Published in final edited form as:
Int J Hyg Environ Health. 2020 June 01; 227: 113521. doi:10.1016/j.ijheh.2020.113521.

Development and reliability of a quantitative personal hygiene assessment tool

Maryann G. Delea[a,*], **Jedidiah S. Snyder**[a], **Mulat Woreta**[b], **Kassahun Zewudie**[b], **Anthony W. Solomon**[c], **Matthew C. Freeman**[a]

[a]Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, USA [b]Emory Ethiopia, Bahir Dar, Ethiopia [c]Department of Control of Neglected Tropical Diseases, World Health Organization, Geneva, Switzerland

Abstract

Personal hygiene practices, including facewashing and handwashing, reduce transmission of pathogens, but are difficult to measure. Using color theory principles, we developed and tested a novel metric that generates quantitative measures of facial and hand cleanliness, proxy indicators of personal hygiene practices. In this cross-sectional study, conventional qualitative cleanliness metrics (e.g., presence or absence of nasal and ocular discharge, dirt under nails or on finger pads and palms) were also recorded. We generated Gwet's agreement coefficients to determine the inter-rater reliability of novel and conventional metrics between various rating groups, where appropriate, including two non-blinded raters, non-blinded vs. blinded raters, three blinded raters, and blinded vs. computer raters. Inter-rater reliability of the novel metric was high across all rating groups, ranging from 0.98 (95% CI: 0.97, 0.99) to 0.90 (95% CI: 0.90, 0.91) for facial cleanliness, and 0.97 (95% CI: 0.96, 0.98) to 0.92 (95% CI: 0.91, 0.93) for hand cleanliness. Our novel metric generates more nuanced data than conventional qualitative metrics, and allows for quantifiable assessments of facial and hand cleanliness.

Keywords

Personal hygiene; Hygiene; Water, sanitation, and hygiene; WASH; Behavioral outcomes

[*]Corresponding author. Emory University, Rollins School of Public Health, Gangarosa Department of Environmental Health, Claudia Nance Rollins Building, 1518 Clifton Road, Atlanta, GA, 30322, USA. mdelea@emory.edu (M.G. Delea).

Author contributions

AWS conceived of the idea of using a color scale to rate cleanliness; MGD conceived of employing color theory to create an appropriate color scale; JSS conceived of employing densitometry to obtain computer-simulated cleanliness ratings. MCF and MGD obtained funding for the study. MGD designed and managed the study. MCF was responsible for study administration. MW was involved in the development and testing of qPHAT methodology, and KZ was involved in training, supervision, and field management of the study. JSS served as the master rater and carried out the densitometric analyses. MW and KZ served as the two additional blinded raters. MGD cleaned, curated, and analyzed the data, and wrote the first draft of the manuscript. All authors contributed to the development of the manuscript, and reviewed and approved of the final version.

Declaration of competing interest

The authors declare no competing financial interests. The authors alone are responsible for the views expressed in this article and they do not necessarily represent the views, decisions or policies of the institutions with which they are affiliated.

1 Introduction

Inadequate personal hygiene contributes to the global burden of disease (Prüss-Ustün et al., 2014). Improved facial and hand cleanliness might prevent ocular diseases such as trachoma, respiratory infections such as pneumonia and influenza, and enteric infections such as soil-transmitted helminthiases, shigellosis, and cryptosporidiosis (Aiello and Larson, 2002; Esrey et al., 1985; Freeman et al., 2014; Rabie and Curtis, 2006; Strunz et al., 2014; Utsi et al., 2016; West et al., 1995). Personal hygiene practices are believed to represent intermediate behavioral factors along the causal pathways that lead to these diseases. Consequently, many public health programs promote the adoption of improved personal hygiene practices for disease prevention and control. In low and middle-income settings, interventions promoting personal hygiene are often undertaken within community-based water, sanitation, and hygiene (WASH) and neglected tropical disease (NTD) programming (Boisson et al., 2016). The intervention techniques used in many of those efforts, however, may not be sufficient to produce sustainable change (Delea et al., 2018); more evidence is needed to clarify which intervention approaches bring about sustained personal hygiene improvements (Dodson et al., 2018; Ejere et al., 2015). Valid, reliable measurement of personal hygiene practices would be useful for monitoring, evaluation, and development of better interventions.

An array of hygiene metrics – which seek to reflect personal hygiene behaviors – are available. Some metrics aim to measure actual behavior through respondent reports (Edwards et al., 2008; Manun’Ebo et al., 1997), direct observation (Ram et al., 2010), or video surveillance (Pickering et al., 2014). Other metrics are proxy indicators, including observations of environmental conditions (Biran et al., 2008; Luby and Halder, 2008); the cleanliness of an individual’s hands, face, or body (Halder et al., 2010; West et al., 2017); the presence or absence of pathogens or indicator organisms on relevant body parts (Burr et al., 2013; Parvez et al., 2019; Ram et al., 2011); and sensor-recorded measurement of materials required to execute hygiene practices (Ram et al., 2010).

Several of these methods are limited with regard to their validity, granularity, and cost (Biran et al., 2008; King et al., 2011). Respondent reports are subjective and prone to systematic biases (e.g., courtesy, recall, and social desirability biases) (Curtis et al., 1993; Manun’Ebo et al., 1997); measures captured through direct observation and video surveillance are often prone to reactivity (Pickering et al., 2014; Ram et al., 2010); and proxy indicators may not be valid measures of recent cleansing (Biran et al., 2008; Curtis et al., 1993; King et al., 2011; Ram et al., 2011). In addition, several of these methods yield dichotomous outcomes (e.g., presence/absence of: soap, washing station, water for washing [environmental conditions]; nasal and ocular discharge [facial cleanliness], dirt under nails or on finger pads and palms [hand cleanliness]) (Halder et al., 2010; West et al., 2017). Dichotomous data (or coarse polychotomous data (Pickering et al., 2010)) may not be nuanced enough to detect incremental changes in personal hygiene practices. Conventional qualitative metrics also impede the examination of dose-response relationships. Therefore, these conventional metrics may limit the type of evidence available for monitoring behavior and evaluating the effectiveness of personal hygiene interventions on behavioral outcomes and downstream health and well-being.

To enable better assessment of personal hygiene practices and enhance evaluations of personal hygiene interventions, we sought to create a metric that could generate quantitative data on facial and hand cleanliness. The aim of this study was to develop and pilot this novel quantitative metric, assess its reliability, and compare its reliability estimates and measurement attributes to those of conventional qualitative facial and hand cleanliness metrics.

2 Material and methods

We embedded this cross-sectional study of the inter-rater reliability (IRR) of our novel, quantitative hygiene metric within a cluster-randomized trial called the Andilaye Trial. This trial aimed to evaluate the impact of a demand-side sanitation and hygiene intervention on behavior and health in NTD-endemic Amhara, Ethiopia (registered as NCT03075436 on clinicaltrials.gov). Details of the Andilaye Trial are published elsewhere (Delea et al., 2019). In summary, we randomly selected and assigned 50 sub-district (kebele) clusters within three purposively selected districts (woredas) to receive either an enhanced community-based demand-side sanitation and hygiene intervention (i.e., Andilaye intervention) or the standard of care intervention (i.e., community-led total sanitation and hygiene). The trial's outcomes included NTD-preventive sanitation and hygiene behaviors and mental well-being. As such, it was important for us to identify and employ a personal hygiene metric that did not rely on reported behavior and would be sensitive enough to detect incremental changes in personal hygiene practices over time, particularly given that the Andilaye intervention promoted incremental change.

2.1 Development of the quantitative personal hygiene assessment tool (qPHAT)

During the Andilaye Trial's formative phase, we used color theory principles to create a color scale to assess personal hygiene via standardized cleanliness assessments. This scale depicted a spectral light (i.e., red/green/blue - RGB) color model array, ranging from 10 (i.e., all hues full strength saturation; white) to 0 (i.e., single hue saturation; darkest brown). Each step (i.e., color and accompanying number) along the array represented a 10% increase in saturation compared to the preceding step. This yielded the 11-point qPHAT color scale (Fig. 1A).

We pre-tested the qPHAT color scale in Andilaye formative research communities by collecting wipes of children's faces and hands, and comparing these wipes to the color scale to determine whether it provided an appropriate array against which the wipes could be compared. Once we finalized the color scale, we used the qPHAT methodology to obtain quantitative proxy measures of personal hygiene practices via facial and hand cleanliness assessments, as indicated below.

2.2 Data collection

For the Andilaye Trial, we randomly selected 50 clusters from the list of eligible kebeles in Farta, Fogera, and Bahir Dar Zuria Woredas of South Gondar and West Gojjam Zones in Amhara. All rural and peri-urban kebeles within Bahir Dar Zuria, Fogera, and Farta woredas that were accessible throughout the course of the year were eligible for random selection.

During March-April 2017, the end of the local dry season and beginning of the small rains, trained enumerators enrolled approximately 30 households that were randomly selected from each study cluster's household census register. At each of these households, an enumerator collected data on household characteristics and sanitation and hygiene practices from one adult member of the household, amongst other data relevant to the Andilaye Trial (Delea et al., 2019), and observed and recorded the cleanliness of this individual's hands using conventional qualitative hand cleanliness metrics. The youngest 1–9-year-old living in the household was identified as the index child, and the enumerator observed and recorded the cleanliness of this child's hands and face, first using conventional qualitative facial and hand cleanliness metrics, and then using the qPHAT methodology. The enumerator also conducted spot checks of the household's compound, latrine, and washing stations. All data were collected electronically on encrypted, password-protected mobile phones using Open Data Kit (http://opendatakit.org/), and uploaded to and stored on a secure server.

2.3 Conventional qualitative facial and hand cleanliness metrics

Conventional facial and hand metrics reflected those commonly captured as measures of facial and hand cleanliness by the WASH and NTD communities (Halder et al., 2010; King et al., 2011; Parvez et al., 2019; West et al., 2017). Specifically, enumerators captured data on the presence or absence of ocular discharge; wet and dry nasal discharge; dirt, dust, or other debris on the face; and the number of fly-face contacts during a 1-min observation period. Conventional qualitative facial cleanliness metrics represented epidemiological associations of active trachoma and ocular Chlamydia trachomatis infection (Ngondi et al., 2008; West et al., 1991) and other signs of poor facial hygiene. Enumerators assessed hand cleanliness by capturing data on the presence or absence of dirt, mud, or debris under or on the finger nails, finger pads, and palms of each hand (assessed separately, given hand dominance and norms regarding the use of different hands for certain activities, such as eating and cleaning). Evidence is mixed regarding whether and to what extent these individual conventional hand cleanliness metrics represent epidemiological risk factors for hand contamination with enteric organisms (Morrill et al., 2018; Parvez et al., 2019; Pickering et al., 2010) or health outcomes such as gastrointestinal illness or respiratory symptoms (Pickering et al., 2010). Field supervisors also collected observational data on a validation sample that reflected approximately 10% of households from each study cluster. Field supervisors entered their own assessments, independently and without knowledge of the enumerator's assessments, on their own data collection device at the same time the enumerator was entering their assessments. Capturing data on conventional metrics allowed us to examine adjusted inter-rater reliability estimates and measurement attributes between these metrics and the qPHAT methodology.

2.4 qPHAT methodology

To obtain quantitative cleanliness data via the qPHAT methodology, enumerators used gauze pads pre-moistened with sterile saline (Hygea), with excess solution removed, to collect one wipe from the skin around the eyes and one wipe from the skin of the inside of the index child's right hand. Enumerators employed standardized procedures to trace the skin along the index child's eyes and hand. After taking each wipe, the enumerator sealed it in its own labelled plastic bag. These wipes were then scored against the qPHAT color scale by a rater who was not blinded to the conditions of the child or the child's household (i.e., the enumerator) and one master rater who was blinded to both conditions. A sub-set of the wipes was rated by two additional blinded raters; this sub-set also underwent densitometric analyses to produce computer-simulated ratings. Engaging multiple raters, of different types, allowed us to examine whether the qPHAT measures were reliable, and whether inter-rater reliability estimates differed meaningfully across different types of raters (Fig. 2).

2.5 Non-blinded rater assessment

Within each household, face and hand wipes were obtained from the index child shortly after data collection for that household commenced. At the end of data collection for that household, the enumerator was prompted by the data collection program to rate the wipes by matching the color of the darkest point within the darkest square half-inch of the gauze pads (i.e., roughly the size of a fingernail) to a color along the qPHAT color scale (Fig. 1B); the time between collection and rating averaged approximately 1 h. All rating was conducted under natural lighting outside of study participants' homes. We did not obtain any quantitative data on adult hand cleanliness via qPHAT measures for this study.

2.6 Blinded rater assessment

Enumerators collected data on the child's facial and hand cleanliness, and assessed other sanitation and hygiene-related data from the household itself. They were consequently not blinded to these conditions, which has the potential to introduce bias. Therefore, one master rater blinded to the conditions of the study households and children rated all wipes using the qPHAT methodology. These blinded ratings of the wipes typically occurred within one to five days of wipe collection, and were performed outside, under similar lighting conditions. The purpose of the blinded rater assessment was to ascertain the IRR between non-blinded and blinded raters using the qPHAT methodology.

2.7 Blinded rater vs. computer-simulated rater assessments

An additional two raters, also blinded to the conditions of study households and children, used the qPHAT methodology to rate a sub-set of wipes (n = 87 face wipes, n = 87 hand wipes; all wipes obtained from the three study clusters for which data collection was scheduled the day before the three blinded rater assessments were scheduled). For this subset of wipes, we captured a high-resolution photograph of each wipe under natural lighting, without flash (to mimic rating conditions using the unaided human eye) alongside a negative control (i.e., unused wipe) that was used for calibration. We then employed densitometry, as indicated below, to generate computer-simulated ratings of this sub-set of face and hand wipes.

2.8 Analytical methods

We produced descriptive statistics related to all metrics. For conventional facial and hand cleanliness metrics, we generated data on the prevalence of each individual sign of facial/ hand cleanliness. We also generated data on the prevalence of two measures of clean face (i.e., absence of ocular and nasal discharge [often used in trachoma research] and absence of all signs of an unclean face [composite of all observed signs]) and one measure of clean hand (i.e., absence of all observed signs of an unclean hand). For the qPHAT methodology, we generated data related to the distribution of rating scores for each type of rater.

2.9 Chance-corrected reliability

We generated statistics on IRR, which indicate the consistency of ratings across two or more independent raters (Kozlowski and Hattrup, 1992), for both conventional facial and hand cleanliness metrics as well as the qPHAT methodology. Basic measures of agreement, such as comparisons of prevalence and non-corrected IRR estimates, do not correct for agreement due to chance, and can therefore yield misleading results when presented on their own (Gwet, 2014; Klein, 2018). Therefore, we produced chance-corrected IRR coefficients for both conventional and novel cleanliness metrics. Given critiques of kappa coefficients (Gwet, 2008; Klein, 2018), we employed Gwet’s agreement coefficient, which represents a chance-corrected coefficient that adjusts for the number of rating categories and the frequency with which rating categories are used by raters (Gwet, 2014; Klein, 2018) (Appendix A. Supplementary material).

We assessed partial agreement of qPHAT measures through the application of quadratic weights (Gwet, 2014; Klein, 2018; Lin et al., 2007). To minimize misinterpretations of reliability results that may occur when deterministic benchmarking is used (i.e., potential over-estimations of the magnitude of agreement), we employed a probabilistic benchmarking approach and the Landis and Koch scale to interpret agreement coefficient estimates (Gwet, 2014; Klein, 2018; Landis and Koch, 1977). We generated IRR estimates of qPHAT measures, indicated via Gwet’s agreement coefficients for the following rating pairs: 1) non-blinded raters (i.e., enumerators collecting data) and a master rater blinded to the conditions of households and children; 2) three raters, all blinded to the conditions of households and children; and 3) blinded raters and computer-simulation (i.e., densitometric analyses).

2.10 Densitometric analyses comparing human-generated and computer- simulated ratings

Computer-simulated ratings of face and hand wipes were generated through the application of densitometric analyses conducted on photographs of the sub-set of wipes rated by the three blinded raters. The purpose of these IRR assessments was to determine whether there were meaningful differences between blinded human-generated and objective computer-simulated ratings.

We performed the densitometric analyses on face and hand wipes by adapting a quantification protocol developed for ImageJ, an open access image processing software developed by the National Institutes of Health (NIH) (Davarinejad, 2017; Schneider et al., 2012). Our protocol specified scanning the high-resolution photograph of each wipe for the square half-inch frame with the highest optical density (i.e., intensity/ area). This was the same method used by the human raters (Fig. 1B). We used histogram values of the darkest fifth percentile of this area in our densitometric analyses to simulate the human rater protocol of scoring the darkest point within the darkest square half-inch of wipes (Fig. 1C). We conducted sensitivity analyses to determine whether mean histogram values produced meaningfully different results (Appendix A. Supplementary material, Table A1). We normalized densitometry values for lighting by deducting the background (i.e., an unused wipe) captured within the same photographic image as each wipe. We used this same process to capture densitometry values for each color in the qPHAT color scale. Computer-simulated rating scores were generated for each wipe by comparing normalized densitometry values of: 1) images of wipes, and 2) the 11 colors in the qPHAT scale.

2.11 Ethical approval

The Andilaye Trial and its sub-studies received ethical approval from Emory University's Institutional Review Board (IRB00076141) and the Amhara Regional Health Bureau Research Ethics Review Committee (HRTT0135909). Fieldworkers provided study participants details regarding the study prior to requesting consent to participate, and took steps to ensure confidentiality.

3 Results

3.1 Analytical sample and characteristics

We collected data from 1332 of 1333 index children (hand and facial cleanliness data generated via conventional qualitative metrics and qPHAT methodology) and 1332 of 1333 adults (hand cleanliness data generated via conventional qualitative metrics). Field supervisors completed independent quality control observations of facial and hand cleanliness using conventional qualitative cleanliness metrics on 124 (9%) of each of these index children and adults. The majority of adult respondents were either the mother of the index child (88%, n = 1168) or female caregivers (4%, n = 59; Table 1). Approximately half of index children were girls (49%, n = 658), and the average age of these children was 4 years (IQR: 2, 6).

3.2 Conventional qualitative facial cleanliness ratings

Prevalence of signs of a clean face differed by individual qualitative cleanliness metric (Table 2), and ranged from 63% for absence of ocular discharge to 21% and 18% for absence of flies on face amongst the 123 children independently observed by enumerators and field supervisors, respectively. IRR of facial cleanliness differed across qualitative facial cleanliness metrics (Table 2). These results suggest only moderate IRR for some individual qualitative metrics of facial cleanliness.

3.3 Conventional qualitative hand cleanliness ratings

Prevalence of signs of a clean hand also differed by individual qualitative hand cleanliness metric (Table 2). Amongst 124 children, qualitative signs of cleanliness measured by enumerators and supervisors, respectively, ranged from 17% to 19% for absence of dirt on the palm (right hand) to 7% and 6% for absence of dirt under all fingernails (left hand). Amongst 124 adults, prevalence of these signs of cleanliness ranged from 42% to 37% for absence of dirt on the palms (right hand and left hand, respectively) to 21% for absence of dirt under all fingernails (left hand). As indicated in Table 2, IRR of all conventional qualitative hand cleanliness metrics was lower for observations of adults than children, but the prevalence of hand cleanliness, as indicated by all qualitative cleanliness metrics, was higher amongst adults than children. Our data suggest that, when deployed amongst adults, IRR of some individual qualitative metrics of hand cleanliness is only moderate.

3.4 qPHAT facial cleanliness ratings

For the 1332 wipes rated by non-blinded enumerators and the blinded master rater, the average and distribution of the facial cleanliness scores were the same across rater type (6 [IQR: 5, 7]; Table 3). While the average of the facial cleanliness scores was the same for all three blinded raters and computer-simulated ratings, the variance differed slightly between raters (5 [IQR: 4, 6]; 5 [IQR: 4, 8]; 5 [IQR: 4, 6]; and 5 [IQR: 5, 7], respectively; Table 3). IRR of qPHAT facial cleanliness ratings was almost perfect across all rating groups, according to the Landis and Koch scale (Table 3).

3.5 qPHAT hand cleanliness ratings

For the 1332 wipes rated by both non-blinded enumerators and the blinded master rater, the average of the hand cleanliness scores was the same, but the variance differed slightly between raters (3 [IQR: 2, 5] vs. 3 [IQR: 2, 4], respectively; Table 3). For the sub-set of wipes rated by three blinded raters, the median hand cleanliness score was the same for the master rater and rater 2 (3 [IQR: 2, 4]), but differed slightly for rater 3 (4 [IQR: 3, 5]) and computer-simulated ratings (4 [IQR: 3, 4]). IRR of qPHAT hand cleanliness ratings was also almost perfect across all rating groups, according to the Landis and Koch scale (Table 3).

4 Discussion

The purpose of this study was to develop and assess the reliability of a novel personal hygiene metric that yields quantitative facial and hand cleanliness data. We determined that the qPHAT methodology produces highly reliable estimates of facial and hand cleanliness across all types of rater comparisons. While our results indicated IRR was higher across blinded and computer-simulated raters than the non-blinded vs. blinded rating pairs, all IRR and probabilistic benchmarking estimates fell within a range interpreted as almost perfect, according to the Landis and Koch scale. We conclude that the qPHAT methodology yields reliable measures of facial and hand cleanliness.

Our IRR estimates of two conventional clean face measures (i.e., absence of ocular and nasal discharge and absence of all signs of an unclean face) also fell in the almost perfect range. However, the individual conventional qualitative metrics that comprise these composite measures did not perform well on their own. Two of five conventional facial cleanliness metrics assessed in this study (wet nasal discharge; dirt/debris on the face) had IRR estimates in the moderate range and two (ocular discharge; dry nasal discharge) had IRR estimates in the substantial range. Our results corroborate existing evidence that suggests conventional qualitative metrics assessing ocular discharge, nasal discharge, and flies on the face perform better than qualitative metrics assessing dirt/debris on the face (West et al., 1991; Zack et al., 2008).

When interpreting data on the two conventional clean face measures – the absence of ocular and nasal discharge and the absence of all signs of an unclean face – one should consider that there is no standardized guidance for assessing or defining facial cleanliness

(International Trachoma Initiative & Neglected Tropical Diseases Support Center, 2019). For instance, some trachoma researchers and implementers consider clean face as the absence of ocular and nasal discharge (West et al., 2017), while others include other signs of facial cleanliness such as fly-eye contact and dirt/debris on the face (Burr et al., 2013; King et al., 2011). There is also no clear guidance on whether wet and dry nasal discharge should be assessed separately; some do (King et al., 2011) while others do not (West et al., 2017). Nasal discharge that ends up sticking to the face does not do so when already dry: the wet versus dry distinction is an artifact of the timing of observation and is subjective, since the state of dryness is continuous rather than binary. Our findings suggest that any nasal discharge would be a better marker than using wet and dry sub-categories, given the low reliability of those two individual measures. More critically, though, qPHAT ratings by blinded raters yielded more reliable cleanliness measures than any of the individual qualitative cleanliness metrics, when assessed either in this study or by others (King et al., 2011; West et al., 2017) in previously published work.

The conventional qualitative signs of hand cleanliness that we used here were similar to those used elsewhere (Halder et al., 2010; Pickering et al., 2010; Webb et al., 2006). Prevalence of qualitative signs of hand cleanliness observed in our study indicated that adults generally had cleaner hands than their children, yet assessments of adult hands were less reliable than assessments of children’s dirtier hands. Our results align with evidence from a study conducted in Guatemala that investigated the repeatability of hygiene measures, which also found that mother’s/caregiver’s hands were cleaner (per qualitative cleanliness metrics) and yielded less reliable results than assessments of their children’s hands (Webb et al., 2006). These findings could imply that when hands are cleaner, conventional qualitative metrics generate less reliable cleanliness data. Lower IRR estimates amongst adults with cleaner hands may be an artifact of a qualitative metric that forces raters to choose between only two options – presence or absence of a particular sign of cleanliness. Even after training and standardization, rater assessments may be somewhat subjective, and when hands or faces are cleaner, it may be harder to make dichotomous distinctions.

In addition to being at least as reliable, if not more reliable, than conventional qualitative cleanliness metrics, the qPHAT methodology generates more nuanced data, which may serve to improve the measurement of personal hygiene behavior and changes therein. It should be noted that the qPHAT methodology is intended to serve as a metric that generates quantitative measures of facial and hand cleanliness, and is not intended to capture data on all facial and hand cleanliness conditions (e.g., presence/absence of flies on one’s face, dirt under one’s nails). The quantitative cleanliness data generated by qPHAT permits examinations of dose-response relationships not previously available. The qPHAT methodology also allows raters to be blinded to the conditions of the household, the physical appearance of the subject, and exposure to interventions, which may further minimize bias. The measurement attributes of qPHAT may, in turn, facilitate enhanced monitoring of personal hygiene behavior, and improve the type of data available for evaluating the effectiveness of personal hygiene behavior change interventions on downstream health and well-being.

This study has three key limitations. First, we focused on assessing facial and hand cleanliness amongst children. We did not deploy the qPHAT methodology amongst adults. While this limitation prevented us from ascertaining the reliability of qPHAT hand cleanliness measures amongst seemingly cleaner adults, it should not discredit the reliability of qPHAT hand cleanliness measures amongst children. Second, this study was designed to create a novel quantitative hygiene metric and determine its reliability. It was outside the scope of this study to determine whether qPHAT methodology provides valid measures of recent cleansing (King et al., 2011) or would be sensitive enough to detect incremental changes in these personal hygiene practices. More work in this direction is needed. Third, we did not aim to determine whether facial and hand cleanliness measures were associated with contamination of these body parts with disease-causing pathogens or indicator organisms. Theories of change are critical here, since assessments of cleanliness are approximations for actual pathogen transmission potential.

5 Conclusions

The qPHAT methodology yielded highly reliable measures. It allows for quantifiable assessments of facial and hand cleanliness, providing more nuanced assessments of the degree of facial and hand cleanliness than conventional metrics. Additional research is needed to determine intra- and inter-rater reliability outside of Amhara, Ethiopia, and to ascertain whether the metric is a valid measure of recent cleansing and incremental changes in behavior. If further evaluations are supportive, the qPHAT methodology could facilitate future evaluations of personal hygiene interventions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by a grant from The World Bank Group’s Strategic Impact Evaluation Fund (7175829). The funder had no role in the design of this sub-study; the collection, analysis, and interpretation of data; the writing of this article; or the decision to submit the article for publication. The authors would like to thank our Andilaye study participants, who generously gave their time to participate in our formative work, pilot testing, and household survey. We are grateful for the support we received from numerous partners at the Democratic Republic of Ethiopia’s Federal Ministry of Health; the Amhara Regional Health Bureau; South Gondar and West Gojjam Zonal Health Departments; and the Bahir Dar Zuria, Farta, and Fogera Woreda Health Offices. We would also like to acknowledge support provided by the Health Extension Workers, Women’s Development Army Leaders, and Health Development Army members from our study kebeles. We thank the cadre of field supervisors (Destaw Asnakew, Tigist Abebe, and Yeserash Gashu) and enumerators (Adanech Admasu, Asayech Bimrew, Balemlaye Addisu, Eleni Nebiyu, Mahider Adamu, Maritu Yibrie, Mulubirhan Shitu, Rahel Tsegaye, Selamawit Abebe, Senait Mulualem, Sintayehu Wasihun, Tibeltalech Mihiret, Tigist Bitew, Tirusew Alayu, Tiruzer Engidaw, Woyneshet Genetu, and Yeworkwuha Abay) who captured these data. We acknowledge Resom Berhe for his support in pilot testing the qPHAT methodology, and Mulusew Belew for his management support.

References

Aiello AE, Larson EL. What is the evidence for a causal link between hygiene and infections? The Lancet Infect Dis. 2002; 2:103–110. [PubMed: 11901641]

Biran A, Rabie T, Schmidt W, Juvekar S, Hirve S, Curtis V. Comparing the performance of indicators of hand-washing practices in rural Indian households. Trop Med Int Health. 2008; 13:278–285. [PubMed: 18304276]

Boisson S, Engels D, Gordon BA, Medlicott KO, Neira MP, Montresor A, Solomon AW, Velleman Y. Water, sanitation and hygiene for accelerating and sustaining progress on neglected tropical diseases: a new Global Strategy 2015–20. Int Health. 2016; 8:i19–i21. [PubMed: 26940305]

Burr SE, Hart JD, Edwards T, Baldeh I, Bojang E, Harding-Esch EM, Holland MJ, Lietman TM, West SK, Mabey D, Sillah A, et al. Association between ocular bacterial carriage and follicular trachoma following mass azi-thromycin distribution in the Gambia. PLoS Neglected Trop Dis. 2013; 7:e2347.

Curtis V, Cousens S, Mertens T, Traore E, Kanki B, Diallo I. Structured observations of hygiene behaviours in Burkina Faso: validity, variability, and utility. Bull World Health Organ. 1993; 71:23–32. [PubMed: 8440034]

Davarinejad H. Quantifications of western blots with Image J. 2017

Delea MG, Solomon H, Solomon AW, Freeman MC. Interventions to maximize facial cleanliness and achieve environmental improvement for trachoma elimination: a review of the grey literature. PLoS Neglected Trop Dis. 2018; 12

Delea MG, Snyder JS, Belew M, Caruso BA, Garn JV, Sclar GD, Woreta M, Zewudie K, Gebremariam A, Freeman MC. Design of a parallel cluster-randomized trial assessing the impact of a demand-side sanitation and hygiene intervention on sustained behavior change and mental well-being in rural and peri-urban Amhara, Ethiopia. Andilaye Study Protoc BMC Publ Health. 2019; 19:801.

Dodson S, Heggen A, Solomon AW, Sarah V, Woods G, Wohlgemuth L. Behavioural change interventions for sustained trachoma elimination. Bull World Health Organ. 2018; 96:723–725. [PubMed: 30455520]

Edwards T, Harding-Esch EM, Hailu G, Andreason A, Mabey DC, Todd J, Cumberland P. Risk factors for active trachoma and Chlamydia trachomatis infection in rural Ethiopia after mass treatment with azithromycin. Trop Med Int Health : TM & IH. 2008; 13:556–565. [PubMed: 18282237]

Ejere HOD, Alhassan MB, Rabiu M. Face washing promotion for preventing active trachoma. Cochrane Database Syst Rev. 2015; (2)doi: 10.1002/14651858.CD003659.pub4

Esrey SA, Feachem RG, Hughes JM. Interventions for the control of diarrhoeal diseases among young children: improving water supplies and excreta disposal facilities. Bull World Health Organ. 1985; 63:757–772. [PubMed: 3878742]

Freeman MC, Stocks ME, Cumming O, Jeandron A, Higgins JP, Wolf J, Pruss-Ustun A, Bonjour S, Hunter PR, Fewtrell L, Curtis V. Hygiene and health: systematic review of handwashing practices worldwide and update of health effects. Trop Med Int Health : TM & IH. 2014; 19:906–916. [PubMed: 24889816]

Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008; 61:29–48. [PubMed: 18482474]

Gwet KL. Handbook of inter-rater reliability. The Definitive Guide to Measuring the Extent of Agreement Among Raters. 2014

Halder AK, Tronchet C, Akhter S, Bhuiya A, Johnston R, Luby SP. Observed hand cleanliness and other measures of handwashing behavior in rural Bangladesh. BMC Publ Health. 2010; 10:545–545.

International Trachoma Initiative and Neglected Tropical Diseases Support Center. F in SAFE Strategic Technical Meeting Report. Atlanta GA, , USA: 2019.

King JD, Ngondi J, Kasten J, Diallo MO, Zhu H, Cromwell E, Emerson PM. Randomised trial of face-washing to develop a standard definition of a clean face for monitoring trachoma control programmes. Trans R Soc Trop Med Hyg. 2011; 105:7–16. [PubMed: 21036378]

Klein D. Implementing a general framework for assessing interrater agreement in stata. STATA J. 2018; 18:871–901.

Kozlowski S, Hattrup K. A disagreement about within-group Agreement: disentangling issues of consistency versus consensus. J Appl Psychol. 1992; 77:161–167.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33:159–174. [PubMed: 843571]

Lin LI, Hedayat AS, Wu W. A unified approach for assessing agreement for continuous and categorical data. J Biopharm Stat. 2007; 17:629–652. [PubMed: 17613645]

Luby SP, Halder AK. Associations among handwashing indicators, wealth, and symptoms of childhood respiratory illness in urban Bangladesh. Trop Med Int Health. 2008; 13:835–844. [PubMed: 18363587]

Manun'Ebo M, Cousens S, Haggerty P, Kalengaie M, Ashworth A, Kirkwood B. Measuring hygiene practices: a comparison of questionnaires with direct observations in rural Zaïre. Trop Med Int Health. 1997; 2:1015–1021. [PubMed: 9391503]

Morrill V, de Aceituno A, Bartz F, Garcia N, Shumaker D, Grubb J, Arbogast J, Leon J. Visible soil as an indicator of bacteria concentration on farmworkers' hands. Food Protect Trends. 2018; 38:122–128.

Ngondi J, Gebre T, Shargie EB, Graves PM, Ejigsemahu Y, Teferi T, Genet A, Mosher AW, Endeshaw T, Zerihun M, Messele A, et al. Risk factors for active trachoma in children and trichiasis in adults: a household survey in Amhara Regional State, Ethiopia. Trans R Soc Trop Med Hyg. 2008; 102:432–438. [PubMed: 18394663]

Parvez SM, Azad R, Pickering AJ, Kwong LH, Arnold BF, Rahman MJ, Rahman MZ, Alam M, Sen D, Islam S, Rahman M, et al. Microbiological contamination of young children's hands in rural Bangladesh: associations with child age and observed hand cleanliness as proxy. PloS One. 2019; 14

Pickering AJ, Davis J, Walters SP, Horak HM, Keymer DP, Mushi D, Strickfaden R, Chynoweth JS, Liu J, Blum A, Rogers K, et al. Hands, water, and health: fecal contamination in Tanzanian communities with improved, non-networked water supplies. Environ Sci Technol. 2010; 44:3267–3272. [PubMed: 20222746]

Pickering AJ, Blum AG, Breiman RF, Ram PK, Davis J. Video surveillance captures student hand hygiene behavior, reactivity to observation, and peer influence in Kenyan primary schools. PloS One. 2014; 9:e92571–e92571. [PubMed: 24676389]

Prüss-Ustün A, Bartram J, Clasen T, Colford JM Jr, Cumming O, Curtis V, Bonjour S, Dangour AD, De France J, Fewtrell L, Freeman MC, et al. Burden of disease from inadequate water, sanitation and hygiene in low- and middle-income settings: a retrospective analysis of data from 145 countries. Trop Med Int Health. 2014; 19:894–905. [PubMed: 24779548]

Rabie T, Curtis V. Handwashing and risk of respiratory infections: a quantitative systematic review. Trop Med Int Health : TM & IH. 2006; 11:258–267. [PubMed: 16553905]

Ram PK, Halder AK, Granger SP, Jones T, Hall P, Hitchcock D, Wright R, Nygren B, Islam MS, Molyneaux JW, Luby SP. Is structured observation a valid technique to measure handwashing behavior? Use of acceleration sensors embedded in soap to assess reactivity to structured observation. Am J Trop Med Hyg. 2010; 83:1070–1076. [PubMed: 21036840]

Ram PK, Jahid I, Halder AK, Nygren B, Islam MS, Granger SP, Molyneaux JW, Luby SP. Variability in hand contamination based on serial measurements: implications for assessment of hand-cleansing behavior and disease risk. Am J Trop Med Hyg. 2011; 84:510–516. [PubMed: 21460002]

Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods. 2012; 9:671–675. [PubMed: 22930834]

Strunz EC, Addiss DG, Stocks ME, Ogden S, Utzinger J, Freeman MC. Water, sanitation, hygiene, and soil-transmitted helminth infection: a systematic review and meta-analysis. PLoS Med. 2014; 11

Utsi L, Smith SJ, Chalmers RM, Padfield S. Cryptosporidiosis outbreak in visitors of a UK industry-compliant petting farm caused by a rare Cryptosporidium parvum subtype: a case-control study. Epidemiol Infect. 2016; 144:1000–1009. [PubMed: 26424385]

Webb AL, Stein AD, Ramakrishnan U, Hertzberg VS, Urizar M, Martorell R. A simple index to measure hygiene behaviours. Int J Epidemiol. 2006; 35:1469–1477. [PubMed: 17023500]

West SK, Congdon N, Katala S, Mele L. Facial cleanliness and risk of trachoma in families. Arch Ophthalmol. 1991; 109:855–857. [PubMed: 2043075]

West S, Munoz B, Lynch M, Kayongoya A, Chilangwa Z, Mmbaga BB, Taylor HR. Impact of face-washing on trachoma in Kongwa, Tanzania. Lancet. 1995; 345:155–158. [PubMed: 7823670]

West SK, Ansah D, Munoz B, Funga N, Mkocha H. The "F" in SAFE: reliability of assessing clean faces for trachoma control in the field. PLoS Neglected Trop Dis. 2017; 11

Zack R, Mkocha H, Zack E, Munoz B, West SK. Issues in defining and measuring facial cleanliness for national trachoma control programs. Trans R Soc Trop Med Hyg. 2008; 102:426–431. [PubMed: 18346769]

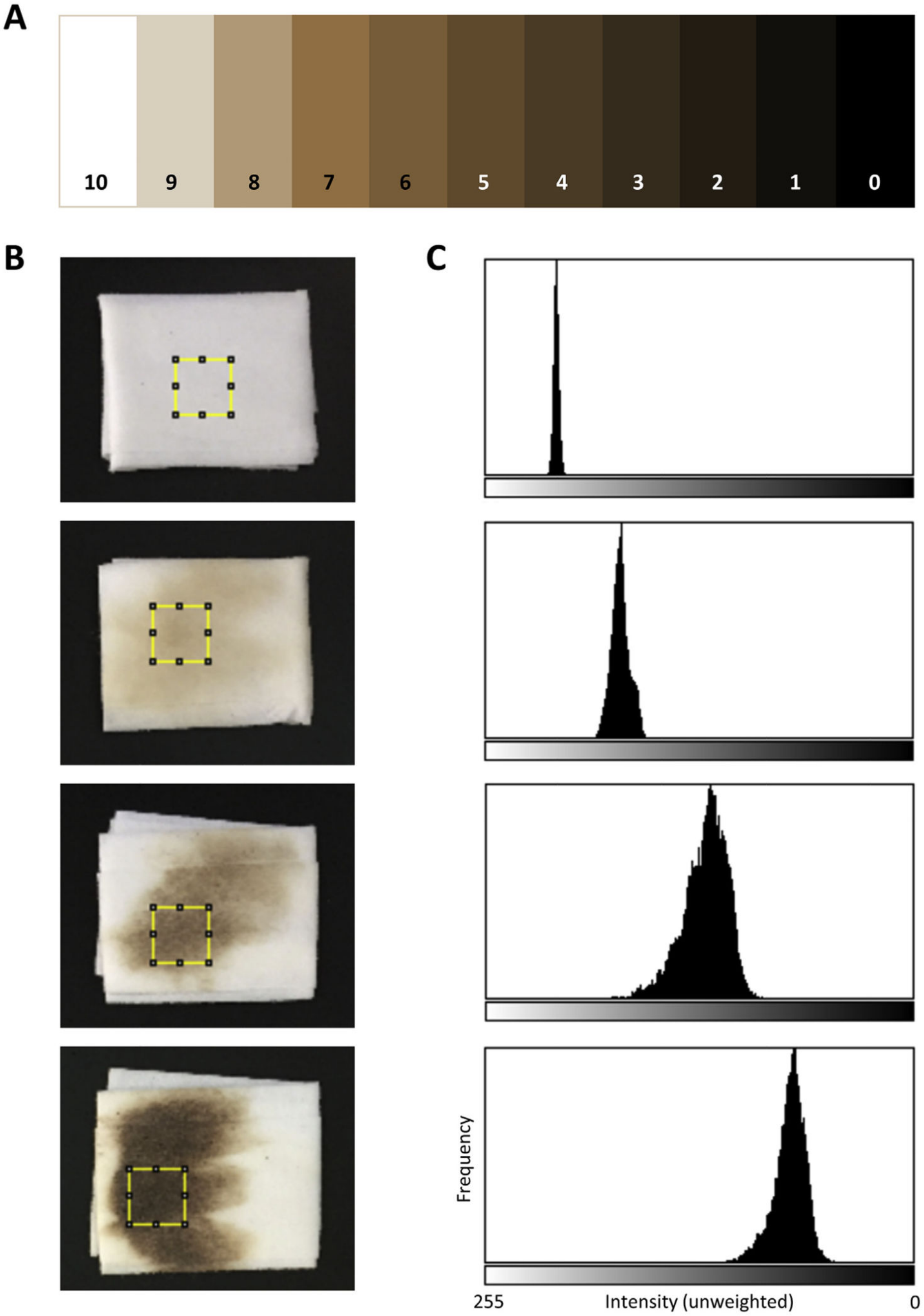

Fig. 1.

Face and hand wipes obtained via the qPHAT methodology, densito-metric data. (A) 11-point qPHAT color scale representing a color model array with 10% step-changes in saturation along the array. (B) Photographs of face and hand wipes used to obtain quantitative measures of cleanliness, including an unused wipe (top) used for calibration (computer-simulated ratings only). Human raters generated quantitative cleanliness data by matching the color of the darkest point within the darkest square half-inch of the wipe to a color represented in the qPHAT color scale. (C) ImageJ-derived histograms of unweighted intensity (minimum of 255, maximum of 0) generated from densitometric analyses of the area on the wipes with the highest optical density (i.e., darkest square half-inch, as indicated by the yellow boxes in panel B), where density is defined as intensity over area. This illustrative set of wipes scored 10, 9, 7, and 4 (top to bottom), per computer-simulated qPHAT ratings.

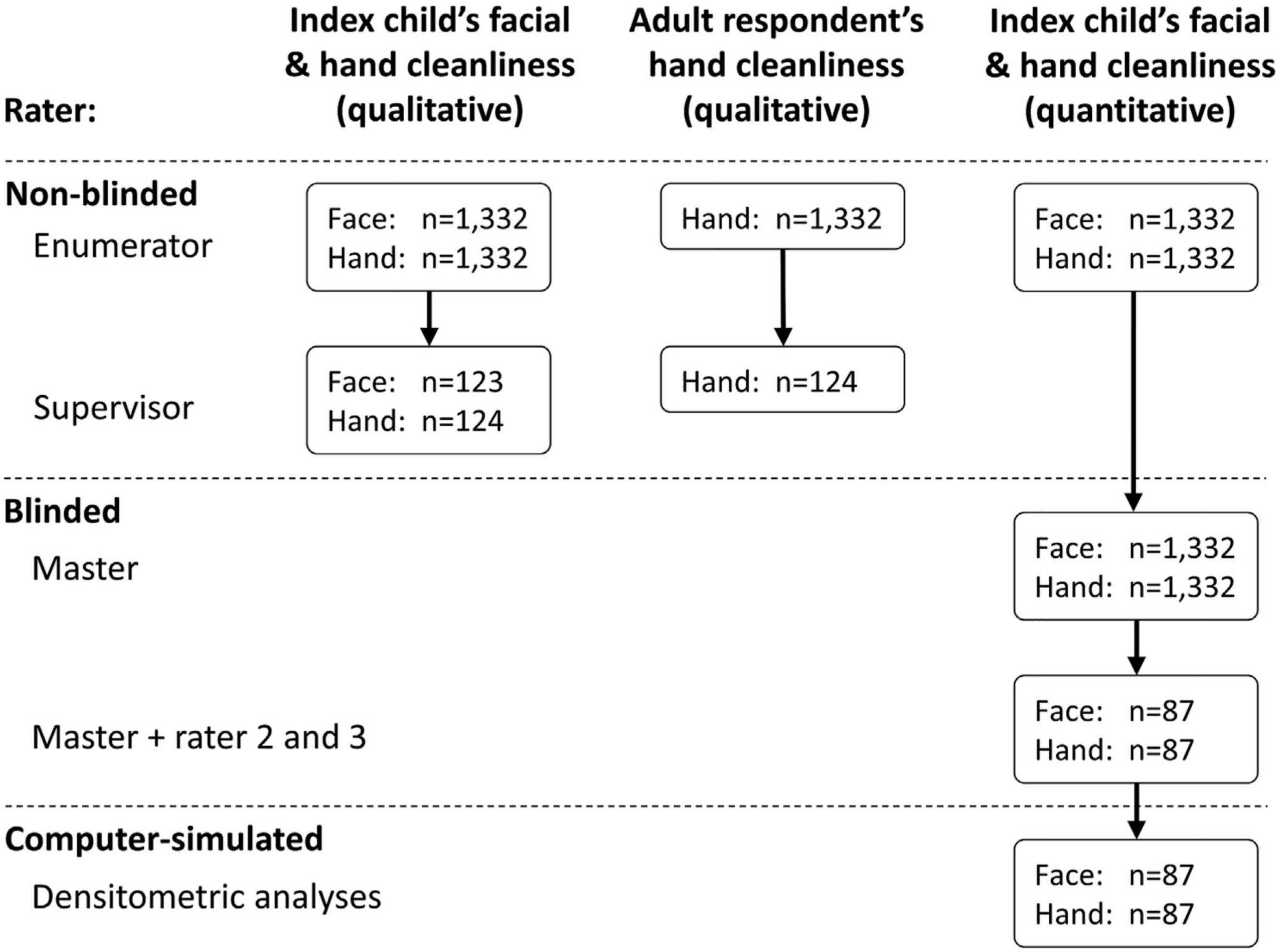

Fig. 2.

Flow of conventional qualitative and novel quantitative (qPHAT) cleanliness data.

Table 1

Sample characteristics and demographics.

Study subject	n (%)
ADULT RESPONDENT (N = 1332)	
Female	1231 (92%)
Caregiving responsibilities	
Mother of index child	1168 (88%)
Other female caregiver	59 (4%)
Male caregiver	96 (7%)
Other adult household member	9 (1%)
Age[*]	32 (IQR: 27, 38)
INDEX CHILD (N = 1332)	
Female	658 (49%)
Age[*]	4 (IQR: 2, 6)

Notes.

[*] Median age presented in years along with the inter-quartile range (IQR).

Table 2
Prevalence and IRR of conventional qualitative facial and hand cleanliness metrics.

Cleanliness metric	Enumerator n (%)	Supervisor n (%)	IRR[b] (95% CI)	Probabilistic benchmarking	Landis Koch interpretation
INDEX CHILD'S FACIAL CLEANLINESS (N = 123)[a]					
Absence of ocular discharge	78 (63%)	78 (63%)	0.76 (0.64, 0.87)	0.60–0.80	Substantial
Absence of ANY nasal discharge	27 (22%)	23 (19%)	0.86 (0.77, 0.94)	0.60–0.08	Substantial
Absence of wet nasal discharge	62 (50%)	57 (46%)	0.56 (0.41, 0.71)	0.40–0.60	Moderate
Absence of dry nasal discharge	39 (32%)	36 (29%)	0.73 (0.61, 0.85)	0.60–0.80	Substantial
Absence of dirt/debris on face	41 (33%)	26 (21%)	0.66 (0.53, 0.80)	0.40–0.60	Moderate
Absence of flies on face	26 (21%)	22 (18%)	0.93 (0.87, 0.99)	0.80–1.00	Almost perfect
Clean face – No ocular or nasal discharge	23 (19%)	19 (15%)	0.91 (0.84, 0.97)	0.80–1.00	Almost perfect
Clean face[c] - Composite	4 (3%)	2 (2%)	0.98 (0.96, 1.00)	**I**	Almost perfect
INDEX CHILD'S HAND CLEANLINESS (N = 124)					
Absence of dirt under all finger nails on hand					
Left hand	9 (7%)	7 (6%)	0.91 (0.85, 0.97)	0.80–1.00	Almost perfect
Right hand	11 (9%)	11 (9%)	0.90 (0.84, 0.97)	0.80–1.00	Almost perfect
Absence of dirt on all finger pads of hand					
Left hand	17 (14%)	21 (17%)	0.80 (0.71, 0.90)	0.60–0.80	Substantial
Right hand	15 (12%)	22 (18%)	0.79 (0.70, 0.89)	0.60–0.80	Substantial
Absence of dirt on palm of hand					
Left hand	20 (16%)	22 (18%)	0.82 (0.73, 0.91)	0.60–0.80	Substantial
Right hand	21 (17%)	23 (19%)	0.84 (0.75, 0.93)	0.60–0.80	Substantial
Clean hand[d]					
Left hand	4 (3%)	4 (3%)	0.95 (0.91, 0.99)	0.80–1.00	Almost perfect
Right hand	5 (4%)	6 (5%)	0.96 (0.92, 0.99)	0.80–1.00	Almost perfect
ADULT RESPONDENT'S HAND CLEANLINESS (N = 124)					
Absence of dirt under all finger nails on hand					
Left hand	26 (21%)	26 (21%)	0.73 (0.62, 0.85)	0.60–0.80	Substantial

Cleanliness metric	**Enumerator n (%)**	**Supervisor n (%)**	**IRR[b] (95% CI)**	**Probabilistic benchmarking**	**Landis Koch interpretation**
Right hand	27 (22%)	28 (23%)	0.74 (0.63, 0.86)	0.60–0.80	Substantial
Absence of dirt on all finger pads of hand					
Left hand	37 (30%)	43 (35%)	0.60 (0.45, 0.74)	0.40–0.60	Moderate
Right hand	37 (30%)	33 (27%)	0.67 (0.54, 0.81)	0.40–0.60	Moderate
Absence of dirt on palm of hand					
Left hand	52 (42%)	46 (37%)	0.60 (0.45, 0.74)	0.40–0.60	Moderate
Right hand	52 (42%)	46 (37%)	0.63 (0.49, 0.77)	0.40–0.60	Moderate
Clean hand[d]					
Left hand	20 (16%)	16 (13%)	0.83 (0.74, 0.92)	0.60–0.80	Substantial
Right hand	21 (17%)	19 (15%)	0.85 (0.76, 0.93)	0.60–0.80	Substantial

Notes. Inter-rater reliability (IRR), as indicated by Gwet’s coefficient, was assessed between the enumerator-supervisor rating pair. Enumerators and supervisors were both on site at the study household and observed the conditions of the household compound and overall appearance of the adult and child. Therefore, both raters provided non-blinded ratings. NB: Equivalent proportions of the various signs of cleanliness do not point to consistency in ratings between paired raters. For instance, the 11 children enumerators rated as having an absence of dirt under all fingernails on the right hand were not the same 11 children field supervisors rated as having absence of dirt under all fingernails on the right hand – i.e., there was an inconsistency in the ratings despite equivalent prevalence.

[a] One child was actively crying during observation; therefore, no facial cleanliness data were collected on the child.

[b] IRR reflects chance-corrected inter-rater reliability, as indicated by Gwet’s coefficient and related 95% confidence interval (95% CI).

[c] Absence of all signs of an unclean face (i.e., ocular discharge, wet or dry nasal discharge, other dirt/debris on face, flies on face).

[d] Absence of all signs of unclean hands (i.e., dirt under any finger nail, dirt on any finger pad, dirt on either palm)‖ Perfectly predicted.

Table 3
Distribution and IRR of qPHAT rating scores.

Rating pairs	N	Master rater[a] Median (IQR)	Alt. rater 2[b] Median (IQR)	Alt. rater 3 Median (IQR)	Computer rater Median (IQR)	IRR[c] (95% CI)	Probabilistic benchmarking	Landis Koch interpretation
INDEX CHILD’S FACIAL CLEANLINESS								
Master rater[a] vs. non-blinded rater	1332	6 (5, 7)	6 (5, 7)	-	-	0.90 (0.90, 0.91)	0.80-1.00	Almost perfect
Three blinded raters	87	5 (4, 6)	5 (4, 8)	5 (4, 6)	-	0.96 (0.95, 0.97)	0.80-1.00	Almost perfect
Human vs. computer rater[d]								
Master rater[a] vs. computer rater	87	5 (4, 6)	-	-	5 (5, 7)	0.98 (0.97, 0.99)	0.80-1.00	Almost perfect
Blinded rater 2 vs. computer rater	87	-	5 (4, 8)	-	5 (5, 7)	0.96 (0.94, 0.97)	0.80-1.00	Almost perfect
Blinded rater 3 vs. computer rater	87	-	-	5 (4, 6)	5 (5, 7)	0.97 (0.96, 0.98)	0.80-1.00	Almost perfect
INDEX CHILD’S HAND CLEANLINESS								
Master rater[a] vs. blinded rater	1332	3 (2, 4)	3 (2, 5)	-	-	0.92 (0.91, 0.93)	0.80-1.00	Almost perfect
Three blinded raters	87	3 (2, 4)	3 (2, 4)	4 (3, 5)	-	0.95 (0.93, 0.96)	0.80-1.00	Almost perfect
Human vs. computer raters’[3]								
Master rater[a] vs. computer rater	87	3 (2, 4)	-	-	4 (3, 4)	0.97 (0.96, 0.98)	0.80-1.00	Almost perfect
Blinded rater 2 vs. computer rater	87	-	3(2, 4)	-	4 (3, 4)	0.96 (0.94, 0.98)	0.80-1.00	Almost perfect
Blinded rater 3 vs. computer rater	87	-	-	4 (3, 5)	4 (3, 4)	0.97 (0.96, 0.98)	0.80-1.00	Almost perfect

Notes. Table summarizes inter-rater reliability assessments of quantitative (interval) data generated by the qPHAT methodology across various rating pairs. IQR = inter-quartile range.

[a] The master rater was blinded to the conditions of the child and the child’s household.

[b] For the “Master rater vs. non-blinded rater”, Alt. rater 2 reflects the non-blinded enumerator rating; For the “Human vs. computer raters”, Alt. rater 2 reflects blinded rater 2.

[c] Inter-rater reliability (IRR) reflects chance-corrected inter-rater reliability, as indicated by Gwet’s coefficient and the related 95% confidence interval (95% CI).

[d] Computer-simulated ratings, as generated via the employment of densitometry.