Communication in Physical Sciences, 2024, 11 (4):1266-1280

Comparative Performance Analysis of Mobile Application and Instrumented Wet Bulb Globe Temperature Monitoring: A Field-Based Accuracy Assessment

Oluwaseun Ibuife Oluwaniyi

Received: 19 September 2024/Accepted: 21 December 2024/Published:31 December 2024

Abstract: *The increasing danger of environmental heat stress on outdoor workers, athletics, and vulnerable populations globally is due to the rising intensity and frequency of extreme temperatures of thermal events as climate change worsens. Wet Bulb Globe Temperature (WBGT) index is the accepted gold standard of measuring the exposure to heat stress, but outdated equipment is frequently too expensive and logistically difficult to deploy on a large scale. Mobile applications based on estimating WBGT using meteorological inputs and algorithmic models are also available as the accessible alternatives, yet their measurement accuracy in various field conditions is still not fully defined. In this research, a systematic field comparison was done between calibrated WBGT reference equipment and the American Industrial Hygiene Association (AIHA) Heat Stress App 2.0 on 52 simultaneous measurements at 4 weeks in an outdoor occupational environment. The WBGT measurements were between 19.8°C and 39.4°C, whereas AIHA app was between 16°C and 32°C. The average systematic bias was 3.86°C (WBGT > AIHA), which had a standard deviation of 4.54°C and 95.*

Keywords: *Wet Bulb Globe Temperature, occupational heat stress, mobile health monitoring, environmental measurement accuracy, thermal exposure assessment, digital occupational health.*

Oluwaseun Ibuife Oluwaniyi*
Department of Occupational Risk and Safety Sciences, University of Central Missouri, United States.
Email:
seunibuife.oluwaniyi@gmail.com
http://orcid.org/0009-0008-1960-8442

1.0 Introduction

Heat stress in occupational and environmental settings is an increasingly significant public health concern, driven by the rising frequency, intensity, and duration of extreme heat events associated with climate change. (Alahmad *et al*., 2023; Kjellstrom *et al*., 2016). Workers in outdoor industries such as agriculture, construction, military operations, and emergency response are at elevated risk of heat-related illnesses (HRI), reduced cognitive and physical performance, and, in severe cases, fatal heat stroke (Parsons *et al*., 2022). Physiological responses to heat stress include increased cardiovascular strain, electrolyte imbalance, impaired thermoregulation, and reduced neuromuscular coordination, all of which compromise safety and productivity (Périard *et al*., 2021). In addition to personal health impacts and beyond health impacts, heat stress imposes substantial economic costs through reduced labor productivity, increased workplace injuries, and declines in agricultural and industrial output (Peymaneh *et al*., 2024).

Accurate characterization of environmental thermal conditions is essential for implementing effective heat illness prevention strategies, designing work–rest cycles, and ensuring compliance with occupational health standards in exercising effective heat illness prevention programs, setting up of proper work-rest schedules, and regulatory adherence to occupational health standards. Wet Bulb Globe Temperature (WBGT) index, first created by the military of the United States in the 1950s and since standardized by international protocols (ISO 7243:2017), has become the most widespread environmental screening tool in terms of heat stress violation (Bernard *et al*., 2023). WBGT integrates three key environmental parameters: natural wet-bulb temperature (reflecting humidity and evaporative cooling), black globe temperature (radiant heat load), and dry-bulb air temperature. The composite index has stronger relationships with physiological strain and incidence of heat illnesses compared to any of the single

environmental variables, meaning it is the most desirable to use to set exposure limits and activity discontinuation in relation to heat exposure (Brimicombe *et al*., 2023).

Despite its scientific validity and widespread institutional acceptance, conventional WBGT measurement presents several practical limitations that hinder its large-scale implementation, especially in resource-constrained environments or organizations that do not have committed occupational hygiene facilities installed. Calibrated WBGT instruments are often expensive, require routine maintenance, and must be operated by trained personnel, limiting their accessibility in resource-constrained settings (Kong & Huber, 2024). Measurement procedures allow stabilization time to be between 15 and 30 minutes, depending on the conditions of the environment, which may delay real-time decision-making in situations of rapid thermal changes. It is also true that stationary WBGT measurements could fail to measure spatial heterogeneity in intricate outdoor work environments where employees shift between sunny and shaded regions, come across various surfaces, and microclimatic gradients (Golbabaei *et al*., 2021).

The widespread adoption of smartphones and cloud-based data services has led to the development of mobile applications designed to estimate WBGT using meteorological data and algorithmic models. These applications offer several advantages, including low cost, ease of use, rapid output without stabilization delays, and the ability to collect data across geographically dispersed locations (Dillane & Balanay, 2020; Eggeling *et al*., 2023). An example of this type of technology is the Heat Stress App 2.0 (published in 2024 by the American Industrial Hygiene Association) which estimates WBGT, and provides work-rest guidelines based on the AIHA guidance documents and ISO standards (AIHA, 2024). Despite these advantages, the accuracy and reliability of app-based WBGT estimates under real-world field conditions remain insufficiently validated. However, there are still some critical questions concerning the accuracy of measurement and the operational reliability of app-based WBGT estimation in comparison with reference instrumentation in real field conditions. Despite these advantages, the accuracy and reliability of app-based WBGT estimates under real-world field conditions remain insufficiently validated (Liljegren *et al*., 2008). This modeling method presents possible systematic biases and random errors, especially where there is a localized microclimatic situation that does not conform to the spatial scales of the weather station network or the meteorological products of satellites. Recent comparative studies have identified instances where app-based technologies consistently underestimate instrument-measured Wet Bulb Globe Temperature (WBGT) (Angol, 2024). These inconsistencies also have direct implications on worker safety in case thermal risks are underreported and economic in that case of a conservative set of thresholds, which result in unwarranted work stoppages. Consequently, there is a need for systematic field-based validation studies comparing app-derived estimates with calibrated reference measurements. This study aims to evaluate the accuracy of the AIHA Heat Stress App 2.0 by comparing its WBGT estimates with measurements obtained from calibrated reference instrumentation under real-world outdoor conditions.

The study employs paired observations collected over a four-week period, capturing a range of environmental conditions, diurnal cycles, and solar exposure levels. . The analysis framework combines traditional methods of comparison of measurements, such as Bland Altman agreement analysis and correlation analysis, in order to fully describe systematic bias, random measurement error, and environmental variables related to app-instrument discrepancies.

The work is of value to the existing literature on occupational health and applied meteorology by offering empirical evidence on the precision limits of a commonly used mobile application, finding environmental factors that worsen the accuracy of apps, and

making practical recommendations on what hybrid approaches to monitoring can be used to trade off accessibility and measurement fidelity. The results are directly applicable to occupational safety specialists in charge of making real-time heat safety decisions, regulatory bodies in charge of defining compliance solutions, app developers interested in streamlining algorithm methods, and standards bodies revising guideline reports on what heat stress monitoring technologies should be allowed to use. The findings will support evidence-based decision-making in selecting appropriate heat stress monitoring tools for occupational safety management. Unlike previous studies that rely on controlled or short-term observations, this study provides extended field-based evidence under dynamic outdoor conditions.

The following sections detail the field methodology and analytical procedures that were used to collect and compare the data, provide the empirical findings of the analysis of the differences in the agreement of measurements and support them by the statistical tests and graphical illustrations, comment on the implications of the observed differences on the operational aspects of the occupational thermal safety programs in terms of the existing literature and the practical occupation of health care, and present the recommendations on the evidence-based use of mobile heat stress monitoring devices as the part of the comprehensive thermal safety program implementation.

2.0 Materials and Methods

2.1 Study Design and Setting

A field-based comparative measurement study was conducted over a four-week period from 9 September to 6 October 2024 in Warrensburg, Missouri, United States (38.76°N, 93.74°W; ~280 m above sea level). The study was conducted on the campus of the University of Central Missouri, which provided an outdoor environment representative of typical occupational settings with direct solar exposure, variable wind conditions, and mixed surface materials such as grass, concrete, and asphalt. The late summer to early autumn period provided a wide range of thermal conditions, from moderate temperatures with low solar load to peak afternoon heat stress typical of outdoor occupational environments. The measurements were made in open outdoor areas with no major artificial sources of shading or heating which will interfere with the measurements, in line with the ISO 7243 guidelines of a representative WBGT measurement in an outdoor workplace. The measurement protocol has chosen to sample at different times of the day to obtain diurnal variation of solar radiation, ambient temperature and humidity that affect both the direct measuring protocol and algorithmic WBGT estimation. Sampling was conducted across three daily time periods: morning (08:00–11:00), midday (11:00–14:00), and afternoon (14:00–17:00), across multiple days, resulting in 52 paired concurrent observations.

2.2 Instrumentation and Mobile Application

Reference WBGT Monitor

*T*he reference standard for comparison was a calibrated thermal environment monitoring system compliant with ISO 7243 requirements for WBGT measurement. The instrument consists of three sensors: a natural wet-bulb thermometer (with wetted wick exposed to natural airflow), a black globe thermometer (150 mm matte-black copper sphere with an internal thermistor), and a dry-bulb air temperature sensor. The instrument was factory-calibrated within the past 12 months and verified against NIST-traceable reference thermometers before field deployment. The resolution of the measurement was ±0.1°C, and the accuracy of the manufacturer was 0.5°C in the operating range.

Before each measurement session, the WBGT monitor was positioned at approximately 1.1 m above ground level, corresponding to the center of mass of a standing worker, in accordance with ISO 7243 recommendations. The instrument was allowed to stabilize for at least 20 minutes before recording measurements to ensure thermal equilibrium of the wet-bulb and globe components.

2.3 AIHA Heat Stress App 2.0

The AIHA Heat Stress App 2.0 was installed on an iPad (iOS 17.5). On the iOS 17.5 operating system. The app calculates estimates of WBGT by using a combination of parameters entered by the user (location, types of clothing, workload classification) and weather data (meteorological data) received through weather service APIs, depending on the GPS position of the device. The app estimates WBGT using empirical regression models based on meteorological inputs including air temperature, relative humidity, wind speed, and solar radiation, as described in the technical documentation and prior literature.

(Liljegren *et al*. 2008).

To ensure consistency, standardized input parameters were used for all measurements: location (GPS-derived coordinates), workload classification (moderate activity), and clothing type (summer work uniform consisting of lightweight cotton clothing). Weather data retrieval was initiated one minute after recording the reference WBGT measurement to ensure temporal alignment of environmental conditions. The WBGT estimate generated by the app to the main results screen was manually entered by research personnel.

2.4 Data Collection Protocol

The research personnel conducted all measurements using a standardized protocol to ensure consistency across all sampling sessions. At each site, the WBGT instrument was installed, oriented correctly, and the wet-bulb reservoir was filled with distilled water. The instrument was allowed to stabilize before measurements were recorded. The instrument was then left to stabilize and the environmental conditions visually assessed and recorded.

The researcher recorded the stabilized WBGT value (to 0.1°C resolution), along with the corresponding date and time. Simultaneously, the researcher operated the AIHA app, enabled GPS synchronization, entered standardized inputs and recorded the resulting WBGT estimate. To the accuracy given (usually in whole degrees Celsius), the WBGT value was entered into the app.

The National Institute of Occupational Safety and Health (NIOSH) Heat Stress Index, also generated within the app using a separate algorithm, was recorded for contextual analysis but excluded from the primary comparison. All the data collected were recorded and then plotted in a structured field data sheet and then transferred into a digital spreadsheet to be analyzed.

To assess reproducibility and detect potential temporal drift, 19 measurement sessions were conducted multiple times within the same day at intervals of 1–4 hours. The remaining 33 sessions were conducted on separate days.

2.5 Statistical Analysis

All statistical analyses were performed using R (v4.3.0) and Python (v3.11) with standard scientific computing libraries.

Both datasets of WBGT monitor and AIHA app have been analyzed using descriptive statistics such as mean, standard deviation, minimum, maximum and interquartile ranges. The Bland–Altman method was used to assess agreement between the two measurement techniques. (Bland & Altman, 1986).

For each paired observation, the mean and difference were calculated using equations 1 and 2

$$\overline{x_i = \frac{(WBGT_i + AIHA_i)}{2}} \quad (1)$$

$$d_i = WBGT_i - AIHA_i \quad (2)$$

The systematic bias was determined as the average of all differences $\bar{d}$, and positive values would mean that the WBGT monitor values were higher than the app values. Random measurement error was determined by the standard deviation of differences s_d and 95% limits of agreement (LoA) were computed as $\bar{d} \pm 1.96 s_d$, in which any difference between measurements is likely to be within in the case of constant systematic and random error.

Pearson's correlation coefficient (r) was used to assess linear association between the two methods. However, correlation alone does not imply agreement, as systematic bias may still be present. Even though the correlation coefficients outline the degree of linear association between two variables, it is necessary to note that, a high level of correlation does not always mean that there is

an agreement in measurement as it can be expected that there can be systematic bias even when the two variables comparably follow each other. Thus, the interpretation of the results of correlation and Bland-Altman analysis was done together to present a complementary view of measurement performance.

Further exploratory tests were done to determine whether the difference in measurements was systematic with environmental condition. The dataset was stratified into quartile WBGT monitor reading in order to determine whether the underestimation in the app rose with higher levels of thermal stress. Each pair of measurements were plotted and linear regression models of variables were created to visualize a relationship between the measurements and additionally a trend of variables across days were plotted using scatterplots.

To examine whether measurement differences varied with environmental conditions, data were stratified into WBGT quartiles. Linear regression and scatterplots were used to explore relationships between measurement differences and environmental variables. A significance level of $\alpha = 0.05$ was adopted. Given the field-based nature of the study, greater emphasis was placed on identifying systematic bias and practical measurement differences rather than relying solely on statistical significance.

3.0 Results and Discussion

3.1 Descriptive Statistics and Environmental Conditions

Across the 52 paired measurements, environmental thermal conditions were well distributed and representative of typical occupational heat stress scenarios in temperate outdoor environments. This range spans moderate thermal conditions, where heat stress risk is low for acclimatized workers, to extreme conditions approaching or exceeding 32°C, where work restrictions or engineering controls are typically required (ISO 7243:2017). (Table 1). This temperature span includes moderate thermal conditions where the heat stress is small at workers acclimated to performing light to moderate work, to extreme conditions near or beyond the 32 C limit which normally elicits work constraints or engineering controls on even acclimatized workers who perform light work (ISO 7243:2017).

Table 1: Descriptive statistics of the measurements of heat stress.

Parameter	WBGT Monitor (°C)	AIHA App (°C)
Minimum	19.8	16.0
Maximum	39.4	32.0
Mean	27.7	23.8
Standard Deviation	4.9	4.1
Median	27.3	24.0
25th Percentile	24.3	21.0
75th Percentile	29.7	27.0

In contrast, AIHA Heat Stress App estimates ranged from 16.0°C to 32.0°C, with a mean of 23.8°C and a standard deviation of 4.1°C. This narrower range and lower mean indicate systematic underestimation by the app relative to the reference instrument. Notably, the app did not report values above 32°C, even when the reference instrument recorded temperatures as high as 39.4°C, suggesting a potential ceiling effect or algorithmic limitation in extreme conditions.

As shown in Figure 1, both measurement systems captured similar diurnal patterns, with values increasing from late morning to early afternoon and declining thereafter. This similarity in behavior validates the fact that the meteorological data inputs and algorithmic transformation behind the app are experiencing the first-order time variation of environmental cooling and heating. However, a consistent magnitude difference between the two methods was observed across all time

periods and environmental conditions. The two systems also monitor diurnal trends but have a systematic imbalance where WBGT monitor records higher values always.

3.2 Correlation Analysis

Pearson's correlation analysis between the reference WBGT monitor and AIHA app measurements yielded a coefficient of r = 0.384 (p = 0.005), indicating a statistically significant but moderate linear relationship (Fig. 2). This result has important implications for the practical use of mobile heat stress applications.

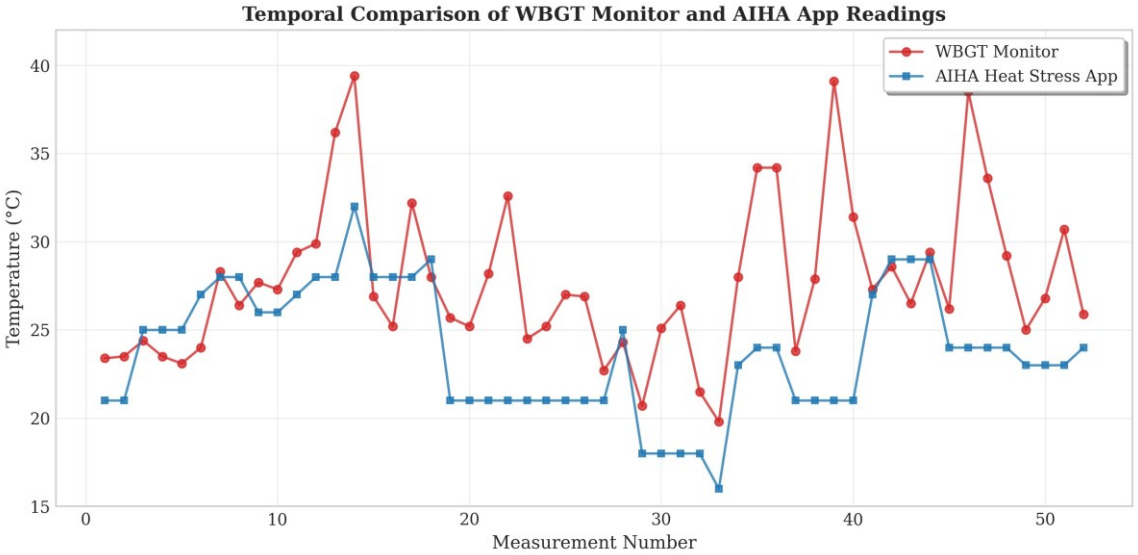

Fig. 1: Comparison of readings of both WBGT monitor and AIHA Heat Stress App over time in 52 simultaneous measurements.

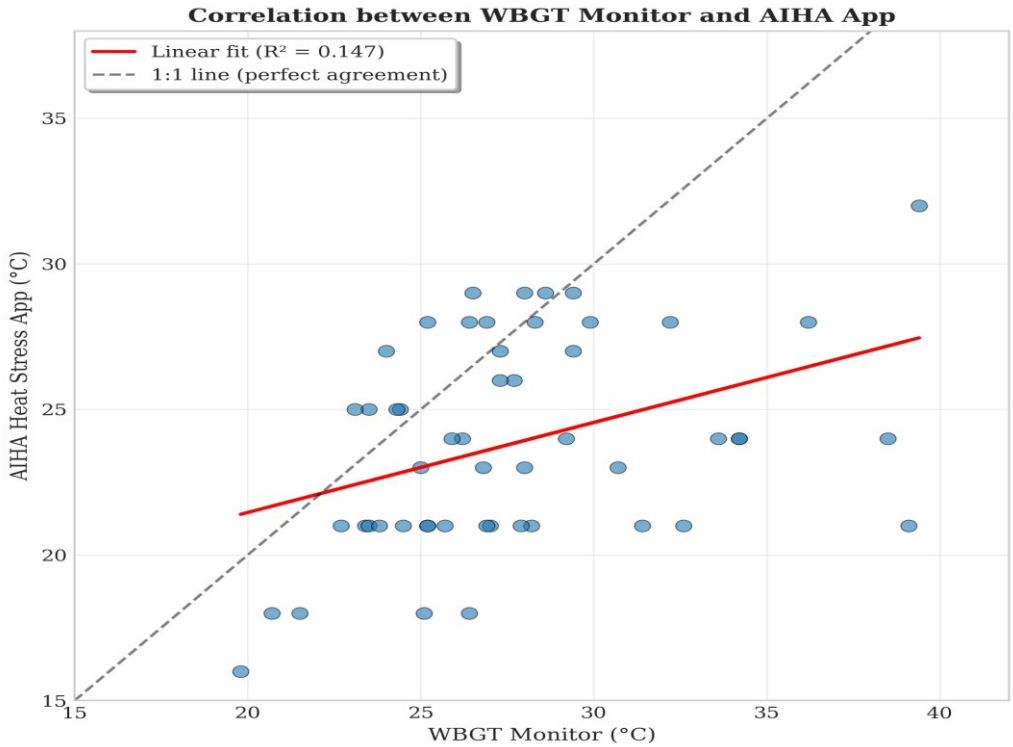

Figure 2: Scatterplot of the results of the correlation of the WBGT monitor with the AIHA application (r = 0.384, p = 0.005). The dashed line would be the ideal 1:1 agreement; the solid regression line would show the systematic under-estimation of the app.

The positive correlation indicates that increases in instrument-measured WBGT are generally associated with increases in app estimates, suggesting that the app can track directional changes in thermal conditions.

This directional consistency implies that the app is also capable of telling in a reliable manner whether the thermal conditions are getting better or worse as time goes on, which can justify its use in trend monitoring and also indicate when heat stress might be becoming apparent. The employees and managers operating the app might reasonably assume that increased app values reflect worsening temperatures and demand greater vigilance and possible countermeasures. However, the relatively low correlation (r = 0.384), explaining only about 15% of the variance, indicates substantial scatter and limited predictive accuracy.

The regression line (Fig. 2) deviates from the 1:1 line of perfect agreement, confirming systematic underestimation and the presence of both bias and random error.

This deviation continues throughout the measurement range and there are no discernable inflection points that would indicate a change of regime behavior. This is because the slope of the regression relationship is not equal to one, which mathematically shows systematic underestimation, and the large scatter around the regression line is evidence of significant random measurement error being superimposed upon the systematic bias

Operationally, this level of correlation suggests that the app is unsuitable for accurate quantification of WBGT or classification into risk categories, but may still be useful for tracking relative changes over time. Two app readings within a range of 2-3°C could be instrumented WBGT difference values ranging between 0-10°C between coincident environmental influences and random variations in measurements

3.3 Bland-Altman Agreement Analysis

Bland–Altman analysis provides a more direct assessment of agreement by quantifying systematic bias and limits of agreement between the two methods (Fig. 3).. The mean difference (WBGT monitor – AIHA app) was 3.86°C, with a standard deviation of 4.54°C. The 95% limits of agreement ranged from −5.03°C to 12.75°C.

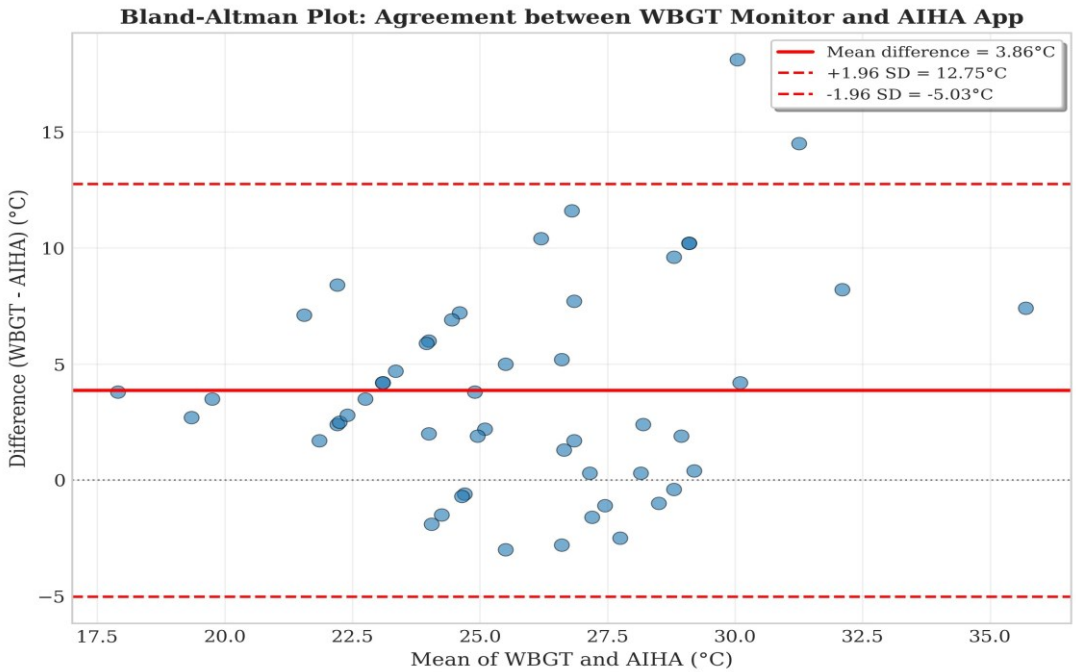

Fig. 3: Bland-Altman plot of the agreement between WBGT monitor and AIHA app. The mean bias is 3.86°C (solid red line) with the lower limit of agreement of -5.03°C (dashed red lines) and the upper limit of agreement of 12.75°C (dashed red lines).

This positive mean bias confirms systematic underestimation by the app, with values averaging nearly 4°C lower than the reference measurements.

Given that WBGT thresholds for work-rest decisions typically differ by only 2–3°C, this magnitude of error is operationally significant and may lead to incorrect risk classification. In reference, the typical increment in WBGT used in the ISO 7243 and AIHA guidelines to change between work-rest regimens is 2-3°C to move up or down the scale, and a 4°C error would likely shift a range of risk bands in exposure classification.

This is reflected by the broad range of the 95 percent range of agreement, which is about 18°C as the true values can vary by significant values around the mean bias as a result of random error. Practically, an WBGT monitor reading of 30°C could be estimated by AIHA app as between about 18°C and 35°C depending on the variation seen, though the values near either end of this range are less likely. Such uncertainty compromises trust on app readings and use of the readings to make consequential operational decisions like work stoppage or emergency interventions.

As the analysis of Bland-Altman plot (Figure 3) shows, the differences in measurements are not observed to be totally arbitrary over the range of temperatures. It is hinted at more underestimation at high mean temperatures, with the points in the right side of the plot clustering about the right side of the difference being larger. The pattern suggests possible heteroscedasticity and proportional bias, with greater underestimation at higher temperature ranges.

3.4 Measurement Distributions and Extremes

The distributions of WBGT monitor and AIHA app measurements are compared using box plots in Fig. 4. The WBGT monitor shows a wider distribution and higher upper values, while the app exhibits a compressed distribution with a lower median and truncated upper range. There is a more compressed distribution of the app, a lower median, and the upper tail is lower, indicating that the app fails to capture the peak thermal conditions.

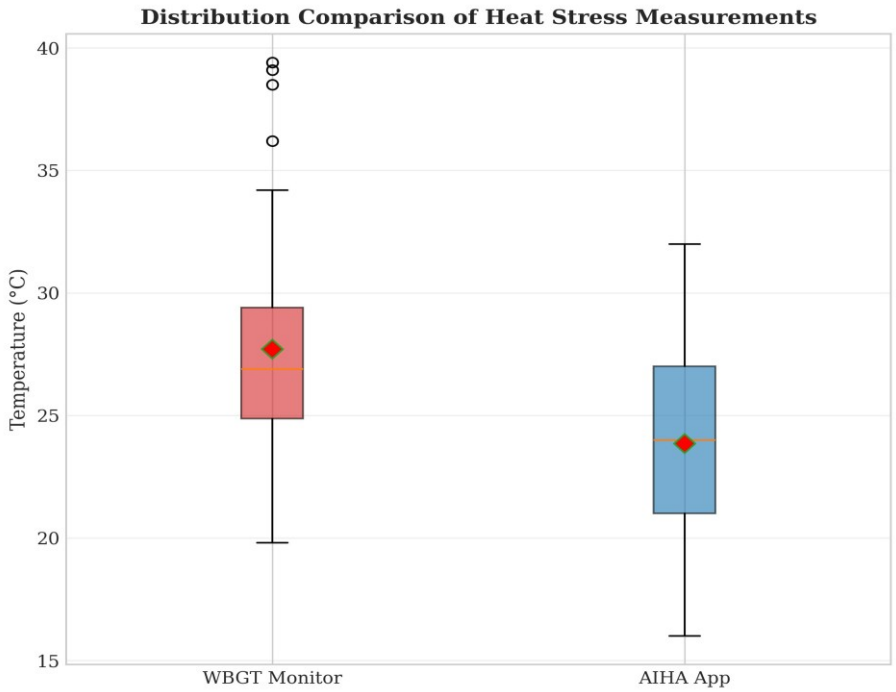

Figure 4: Comparison of box plot with distribution of measure of WBGT monitor and AIHA app. The means are shown by red diamonds, the medians by the horizontal lines, the boxes reflect the interquartile ranges, and the extremes by the whiskers within the range of 1.5 IQR.

The performance of the app in the harshest thermal conditions that were recorded in this study is of particular interest as far as risk management is concerned. The highest WBGT value recorded was 39.4°C, while the app reported only 32°C under the same conditions, resulting in a substantial underestimation of 7.4°C. This reading is in the extreme category of hazard, where even the shortest exposures can trigger a heat stroke in those who are prone to it. During the same period and place, the AIHA application recorded the temperature standing at 32°C, which is below the actual temperature of 7.4° C. Although 32°C is still a high degree of heat stress that precautions are necessary, the magnitude of the underestimation indicated that the app was not able to give the actual degree of the harshness of thermal hazard faced. Such a maximum underestimation coincidence with maximum instrumented WBGT hints to the possibility of the algorithmic models of the app saturating or using damping factors to compress extreme predictions, possibly to prevent false alarms or due to model uncertainty in the tails of the distribution. This suggests that the app may fail to adequately represent extreme thermal stress conditions, which are critical for risk management.

3.5 Distributional Analysis of Measurement Discrepancies

The histogram (Fig. 5) of measurement differences (WBGT − AIHA) shows a right-skewed distribution, indicating that the app more frequently underestimates WBGT. Fig. 5 of the histogram of the changes in measurements indicates that there was a skew to the right, which validates that there were more cases where WBGT was higher than AIHA values. The difference in modal is within the range of 2-4°C which is in agreement with the calculated mean bias. This asymmetry suggests that the app is not conservatively biased and may underestimate thermal risk in most cases.

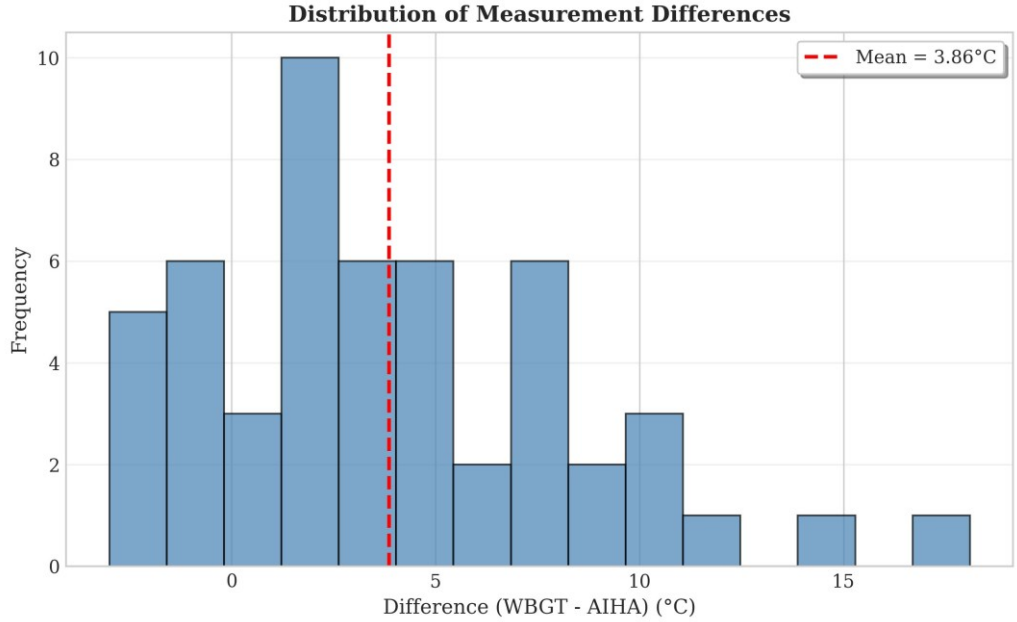

Fig. 5: Histogram of distribution of difference in measurements (WBGT-AIHA). The positive values represent underestimation of the app. Mean difference 3.86°C indicated by red dashed line.

About 90 percent of the observed differences were within the range of 0°C to 10°C, that is, the app displayed lower readings nearly always in comparison with the instrument, but the difference was quite variable. There were only two tests with AIHA readings above WBGT by more than 2°C and even the exceptions were only small deviations of 3-4°C. This effect of asymmetry of the distribution of the errors has an implication on the perception of risk and in decision-making. It could be expected by users that the readings

of the apps would be conservative (i.e., over-cautious) estimates of likelihood of excessive precaution, when in the actuality the opposite occurs: the app consistently underestimates the thermal hazards.

3.6 Environmental and Temporal Factors Influencing Discrepancies

Although not all microclimatic variables were directly measured, qualitative observations provided insight into environmental factors influencing discrepancies.

Also, it was not possible to instrument all the microclimatic variables of interest in this field comparison, qualitative evaluation of the environmental conditions during measurement sessions made some insights on factors that were related to higher app-instrument differences. The highest underestimations were always achieved on clear days with minimal wind and during the times of the day when the sun was at its highest point (11:00-15:00). For example, on 13 September 2024 at 16:07, WBGT was 36.2°C while the app reported 28.0°C, a difference of 8.2°C. Conversely, sessions in the morning (08:00-10:00) and the late afternoons (after 16:00) when the solar elevation was lower and the radiant heat load was smaller were usually found to have smaller differences, usually within the 1-3°C range. The core physics behind this trend is that WBGT temperature component of the globe reacts directly to the amount of radiant heat collected by the solar and reflected sources, but generally app algorithms indirectly compute radiant effects based on the solar elevation, cloud cover, and presumptive surface reflectance that is retrieved by meteorological models (Liljegren et al., 2008 *et al*., 2008).

Surface properties also emerged to have a bearing but were not directly tested. Measurement sessions that took place on concrete or asphalt areas which reflect the absorbed energy of the sun, at times, gave greater WBGT-AIHA gaps compared to those measured on grass or vegetated areas. This observation implies that app algorithms can make use of generalized surface assumptions which might not capture site-dependent radiant exchanges especially of high-albedo or high-emissivity materials frequently found in workplaces (building materials, vehicles, machinery).

When sufficiently calm winds (subjectively estimated below 1 m/s) were observed, the difference in measurements was sometimes greater, which is consistent with the idea that the natural wet bulb element of a reference WBGT monitor is optimally sensitive to the presence of moisture in the air when wind speeds are low and app models might assume a standardized wind speed in the absence of actual data, or with unreliable weather station networks (Golbabaei *et al*., 2021).

3.7 Comparison with Prior Literature

The results of this paper are in line with and expand the findings of other recent studies that assessed the accuracy of mobile heat stress applications versus reference instrumentation. Angol, (2024) compared smartphone application WBGT estimates with on-site WBGT instruments in 26 high school athletic facilities in 11 states of the United States and found a correlation of r=0.89 but systematic underestimation in WBGT (especially at temperatures above 90°F (32.2°C) WBGT. Although their observed correlation was stronger than the r = 0.384 of the present study, presumably due to the fact that their measurement settings were more homogeneous and single-app version, the directional bias and magnitude-related error patterns are similar to those of our results.

Giraldo, (2024) reported the mean deviations of 2.4°C in estimated app wet bulb temperature in tropical field cases, which is similar to the deviation of 3.86°C in this case because WBGT combines other temperature variables other than a wet bulb. Their investigation also found solar radiation as a major cause of app-instrument variances, which points to the mechanistic explanation that algorithms are difficult to tune to the behavior of radiant heat exchanges in complex exterior conditions.

Dillane & Balanay (2020) tested several applications of heat stress, such as previous versions of AIHA and found correlation coefficients of 0.70 to 0.85 based on the apps

and the environment. They observed depressed ranges in app output as compared to instrumented measurement, which was also observed in the present study of the AIHA app 2.0 not registering a value above 32°C even though instrumented values were 39.4°C.

This growing body of literature is capable of producing a general trend: mobile applications can offer reasonably good monitoring of changes in temporal patterns and of relative change in heat stress, although there is systematic underestimation in absolute values of WBGT, with the magnitude of bias growing in high solar radiation, complicated microclimates, and extreme thermal conditions. Collectively, these studies indicate that while mobile applications can track relative trends in heat stress, they tend to systematically underestimate absolute WBGT values, particularly under high solar radiation and complex environmental conditions.

3.8 Operational Implications and Activity Modification Misclassification

The systematic underestimation that is found in the AIHA application has direct implications for occupational safety decisions based on categorical WBGT limits. As an example, the ISO 7243 work-rest recommendation of an acclimatized worker working in moderate intensity labor and using a normal working garment can be taken into consideration. The standard specifies WBGT thresholds of approximately 28°C for continuous work, 26°C for 75% work / 25% rest each hour, and 24°C for 50% work / 50% rest hourly cycles.

On the 52 measurement sessions used in this research, the reference WBGT gauge registered temperatures over 28°C on 31 occasions (60 percent of measurements), which suggests conditions under which work-rest cycle should be applied even to acclimatized moderate-intensity workers. The AIHA app, however, recorded values lower than 28°C 45 times (87% of measurements), with 14 of these values below 28°C yet the app still read lower than 28°C. This is a miscarriage rate of 45 percent (14/31) at the critical temperature of 28°C in misclassification in which the failure to apply rest breaks might result in cumulative heat strain and high Illness risk.

This was exceeded by the reference monitor in 8 instances at the more stringent 32°C threshold, which represents the extreme hazard conditions where even the acclimated workers working on light labor would need considerable work restrictions. The AIHA app only predicted the extreme conditions of high hazards once, registering a maximum of precisely 32°C, and no longer alerted of the high hazard conditions, in the other seven cases. This 88 percent misclassification (7/8) at the most hazardous threshold shows that the app is most inaccurate at the extreme locations of the most hazardous thermal situations where action is most vital.

These inaccurate rates of misclassification are translated into real-life safety implications. Employees or managers who only use the app during the most risky measurement period (WBGT 39.4°C, app indicating 32.0°C) may use work-rest-regimens that suit 32°C when in fact the conditions warrant virtually complete work termination or engineering interventions. The difference between 32°C and 39°C WBGT shows significant physiological strain difference: The rates of rise of core temperature, cardiovascular demand, and probability of heat illness increase significantly within this range of 7°C (Parsons *et al*., 2022).

3.9 Limitations and Sources of Uncertainty

These findings have many methodological limitations that should be recognized when interpreting these findings. To begin with, one reference WBGT device and one mobile phone with the AIHA application were used in the study. Although the reference monitor was calibrated and within ISO standards, instrument-to-instrument differences between nominally similar WBGT devices could be as large as 1-2°C in the field, which could be a contributing factor to observed differences. Equally, differences in smartphone models, iOS version disparities, or weather data quality in a location could have other impacts on the functionality of the apps that were not evaluated in this case.

Second, the nature of perfection concurrent measurements means that there is uncertainty about time. Whereas there was an attempt to capture WBGT monitor and app measurements within a minute of one another, the changing environments during passing clouds, wind gusts or sun angles swings would create slight temporal discrepancies. The 20 minutes WBGT monitor stabilization criterion also implies that the instrument reading shown is an average of the 20 minutes of equilibration, as opposed to the app to calculate a real-time estimate based on the observed meteorological parameters.

Third, the geographic and temporal coverage of the study though representing a variety of diurnal and day-to-day variation is limited in one location in the late-summer-early-autumn period of the climate zone of North American mid-continent. Their generalization to new climates (tropical humid, arid desert, high altitude), seasons (winter cold stress, summer extreme heat), and geographic settings (coastal marine, urban heat islands, industrial hot processes) should be further validated. The range of WBGT that was covered (19.8-39.4°C) is typical in most work settings but does not go to extreme conditions in deserts or the harshest furnaces in industries where WBGT may be greater than 45 °C.

Fourth, the values of the user input parameters used in the AIHA app were kept the same in each of the sessions (moderate work load, summer working clothes) to remove the effects of input selection on the environmental measurements. When these inputs are varied to achieve the actual workload and clothing in the operation, this would be anticipated in the operational deployment, which may have different outputs in the app depending on the path taken by the algorithms. The current results hence describe the performance of apps in controlled input parameters as opposed to real field application where inputs may be mis-specified or have uneven application.

Lastly, the small size of the sample used (n = 52 paired measurements) has sufficient statistical ability to find the large systematic biases it found, but might not be able to fully describe the nonlinear relationships or detect rare outlier situations. The addition of multi-site, multi-season campaigns would enhance the view on the validity of the generalizability of results.

3.10 Synthesis and Contextualization

Irrespective of such limitations, the convergent evidence is in the form of the correlation analysis, the agreement test of Bland-Altman, distributional comparisons, and examination of the temporal patterns, all indicate a consistent finding that does show the AIHA Heat Stress App 2.0 systematically underestimates environmental heat stress, as measured by reference WBGT instrumentation, with the magnitude of such underestimation averaging about 4°C and growing with high solar radiance and extreme thermal conditions. The practical implication is that the app, when used as a separate decision tool, will be prone to categorize thermal environments as being less dangerous than they really are, and with similar risks of insufficient protection measures.

Such a systematic underestimation is opposite of the possible error direction that a screening tool would prefer to make in the first step. Intuitively, it could be claimed that conservative (over-estimating) apps that are prone to be overly cautious would be safer, although inefficient economically by imposing unwarranted work restrictions. Nevertheless, the falsely positive effect of apps gives rise to an alternative profile of risk: through their use, a sense of false assurance and a timely postponed response when the situation is dangerous is created. Employees who are used to taking up what the app tells them to do might become complacent when that telling always points to less risk of danger than the reality, and thus negates the purpose of having heat stress warning.

The moderate correlation ($r = 0.384$) that is observed implies that there is certain environmental variability in the app that is captured and that lends credence to its use in the determination of qualitative trends. It was reasonable that users would use the app to track the state of conditions at a work shift getting better or worse, the relative thermal loads at work at different work locations or

times of the day, or to alert users about when heat stress was significant enough to warrant increased attention. These applications do not require the absolute measurement accuracy but rather take advantage of the directional accuracy of the app.

Nonetheless, the app is not to be entrusted with making the quantitative decisions related to the special WBGT levels, including regulatory compliance checks, the initiation of mandatory work-rest cycles according to ISO or OSHA standards, and the identification of situations when the conditions need emergency heat illness procedures. In these high-stake applications, systematic bias and broad limits of agreement reported here present intolerable uncertainty and the risk of misclassification

4.0 Conclusion

This field-based comparative study demonstrates that the AIHA Heat Stress App 2.0 exhibits systematic measurement bias and limited absolute accuracy when compared with calibrated WBGT reference instrumentation across a range of outdoor thermal conditions. On average, the app underestimated environmental heat stress by 3.86°C, with wide variability reflected in the 95% limits of agreement (−5.03°C to 12.75°C). Although a statistically significant correlation ($r = 0.384$) indicates that the app captures general temporal trends in thermal conditions, the relatively weak association and substantial scatter limit its reliability for precise quantitative assessment or threshold-based occupational health decision-making.

From a practical perspective, these findings highlight the importance of aligning tool selection with intended use. Mobile heat stress applications offer clear advantages in accessibility, ease of use, and broad spatial and temporal coverage. They can support awareness, worker education, and preliminary screening across dispersed worksites. However, the observed systematic underestimation and reduced sensitivity to extreme thermal conditions limit their suitability as standalone tools for regulatory compliance, high-risk exposure assessment, or critical activity modification decisions. The high misclassification rates observed at key WBGT thresholds (45% at 28°C and 88% at 32°C) underscore the potential for under-protection of workers if such applications are used in isolation.

An evidence-based implementation strategy should therefore adopt a hybrid approach. Mobile applications can be effectively used for continuous monitoring and early identification of potentially hazardous conditions, while validated WBGT instrumentation should be employed to confirm exposure levels, guide work–rest decisions, and ensure regulatory compliance during high-risk periods. This complementary framework balances accessibility and cost with the accuracy required for safeguarding worker health.

Future advancements should focus on improving the accuracy of app-based estimates through enhanced modeling of microclimatic variability, integration of low-cost environmental sensors particularly for radiant heat and incorporation of uncertainty indicators to inform users of potential estimation errors. Until such improvements are realized and independently validated, occupational health practitioners and regulatory agencies should avoid relying solely on mobile application outputs for critical safety decisions and instead prioritize reference-grade measurements where accuracy is essential.

5.0 References

American Industrial Hygiene Association (2024). *AIHA Heat Stress App 2.0 User Guide*. AIHA, Falls Church, VA. https://www.aiha.org/resources/ heat-stress-app

Alahmad, B., Khraishah, H., Royé, D., Vicedo-Cabrera, A.M., Guo, Y., Papatheodorou, S.I., Achilleos, S., Acquaotta, F., Armstrong, B., Bell, M.L. & Others (2023). Associations between extreme temperatures and cardiovascular cause-specific mortality: Results from 27 countries. *Circulation*, 147(1), 35-46. https://doi.org/10.1161/ CIRCULATIONAHA.122.061832

Angol, B. (2024). Comparison between Wet Bulb Globe Temperature (WBGT) App Prototype and WBGT Monitor to Assess Heat Stress Risk in Groundskeeping in an Eastern North Carolina University Setting.

Bernard, T.E., Caravello, V., Schwartz, S.W., & Ashley, C.D. (2023). WBGT clothing adjustments for four clothing ensembles and the effects of metabolic demands. *Journal of Occupational and Environmental Hygiene*, 20(1-2), 82-89. https://doi.org/10.1080/15459624.2022.2140952

Bland, J.M., & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476), 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8

Brimicombe, C., Lo, C.H., Pappenberger, F., Macleod, D., Cornforth, R., & Cloke, H.L. (2023). Wet bulb globe temperature: Indicating extreme heat risk on a global grid. *GeoHealth*, 7(2), e2022GH000701. https://doi.org/10.1029/2022GH000701

Dillane, D., & Balanay, J.A.G. (2020). Heat stress evaluation of outdoor workers: A case study in the southern United States using different app-based WBGT meters. *Safety and Health at Work*, 11(3), 268-273. https://doi.org/10.1016/j.shaw.2020.04.003

Eggeling, M., Fischer, J.E., & Debus, F. (2023). Digital solutions for occupational heat stress: A scoping review and research agenda. *International Archives of Occupational and Environmental Health*, 96(8), 1067-1084. https://doi.org/10.1007/s00420-023-01989-x

Giraldo, A. (2024). Interrelationships Among Local Values of Wet Bulb Globe Temperature, Heat Index, and Adjusted Temperature.

Golbabaei, F., Heidari, H., Shamsipour, A., Forushani, A.R., & Gaeini, A. (2021). The role of environmental parameters and anthropometric characteristics in evaluation of heat stress using PHS model. *Journal of Environmental Health Science and Engineering*, 19(1), 1067-1077. https://doi.org/10.1007/s40201-021-00672-4

International Organization for Standardization (2017). *ISO 7243:2017 Ergonomics of the thermal environment - Assessment of heat stress using the WBGT (wet bulb globe temperature) index*. ISO, Geneva, Switzerland.

Kjellstrom, T., Briggs, D., Freyberg, C., Lemke, B., Otto, M., & Hyatt, O. (2016). Heat, human performance and occupational health: A key issue for the assessment of global climate change impacts. *Annual Review of Public Health*, 37(1), 97-112. https://doi.org/10.1146/annurev-publhealth-032315-021740

Kong, Q., & Huber, M. (2024). A new, zero-iteration analytic implementation of wet-bulb globe temperature: Development, validation, and comparison with other methods. *GeoHealth*, 8(10), e2024GH001068. https://doi.org/10.1029/2024GH001068

Liljegren, J.C., Carhart, R.A., Lawday, P., Tschopp, S., & Sharp, R. (2008). Modeling the wet bulb globe temperature using standard meteorological measurements. *Journal of Occupational and Environmental Hygiene*, 5(10), 645-655. https://doi.org/10.1080/15459620802310770

Parsons, L.A., Masuda, Y.J., Kroeger, T., Shindell, D., Wolff, N.H., & Spector, J.T. (2022). Global labour loss due to humid heat exposure is underestimated for outdoor workers. *Environmental Research Letters*, 17(1), 014050. https://doi.org/10.1088/ 1748-9326/ac3dae

Périard, J.D., Racinais, S., & Sawka, M.N. (2021). Adaptations and mechanisms of human heat acclimation: Applications for competitive athletes and sports. *Scandinavian Journal of Medicine & Science in Sports*, 25(S1), 20-38. https://doi.org/10.1111/ sms.12408

Peymaneh, H., Jaleh, R., Amirhossein, M., Farank, M., Saeed, F., & Ahad, H. (2024).

Climate change and heat stress resilient outdoor workers: Findings from systematic literature review. *BMC Public Health*, 24(1), 1711. https://doi.org/10.1186/ s12889-024-19212-3

Declaration

Consent for publication

Not Applicable

Availability of data

Data shall be made available upon request

Ethical Considerations

Not applicable

Competing interest

The authors report no conflict or competing interests

Funding

The authors declared no source of funding

Authors' Contributions

All components of the work were carried out by the author