EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE

Open Access

Human biases in body measurement estimation

Kirill Martynov[1], Kiran Garimella[2] and Robert West[3]*

*Correspondence:
robert.west@epfl.ch
[3]EPFL, Lausanne, Switzerland
Full list of author information is
available at the end of the article

Abstract

Body measurements, including weight and height, are key indicators of health. Being able to visually assess body measurements reliably is a step towards increased awareness of overweight and obesity and is thus important for public health. Nevertheless it is currently not well understood how accurately humans can assess weight and height from images, and when and how they fail. To bridge this gap, we start from 1,682 images of persons collected from the Web, each annotated with the true weight and height, and ask crowd workers to estimate the weight and height for each image. We conduct a faceted analysis taking into account characteristics of the images as well as the crowd workers assessing the images, revealing several novel findings: (1) Even after aggregation, the crowd's accuracy is overall low. (2) We find strong evidence of contraction bias toward a reference value, such that the weight of light people and the height of short people are overestimated, whereas the weight of heavy people and the height of tall people are underestimated. (3) We estimate workers' individual reference values using a Bayesian model, finding that reference values strongly correlate with workers' own height and weight, indicating that workers are better at estimating people similar to themselves. (4) The weight of tall people is underestimated more than that of short people; yet, knowing the height decreases the weight error only mildly. (5) Accuracy is higher on images of females than of males, but female and male workers are no different in terms of accuracy. (6) Crowd workers improve over time if given feedback on previous guesses. Finally, we explore various bias correction models for improving the crowd's accuracy, but find that this only leads to modest gains. Overall, this work provides important insights on biases in body measurement estimation as obesity-related conditions are on the rise.

Keywords: Crowdsourcing; Bias; Human measurement; Weight and height; Visual estimation

1 Introduction

Height and weight play a key role as indicators of health [1]. Population-level height and weight statistics open a window onto the general health of a region. With obesity and other lifestyle diseases on the rise, it is important to understand how humans perceive height and weight. For instance, research has shown that exposure to obesity changes the perception of the weight status of others [2], and that perceived deviations of one's own weight

Springer

from an "ideal" weight often lead to issues such as social anxiety, depression, peer victimization, and lower self-worth [3]. In terms of action, understanding human perception of body measurements should therefore play a critical role for public health agencies when striving to design effective strategies for raising awareness about the prevalence and risks of overweight and obesity. For all these reasons, the estimation of human body measurements is an active area of interest in medical research [2, 4–7].

While a thorough understanding of the perception biases in height and weight estimation is important in and of itself, it is also important for several potential downstream applications. For instance, crowdsourcing systems for inferring body measurements from social-media images [8] could be leveraged to estimate regional weight and height distributions across the world, but such systems require accurate crowd labels. Understanding humans' accuracy and biases is thus essential for establishing the bounds of what one might hope to be achieved by crowd-based weight and height sensing systems. Crowdsourced estimates could also be useful for training machine learning models that could ultimately produce height and weight estimates in an automated fashion [9–14]. Such methods require large numbers of labeled training images, and crowdsourcing could potentially provide an efficient way of collecting such labeled samples. Machine learning models can only be as good as the data they are trained on, so here too, quantifying how well crowds perform at height and weight estimation establishes upper bounds for machine learning systems that use crowdsourced training data, and understanding human biases opens up avenues for addressing the crowd's specific shortcomings.

Existing studies of body measurement estimation [2, 4–7, 15, 16] are typically based on surveys that reach only small numbers of people and thus tend to lack the scale to provide generalizable insights. The present paper contributes to bridging this gap through a large-scale, systematic analysis of how humans estimate others' weight and height. We designed and conducted a large study on an online crowdsourcing platform, where we asked participants ("workers") to estimate the height and weight of people in 1682 images that had been collected from an online weight-loss forum and labeled with ground-truth height and weight values by the persons in the images themselves. In addition to a total of over 100,000 guesses, our pool of 1767 crowd workers also provided their own height, weight, and demographic variables. This data allowed us to conduct a faceted analysis of the accuracy of crowd workers as well as their biases—the systematic ways in which they succeed and fail.

Our main contributions are as follows. In a *large-scale crowdsourced study* (Sect. 3), we collected over 100,000 height and weight estimates from a diverse set of crowd workers, showing that the *crowd's accuracy* is low (Sect. 4), with an overall mean absolute error of 15.5 kg (6.3 cm), and 8.8 kg (5.2 cm) on images subsampled to match the weight (height) distribution of workers. A small number of workers (20 or 30) sufficed to achieve low variance, but the error remained large due to contraction bias toward a reference value, such that the weight of light people and the height of short people were systematically overestimated, whereas the weight of heavy people and the height of tall people were underestimated. Also, the taller a person, the more their weight was underestimated.

We then analyzed the *dependency of accuracy on worker characteristics* (Sect. 5). Although the height and weight of females were found to overall be easier to guess than those of males, female and male workers were no different in terms of accuracy. Workers were, however, better at estimating bodies similar to their own in terms of height and weight.

We explain this finding with a Bayesian model that assumes every worker to be biased toward a worker-specific reference value. By inferring a worker's reference values from the worker's estimates and from ground-truth image labels alone (without using the worker's own measurements), we found that the inferred reference values correlated strongly with workers' own measurements.

Based on these insights, we explored paths *towards more accurate crowdsourcing* (Sect. 6). We found that various post-hoc bias correction models only led to modest gains, whereas varying the setup of the crowdsourcing task itself is more promising: the accuracy could be improved by giving workers feedback on the quality of their guesses, effectively teaching them to make better guesses in the future. Another setup, however, where we gave workers access to the height of the person to be estimated, decreased the weight error only mildly.

We conclude the paper (Sect. 7) by discussing limitations of our work and by revisiting related work (cf. Sect. 2) in the context of our results. Finally, hoping to inform future research in this area, we digest our findings into a set of implications and best practices for crowdsourcing and machine learning tasks involving body measurements.

2 Related work

The study of how humans estimate body measurements is an active area of research spanning multiple diverse domains, including psychology, medicine, and computer science. The most relevant pieces of prior work pertain to human perception biases and to the usage of crowdsourcing. We review these two directions in turn.

Perception biases in body measurement estimation Biases in weight and height estimates have been studied in medicine and behavioral psychology [15–17]. When asked to judge their own weight, people tend to systematically under- or overestimate. For instance, it has been established through lab studies that people show a progressive underestimation of overweight and obese bodies [16], attributed to a psychological phenomenon known as contraction bias [17]. According to these studies, such underestimation may compromise people's ability to recognize weight gain and undertake compensatory weight control behaviors.

Other psychological experiments in lab settings have tried to explain the aforementioned biases by constructing mental models for how people judge weight and height. Facial adiposity [4] was found to constitute an important signal used in weight estimation, whereas features such as dominance and facial maturity [5], as well as head-to-shoulder ratio [18], are prominent in height estimation. Robinson [7] proposed visual normalization theory, a framework for explaining why people under- or overestimate certain body shapes. This theory is based on the notion that weight is judged relative to the body-size norms prevalent in the respective society. As larger body sizes are common in the United States, participants in Robinson's [7] study assessed overweight individuals as normal. Robinson et al. [6] studied a similar phenomenon for young adults from the United States (high obesity prevalence), the United Kingdom, and Sweden (lower obesity prevalence) to understand cultural differences between perceptions. They did not find cross-cultural differences in judging overweight males.

Based on questions and answers from Yahoo! Answers, Yom-Tov [19] found that men estimate their weight status fairly well, whereas women do not. This could potentially be

explained by visual normalization theory (see above), where, due to societal norms on body shapes of women, being slightly over the median might be considered "overweight" for women but not for men. Apart from gender, weight perception was also shown to be subject to generational shifts [20].

Humans tend to update their perceptions when shown evidence that contradicts their mental model. For instance, Robinson and Kirkham [2] showed that people exposed to photographs of obese young males changed their visual judgments of whether an overweight young male was of healthy weight.

Researchers have also worked on statistical models for correcting human estimation biases, usually based on simple regression techniques [21, 22]. Finally, Martynov et al. [23] have investigated the impact of clothing on body-size perception, showing that dark clothes have a small but statistically significant slimming effect.

Crowdsourcing for body measurement estimation   Some pieces of previous work have leveraged crowdsourcing to estimate height and weight. Indeed, historically speaking, body measurement and crowdsourcing are tightly intertwined. In his 1907 paper *Vox populi* [24]—probably the first paper on the "wisdom of crowds"—Francis Galton evaluated data from a weight-judging competition held at a Plymouth county fair, where 787 participants guessed the weight of an ox. Galton observed that the median guess was surprisingly accurate, only 9 pounds (0.8%) above the true weight of the ox (1198 pounds). One might hope that, if humans excel at estimating the weight of an ox, they would perform even better when the ox is replaced with a fellow human. As we will see, though, this is unfortunately not the case.

In the context of humans, rather than oxen, crowdsourcing has been shown to be a useful tool for assessing childhood predictors of adult obesity [25] and for estimating weight from social media profile images [8].

It has been established that, when using crowds as estimators, perception biases [26] and social influence [27] can lead to severe biases. In the present work, we turn the vice into a virtue: our primary goal is not to build accurate weight and height estimators based on crowds, but rather to assess the very biases inherent in crowds as estimators.

Present work   Our work is situated at the intersection of the above two strands of research. On the one hand, we replicate findings about biases in weight and height perception from small-scale lab studies in a much larger setting with thousands of crowd workers, revealing similar biases. On the other hand, by recruiting a global and diverse pool of crowd workers, we can go beyond the scope of previous studies and investigate how biases vary across workers' countries of residence and demographic strata: by collecting workers' own weight, height, age, gender, and country, we can systematically study the impact of those factors and their interplay with the image whose weight is to be estimated.

3 Data
3.1 Height- and weight-labeled body images
The original data was provided by Kocabey et al. [13] and includes about 10k samples collected from a Reddit forum ("subreddit") called *r/progresspics*,[a] where users who intend to lose (or sometimes gain) weight post pictures of themselves before and after their weight transformation. Each sample contains the ID of the Reddit post, the height and gender of
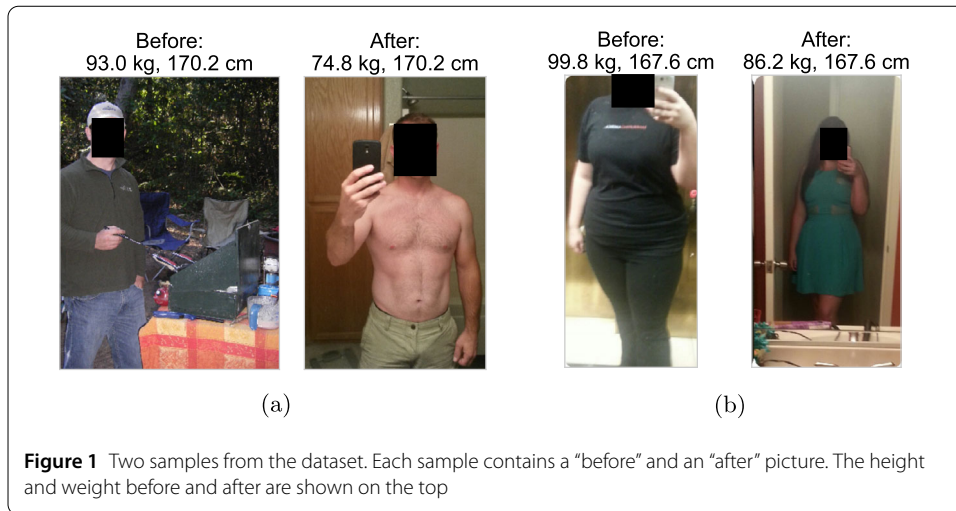
Figure 1 Two samples from the dataset. Each sample contains a "before" and an "after" picture. The height and weight before and after are shown on the top
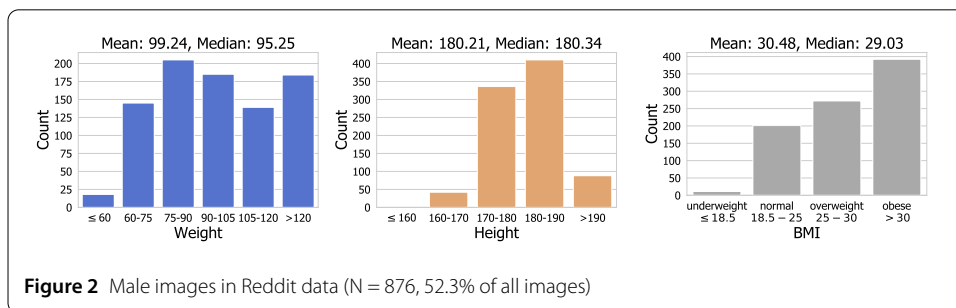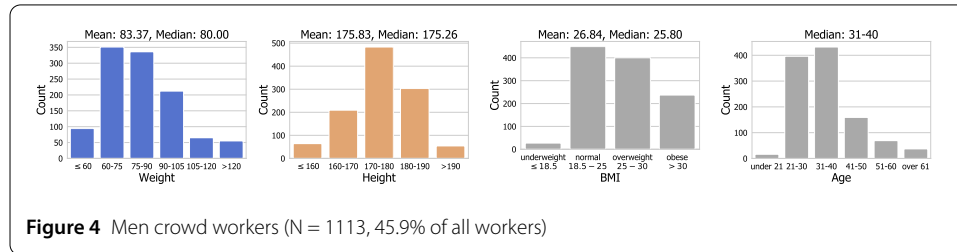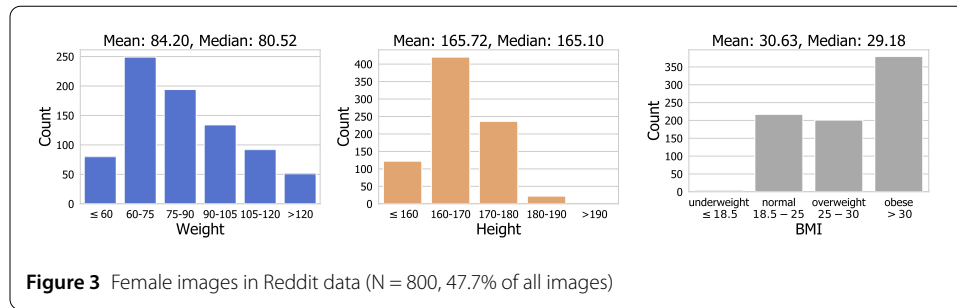

Figure 2 Male images in Reddit data (N = 876, 52.3% of all images)

the user, their weight before and after the transformation, as well as one or several images. Note that the weight and height labels were provided by the Reddit-users themselves and might not always be exact or accurate. However, throughout the paper, we refer to these labels as the *(weak) ground-truth* values, because we are not aware of a method that could check or improve their quality.

Several preprocessing steps were required to prepare the data for our weight and height estimation task. In many cases, users combined the "before" and "after" pictures into a single image file, so our first processing step consisted in automatically splitting such collages into its component images, such that a single weight could be associated with each image. To avoid ambiguous assignments, collages with more than two images were excluded from the dataset. Furthermore, we dropped images that contain weight or height labels as text or that depict multiple people. Finally, low-quality photos, as well as images that contain only faces, were omitted.

As a result, we obtained 841 samples (431 and 410 posts from male and female users, respectively), for a total of 1682 images (two images per person, each with a weight and a height label). Two samples from the dataset are shown in Fig. 1. We manually inspected the dataset and characterized the different types of images present in our data. Only around 10% of the images are full-body pictures, the rest lack part of the body (e.g., head or legs); 80% depict fully dressed people; around 50% are selfies taken in front of a mirror; and in around 75% we can see the head. Figure 2 (men) and Fig. 3 (women) show distributions of weight, height, and body-mass index (BMI) for the collected images.

Figure 3 Female images in Reddit data (N = 800, 47.7% of all images)

Figure 4 Men crowd workers (N = 1113, 45.9% of all workers)

## 3.2 Crowdsourced height and weight estimates

We used Amazon Mechanical Turk to collect weight and height estimates from diverse groups of workers. For this purpose, the dataset described in Sect. 3.1 was split into 168 tasks with 10 images in each task. For each image, crowd workers were asked to guess the weight and height of the depicted person. Both input fields were located under the image, with the field for weight estimates coming first. The workers could choose between "kg" and "lbs" units for weight estimation, and "cm" and "ft/in" for height estimation. Conversions between different units were automatically performed on the fly, i.e., the other field was updated at the same time as workers were typing the guesses in their preferred format. To encourage high-quality estimates, 25% of the most accurate (on a per-task basis) workers were awarded with a bonus that doubled their usual reward.

Each task (a collection of 10 images) was independently performed by 45 workers who estimated both weight and height. In this way, we gathered 45 weight and height labels per image, or 75,600 estimates from 1767 unique workers overall. Sections 4 and 5 focus mostly on this main dataset.

Apart from the main experiment described above, we explored two additional setups, where workers estimated (1) the weight while being shown the true height (Sect. 6.2) or (2) the weight and height while given the accuracy of their previous guess (Sect. 6.3). In each of these experiments, we collected another 20k estimates.

Finally, for each experimental setup, we collected the following personal information from all participating workers: their own height and weight, age, gender, and country of residence. The numerical characteristics are summarized in Fig. 4 (men) and Fig. 5 (women) for the 2426 crowd workers who performed at least one of our tasks. Regarding the country of residence, most of the crowd workers are residents of the United States (74%) or India (17%).

The raw weight and height estimates obtained from the crowd workers were filtered and preprocessed in several ways. In particular, we removed data from workers who appear to have used scripts to generate random guesses, as well as estimates from those who did not provide their personal information. Furthermore, several kinds of erroneous guesses, such

Figure 5 Women crowd workers (N = 1313, 54.1% of all workers)
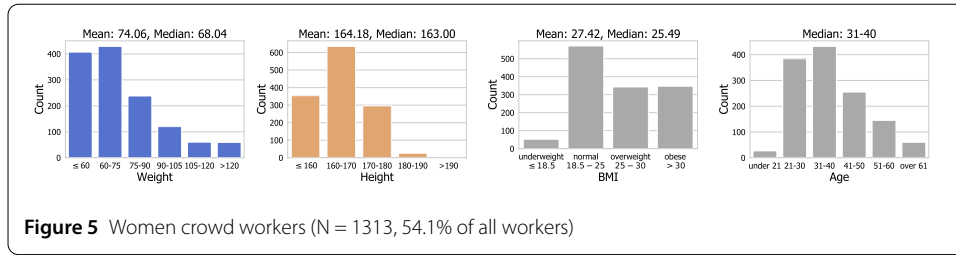
as obvious typos or usage of wrong measurement units (e.g., a worker accidentally chose lbs instead of kg or vice versa), were detected. After these preprocessing steps, roughly 75% of the initial estimates remained in each of the experiments.

4 Crowd-worker accuracy

The basic idea underlying the "wisdom of crowds" is that, for certain estimation tasks, aggregating local estimates from multiple people can lead to a much more accurate global estimate. Aggregation can be performed in multiple ways, e.g., via the mean or the median. As we will see later (Table 1), the two yield similar results, so in the rest of the paper, we will use the mean of all individual guesses collected for an image i as an estimate of the true weight of i:

$$w_{\text{est}}^i = \frac{1}{n} \sum_{j=1}^{n} w_j^i, \tag{1}$$

where $w_1^i, \ldots, w_n^i$ are the weight guesses collected for image i. The error for image i can then be defined as

$$w_{\text{err}}^i = w_{\text{true}}^i - w_{\text{est}}^i, \tag{2}$$

where w_{true}^i is the true weight for image i. That is, *positive (negative) errors signify underestimation (overestimation)*. Similarly, we define the true height h_{true}^i, the estimated height h_{est}^i, and the height error h_{err}^i.

Remember that, for each sample of the dataset from Sect. 3.1, two weights and two images ("before" and "after") need to be matched. This is done in an intuitive manner that also minimizes the error: the image with the higher value of w_{est}^i also is assigned the higher ground-truth label w_{true}^i. The number of wrong assignments is very low with this strategy: among 70 samples (140 images) that we manually inspected, only two were assigned incorrectly. Thus, the resulting accuracy of the ground truth weight labels is expected to be above 95%.

4.1 Height and weight estimation errors

Figure 6 shows the dependence of the mean height and weight error on the number of collected guesses. We see that the 95% confidence intervals around the mean error become very small after around 30 guesses both for height and weight. As a side note, this fact justifies, *post hoc*, our choice of collecting 45 guesses per image.

The distributions of errors in terms of mean error (i.e., the mean of w_{err} and h_{err}, respectively; ME) and mean absolute error (i.e., the mean of the absolute value of w_{err} and h_{err},

Figure 6 Convergence of mean error on all 1,682 images

Figure 7 Distributions of weight and height errors
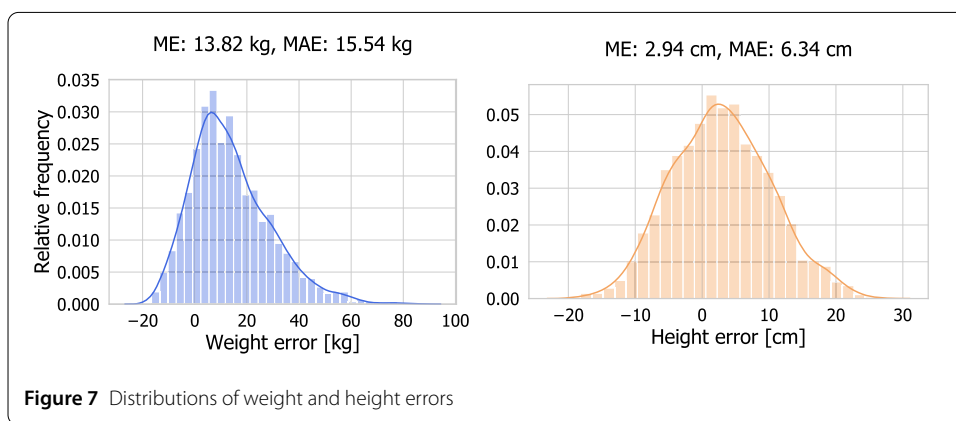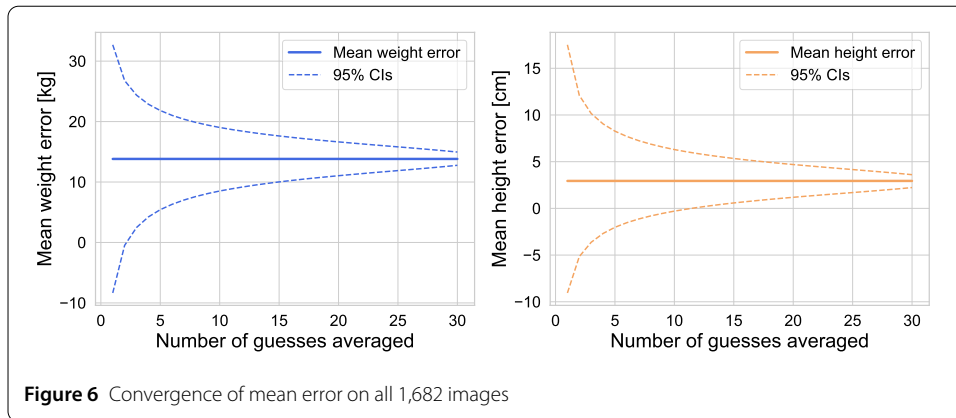
Table 1 Mean error (ME) and mean absolute error (MAE) when aggregating individual estimates via mean and median, with 95% confidence intervals

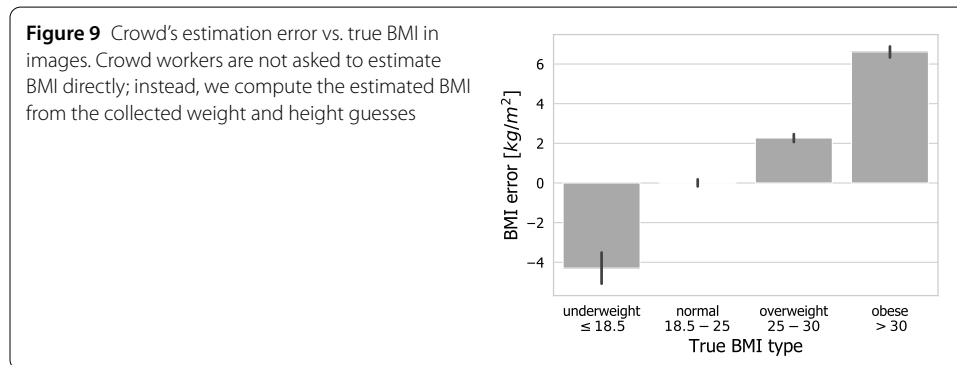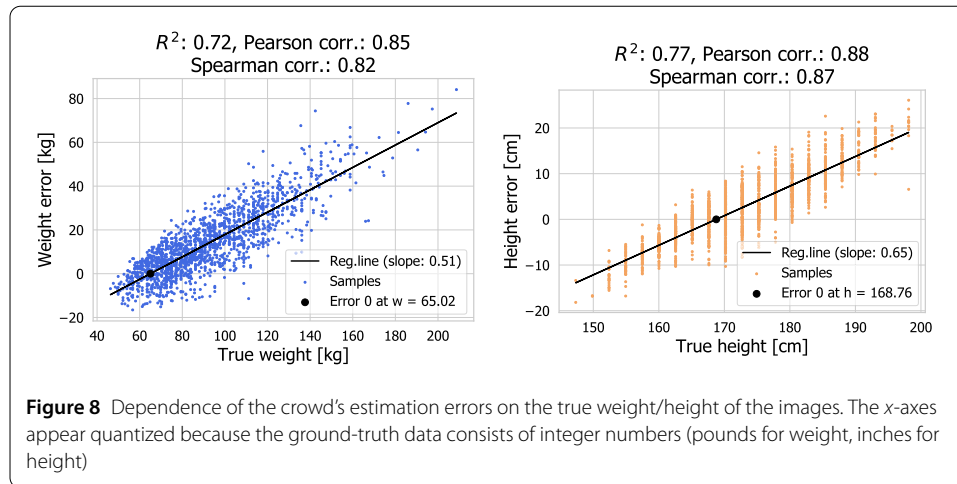|	Mean	Median
Weight ME [kg]	13.82 ± 0.7	14.47 ± 0.7
Weight MAE [kg]	15.54 ± 0.65	16.09 ± 0.7
Height ME [cm]	2.94 ± 0.35	2.55 ± 0.35
Height MAE [cm]	6.34 ± 0.25	6.20 ± 0.2

respectively; MAE) are shown in Fig. 7. We see that the quality of the obtained estimates is low and that the errors of these "raw" crowdsourced estimates are too large to be of immediate use for any practical applications.

According to the original idea of the "wisdom of crowds" as introduced by Galton [24], we could also use the median of all collected guesses as the aggregate estimate w_{est}^{i}. Table 1 shows that the difference between both methods is marginal. Given the similar accuracy, we use the mean value as the aggregate estimate in the rest of the paper.

4.2 Dependence of errors on true measurements

From Figs. 6 and 7, it becomes clear that, while the estimates converge, the resulting w_{est} differ strongly from the ground truth labels w_{true}. Such behavior indicates that there is a systematic error associated with visual perception of weight and height. Figure 8 shows how the errors depend on the true measurements. We can see that deviations from average weight or height are largely underestimated, such that, e.g., the weight of light people

Figure 8 Dependence of the crowd's estimation errors on the true weight/height of the images. The *x*-axes appear quantized because the ground-truth data consists of integer numbers (pounds for weight, inches for height)

Figure 9 Crowd's estimation error vs. true BMI in images. Crowd workers are not asked to estimate BMI directly; instead, we compute the estimated BMI from the collected weight and height guesses
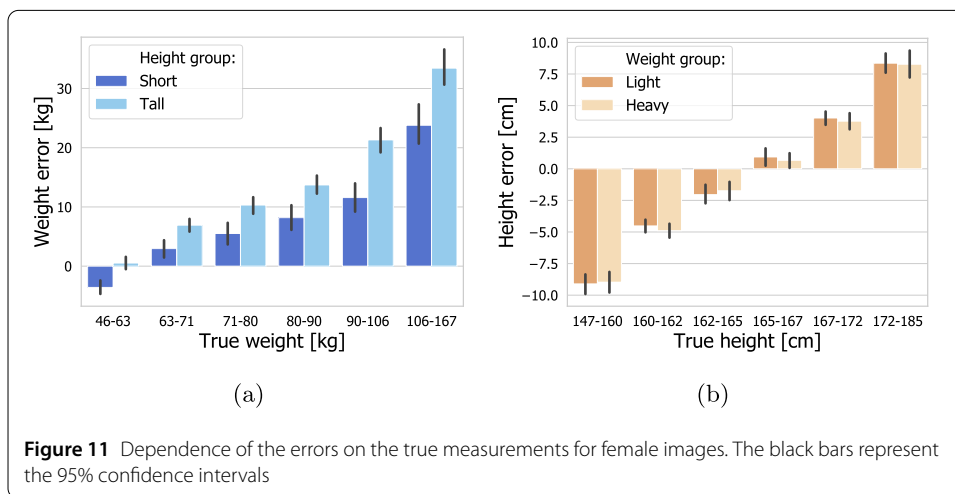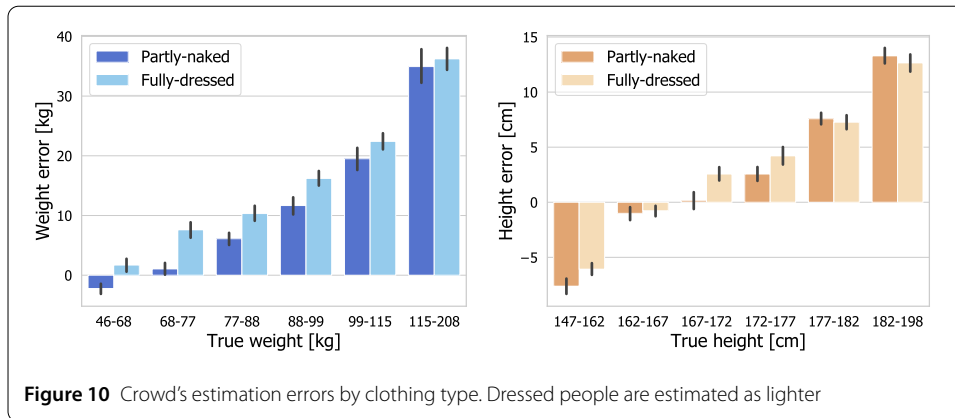
is overestimated (negative errors), whereas the weight of heavy people is underestimated (positive errors). This observation explains the large positive mean error that we observe in Fig. 7: most images in the dataset represent obese people whose weight is clearly above average (cf. Figs. 2 and 3). This is an important finding, as it implies that the BMI of overweight and obese people is strongly underestimated, as also confirmed by Fig. 9, which demonstrates that we cannot rely on humans' visual perception of weight in order to recognize obesity.

These findings are in agreement with the literature. The results of Winkler and Rhodes [28] suggest that people adapt their estimate to a *reference value* of weight and height determined by all the bodies that people have seen, with a particularly strong influence of the most recent observations. In simplified terms, we can view crowd workers' reference values as their idea of a normal human weight or height. Cornelissen et al. [16] suggest that systematic visual biases in the estimation of weight can be explained by combining the assumption about reference values with a phenomenon known as "contraction bias" [17], which shifts the estimated value towards the guesser's reference value. Thus, the trend in Fig. 8 can be explained by contraction bias if we assume that workers' reference values are close to the average human weight and height. In Sects. 5.2 and 5.3, we formulate a statistical, Bayesian model to derive individual reference values in a data-driven fashion, confirming this intuition.

Clearly, the estimates of the crowd workers are influenced by other factors apart from the true measurements of people in the images. For example, Fig. 10 shows how clothing

Figure 10 Crowd's estimation errors by clothing type. Dressed people are estimated as lighter

Figure 11 Dependence of the errors on the true measurements for female images. The black bars represent the 95% confidence intervals

can impact the guesses: while the height guesses do not depend on clothing, the weight of fully-dressed people is consistently estimated lower than the weight of similar (in terms of weight) partly-naked people. Exploring this and other possible factors goes beyond the scope of this paper, and in the following we mainly focus on the bias caused by contraction towards the reference value.

4.3 Dependencies between height and weight estimation

In our experiment, workers estimated both the weight and height of people shown in the images. This setup allows us to gain further insights into how people approach this task. In particular, we are interested in understanding whether a worker's weight estimate of a given person is also influenced by that person's height, rather than only by their weight (and analogously when swapping weight and height). Since weight and height are correlated, we need to stratify the data in order to answer this question, as follows. We first partition the set of all images into groups based on the quantiles of the weight distribution and further split each group into two subgroups: short and tall (i.e., height below vs. above group median). We then plot the errors of the aggregated estimates for each subgroup of images (Fig. 11(a)).[b] With a similar procedure, but partitioning first on height and then on the weight, we obtain Fig. 11(b).

Figure 11(b) shows that, while the height error grows linearly with the true height, there is no significant difference between the two weight groups within each height group.[c] The

Figure 12 Causal diagram [29] capturing the dependence of weight and height estimates on true weight and height. The dashed line between true weight and height signifies that the two are correlated
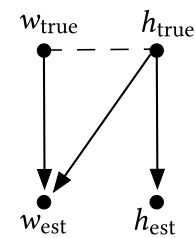
pattern in Fig. 11(a) is markedly different: in each weight group, the weight of tall people (whose height is larger than the group's median) is significantly more underestimated compared to the shorter half of the group.[d] This implies that height estimates are conditionally independent of the true weight, given the true height (Fig. 11(b)), whereas weight estimates depend on the true height even conditioned on the true weight (Fig. 11(a)). For clarity, Fig. 12 depicts this dependence structure as a causal diagram [29].

This finding has practical as well as theoretical implications. In practical terms, we shall later (Sect. 6.2) attempt to exploit the dependence of weight estimates on ground-truth height in order to improve weight estimates by supplying the ground-truth height (which is more stable and thus easier to obtain than the true weight) to workers at guessing time. In theoretical terms, the unidirectional dependence can help us hypothesize about the mental models at work during weight and height estimation. We will return to this point in Sect. 7.1.
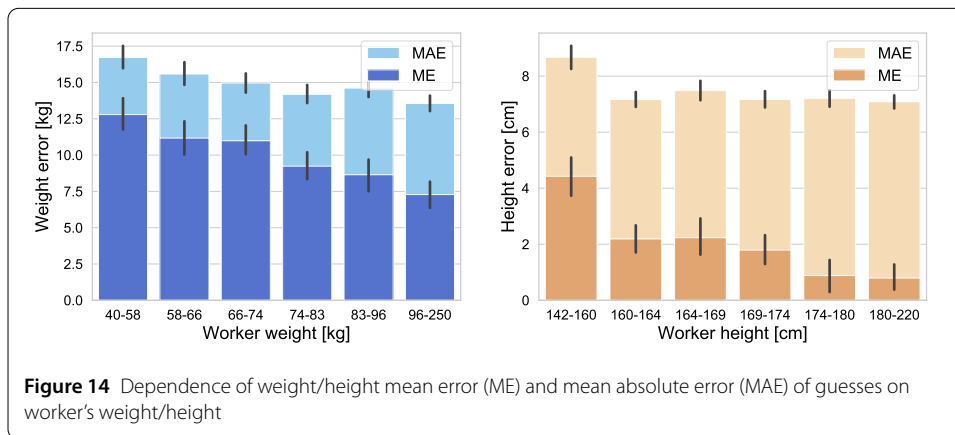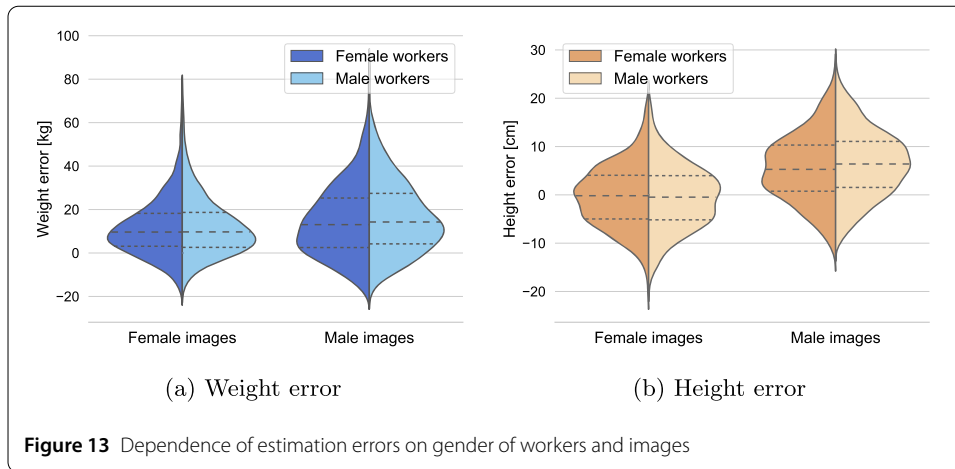
5 Effect of worker characteristics

So far, our discussion has focused on the results obtained by aggregating multiple guesses and did not take into account worker-specific characteristics. We now proceed to using the information collected from the workers in order to evaluate the impact of gender, age, country of residence, and, most important, the worker's own measurements on the produced guesses.

5.1 General observations

We begin our discussion by looking at how the gender of workers and of people in the images influences the estimates. One might assume that workers would guess the measurements of people of their own gender more accurately. But, as Fig. 13 shows, accuracy is essentially identical for male and female workers: after separate aggregation of guesses from male and female workers, the weight and height errors made by male workers follow similar distributions as those made by female workers, for both male and female images. (Student's *t*-tests yield $p = 0.023$ for weight errors and $p = 0.12$ for height errors; i.e., although the differences are minuscule, they may still be real, due to the large sample size.)

On the other hand, we can discern a clear difference in errors for male vs. female *images,* with considerably larger mean height and weight errors for male than for female images ($p = 1.7 \times 10^{-7}$ for weight errors and $p = 1.0 \times 10^{-79}$ for height errors, according to Student's *t*-tests), which is unexpected in the sense that the body-mass index (BMI) distributions are similar for males and females in our image data. It might, however, be explained by the fact that the height and weight is larger for male images, which are therefore more underestimated (cf. Fig. 8). Also, the weight error distribution (Fig. 13(a)) has a larger variance for male images, which may be explained by the larger variance of the true

(a) Weight error   (b) Height error

Figure 13 Dependence of estimation errors on gender of workers and images

Figure 14 Dependence of weight/height mean error (ME) and mean absolute error (MAE) of guesses on worker's weight/height

weight of male images in the dataset (standard deviation 25.6 kg for men, vs. 21.7 kg for women).

We next look at the dependence of the guesses on the workers' own body measurements. For this purpose, we split the workers into six bins based on their weight or height quantiles, and compute the mean error and the mean absolute error of all guesses provided by the workers from each bin. Figure 14 shows the relationships between the workers' weight/height and their errors. The fact that all mean errors (ME, dark bars) are positive implies that, regardless of their own weight and height, workers tend to underestimate weight and height in images. Additionally, we can see a clear monotonic decrease of the mean error as the worker's weight or height, respectively, decreases. This means that taller (heavier) workers tend to guess larger values for height (weight), leading to lower positive errors (i.e., less underestimation).

Figure 14 also shows a decrease in mean *absolute* error (MAE, light bars). *A priori*, this could be caused by heavier and taller workers being more accurate in general. However, this conclusion is invalidated by Fig. 15, which shows that heavier (taller) workers are more accurate than lighter (shorter) workers specifically on images of heavier (taller) people. We therefore conclude that the decrease of errors with increasing worker weight (Fig. 14) is simply an artefact of the overrepresentation of heavy images in our dataset. (Similar conclusions hold for height, although the performance difference between short and tall workers is less drastic.)

Figure 15 Dependence of weight/height mean absolute error (MAE) of guesses on true weight/height of the person in the image (*x*-axis) and worker's weight/height
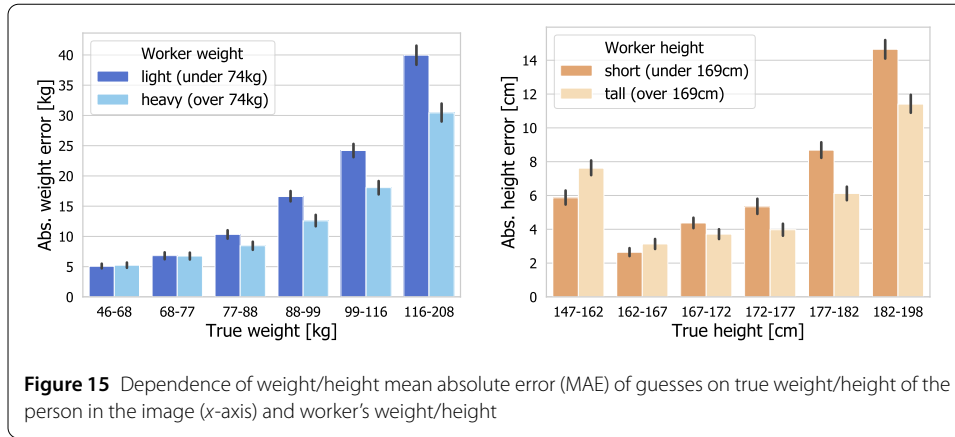
5.2 Model for inferring reference values

Based on the results from Sect. 5, in combination with the prior literature on reference values and contraction bias (Sect. 2), we hypothesize that heavier and taller workers have larger reference values. To support this hypothesis, we now develop a simple generative model that describes the process of guessing and is based on the notions of reference values and contraction bias introduced in Sect. 4. We describe our model for the case of weight; the case of height is fully analogous.

In our model, we assume that there are two factors that influence a worker's guess: the worker's personal reference value and the true weight of the person in the image. More specifically, a worker j produces a weight guess w_j^i for an image i with true weight w_{true}^i according to the following equation:

$$w_j^i = \alpha_j w_j^{\text{ref}} + (1 - \alpha_j) w_{\text{true}}^i + \epsilon, \tag{3}$$

where w_j^{ref} represents a worker-specific reference value towards which the worker's guesses are skewed, $\alpha_j \in [0, 1]$ represents the worker-specific contraction coefficient and describes the power of the contraction effect (larger α_j implies stronger averaging due to contraction bias), and ϵ denotes the error and accounts for all remaining complexity of the visual perception task, such as various view angles, body positions, etc. As the worker j is fixed in the following derivations, we drop the worker index j from here on for readability.

In order to infer the weight reference value w^{ref} and contraction coefficient α for a given worker, we use Bayes' rule (Eq. (4)) and leverage two independence assumptions: first, that a given worker's reference value and contraction coefficient are independent of the images shown to the worker (Eq. (5); this independence is implied by modeling the parameters as fixed for each worker); and second, that the reference value and the contraction coefficient are independent of one another (Eq. (7)):

$$p\big(w^{\text{ref}}, \alpha | w, w_{\text{true}}\big) = \frac{p(w|w_{\text{true}}, w^{\text{ref}}, \alpha)\, p(w^{\text{ref}}, \alpha | w_{\text{true}})}{p(w|w_{\text{true}})} \tag{4}$$

$$= \frac{p(w|w_{\text{true}}, w^{\text{ref}}, \alpha)\, p(w^{\text{ref}}, \alpha)}{p(w|w_{\text{true}})} \tag{5}$$

$$\propto p\big(w|w_{\text{true}}, w^{\text{ref}}, \alpha\big)\, p\big(w^{\text{ref}}, \alpha\big) \tag{6}$$

$$= p\big(w|w_{\text{true}}, w^{\text{ref}}, \alpha\big)\, p\big(w^{\text{ref}}\big)\, p(\alpha), \tag{7}$$

where, w is a vector that contain all guesses collected from the given worker, and w_{true} is a vector of the corresponding ground-truth labels.

We assume wide priors for reference values: $w^{\text{ref}} \sim \mathcal{N}(\mu, \sigma)$, where $\mu = 70$ kg is an estimate of average human weight (taking into account the workers' countries of residence), and where $\sigma \approx 25$ kg is a large standard deviation, such that all reasonable values of reference weight lie within one standard deviation from μ and more influence is given to the data evidence. As the power of the contraction effect is hard to estimate *a priori*, a uniform distribution on $[0, 1]$ is used: $\alpha \sim \mathcal{U}(0, 1)$. Finally, it is assumed that the errors follow a normal distribution: $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$, where $\sigma_\epsilon \approx 15$ kg is the empirical standard deviation of weight errors among the crowdsourced weight estimates.[e]

With these assumptions, the following distributions from Eq. (7) can be written explicitly:

$$p\big(w|w_{\text{true}}, w^{\text{ref}}, \alpha\big) = \prod_i \frac{1}{\sqrt{2\pi}\sigma_\epsilon} \exp\left(-\frac{(w^i - \alpha w^{\text{ref}} - (1-\alpha)w^i_{\text{true}})^2}{2\sigma_\epsilon^2}\right), \quad (8)$$

$$p\big(w^{\text{ref}}\big) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(w^{\text{ref}} - \mu)^2}{2\sigma^2}\right), \quad (9)$$

$$p(\alpha) = \begin{cases} 1 & \text{if } 0 \le \alpha \le 1, \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

To simplify the notation, from now on we assume that only the meaningful range $0 \le \alpha \le 1$ is considered. Inserting Eq. (8) and (9) into the main Eq. (7), we have

$$p\big(w^{\text{ref}}, \alpha|w, w_{\text{true}}\big) \propto \exp\left(-\frac{(w^{\text{ref}} - \mu)^2}{2\sigma^2}\right) \prod_i \exp\left(-\frac{(w^i - \alpha w^{\text{ref}} - (1-\alpha)w^i_{\text{true}})^2}{2\sigma_\epsilon^2}\right). \quad (11)$$

For computing the maximum *a-posteriori* (MAP) estimates of w^{ref} and α, we take logarithms and multiply with $-2\sigma_\epsilon^2$, thus obtaining the loss function

$$L\big(w^{\text{ref}}, \alpha\big) = \frac{\sigma_\epsilon^2}{\sigma^2}\big(w^{\text{ref}} - \mu\big)^2 + \sum_i \big(w^i - \alpha w^{\text{ref}} - (1-\alpha)w^i_{\text{true}}\big)^2. \quad (12)$$

The MAP estimates of w^{ref} and α are then obtained by minimizing L via gradient descent under the constraint $0 \le \alpha \le 1$.[f]

We emphasize that reference values w^{ref} and contraction factors α are computed solely based on the respective worker's guesses and the ground-truth weight of the images they provide guesses on, and not based on the worker's own weight. Next, we will show that, nonetheless, reference values correlate naturally with the worker's own weight.

5.3 Analysis of reference values

In this section, we analyze the reference values and contraction coefficients inferred in Sect. 5.2 in order to gain further insights into the crowd workers' biases during weight and height estimation.

The dependence of the reference values on the worker's own weight and height is shown in Fig. 16. The plots support our previous hypothesis that the reference values grow with
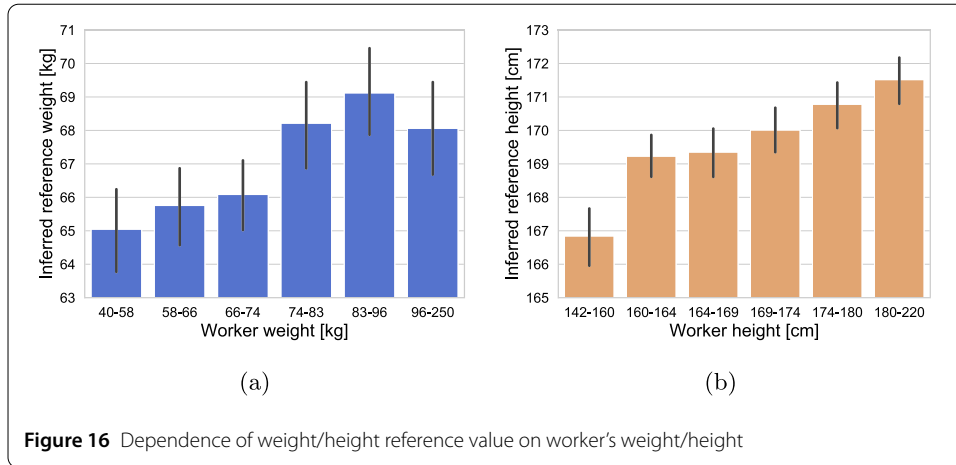
Figure 16 Dependence of weight/height reference value on worker's weight/height
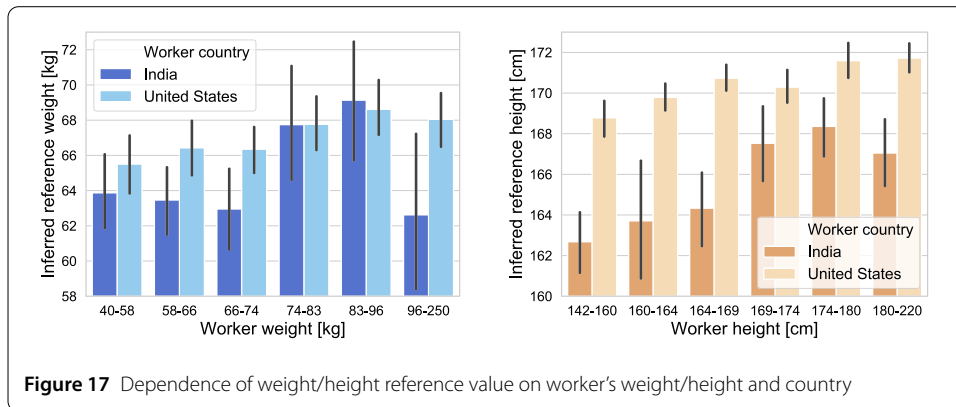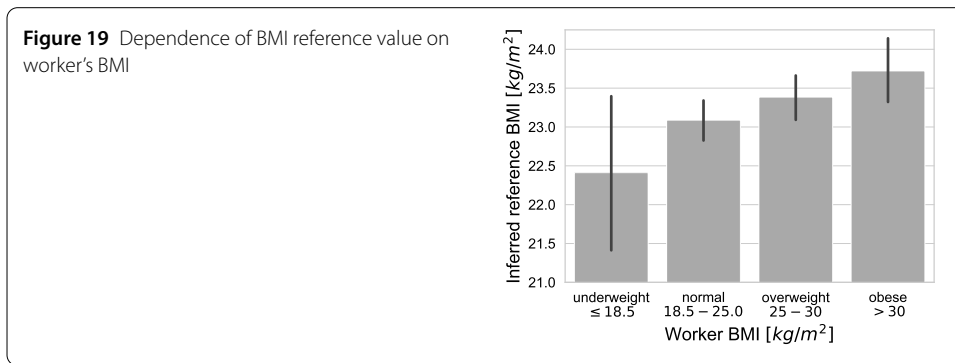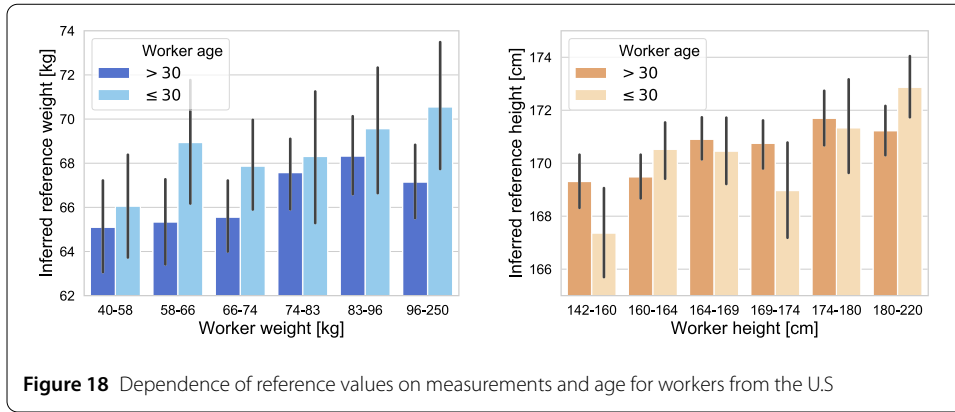

Figure 17 Dependence of weight/height reference value on worker's weight/height and country

the worker's own weight and height. Note how the dependence for weight breaks with the last bin in Fig. 16(a). It is possible that workers from this group provided wrong information (whether due to typos or on purpose) about their own weight, as we received suspiciously high values (up to 250 kg) from several workers.

Next, we look at differences in reference values depending on where the crowd workers are from. (Since about 90% of all our workers are from India or the U.S., we focus on these two countries.) As indicated by Fig. 17, the reference values for weight and especially height of workers from India are considerably smaller than those of workers from the U.S. This fact may be explained by the large difference in height between residents of both countries: the average height of both men and women from India is about 10 cm smaller compared to the U.S.[g] Thus, Fig. 17 shows how both the worker's own characteristics and the standards of their society impact the reference values.

The reference values vary not only with geography, but also with the worker's age. Figure 18 demonstrates this dependence for workers from the U.S. While the height values vary rather randomly among both groups, there is a clear trend for the weight values: younger workers have higher reference values in all weight groups. A possible explanation for this trend could be the increasing obesity rates in the U.S. [20, 30], due to which young people build their reference values while growing up in a society where obesity becomes more common, thus shifting their perception of normal weight towards higher values. The

Figure 18 Dependence of reference values on measurements and age for workers from the U.S

Figure 19 Dependence of BMI reference value on worker's BMI

Figure 20 Dependence of weight/height contraction coefficients on worker's weight/height

lack of such a trend for height, rather than weight, may be explained by the fact that, across time, height distributions are much more stable than weight distributions [31–33].

Finally, as the dependence of the reference values on workers' own measurements holds for both weight and height, it is also reflected in the BMI: Fig. 19 shows that the reference BMI increases together with the worker's own BMI.

Apart from the reference values, the model from Sect. 5.2 also contains the contraction coefficient α. Figure 20 shows the coefficients for various worker weight and height groups. The contraction coefficient relates to the quality of the estimation: an α close to 1 mean that the worker tends to guess their reference weight/height most of the times,

whereas a small α indicates that the guesses closely follow the true labels. From Fig. 20(b) we can observe that α_{height} stays roughly constant across groups, whereas Fig. 20(a) shows that α_{weight} decreases with increasing worker weight (and therefore with increasing reference weight). As previously discussed (Sect. 3.1, Figs. 4 and 5), this is mainly caused by the specific dataset, which contains a lot of images of obese people and thus favors high reference values for weight. When the data is downsampled to represent the distribution of weight in the general population, the differences between groups become negligible. Another observation is that α_{height} is in general larger than α_{weight}, which means that the averaging effect is stronger for height estimation.

6 Towards more accurate crowdsourcing

The main contribution of this paper is to study human biases in height and weight estimation using crowdsourced data. Moving beyond this scientific goal, if crowdsourced height and weight labels were accurate, they could potentially also be harnessed for engineering practically useful applications. For instance, for monitoring obesity levels across time and space, it would be tremendously useful to have access to low-cost, low-latency estimates of population height and weight. Such "sensors" could be built by feeding representative images sampled from a population (e.g., via social media) to a crowdsourcing system. The data thus collected could further be used to train fully automated height and weight models based on machine learning [8].

As we saw in the previous sections, the labels collected using our simple crowdsourcing method are rather inaccurate. The present section explores simple ways of improving the accuracy, with the goal of establishing whether crowdsourcing is a promising candidate not only for measuring human biases (the main contribution of this paper), but also for collecting high-quality labels for downstream tasks.

We proceed in two directions: first, via statistical corrections to remove human biases (Sect. 6.1), and second, by giving crowd workers access to more information when guessing (Sects. 6.2 and 6.3). Neither of these two directions was found to increase crowd workers' accuracy sufficiently in order to directly enable a crowd-powered obesity monitoring system as sketched above. We shall reflect further on this negative result in Sect. 7.3, where we discuss lessons learned from the present research for potential future crowdsourcing and machine learning applications.

6.1 Statistical correction models

We are interested in the performance of crowdsourcing and correction models on images representing the general population, whereas the dataset contains a disproportional number of images of obese people. Hence, we downsample the data to 500 images in such a way that the distribution of weight among workers and samples becomes similar. Furthermore, for performance evaluation, the set is split into 400 training and 100 validation images.

Correction via reference values   First, we try to use the previously derived reference values for correction, by rewriting Eq. (3) as

$$w_{\text{true}}^i = \frac{w_j^i - \alpha_j w_j^{\text{ref}} - \epsilon}{1 - \alpha_j}. \tag{13}$$

Under the assumption that the noise ϵ follows a zero-centered normal distribution, we obtain a maximum *a-posteriori* (MAP) estimate of the ground-truth label w_{true}^i by evaluating Eq. (13) with $\epsilon = 0$. This MAP estimate represents a corrected weight label. Tables 2 and 3 show that this approach leads to poor results. The main problem is the poor performance of several workers whose $\alpha_j \approx 1$ makes the denominator of Eq. (13) very small, which in turn amplifies the noise in the guess w_j^i.

Linear regression   Next, we introduce several linear regression models that differ with respect to the features they use. The output of each model is a corrected weight or height estimate. The following models are considered:

- Global: a single model fitted after aggregating all individual guesses for the same image via the mean. The features we consider include the mean of all collected weight and height guesses (MW and MH, respectively) and the gender of the person in the image (GN).
- Per-worker: a separate model for each worker, taking as inputs individual guesses (i.e., each model learns different worker-specific parameters). The features we consider include the estimated weight and height (EW and EH, respectively) and the gender of the person in the image (GN). The outputs of all worker-specific models are averaged to get a single corrected weight label for each image.
- Mixed: a global correction model for individual guesses that also takes into account worker parameters. The inputs are individual weight and height estimates, the gender of the person in the image, as well as the worker's weight, height, and gender. Again, the average of all collected guesses is used as a final corrected estimate for each image.

The performance of raw crowdsourcing and different correction models is summarized in Tables 2 and 3.[h] For space reasons, we focus on weight (Table 2) in our discussion. The best performance on the test set is obtained by the global correction models. We believe that these models can better leverage the wisdom of crowds and are thus more

Table 2 Correction models for weight estimation

Model type	MAE per image train/test [kg]	MAE per guess train/test [kg]
raw guesses	8.83 / 8.65	11.39 / 11.38
ref. val. correction	10.04 / 10.21	14.14 / 15.40
global: MW	8.44 / 7.43	–
global: MW, MH, GN	7.98 / 7.02	–
worker: EW	7.46 / 7.41	9.15 / 10.07
worker: EW, EH, GN	6.38 / 7.58	8.19 / 10.94
mixed	8.81 / 7.99	10.61 / 10.16

Table 3 Correction models for height estimation

Model type	MAE per image train/test [cm]	MAE per guess train/test [cm]
raw guesses	5.28 / 4.91	7.01 / 6.82
ref. val. correction	6.99 / 6.42	12.05 / 12.21
global: MH	5.24 / 4.70	–
global: MH, MW, GN	4.91 / 4.64	–
worker: EH	4.93 / 5.14	5.50 / 5.93
worker: EH, EW, GN	4.05 / 4.93	4.40 / 5.81
mixed	5.21 / 5.09	5.24 / 5.13

suited for crowdsourcing tasks, because the global models work on the aggregated guesses of multiple workers, whereas the worker-specific models try to correct single estimates that are more noisy by their nature. In particular, the worker-specific models fit both the relevant signal and the noise of separate guesses, and do not generalize well (as can be seen when comparing training and testing errors: the difference between the two is smaller for the worker-specific models). An interesting effect is observed for the mixed models: while the error of single estimates decreases compared to the raw guesses, this improvement is lost once the corrected guesses are aggregated. For example, on the training set, the MAE of single estimates drops from 11.39 to 10.61 after correction (Table 2), but the MAE for images remains almost the same.

Limits of correction   The simple correction models discussed so far demonstrate that correction of the collected estimates poses a challenging task. In this section, we provide evidence that more sophisticated statistical models would also have limitations resulting from the collected data.

The initial dataset contains pairs of images ("before" and "after") and weight labels (cf. Sect. 3.1). Each of the two weight labels in a pair needs to be assigned to one of the two images in the pair. An obvious solution, discussed in Sect. 4, is to aggregate the guesses (e.g., by averaging, as we did throughout the paper), and to assign the larger ground-truth label to the image with the higher aggregated value.

Though the resulting assignment is correct for almost all samples, it is wrong in some rare cases. Figure 1(a) contains an example of a wrong assignment, for an image pair capturing a particularly large weight transformation. It shows how after a major weight loss of almost 19 kg the person is labeled as slightly heavier than before: the mean of "before" guesses is 79.4 kg, whereas the mean of "after" guesses is 81 kg. Furthermore, we observed that the guesses collected for both images follow similar distributions, and even come from similar pools of workers (in terms of the workers' own weight and height distributions). Under such conditions, no statistical model can correct the results. The important difference here is purely visual: the images are taken from different angles and under very different circumstances (most importantly, fully dressed vs. bare-chested), which makes the task unequally hard across images.

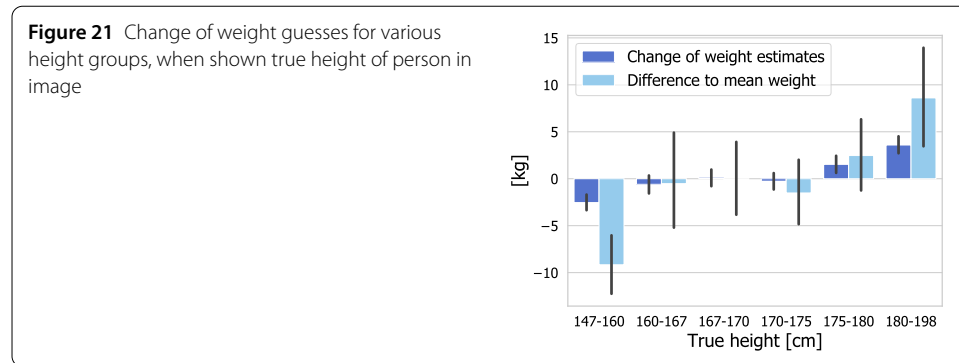## 6.2 Crowdsourcing variation 1: guessing weight for known height

Above, we showed that correcting the estimates *post hoc* is challenging. Next, we suggest two different setups that could simplify the task for crowd workers and lead to more accurate results. The experiments in these modified setups are again performed on the reduced set of 500 images that were chosen to represent the general population, as discussed in Sect. 6.1.

Given that height does not change significantly during adulthood, whereas weight may change widely and might therefore have to be frequently re-estimated for the same person, one might provide workers with the true height labels for the images and collect only weight guesses.

For this setup, we apply an additional filter based on workers' countries of residence: as previously shown, reference values of height for people in Asian countries do not match the dataset at hand. Thus, providing the actual height of the person in the image can further confuse such workers (indeed, the quality of their guesses slightly decreases when

Table 4 Comparison of weight-guessing results when showing vs. not showing true height labels to workers, with bootstrapped 95% confidence intervals (only workers from Europe and U.S)

	Height not given	Height given
ME [kg]	4.54 (3.54, 5.57)	4.22 (3.24, 5.23)
MAE [kg]	8.70 (7.96, 9.46)	8.47 (7.73, 9.21)

Figure 21 Change of weight guesses for various height groups, when shown true height of person in image

the true height is revealed). Hence, we focus on the performance of crowd workers from Europe and the U.S. The results in Table 4 summarize the error distributions for both scenarios (i.e., with and without height given), indicating only small improvements of the mean error (ME) and mean absolute error (MAE) for the new setup. Moreover, the improvements are not significant: both a t-test for the mean and a Bartlett test for the variance show no significant difference between the distributions of errors for two scenarios ($p = 0.66$ and 0.72 respectively).
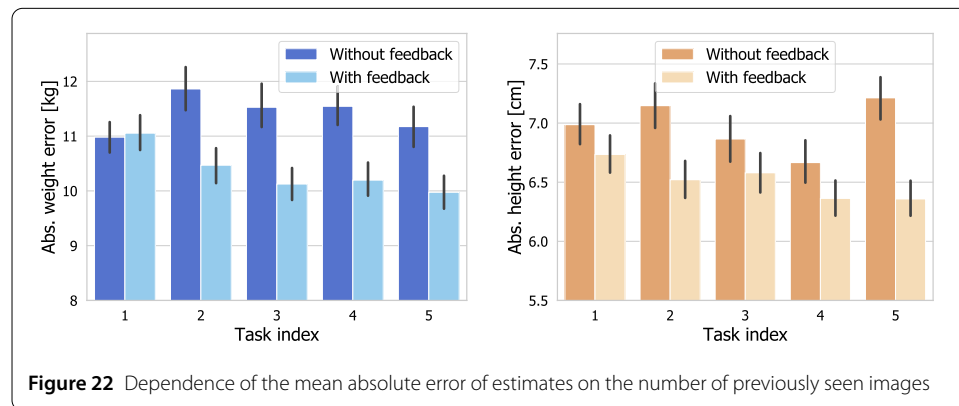
The comparison in Fig. 21 shows how the weight estimates are affected by the known true height (dark bars) and how the true weight differs from the overall average of the true weight (across all images in the dataset; light bars) as a function of the true height. The figure provides a possible explanation for the small performance improvement. First, while workers generally shift their estimates into the right direction (add or subtract a few kilograms based on the actual height), the magnitude of this shift seems to be too small, in particular for the shortest and tallest images. Furthermore, the bars that show difference to the mean weight have wide confidence intervals and, thus indicate large variations of weight within each group. Indeed, the growth of weight as a function of height, as depicted in Fig. 21, appears only after aggregating a large number of images. The real relationship between weight and height is complicated and nonlinear [34]. This is further supported by statistics of the dataset at hand: the correlation between weight and height is weak according to both Pearson's (0.27) and Spearman's (0.28) correlation coefficient.

6.3 Crowdsourcing variation 2: guessing weight and height with feedback
Next, we consider a second variation: instead of making the task itself simpler, we suggest a design that gives workers an opportunity to learn, by providing immediate feedback to the workers after each guess. The feedback contains the true weight and height of the person in the image, as well as the errors of their guess. Of course, if we already knew the ground truth for every image, this would defeat the purpose of collecting it via crowdsourcing. We might, however, have a few labeled images that could be used in this setup for teaching workers early on.

Table 5 Comparison of weight- and height-guessing results with vs. without feedback on the previous guess, with bootstrapped 95% confidence intervals (only workers from Europe and U.S)
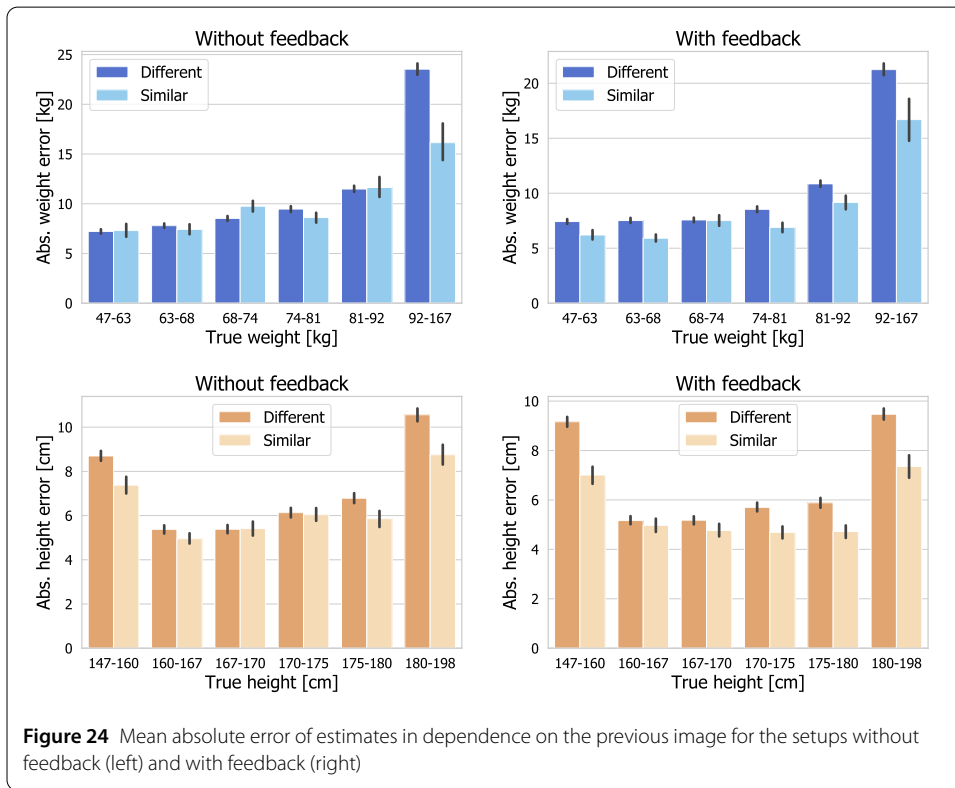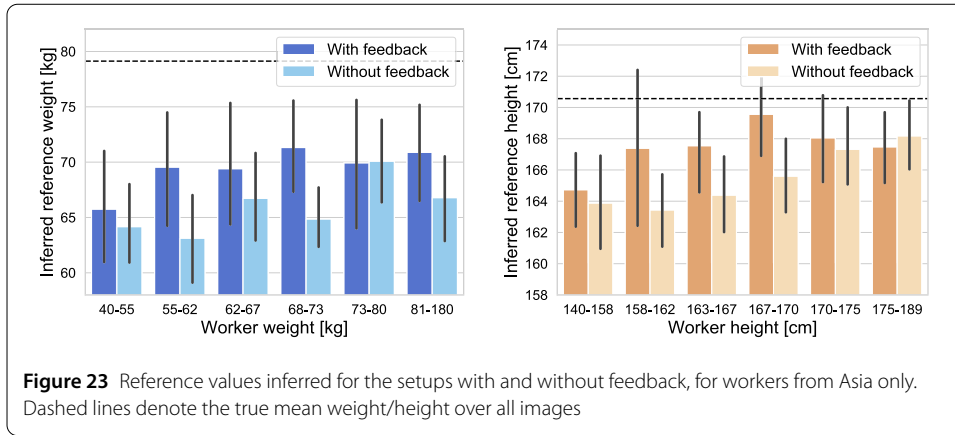
	Without feedback	With feedback
Weight ME [kg]	4.90 (3.90, 5.95)	3.25 (2.33, 4.24)
Weight MAE [kg]	8.80 (8.05, 9.57)	8.11 (7.48, 8.82)
Height ME [cm]	0.47 (−0.12, 1.07)	−0.26 (−0.8, 0.34)
Height MAE [cm]	5.20 (4.86, 5.55)	5.10 (4.76, 5.45)

Figure 22 Dependence of the mean absolute error of estimates on the number of previously seen images

The results in Table 5 show that the accuracy slightly improves in the setup where workers receive feedback. The values in Table 5 were computed after aggregating all guesses per image, and the improvements appear to be minor. However, the quality of individual guesses (before aggregating per image) improved significantly with feedback: the mean absolute error for weight decreased from 11.4 kg in the basic setup to 10.4 kg in the feedback setup, and similarly for height, where it decreased from 7.0 cm to 6.5 cm ($p < 10^{-17}$ according to Student's t-tests, for both errors and absolute errors). Further evidence of the increase in accuracy due to training effects can be seen in Fig. 22. Here, single estimates are split into bins according to the number of preceding guesses (and, thus, instances of feedback received). In this way, the estimates in later bins are produced by more "experienced" workers who have previously received more feedback. The plots in Fig. 22 show how the MAE of the weight and height guesses decreases with progressing bin indices in the experiment with feedback, while no improvement is observed for the initial setup.

The benefits of training can also be observed when we look at the change of the reference values. For this purpose, it is interesting to consider crowd workers from Asian countries (mostly India). As previously discussed, their reference values tend to deviate from the reference values of Europeans and Americans, and thus hinder their performance on the given dataset. Figure 23 shows how the reference values for both weight and height are changing in the experiment with feedback. Since the mean weight and height of the images from the dataset (shown with a dashed line) exceeds all initial reference values, crowd workers shift their references values towards higher weights and heights as they learn these new standards and adapt to the dataset. On the other hand, we did not observe any considerable change of the reference values for European or American workers.

The previous discussion was centered around long-term training effects, i.e., performance improvement as a result of feedback on multiple previous guesses. Now, we focus on the short-term effects of getting feedback. Figure 24 shows how the MAE depends on

Figure 23 Reference values inferred for the setups with and without feedback, for workers from Asia only. Dashed lines denote the true mean weight/height over all images

Figure 24 Mean absolute error of estimates in dependence on the previous image for the setups without feedback (left) and with feedback (right)

the previous image seen during a task (remember that each task consisted of 10 images shown one after the other). In particular, we distinguish between two cases:

- Similar: the previous image shows a person of the same gender, whose weight (height) differs from the weight of the person in the current image by at most 7.5 kg (7.5 cm).
- Different: the previous image does not satisfy the above condition.

Figure 24 shows how the MAE depends on the previous image across weight and height groups, and compares the two versions of the experiment (with vs. without feedback). Clearly, the impact of the previous image is marginal when no feedback is shown: for most groups, the MAE remains roughly the same for similar and different images. The previous image, however, appears to be important when workers start to receive feedback: in almost

all cases, the average accuracy of the guesses improved when the new image was similar to the previous one.

7 Discussion

We conclude the paper by discussing relations to prior work (Sect. 7.1; also cf. Sect. 2), limitations of our methodology (Sect. 7.2), and implications and best practices for crowd-sourcing and machine learning tasks involving body measurements (Sect. 7.3).

7.1 Relation to prior work

We observe that crowd estimates are skewed toward an intermediate value close to the population-wide average, a phenomenon known as *contraction bias* [17]. This bias has been shown to in particular lead to the underestimation of overweight and obese bodies [16], which may in turn compromise people's ability to recognize weight gain and undertake compensatory weight-control behaviors.

Taking a closer look, we observe that the contraction does not, however, skew all estimates to a global constant, but rather to the typical height and weight of the environment in which the respective crowd worker lives (Fig. 17), as well as to their own height and weight (Fig. 16). We confirm this effect via a simple model, which elicits individual workers' reference values (Sect. 5.2) from their estimates in combination with the ground-truth height and weight of the images they judge. The fitted model parameters indicate, e.g., that workers from India have lower reference heights and weights than workers from the U.S., reflecting lower average measurements in India. Similarly, older U.S. workers have lower reference weights than younger U.S. workers, potentially caused by an increasing American average weight. This echoes the claims of *visual normalization theory* [2, 6, 7], which states that weight status is judged relative to body-size norms prevalent in society, leading to the systematic misestimation of certain body shapes.

We also provide evidence that users anchor their weight estimates in height estimates, but less so in the reverse direction (Fig. 11). Regarding height, prior research has shown that estimates are also heavily anchored in facial features [4, 5] and head-to-shoulder ratio [18].

In his seminal 1907 paper about human accuracy at guessing the weight of an ox [24], Francis Galton wrote: "I have not sufficient knowledge of the mental methods followed by those who judge weights [. . .]." Here Galton raises the question of mental models. Although our work is not primarily concerned with eliciting mental models, our findings can lead the way to a deeper scrutiny of this aspect. For instance, in Sect. 4.3 we showed that height estimates are independent of true weight for a fixed true height, whereas weight estimates still depend on true height even for a fixed true weight (cf. Fig. 12). This asymmetry might hint at a mental process where weight and height estimation are performed sequentially, rather than simultaneously. Future work should investigate the question of mental models further, e.g., with experiments for discerning whether height and weight estimation indeed happen sequentially.

7.2 Limitations

Our study is limited in several ways. First, as our dataset of weight- and height-labeled images was collected from an online forum that is primarily concerned with weight loss, the data is biased towards heavier people. Additionally, the crowd workers who provided

estimates came primarily from the United States and India, leading to a biased sample of the global population. Future work should therefore validate our findings on further datasets and with further worker demographics.

The above two sources of bias, stemming from images and workers, respectively, are distinct from one another. Where required, we circumvented this issue by subsampling images to reflect the demographics of the worker population.

A final limitation stems from the fact that users posting their images and measurements on Reddit, as well as crowd workers, may in principle have indicated their own weight and height incorrectly, e.g., due to social desirability, lacking information on true measurements, or sheer malice. While we cannot rule out such behavior, neither users labeled in images nor crowd workers had an obvious incentive to wilfully act this way. Nevertheless, our findings could be strengthened by follow-up studies with images and estimates collected in a more controlled environment, e.g., where researchers themselves measure the height and weight of both estimated and estimating participants.

7.3 Implications for crowdsourcing and machine learning

Based on the studies of this paper, numerous lessons can be drawn that can inform the design of future crowdsourcing and machine learning systems for height and weight estimation. We conclude this section by summarizing these lessons.

The cause for the rather low crowd accuracy (Sect. 4) is not that workers' estimates are widely dispersed; in fact 20 to 30 independent estimates suffice for the mean to converge (Fig. 6). The issue is rather that estimates are systematically off due to the contraction bias toward workers' reference values. Hence, if the images to be annotated are known to be drawn from a certain subpopulation, workers should be recruited to match that subpopulation. If, on the contrary, the images depict bodies from a wide variety of backgrounds, one should strive to assemble a worker pool reflecting that variety.

To raise the accuracy, we explored the option of facilitating weight estimation by providing workers with known height labels, finding, however, that this makes guessing the weight only mildly easier (Table 4), presumably because the relationship between height and weight is nonlinear and complex [34], such that workers cannot accurately incorporate the height information (Fig. 21).

What does help, on the other hand, is giving workers feedback on their previous guess. We therefore conclude that it is advisable to train workers on a set of ground-truth images if available. Our results also indicate that workers are slightly more accurate when subsequent images depict people of similar measurements. The latter finding points to a promising direction: instead of simply sampling a random image at every step, future work should design more intelligent crowd algorithms for cleverly assigning images to the workers most likely to make accurate guesses, given their personal background and task history. Starting from random assignments, such algorithms could learn more about workers with every guess and could later on route tasks more intelligently [35, 36].

Orthogonally, we explored ways to correct the contraction bias in a postprocessing step (Sect. 6.1). In line with previous work [21, 22], accounting for worker height and weight allows us to significantly reduce the error in individual workers' estimates, but interestingly, these gains are lost once we aggregate all workers' estimates, such that simply averaging raw guesses and shifting and scaling the result by global constants achieves better results than first correcting on a per-worker basis and then averaging (Tables 2 and 3). We

conclude that aggregating many potentially diverse guesses ("wisdom of crowds") is more essential than correcting individual guesses. Overall, bias correction remains challenging, with even our best models lowering the error only marginally, from 8.7 kg to 7.0 kg (Table 2).

This negative result has important practical implications. Given the large errors, we expect a crowd-labeled image set to be too noisy for training accurate machine learning models, even with correction. Human performance is, of course, not necessarily an upper bound for machine performance. It is well conceivable that machine learning algorithms trained on ground-truth, rather than crowdsourced, height and weight labels may surpass human performance.

That said, noisy, crowdsourced labels might still become useful via transfer learning [37]: deep learning needs large amounts of training data, so one may start by training a weak model with many noisy labels and then fine-tune it with more accurate training data. We hope that future work will build on our insights to develop better health monitoring solutions based on crowdsourcing and machine learning.

8 Conclusion

We investigated human performance at estimating body measurements by using a novel large-scale dataset of height- and weight-labeled images from the Web as input to an estimation task deployed to a diverse set of human guessers via crowdsourcing. We find that human estimates are overall of low accuracy, with mean absolute errors of 15.5 kg for weight and 6.3 cm for height (8.8 kg and 5.2 cm, respectively, when subsampling images to represent the height and weight distributions among participating crowd workers). Estimates are biased in distinct ways, such that errors vary systematically with properties of both the estimated and the estimating person. Future work should extend this research by shedding further light on the mental mechanisms at play during estimation and by building tools for improving people's weight awareness.

Availability of data and materials
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
KM conducted all the analysis. KM and KG collected the data. KM, KG, and RW designed the analysis and experiments and wrote the manuscript. KM and KG performed the work mostly while at EPFL. All authors read and approved the final manuscript.

Author details
[1]Google, Munich, Germany. [2]Massachusetts Institute of Technology, Cambridge, MA, USA. [3]EPFL, Lausanne, Switzerland.

Endnotes
[a] https://www.reddit.com/r/progresspics

[b]   Since males and females differ with respect to their weight and height distributions, we perform this analysis separately for each gender. Figure 11(a) pertains to females. Similar results are obtained for males, but we omit them for space reasons.

[c]   Student's t-tests yield the following p-values for the null hypothesis of no difference in mean height errors between the light and heavy weight groups for fixed true height (from left to right in Fig. 11(b)): 0.81, 0.34, 0.56, 0.55, 0.56, 0.90.

[d]   Student's t-tests yield the following p-values for the null hypothesis of no difference in mean weight errors between the short and tall height groups for fixed true weight (from left to right in Fig. 11(a)): 2.1×10^{-6}, 5.9×10^{-5}, 7.7×10^{-5}, 7.0×10^{-5}, 1.3×10^{-8}, 8.9×10^{-5}.

[e]   For height, rather than weight, we use $\mu = 170$ cm, $\sigma = 15$ cm, $\sigma_\epsilon = 10$ cm.

[f]   As a technical detail, we note that, although L is convex for each variable separately, it is not convex in $(w^{\mathrm{ref}}, \alpha)$ jointly, due to the product αw^{ref}. Gradient descent is hence not guaranteed to find a global, but merely a local, minimum. As a sanity check, we hence also minimized L by splitting the α interval $[0, 1]$ into 100 pieces and performing constrained gradient descent on each piece separately, starting from 5 random initializations per piece. Comparing to the results from vanilla gradient descent on the full interval $\alpha \in [0, 1]$, the loss improved for only 1 out of about 300 workers, so we conclude that vanilla gradient descent is good enough for this setting.

[g]   https://en.wikipedia.org/wiki/List_of_average_human_height_worldwide

[h]   Errors are lower than in Table 1 because the dataset used here was downsampled to match the worker population and thus contains fewer overweight images.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References
1. WHO Multicentre Growth Reference Study Group (2006) WHO child growth standards based on length/height, weight and age. Acta Pædiatr 95:76–85
2. Robinson E, Kirkham T (2014) Is he a healthy weight? Exposure to obesity changes perception of the weight status of others. Int J Obes 38(5):663
3. Lanza HI, Echols L, Graham S (2012) Deviating from the norm: body mass index (BMI) differences and psychosocial adjustment among early adolescent girls. J Pediatr Psychol 38(4):376–386
4. Coetzee V, Perrett DI, Stephen ID (2009) Facial adiposity: a cue to health? Perception 38(11):1700–1711
5. Burton CM, Rule NO (2013) Judgments of height from faces are informed by dominance and facial maturity. Social Cogn 31(6):672–685
6. Robinson E, Hogenkamp PS (2015) Visual perceptions of male obesity: a cross-cultural study examining male and female lay perceptions of obesity in caucasian males. BMC Public Health 15(1):492
7. Robinson E (2017) Overweight but unseen: a review of the underestimation of weight status and a visual normalization theory. Obes Rev 18(10):1200–1209
8. Weber I, Mejova Y (2016) Crowdsourcing health labels: inferring body weight from profile pictures. In: Proceedings of the 6th international conference on digital health conference. ACM, New York, pp 105–109
9. BenAbdelkader C, Yacoob Y (2008) Statistical body height estimation from a single image. In: Automatic face & gesture recognition, 2008. FG'08. 8th IEEE international conference on. IEEE Press, New York, pp 1–7
10. Dey R, Nangia M, Ross KW, Liu Y (2014) Estimating heights from photo collections: a data-driven approach. In: Proceedings of the second ACM conference on online social networks. ACM, New York, pp 227–238
11. Günel S, Rhodin H, Fua P (2018) What face and body shapes can tell about height. arXiv preprint. arXiv:1805.10355
12. Jiang M, Guo G (2019) Body weight analysis from human body images. IEEE Trans Inf Forensics Secur 14(10):2676–2688
13. Kocabey E, Camurcu M, Ofli F, Aytar Y, Marin J, Torralba A, Weber I (2017) Face-to-bmi: using computer vision to infer body mass index on social media. In: ICWSM
14. Wen L, Guo G (2013) A computational approach to body mass index prediction from face images. Image Vis Comput 31(5):392–400
15. Coe TR, Halkes M, Houghton K, Jefferson D (1999) The accuracy of visual estimation of weight and height in pre-operative supine patients. Anaesthesia 54(6):582–586
16. Cornelissen KK, Gledhill LJ, Cornelissen PL, Tovée MJ (2016) Visual biases in judging body weight. Br J Health Psychol 21(3):555–569
17. Poulton EC, Poulton S (1989) Bias in quantifying judgements. Taylor & Francis, London
18. Mather G (2010) Head–body ratio as a visual cue for stature in people and sculptural art. Perception 39(10):1390–1395
19. Yom-Tov E (2016) Crowdsourced health: how what you do on the Internet will improve medicine. MIT Press, Cambridge
20. Hansen AR, Duncan DT, Tarasenko YN, Yan F, Zhang J (2014) Generational shift in parental perceptions of overweight among school-aged children. Pediatrics 134(3):481–488
21. Brettschneider A-K, Rosario AS, Wiegand S, Kollock M, Ellert U (2015) Development and validation of correction formulas for self-reported height and weight to estimate bmi in adolescents. Results from the KiGGS study. Obes Facts 8(1):30–42
22. Pérez A, Gabriel KP, Nehme EK, Mandell DJ, Hoelscher DM (2015) Measuring the bias, precision, accuracy, and validity of self-reported height and weight in assessing overweight and obesity status among adolescents using a surveillance system. Int J Behav Nutr Phys Act 12(1):2
23. Martynov K, Garimella K, West R (2020) Darks and stripes: effects of clothing on weight erception. J Soc Comp 1(1)

24. Galton F (1907) Vox populi. Nature 75(7):450–451
25. Bevelander KE, Kaipainen K, Swain R, Dohle S, Bongard JC, Hines PD, Wansink B (2014) Crowdsourcing novel childhood predictors of adult obesity. PLoS ONE 9(2):87756
26. Eickhoff C (2018) Cognitive biases in crowdsourcing. In: Proceedings of the eleventh ACM international conference on web search and data mining. WSDM '18. ACM, New York, pp 162–170
27. Lorenz J, Rauhut H, Schweitzer F, Helbing D (2011) How social influence can undermine the wisdom of crowd effect. Proc Natl Acad Sci USA 108(22):9020–9025
28. Winkler C, Rhodes G (2005) Perceptual adaptation affects attractiveness of female bodies. Br J Psychol 96(2):141–154
29. Pearl J (2009) Causality. Cambridge University Press, Cambridge
30. Mitchell NS, Catenacci VA, Wyatt HR, Hill JO (2011) Obesity: overview of an epidemic. Psychiatr Clin North Am 34(4):717–732
31. (NCD-RisC), NRFC (2016) A century of trends in adult human height. eLIFE
32. Burke MA, Heiland F (2007) Social dynamics of obesity. Econ Inq 45(3):571–591
33. Hammond RA, Ornstein JT (2014) A model of social influence on body mass index. Ann NY Acad Sci 1331(1):34
34. Diverse Populations Collaborative Group (2005) Weight–height relationships and body mass index: some observations from the diverse populations collaboration. Am J Phys Anthropol 128(1):220–229
35. Shahaf D, Horvitz E (2010) Generalized task markets for human and machine computation. In: AAAI
36. Zhang H, Horvitz E, Chen Y, Parkes DC (2012) Task routing for prediction tasks. In: Proceedings of the 11th international conference on autonomous agents and multiagent systems. International foundation for autonomous agents and multiagent systems, pp 889–896
37. Oquab M, Bottou L, Laptev I, Sivic J (2014) Learning and transferring mid-level image representations using convolutional neural networks. In: The IEEE conference on computer vision and pattern recognition (CVPR)