds. Figure 6 depicts the differences between the known co-location structure of our simulated environment and the structure generated by the WS at two different Bluetooth signal cutoffs for two representative sessions. The top image represents the ''true'' co-location structure of individuals during the simulated exercise. The next set of images

represents Bluetooth detections at a RSSI of –90 (most liberal cutoff), followed by the same picture at more and more stringent cutoffs.

Clearly, these pictures show that the decision of which RSSI to use as a cutoff can have a profound influence on the generated co-location structure. In our simulated lab environment, the most liberal cutoff generated too many between-team ties for non–boundary spanners. As may be recalled from Study 1, a Bluetooth signal can sometimes be detected across walls, and hence detections between rooms cannot be ruled out in this case when the threshold is set at the most liberal level. However, at the most stringent cutoff level, the WS data stream failed to generate all the known within-team ties and thus was also problematic.

In this case, a more moderate setting of $\geq$75 generated a social structure that best matched the known co-location network as a whole. This differed from the recommended value set by the manufacturer and also differed from the best setting for the simple detection of boundary spanners per se. Thus, the best cutoff to employ in this context is also contingent on whether the primary interest is in identifying boundary spanners at the individual level or the larger social structure at the team level. One potential explanation for this is that because the correlation is based on variance, the liberal cutoffs generate the largest standard deviation that creates the largest correlation. However, the errors of omission appear to be less of a problem in representing the whole network structure relative to the error of commission (which quickly leads to an overly saturated graph). Thus, again the nature of the research question and environmental factors drives the decisions regarding how to set threshold levels.

Still, although there are differences due to the chosen threshold, it is important to not lose sight of the fact that when the best threshold is identified, the ability of the WSs to detect boundary spanners and the co-location structure of these teams is quite impressive. If a researcher was not on site in this facility on the evenings where this activity took place, that researcher would actually get a highly valid picture when it came to boundary spanning and co-structure just from the appropriately analyzed Bluetooth data streams.

Evidence Regarding Emergent Leadership

Study 3b builds on Study 2 and Study 3a by focusing on using the microphone and, in particular, using measures of verbal activity to predict leadership emergence in a team context. Leadership emergence refers to the recognition of others that a particular individual is distinctive in terms of having a strong degree of social influence within the group. Leadership emergence is of significant interest to behavioral researchers because of the disproportionate influence that such individuals have on the goals, structure, and processes of the group (Zaccaro & Klimoski, 2002).

Leadership emergence can be assessed with self-reports, but social desirability bias often introduces error into such measures because of the societal value placed on being a leader or displaying leadership qualities (DeRue & Ashford, 2010). Thus, measures of leadership emergence often require the collection of reports from other people, but these measures are sometimes colored by halo errors or similar-to-me biases. Still, at this point in time, a measure based on aggregated ratings across a number of different raters who can be shown to agree is considered the best available standard for capturing this construct. In Study 3b, we focus on how the amount of time speaking as captured via a microphone can predict leadership emergence in newly formed teams.

For this study, we gathered data across 15 different sessions from the same research suite described in Study 3a, each lasting between one and two hours. The teams that participated in Study 3a were not the same teams that took part in Study 3b, and hence these data are independent from the data reported previously. In this case, each session included two to three newly formed teams, each consisting of four to five members, working in a computerized simulation activity in a controlled setting. In total, these sessions allowed us to observe 205 individuals nested in 43 teams.

Table 9. Summary Statistics of Speaking and Leadership Emergence.

	Measure	Mean	*SD*	1	2
1.	Perceptions of leadership	3.54	0.69	—	
2.	Manufacturer recommended algorithm speaking time	505.62	195.12	−0.03	—
3.	Optimized algorithm speaking time	229.81	182.05	0.14*	0.30**

Note: $N = 205$.
*$p < .05$. **$p < .01$.

Within this setting, team members worked together to perform a complex, interdependent task called LDX. This task is described in detail elsewhere (see Davison & Hollenbeck, 2012), but for purposes here, we will note that it requires team members to initiate structure, coordinate efforts, and make collective decisions. Team members speaking up in this context often results in their emergence as a leader because as the team grapples for collective understanding and processes, team members often rely on the more vocal members to coordinate activities and make important decisions.

We measured leadership emergence with items from a previously validated scale (Kent & Moss, 1994; Lord, Foti, & De Vader, 1984; Taggar, Hackew, & Saha, 1999). This scale is comprised of four items focusing on whether the individual exerts influence, leads the conversation, and influences the team's goals and directions. Leadership emergence ratings from all of the other team members (provided immediately after completing the session) for the focal individual were then averaged. The raters in this study were shown to agree in the sense that ICC(1), ICC(2), and r_{wg} values were well above the cutoffs typically invoked to justify aggregation. For these data, the ICC(1) was .35, ICC(2) was .71, and r_{wg} was .91 (James, 1982; James, Demaree, & Wolf, 1984).

For measures of speaking, the WS data were collected, and the two speech detection algorithms used in Study 2 were employed here to capture total speaking time (the number of seconds each individual spoke). Thus, we have one measure of leadership emergence and two speaking measures from the WSs.

The summary statistics and correlations for our measures can be found in Table 9. Here we show that the average total speaking time is 505.6 seconds according to the manufacturer recommended algorithm whereas total speaking time is 229.8 seconds according to the optimized algorithm emerging from in Study 2. This equates to approximately 8.4 to 3.8 minutes, respectively, of speaking on average. While the correlation between these two measures is statistically significant ($r = .30$, $p < .05$), suggesting some overlap, the difference in the raw level between these measures is substantial. In addition, we note that correlation differs from our findings in Study 2 that we partially attribute to a firmware upgrade shortly after Study 2 was completed.

Turning to the predictive validity of alternative speaking measures for the criterion of leadership emergence, we see that there was a small but statistically significant correlation between leadership emergence and the optimized algorithm ($r = .14$, $p < .05$). The correlation for manufacturer recommended algorithm was actually opposite the proposed direction ($r = –.03$) but not statistically significant. Thus, we conclude that while the correlation between speaking time as measured by the WS and leadership emergence is not strong, there appears to be some ability of the WSs to predict leadership emergence via speaking time as measured with the optimized algorithm. Collectively, these results suggest that the ability of the WSs and the supporting analytic software to predict leadership emergence is conditional on the algorithm used to assess and assign speaking.

In conclusion, as part of the overall portfolio of studies, Study 3 extends Study 1 and Study 2 by showing how one might use WSs to detect common constructs that might be of interest to behavioral researchers. In addition, the paradigm underlying Study 3 simulates how a WS would perform in a laboratory context with nonscripted human research participants where there are known true scores (or the best alternative, that is, aggregated reports from others shown to agree on a validated scale).

Study 3 is limited, however, in the sense that research participants were observed for only a short time in a small and tightly controlled space. Study 3 does not simulate how WSs might be used in less controlled field studies that might take place over wider space and time intervals. In Study 4, we seek to address these limitations.

Study 4

As previously mentioned, Study 3 attempts to simulate the use of WSs in laboratory contexts. Although less controlled than Study 2, where human participants read scripts, there was sufficient control of time and space to know the true scores for the measures that we were validating. In Study 4, we extend this portfolio of studies to examine how one might employ WSs in a field study where research participants interacted in wide-ranging and uncontrolled space over an extended time period.

In this context, it is impossible to ascertain known true scores, and hence, unlike Study 1 through Study 3, we used subjective self-reports and self-reported schedules as criteria for assessing WS performance. Clearly, for reasons we noted at the outset of this article, these ratings provide a questionable criterion, and if we had faith in these sorts of measures, there would be no need for WSs. However, for comparison purposes, we use self-reports as a helpful proxy for our phenomena of interest. In addition, beyond the issue of detection accuracy, we were also interested in deriving some qualitative experience for a long-term deployment of WSs.

Specifically, Study 4 took place in a field setting and focused on the detection of co-location using infrared and Bluetooth. For this study, the WSs were worn for a 6-week period by 14 individuals working as part of an ongoing research team at a major university. This team included full professors, assistant professors, senior graduate students, junior graduate students, and undergraduate students who were part of two different departments in the same college. In addition to the 14 individuals, we placed 3 WS ''base stations'' in locations where these 14 individuals would be likely to congregate (i.e., conference rooms and laboratories). Base stations are useful when there are a small number of places where participants congregate because they provide triangulation opportunities for assessing co-location.

Co-Location. For the first part of Study 4, we focused on the Bluetooth and infrared sensor to determine the degree to which WS ratings of co-location converged with self-ratings of co-location in an uncontrolled field study where participants wore WSs for six weeks as part of their normal work life on and off campus. In this context, we tested the degree to which weekly self-ratings of co-location were correlated with weekly data from the WSs at various levels of Bluetooth strength. As we noted when discussing the results from Study 1, a conservative Bluetooth signal level threshold led to greater discrimination of proximity. Here we examine multiple levels of Bluetooth RSSI because in this study, we consider a more realistic context for the use of the WSs in a field setting. Understanding the convergence of Bluetooth and surveys of co-location at different levels of Bluetooth RSSI is important in order to understand best practices of using Bluetooth detection data as a measure of co-location.

In this case, we do not know exactly when or where these participants were co-located through the week outside of a regularly scheduled weekly project meeting. To determine the accuracy of the WSs to detect co-location, we triangulate between two other measures. First, we collected self-reports of co-location from the participants on a weekly basis. Second, for the participants of this study, their respective role in the team and their weekly schedules are known by our research team. Thus, while convergence between co-location as measured by the WSs and these two other measures are not entirely free from bias attributable to the limits of the self-reported data, evidence of convergence is still a relevant criterion for assessing the validity of the Bluetooth and infrared data to assess co-location.

Participants in this study completed a weekly survey asking the number of hours they spent with each other member of the team and the chosen locations containing a base station. The values ranged

Table 10. Correlations Between Bluetooth Detections and Self-Reported Co-Location.

Variable	Mean	*SD*	1	2	3	4	5	6	7
1. Self-report	2.20	1.29	—						
2. Bluetooth > –90	75.73	141.12	0.51	—					
3. Bluetooth > –85	51.13	97.78	0.51	0.96	—				
4. Bluetooth > –80	25.91	57.80	0.45	0.83	0.94	—			
5. Bluetooth > –75	13.63	37.01	0.37	0.68	0.81	0.95	—		
6. Bluetooth > –70	7.21	25.53	0.26	0.53	0.64	0.80	0.94	—	
7. Infrared total	3.69	18.88	0.14	0.39	0.47	0.61	0.75	0.91	—
8. Infrared total by minute	0.45	1.53	0.23	0.49	0.57	0.68	0.76	0.83	0.86

Note: $N = 423$. All correlations are significant $p < 0.01$. Numbers presented after the greater than sign represent different radio signal strength indicator (RSSI) thresholds utilized for identification of co-location.

Table 11. Results of Ordinary Least Squares (OLS) Regression Predicting Self-Reported Co-Location.

Self-Reported Co-Location	Model 1	Model 2	Model 3
Bluetooth > –90	0.005** (0.001)	0.005** (0.001)	0.005** (0.001)
Infrared total		−0.005 (0.003)	−0.011* (0.004)
Infrared total by minute			0.091 (0.073)
Constant	1.844** (0.059)	1.843** (0.060)	1.837** (0.059)
R^2	0.265	0.269	0.272
ΔR^2		0.004	0.002

Note: $N = 423$. Standard errors in parentheses. –90 represents the radio signal strength indicator (RSSI) threshold utilized for identification of co-location.
*$p < .05$. **$p < .01$.

from none to over four hours on a 6-point scale. Because the correlation was moderately high ($r = .48$) among the surveys for connections among team members, we averaged the two measures, and when one member failed to complete the survey, we used the other individual's score for that measure of co-location. Because of the lack of interactions during a holiday week, one week was dropped from this analysis. Therefore, the number of observations for this study is 423, which is an average of 85 dyads across 5 weeks.

As shown in Table 10, there was convergence among all these measures of co-location, although the magnitude of this varied substantially. Less stringent cutoffs generated higher convergent validities ($r = .51$, $p < .05$), and these validities decreased steadily as the cutoff became more stringent, bottoming out at $r = .26$, $p < .01$. Self-reported co-location correlated higher with Bluetooth than infrared across all thresholds investigated.

Triangulation is another approach to evaluate the WSs ability to measure co-location. To assess the combination of both infrared and Bluetooth to measure co-location, we conducted a regression analysis, which is reported in Table 11. In this analysis, we predict the survey measures of co-location, and in Model 1, we include the count of Bluetooth detections of –90 or greater RSSI. This coefficient is positive and significant ($\beta = 0.005$ $p < .01$), as expected, and accounts for 26.5% of the variance in self-reported co-location. In Model 2, we include the total count of infrared detections. Finally, in Model 3 we also include infrared total by minute. This coefficient is not significant ($p = .21$), and the inclusion of this variable explains an incremental 0.2% of the variance in self-reported co-location, which again is not statistically significant.

In sum, these results suggest that while Bluetooth and infrared supposedly offer complementary measures of co-location, they in fact do not converge when it comes to assessing co-location. In

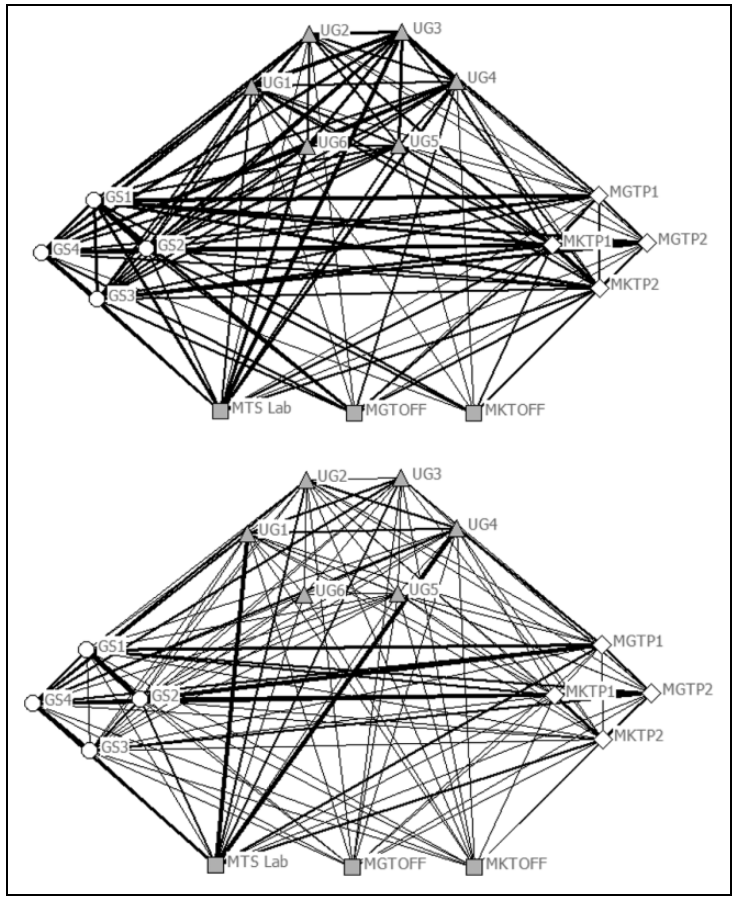

Figure 7. Upper image co-location network based on survey responses. Lower image co-location network based on Bluetooth detections RSSI>-90. Line width indicates the time spent interacting. *Key*: UG: Undergraduate students; GS: Graduate students; MGTP: Management professors; MKTP: Marketing professors; MTS Lab: Research Lab; MGT: Management office; MKT: Marketing office.

addition, the combination of these predictors does not appear to explain substantively more variance in self-reported co-location. Given the space and battery consumption of the infrared sensor, one might question the added value of this sensor relative to just the Bluetooth sensor alone, at least when worn loosely around the neck and chest.

Although correlations such as those reported in Table 10 are informative when it comes to convergence of tie strength among independent actors, this does not consider the structure of a whole network. To understand the convergence of co-location structure measured with self-reports versus Bluetooth, we constructed the co-location networks that would have been generated for this specific set of people via alternative methods, using the $\geq$90 Bluetooth threshold to measure the edges with the individuals and office locations as the nodes. Interestingly, this is a less conservative threshold than Study 3a, which we employed because there were more substantial physical barriers between participants in this study. This further reinforces the importance of configuring sensor calibrations to the research question and physical context.

The upper image of Figure 7 represents the network according to self-reported co-location, and the lower image of Figure 7 indicates the network according to Bluetooth detection. Although there

is some degree of convergence and divergence between these two figures, in general, the co-location networks generated by these alternative sources were very similar. A researcher who did not know this set of people would get a strikingly accurate picture of their co-location network from a set of WSs. For example, the strong connections between the undergraduates (UG) and the MTS laboratory are strong for both networks as well as the connections between the graduate students (GS) and professors within departments (MGTP/MKTP).

Still, there were some discrepancies across sources, and it was instructive for us to examine these at a more molecular level. For example, two graduate students had a class with one of the professors and thus spent a minimum of 3 hours per week co-located as part of this course. The self-report results do not seem to "count" these interactions—almost as if the subjective reporters did not consider the interactions within class as relevant—but the WSs did. In addition, there was an unusually high number of detections between one undergraduate student and the base station set up in the laboratory. After following up on these results, we learned that the undergraduate student's WS was accidently left on in the laboratory without being worn.

Summary. In general, these results suggest that there are significant limitations in using infrared for measuring co-location. Further, these results suggest that while there are some errors regarding both self-reported and Bluetooth measures of co-location, it appears that the WSs offer significant potential for measuring co-location networks via Bluetooth sensor technology.

Verbal Activity. The second part of Study 4 focuses on the convergence between the WS-derived measures of speaking dominance with self-reported measures of speaking dominance. As noted previously, we do not have true score measures of who dominated conversations; however, we do know that all people were together in the same room. Immediately after each meeting, the participants completed a survey indicating the degree to which various participants dominated the conversation relative to not speaking on a scale of 1 to 5. These ratings were then averaged across raters to establish a measure of dominance and rank ordered for each person for a given meeting. To align the measures with the WS data, we took the total speaking time of the participants and rank ordered the participants based on the number of speaking seconds. We then correlated the survey measure with the ranked measures.

The results of this analysis are shown in Table 12. The number of observations for this analysis is 80, based on the fact that it captures activity at 12 separate meetings with attendance ranging from 5 to 9 participants. As illustrated in Table 12, the correlation between the two speech detection algorithms underlying the calculation of verbal dominance reached an encouraging $r = .75, p < .05$. This is not surprising as convergence between algorithms is expected to increase with the duration of the study. Although strictly speaking the number of observations for generating the correlation reported in Table 12 is 80, in reality, because most of the WSs fire several times a minute (or second), the number of observations (or "items" in psychometric parlance) across the 12 roughly one hour meetings is well into the thousands. Therefore, the number of observations greatly increases the

Table 12. Summary Statistics of Ranked Dominance.

	Measure	Mean	*SD*	1	2
1.	Survey ranking	3.95	2.11	—	
2.	Manufacturer recommended algorithm speaking rank	3.95	2.11	0.53**	—
3.	Optimized algorithm speaking rank	3.96	2.11	0.53**	0.75**

Note: $N = 80$.
**$p < .01$.

opportunity to converge on the true score of dominance. Table 12 also shows that the two different operationalizations converged when it came to predicting other reports of dominance, with both correlating exactly the same with participant ratings ($r = .53$). Therefore, the two approaches to measuring speech in this study showed similar levels of convergent validity.

Summary. Based on this set of results, taken as a whole, we conclude that although the convergence between the WS and survey measures is not perfect, it appears that the WSs offer some potential to identify dominant speakers in a field study setting. If a researcher was not at these meetings and was not able to survey the participants after each meeting, the researcher would nonetheless have some degree of knowledge about what took place in this room based solely on WS reports. Although the degree of convergence is modest, this fact needs to be considered in light of the limitations of the subjective self-reports, which are far from perfect, and thus place an upper limit on convergence.

Discussion

The advent of "big data" collection and computing is revolutionizing the world in general, and thus it should not be surprising that this would eventually touch the lives of behavioral researchers. There are many different reactions that we as social scientists could take to these developments. First, we could go into denial and hope this is just a fad that will pass over in due time. Certainly, the evidence we have provided here regarding WSs is far from perfect when it comes to assessing behavioral constructs, thus leaving room for denial. This response might be especially attractive because as a discipline, we seem to have become so accepting of the limits of retrospective self-reports and other reports for capturing behavioral measures that we act as if those limitations do not exist. We do not believe that denying the value of WS-generated data is a constructive stance to take toward these technological developments. WSs are here to stay, and the opportunity to use this moment in time to make game-changing adaptations to business as usual should not be squandered.

One might also take the stance that big data as captured by WSs may have value someday, but we should wait for the engineers to perfect these devices before we adopt them in our research. Again, the evidence presented here is not perfect when it comes to supporting the construct validity for some of the measures derived from the WSs. Hence, one could simply ask the engineers to do their psychometric homework and come back to us when the evidence is stronger. We believe this "wait and see" attitude is also unwarranted for two reasons.

First, the technological challenges associated with building platforms such as this, along with the accompanying software and analytic algorithms, are nontrivial. Engineers with the skills to overcome these challenges need to be steeped within their own disciplines. Building effective WSs requires a complex set of skills associated with understanding of the capacity of an ever increasing array of sensors, batteries, and the necessary design and manufacturing acumen to mass produce reliable devices at a feasible cost. WSs have to operate reliably despite being jostled, dropped, covered, and in one case in Study 4, slammed inside a car door. It is a lot to ask of scientists who have the expertise to do all of this to also be experts in psychometrics. For example, within the field of psychometrics, it is well understood that item-total correlations are maximized when the variance for the dichotomous variables approaches the maximum (i.e., base rate value of .50). Thus, one potential means for establishing the threshold value for a sensor is to choose the value that creates this specific base rate. This is an insight that may be more familiar to a psychometrician relative to an engineer who is unfamiliar with test construction and construct validation principles.

Second, even if these individuals were to develop the psychometric skills we take for granted in the behavioral sciences, the potential that they might derive measures and constructs that differ from what already exists within our literature would be a risk that is too great to tolerate. That is, we cannot also expect this set of people to be experts in the extant scientific literature on individuals,

interpersonal dyads, groups/teams, and larger organizational structures. Yet, this would have to be the case to prevent a situation where WS engineers start generating new constructs on their own that overlap little with the constructs upon which our theories and knowledge base have developed over the last one hundred years in behavioral sciences. The ability to employ WSs as a game-changing development in the sciences will require that we integrate the constructs and measures that emerge from this measurement technology with the extant knowledge base. It would be a loss if we were unable to leverage the current knowledge base and then have to "start over from scratch" with these new constructs and measures. Thus, we do not feel that a "wait and see" attitude is an appropriate reaction to these developments.

Finally, a third reaction that one might have toward this new technology is to simply accept it as is, place WSs on research participants, cross our fingers, and then hope for best. There is very little in the evidence that we present here that would warrant this course of action. Uninformed and uncritical use of this technology is particularly hazardous because of the highly interdependent nature of decisions related to construct (e.g., level of analysis), technology (e.g., sensor type), and analytical approaches (e.g., algorithms). These challenges are exacerbated by the nontransparent nature of the proprietary analytical process for creating measures of behavioral constructs generated by WS data.

In our opinion, the stance that we, as behavioral and social scientists, should be taking toward these new technological developments is to work alongside engineers in order to help improve the measures derived from WSs and integrate them into the extant knowledge base. That is, starting with the extant knowledge base and validated measures as the initial groundwork, we need to ask questions such as "How can boundary spanning, as traditionally defined and measured, be effectively captured with data generated with WS technology?" "How can cohesiveness as traditionally defined and measured be effectively captured with data generated with WS technology?" and "How can 'tie strength' as traditionally defined and measured be effectively captured with data generated with WS technology?"

Although this entails a great deal more effort than "denial," or "wait and see," or "blindly and uncritically hoping for the best," we believe this challenge is deserving of our efforts. In some cases that we document here, the current generation of WSs actually performs with almost striking accuracy. We were impressed by the strong convergence in the co-location networks generated by the Bluetooth data in Study 3 and Study 4. To think that someone who did not know or survey this set of individuals could so accurately know their social structure was remarkable. Of the basic measures that are generated by the WSs in this study, in our opinion they identify co-location very well *as is*, especially when the data are collected over long durations.

As a means of detecting verbal activity, the WSs we studied performed moderately well although not nearly as well as was the case for co-location. Although we did not have any true scores for Study 4, this was a long duration study where the two different measures of speaking, generated by two different algorithms, converged with each other and showed similar correlations with self- and other reports of speaking frequency. However, this was not the case with Study 2, where the two different measures did not converge (and the optimized algorithm was more accurate for predicting a known true score). The difference here may be attributable to the length of time people were studied, and across all measures we examined here, the longer the time period (i.e., the more items), the better the results.

Recommendations and Best Practices

In terms of going forward, there are several recommendations that can be made based on these results when it comes to planning, conducting, and reporting a WS-based study.

When Planning a Research Project. In the planning phase, when choosing a wearable sensor system, researchers should carefully evaluate whether sensor type, attachment location, and mode

(e.g., lanyard, wristband, etc.) align well with the source and nature of the behavioral signal to be captured. Regardless of the choice of WSs, we strongly encourage allocating time for extensive sensor pretesting to inform decisions related to the observation time and analytical strategy.

For example, in pretesting Bluetooth-based proximity sensors, we recommend that researchers utilize the sensors on subjects with a known co-location network in a physical setting resembling the actual study environment. This approach allows for an informed decision related to what Bluetooth signal strength might constitute a legitimate interaction opportunity.

A similar approach can be used for audio sensors. Building on known conversation patterns, researchers should develop a clear understanding of the joint impact of variation in sensor sensitivity and particular choices of speech detection algorithm on measurement accuracy.

At present, researchers can expect that the components of wearable sensor systems are subject to rapid technological advances. To leverage these advances, manufacturers operate using short development cycles resulting in frequent updates and reduced product longevity. Over the course of a longitudinal study, investigators may have to accommodate multiple pretesting episodes in order to confirm the optimal configuration and measurement consistency.

Contingent on the findings during sensor pretesting, researchers should adjust observation periods. For example, we found that the wearable sensors used in our study exhibited a high level of random error. Longer study windows increase the likelihood that measurements will converge on the true score. Our experience would suggest that one-shot laboratory studies where the observation period is short (one or two hours) may be better off employing digital video recordings.

Finally, in the planning and pretesting phase, it is critical for researchers to scrutinize both the firmware installed on the sensors as well as the software solution used to download and analyze the data. In our studies, we found that performance of the sensors can vary dramatically based on the firmware release and that the options and workflow varied significantly with different software solutions. In the planning phase, it is critical for researchers to pretest the robustness of these factors prior to substantial data collection.

When Conducting Data Collection. When conducting the study, we suggest that researchers pay close attention to subject compliance, procedures for monitoring and maintaining sensor function, and the configuration of sensors in the research site.

For research using WS technology, compliance of the research subjects is paramount. In particular, when sensors capture interaction data, as illustrated in Study 2, the failure of a single sensor or the failure of an individual to wear the sensor has a significant impact on overall network data quality. To minimize this problem, we recommend researchers enhance compliance by educating subjects to properly wear the sensor and regularly monitor data stream integrity to ensure sensor function and subject compliance. This involves downloading WS data on a regular basis and analyzing it near real time to identify anomalies (e.g., sensors not moving or not detecting speech).

Research involving WSs should also include procedures for device charging and synchronization. The WSs used here required regular charging and clock synchronization to avoid temporal drift. Drifting clocks in WSs can lead to the WSs assessing the same phenomenon but at different recorded times, which would greatly affect the accuracy of interaction-based analytics.

Finally, the configuration of the sensors at the research site can be critical to effective data collection. For example, because Bluetooth sensors report some level of random error within a particular timeframe, it can be useful to place base stations at strategic locations (e.g., lunch room) for triangulation purposes when it comes to detecting co-location in well-known meeting areas.

When Reporting and Evaluating Research. When working with WS technology, it is important for researchers to report and justify their choices regarding data analysis and interpretation (e.g., RSSI Bluetooth signal strength threshold, speech detection algorithm setting, and the data aggregation

approach) because the substantive results may vary as a function of these decisions. In terms of Bluetooth-based proximity sensors, the interaction patterns reported by the WS will vary significantly based on the RSSI signal strength threshold adopted as well as what an interaction is deemed to look like. For example, a more liberal approach might count any Bluetooth detection between a dyad as a 60 second interaction while a more conservative approach would require joint detection by both nodes and count that as only a 30 second interaction. These decisions should be made on the basis of the severity of the risk of false positive versus false negative errors when it comes to detecting co-location.

In addition, the choice of speech detection algorithm can significantly alter how the situation is characterized, and therefore, rather than recommending a particular speech detection algorithm or settings, we suggest that through pretesting, researchers choose, justify, and report their choice of speech detection algorithm because a universally optimal setting is unlikely to exist. In addition, we encourage researchers to pay particular attention to the type of algorithm selected (i.e., within- vs. between-individual based measures) in light of their research question. It is critically important to understand that within-person measures assess how much one is talking in one context versus how much that same person is talking in a different context. Between-person measures assess how much one person is talking relative to other people who are all in the same context. These are very different phenomena. Failure to match the measure with the phenomena of interest actually seems to be quite common when we see others present WS-based research at professional meetings.

Relatedly, the high data resolution afforded by WSs provides a myriad of different approaches for aggregating these data to arrive at measures of a given construct. In this work, we aggregated based on simple sums of detections or seconds spoken. For other research, it may be of interest to take a variance or other aggregation approach. We suggest that these aggregation decisions should be explicitly disclosed and justified in future research using WSs.

Finally, it is important for researchers to consider and minimize the risk of non-sensor wearers in a given study. While the central value proposition of wearable sensors relates to the accurate measurement of behavioral aspects related to human interaction, researchers operating in uncontrolled environments need to evaluate the effects of potential signal contamination by non-instrumented subjects.

Conclusion

As one of the first independent studies to examine WSs across the wide variety of contexts in which they might be employed in behavioral research, we obviously have only scratched the surface of what needs to be done to realize the full potential of these devices for all forms of behavioral research. Certainly, we see this as a conversation starter for the field of behavioral research, and this will hardly be the last word given the rapid changes in both sensor diversity and potential platforms for sensors. In particular, with respect to new platforms, it is hard to anticipate all of the potential sensor configurations (accelerometers, microphones, Bluetooth, optical scanners) and attachment modes (badges, wristbands, lapel pins, glasses, and implants). We do believe this is a conversation that is worth having and a conversation that should be led by the scientific community and not one simply left to engineers or to the non-refereed popular press (e.g., Silverman, 2013). We hope as a first effort the studies reported here serve as a catalyst and model for future researchers as we try to radically expand the way we think about measuring behavioral constructs.

Acknowledgment

We would like to thank the Associate Editor Scott Tonidandel and two anonymous reviewers for the extremely helpful and constructive comments in the course of bringing the manuscript to its current form. We are also grateful to Peter Lindeman and Audrey Garneau for their assistance in data collection.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a grant from the National Science Foundation though the Division of Advanced Cyberinfrastructure (Award #1231154).

References

Davison, R. B., & Hollenbeck, J. R. (2012). Boundary spanning in the domain of multiteam systems. In S. J. Zaccaro, M. A. Marks, & L. A. DeChurch (Eds.), *Multiteam systems: An organizational form for dynamic and complex environments* (pp. 323-362). New York, NY: Routledge.

DeRue, D. S., & Ashford, S. J. (2010). Who will lead and who will follow? A social process of leadership identity construction in organizations. *Academy of Management Review*, *35*, 627-647.

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology*, *17*, 245-260.

George, G., Haas, M. R., & Pentland, A. (2014). Big data and management. *Academy of Management Journal*, *57*, 321-326.

Greenberg, Jerald, & Robert, A. Baron (1995). *Behavior in Organizations: Understanding and Managing the Human Side of Work*. 5th edition. Englewood Cliffs, N.J: Prentice Hall College Div.

Hall, E. T. (1990). *The hidden dimension*. New York, NY: Anchor.

Hallberg, J., Nilsson, M., & Synnes, K. (2003). Positioning with Bluetooth. In *10th International Conference on Telecommunications* (Vol. 2, pp. 954-958). Piscataway, NJ: IEEE.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology*, *67*, 219-229.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, *69*, 85-98.

Kent, R. L., & Moss, S. E. (1994). Effects of sex and gender role on leader emergence. *Academy of Management Journal*, *37*, 1335-1346.

Kim, T., McFee, E., Olguin, D. O., Waber, B., & Pentland, A. (2012). Sociometric badges: Using sensor technology to capture new forms of collaboration. *Journal of Organizational Behavior*, *33*, 412-427.

Kozlowski, S. W. J., Chao, G. T., Chang, C.-H., & Fernandez, R. (in press). Team dynamics: Using "big data" to advance the science of team effectiveness. In S. Tonidandel, E. King, & J. Cortina (Eds.), *Big data at work: The data science revolution and organizational psychology*. New York, NY: Routledge Academic.

Lord, R. G., Foti, R. J., & De Vader, C. L. (1984). A test of leadership categorization theory: Internal structure, information processing, and leadership perceptions. *Organizational Behavior and Human Performance*, *34*, 343-378.

Marks, M. A., Zaccaro, S. J., & Mathieu, J. E. (2000). Performance implications of leader briefings and team-interaction training for team adaptation to novel environments. *Journal of Applied Psychology*, *85*, 971-986.

Marrone, J. A. (2010). Team boundary spanning: A multilevel review of past research and proposals for the future. *Journal of Management*, *36*, 911-940.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology*, *85*, 273-283.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.

Olguin, D., & Pentland, A. (2010). Sensor-based organisational design and engineering. *International Journal of Organisational Design and Engineering*, *1*, 69-97.

Olguin, D., Waber, B., 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*, 43-55.

Pentland, A. (2012). The new science of building great teams. *Harvard Business Review*, *90*, 60-69.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*, 879-903.

Podsakoff, P. M., & Organ, D. W. (1986). Self-reports in organizational research: Problems and prospects. *Journal of Management*, *12*, 531-544.

Proakis, J. G. (1999). *Discrete-time processing of speech signals*. New York, NY: Wiley-IEEE Press.

Silverman, R. E. (2013, March 7). Tracking sensors invade the workplace. *Wall Street Journal*. Retrieved from http://online.wsj.com/article/SB10001424127887324034804578344303429080678.html.

Sparrowe, R. T., Liden, R. C., Wayne, S. J., & Kraimer, M. L. (2001). Social networks and the performance of individuals and groups. *Academy of Management Journal*, *44*, 316-325.

Spector, P. E., & Brannick, M. T. (2010). Common method issues: An introduction to the feature topic in *Organizational Research Methods*. *Organizational Research Methods*, *13*, 403-406.

Taggar, S., Hackew, R., & Saha, S. (1999). Leadership emergence in autonomous work teams: Antecedents and outcomes. *Personnel Psychology*, *52*, 899-926.

Titze, I. R. (1994). *Principles of voice production*. Englewood Cliffs, NJ Prentice Hall.

Zaccaro, S. J., & Klimoski, R. (2002). The interface of leadership and team processes. *Group & Organization Management*, *27*, 4-13.

Author Biographies

Daniel Chaffin is finishing his PhD in strategic management at Michigan State University and is currently an assistant professor at the University of Nebraska Kearney. His research interests include demand-side strategy, organizational boundary spanning, and organizational data systems.

Ralph Heidl received his PhD at the University of Washington and his MS from the Pennsylvania State University. He is currently an assistant professor at the University of Oregon, and his research interests include how enterprising firms and individuals engage in multilateral collaboration to create and use new technological resources.

John R. Hollenbeck is a University Distinguished Professor at Michigan State University. His research interests include team decision making, self-regulation theories of work motivation, and employee separation and acquisition processes.

Michael Howe received his PhD from Michigan State University and is currently an assistant professor of management at Iowa State University. His research interests include adaptive performance, interactional dynamics, and methodological refinement.

Andrew Yu is currently a doctoral candidate in the management department at Michigan State University studying organizational behavior. He received his BA in business administration and his MA in economics from California State University, Fullerton. His research program is currently focused around understanding the antecedents and outcomes of how employees collaborate in groups and teams as well as both the within- and between-team factors that may influence these relationships.

Clay Voorhees received his PhD from Florida State University and is an associate professor of marketing at Michigan State University. His research interests include customer loyalty and relationship marketing, service experience management, and return on marketing investments.

Roger Calantone is the Eli Broad Chaired University Professor of Business at the Eli Broad Graduate School of Management at Michigan State University. His research interests include product design and development processes, decision support, pricing, and price perception.