EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

ECON JOURNAL WATCH 18(1)
March 2021: 95–126

Against Standard Deviation as a Quality Control Maxim in Anthropometry

Austin Sandler[1]

LINK TO ABSTRACT

Anthropometry is the study of the measurements and proportions of the human body. It is widely accepted that for practical purposes anthropometry is the most useful tool for assessing the malnutrition status of children (WHO 1986). Malnutrition is responsible for 45 percent of all deaths among children worldwide (Black et al. 2013). In 2017, acute malnutrition (wasting) menaced over 50 million young children while over 150 million young children suffered from chronic malnutrition (stunting) (UNICEF, WHO, and the World Bank 2018). Even a small change in child malnutrition rates can have major consequences in terms of lives saved or lost. The financial and human costs associated with the practice of anthropometry can be enormous. In 2014 alone, global donors disbursed nearly $937 million in nutrition-specific programing (KFF 2016). According to Meera Shekar et al. (2017), to achieve the World Health Assembly global nutrition targets, the world needs to invest $70 billion over 10 years in high-impact nutrition-specific interventions.

The two most widely studied expressions of anthropometric indices are weight-for-height (WHZ) and height-for-age (HAZ) z-scores (de Onis and Habicht 1996; UNICEF et al. 2018). These z-scores express anthropometric measurements in terms of standard deviations below or above a reference population value. A z-score is the difference between a particular child's measure-

1. Graduate student, University of Maryland, College Park, MD 20740. I wish to give thanks to Emily Dacquisto, Julie Silva, Laixiang Sun, and three anonymous referees. Their critical and insightful comments improved the paper greatly.

ments and the mean value of comparable children from a reference population, divided by the standard deviation of that reference population (WHO 1995). Z-scores require a well specified reference population with a normal distribution, a condition which would imply that z-score cutoff values for stunting, wasting, or underweight are stable across different reference populations.

However, many practitioners operate under the assumption that the standard deviation (SD) of a survey's anthropometric indices is a necessary and sufficient measurement for quality control (QC).[2] The practice is particularly persistent for anthropometric surveys within the field of childhood malnutrition, with particularly grievous consequences. In one typical article, the quality control maxim for z-scores states, "summary statistics can be compared with the reference, which has an expected mean Z-score of 0 and a SD of 1.0 for all normalized growth indices" (Mei and Grummer-Strawn 2007, 441). Others suggest that if a survey presents with "an excessive standard deviation…the survey results should be rejected" (Grellety and Golden 2016). The maxim is certainly simple, but does its simplicity compensate for its disadvantages?

Suppose you wish to conduct an anthropometric survey across the Karamoja region of northeast Uganda to assess the health of the region's children. Your well-designed survey includes measurements of height, weight, and age from a sample of children. You combine the measurements to make anthropometric indices of health such as weight-for-height and height-for-age. After performing some rudimentary summary analysis, you discover the sample standard deviations of the survey indices are (for example) 1.3 times greater than those of the 2006 World Health Organization (WHO) reference standards, which is not surprising given that the two groups of children come from two distinctly different populations. However, the quality control maxim used by many anthropometric researchers would dismiss your Karamoja survey as low quality, simply because the standard deviations are 1.3 times greater than the 2006 WHO reference standards.

Anthropometric research generally works with z-scores, however, and the practice that I am objecting to is expressed in terms of z-scores, not sample standard deviations. Couched in terms of z-scores, the nature of the putative quality control requirement is a bit harder to understand. But it is really as simple

2. Exactly how many is up for debate and a potential direction for future research. Suffice it to say the number is large. If one is unfamiliar with this particular body of literature or the day-to-day pragmatics of organizations working in this field, then the SD as QC problem might not seem endemic. But much like dust in the air, to borrow a metaphor, SD as QC seems invisible—even if you're choking on it—until you let the sun in. Then you see it's everywhere. A collection of quotes from this search is provided in Appendix A to help illuminate the extent, certainly representing only a small sample of all the potential articles and reports. Not to mention the many unreported, unknown, and unknowable studies that never saw the light of day because of internal or external suppression for having supposedly overly large standard deviations.

as the Karamoja example: when the ratio of standard deviations (of the sample and a reference) is in excess of a fixed threshold (e.g., 1.3) the study fails the quality control test. It can be shown that an anthropometric survey has a z-score standard deviation of 1.3 (or any other arbitrary cutoff value) *if and only if* the sample standard deviation of the anthropometric index is 1.3 times that of the standard deviation of the reference population. From a mathematical standpoint, a claim about the standard deviation of a z-score is equivalent to a claim about the ratio of an index's sample standard deviation to that of a reference population. For a proof, see Appendix B.

The notion that I wish to challenge is the following: Any anthropometric survey and subsequent z-score index (e.g., height-for-age or weight-for-height) not normally distributed with a standard deviation of approximately 1.0 (e.g., 1.3) indicates a serious problem and should be considered unusable.[3] And I suggest there is neither statistical justification nor scientific evidence that supports the SD as QC maxim.

There are, of course, inaccurate surveys that deserve to be dismissed. Garbage in, garbage out. Wariness is appropriate, but tests and conditions other than a standard deviation threshold must be applied. For example, the United States Agency for International Development (USAID) identify 26 potential indicators that could measure anthropometry data quality during fieldwork (Allen et al. 2019). WHO recommends considering several indicators such as population characteristics, sample size, survey design, measurement methods, and missing

3. Although the maxim is widely practiced, it is not always consistent. WHO suggests the z-score "distribution should be relatively constant and close to the expected value of 1.0 for the reference distribution" (1995, 218). De Onis and Blössner, citing WHO (1995), claim good quality SD ranges of HAZ (1.10 to 1.30), WAZ (1.00 to 1.20) and WHZ (0.85 to 1.10) and state these values are "the expected ranges of standard deviations of the z-score distributions for the three anthropometric indicators" (1997, 51). De Onis and Blössner also state that "[a]ny standard deviation of the z-scores above 1.3 suggests inaccurate data" (ibid.). Golden and Grellety suggest "The spread of the standard deviations…was small; ranging from 0.8 to 1.2 in 95% of the surveys" (2002, 5). Grellety and Golden, citing WHO (1995) and Golden and Grellety (2002), state "the SD for Weight-for-Height (WFH) should be between 0.8 and 1.2 Z-score units in all well-conducted surveys, with about 80% between 0.9 and 1.1Z" (2018, 2). Mei and Grummer-Strawn, citing WHO (1995), present the same example z-score table of HAZ (1.10 to 1.30), WAZ (1.00 to 1.20) and WHZ (0.85 to 1.10) and claim these values are a "recommendation from a WHO expert panel" as the "ranges for data quality assessment" (2007, 445). Mei and Grummer-Strawn (2007) also suggest the ranges for data quality assessment should be wider, given by HAZ (1.35 to 1.95), WAZ (1.17 to 1.46) and WHZ (1.08 to 1.50). We are told by USAID "that high quality anthropometric data should be normally distributed with a standard deviation of approximately 1" (2016, 15). But later USAID informs us that "very large standard deviations, for example greater than 2, might be a sign of poor quality" (ibid.). Bilukha et al., citing WHO (1995) and WHO and UNICEF (2019), give the recommendation that "Absent measurement error, distributions are expected to be approximately normal with a SD close to 1" (2020, 2). However, Bilukha et al. choose the exclusion criteria of "greater than 1.8 or lower than 0.8" (2020, 3).

data (WHO 1995). WHO and UNICEF (2019) suggest performing a seven-point data quality assessment, which interprets and reports: completeness; sex ratio; age heaping; height and weight digit preference; and z-score implausibility, standard deviations, skewness and kurtosis. And Nandita Perumal et al. (2020) have implemented this suggestion to its fullest potential.

Emmanuel Grellety and Michael H. Golden (2016) highlight random measurement, digit preference, and rounding error as potential sources of error. David A. Siegel and Jacob S. Swanson (2004) warn against heaping and digit preference. Researchers should also look out for confounding effects, specification error, non-linearity, bias of the auspices, measurement error, experimental error, and sample selection bias. Others point out that there is not even a consensus in the literature as to what constitutes a usable dataset (Crowe et al. 2014; Waterlow et al. 1977; USAID 2016). Shireen Assaf, Monica T. Kothari, and Thomas W. Pullum (2015) say the need for well-defined quality assessment criteria remains unmet, and they recommend more training and better equipment in the meantime.

In their methodological guidelines for assessing nutrition in crisis situations, the SMART (Standardized Monitoring and Assessment of Relief and Transitions) inter-agency initiative recognized that survey samples do *not* follow reference standards, and that even "the standard population is not normally distributed" (2006, 24 n.9). Later, however, the guidelines rely on the SD as QC maxim, claiming bias "can be estimated from examination of the standard deviation of the WFH, which should always be 0.8–1.2 z-scores" (ibid., 38).

Inspection of surveys for small SD remains in many QC recommendations (e.g., Allen et al. 2019; SMART 2006; WHO and UNICEF 2019) as a necessary if not sufficient condition for acceptance, while for others it is even a sufficient condition (e.g., Bilukha et al. 2020; Grellety and Golden 2016; 2018; Mei and Grummer-Strawn 2007). I propose that SD is neither a necessary nor sufficient indicator of QC. Low-quality surveys can have small SD, and high-quality surveys can have large SD. Errors of commission and omission waste precious resources that are already spread thin. The disregarding of surveys with high standard deviation could result in funds and research being syphoned away from the people most in need. It is my aim to illustrate the archival, statistical, logical, theoretical, and practical evidence that standard deviation should serve as neither a necessary nor a sufficient arbiter of quality control.

Unsound beginnings

It was sculptors and painters who first measured the relative proportions of the human form (Tanner 1981). Scientific study of the measurements of the

human body emerged notably with the work of Adolphe Quételet in 1832. Much like contemporary practitioners, Quételet performed a cross-sectional study of the height and weight of newborns and children, and observed a likeness between the distribution of weight and height to a normal (Gaussian) distribution (Quételet 1832; 1835). This Quételet Index, later redubbed Body Mass Index, is still relevant today. Unlike Quételet, however, contemporary practitioners have transposed his observation, and adopted the quality control practice of judging a survey based on its likeness to a standard normal distribution.

The source of the misconception originates in a presentation at the 15th International Congress of Nutrition in 1993 by Ray Yip. Despite its later impact on the literature, the SD as QC proposal does not even appear in the summary of the workshop, including Yip's abstract (Yip 1993). But two years later the WHO issued a technical report titled *Physical Status: The Use of and Interpretation of Anthropometry* that many have cited as the origin of and authority for the SD as QC maxim.

In less than one page of a 463-page report, some of the most recurrent maxims are found. WHO (1995) outlines several steps involved in assessing the quality of anthropometric data, including the observed standard deviation of the z-score distribution. With accurate measurements, the report claims, the "distribution should be relatively constant and close to the expected value of 1.0 for the reference distribution" (WHO 1995, 218). Citing the 1993 conference abstract, the report presents a table of "the standard deviations of the height-for-age, weight-for-age, and weight-for-height z-score distributions" all ranging "within approximately 0.2 units of the expected value" (WHO 1995, 218). The table of values include: HAZ (1.10 to 1.30), WAZ (1.00 to 1.20), and WHZ (0.85 to 1.10). The expected value of 1.0, the range of plus or minus 0.2 units, and the specific table values have all been widely cited as criteria by which to judge a survey's quality (e.g., Blanton and Bilukha 2013; Bilukha et al. 2020; de Onis and Blössner 1997; Grellety and Golden 2018; Mei and Grummer-Strawn 2007; SMART 2006; WHO and UNICEF 2019).

WHO (1995) presents the table of SD ranges only as an *example* that was observed during multiple large-scale Centers for Disease Control and Prevention (CDC) surveys presented once at a conference. The range of plus or minus approximately 0.2 units is merely a generalization they ascribe to the example surveys. In fact, WHO (1995) goes on to say that in some surveys the observed standard deviations ranged from 1.4 to 1.8, even after excluding extreme outliers. The specific SD values were *not* given in WHO (1995) as QC ranges as many have claimed (e.g., Grellety and Golden 2018; Gupta et al. 2020; Castro Bedriñana and Chirinos Peinado 2014; Kwena et al. 2003; Jacob et al. 2016; Mei and Grummer-Strawn 2007; Wijaya-Erhardt 2019).

The report does suggest a SD > 1 *could* be an indicator of inaccuracy, but the

notion was couched in a larger discussion of indicators, including validity of the reference population, the notorious quality of age estimates, errors of rounding and digit bias, number of missing and improbable values, and overall data compilation and documentation. Standard deviation is but one potential indicator, of many, to flag surveys for further inspection, *not* a sufficient measure of quality (WHO 1995). And the report recommends: "Verification of accuracy is best done by remeasurement of a sub-sample of the original sample by individuals who are fully qualified in anthropometric procedures" (WHO 1995, 216). In other words, standard normal SD is certainly not a sufficient QC condition.

Soon after, Mercedes de Onis and Monika Blössner (1997) echoed the SD as QC maxim as a definitive fact of nutrition surveys in their report *WHO Global Database on Child Growth and Malnutrition*, which many have cited as the source of the idea. In particular, de Onis and Blössner claim:

> If the surveyed standard deviation of the Z-score ranges between 1.1 and 1.2, the distribution of the sample has a wider spread than the reference. Any standard deviation of the Z-scores above 1.3 suggests inaccurate data due to measurement error or incorrect age reporting. (de Onis and Blössner 1997, 51)

The first sentence is referring to the *survey* data compared to the *reference* data. It is only making general statements about how variance and spread can be described for any two distributions of data. The second sentence, however, jumps to the conclusion that a z-score standard deviation above 1.3 "suggests inaccurate data."

Without question, z-score summary statistics can indicate community-wide malnutrition; that is their function. As de Onis and Blössner state earlier "if a condition is severe, an intervention is required for the entire community, not just those who are classified as 'malnourished' by the cut-off criteria" (1997, 50). That is to say, when analyzing z-scores, *if* many observed z-scores are well below the reference, *then* one might conclude that the appropriate intervention mechanism should be aimed at the population, and not the individual level. This is a sensible, if tautological, suggestion. But the inverse is not necessarily true. Namely, if you do not observe a standard normal distribution of z-scores shifted in mean only, then you conclude that none of the population has been affected and the sample is simply of low quality.

It seems obvious that a population by definition will *not* move together as a whole. We know that low-income families are more vulnerable to price volatility and uncertainty because they have fewer options, entitlements, and capabilities (Sen 1984). Calorie elasticity is not zero (Subramanian and Deaton 1996). Low-income families spend a large percentage of income on food, making them more vulnerable, thus skewing the distribution asymmetrically.

Larger z-score SD implies larger spread implies inaccurate data: simple but unsatisfying. I have not found substantiating evidence or theoretical justification for the maxim—in de Onis and Blössner (1997) in particular or the literature in general. But what I have found is a history of citations built upon a shaky foundation.

In my estimation there are really only two studies which one could argue have attempted to show evidence or justification for SD as QC, if only tangentially. The first comes from a conference paper presented at the Standardized Monitoring and Assessment of Relief and Transitions (SMART) Workshop, July 23–26, 2002. At the workshop Michael H. Golden and Yvonne Grellety presented a working paper in which they claim to disprove the assertion that "social heterogeneity would lead to changes in the shape of the distribution curve of acute malnutrition when a population is exposed to famine" (2002, 3). And through their analysis they conclude that "there was no change in the spread of wasting within the population as it became more malnourished" (ibid.).[4]

The findings of the Golden and Grellety (2002) working paper rest largely on Kolmogorov-Smirnov tests. In this case, the null hypothesis claim is that heterogeneity of wasting (i.e., z-score distribution curve) is heteroscedastic and the goal of the test is to falsify that claim. Their objective is to prove distributional spread (i.e., SD) is independent, stable, and standard normal (i.e., close to 1.0) as populations are exposed to starvation and famine (i.e., changes in average z-scores). And as an extension of their Kolmogorov-Smirnov test, they suggest SD is a measure of QC, stating:

> If a survey is observed to differ significantly from normality or have a large standard deviation, then we suggest that either two distinctly different populations may have been included in the sample or there is methodological error. All surveys should be checked for normality and any difference investigated. (Golden and Grellety 2002, 10)

But the specific Kolmogorov-Smirnov tests that Golden and Grellety (2002) devise assume the data are normally distributed from the start. In this case the null hypothesis is not heterogeneity, but that z-score distribution curves are in fact normal. Furthermore, Thomas Bayes (1763) taught that it is incorrect to assume $\Pr(\text{Data} \mid H_0) = \Pr(H_0 \mid \text{Data})$. And testing for normality is not equivalent to testing a unit SD. We are also not provided the power of the tests (i.e., the probability of correctly rejecting the null hypothesis), making it difficult for one to judge a null hypothesis false when it is false.

4. Emmanuel Grellety and Michael Golden (2018) stipulate that these findings confirm that SD should be between 0.8 and 1.2 z-score units in all well-conducted surveys.

Finally, in their figures, they purport that mean and standard deviation are uncorrelated. But if two random variables are statistically uncorrelated, that does not imply they are independent—yet it is independence that they seek. In addition, they show that kurtosis varies from −0.75 to 1.75 decreasing as wasting escalates, and skewness varies from −0.5 to 0.75 increasing as wasting escalates, contradicting the claim that malnutrition prevalence remains fixed and normally distributed.

In my estimation, even if Golden and Grellety (2002) had shown what they intended, it is still a great leap to conclude that therefore standard deviations are a necessary and sufficient quality control measure. The link is missing. Many alternative hypotheses still exist. As Deirdre N. McCloskey and Stephen T. Ziliak point out, "Failing to reject does not of course imply that the null is therefore true. And rejecting the null does not imply that the alternative hypothesis is true: there may be other alternatives which would cause rejection of the null" (1996, 102). And elsewhere, Golden concedes: "Most experimental studies do not include the acutely ill children for ethical reasons; the children are studied after they have recovered from acute infections and other major complications" (2009, S280). The esteemed pediatrician James Tanner knew in 1952 that unhealthy populations could be non-Gaussian and skewed; as such, standard deviations may be biased and not locate the right points (Tanner 1952).

The second study comes from an article by Zuguo Mei and Laurence M. Grummer-Strawn (2007). Mei and Grummer-Strawn claim to "assess whether the SD of height- and weight-based Z-score indicators derived from the 2006 WHO growth standards can still be used as data quality indicators," finding "the SD for all four indicators were independent of their respective mean Z-scores across countries" (Mei and Grummer-Strawn 2007, 441). They conclude that "the SD of Z-scores could still be used as a data quality indicator for evaluation of anthropometric data" (ibid., 445).

Again, WHO (1995, 218) presents a table of z-scores with different ranges of distribution values (i.e., HAZ 1.10 to 1.30, WAZ 1.00 to 1.20, and WHZ 0.85 to 1.10). However, as I hope I have illustrated, the table is presented only as an *example* of observed ranges. And the standard deviation z-score ranges were never meant for data quality assessment, nor has SD ever been shown to be a sufficient QC indicator.

But the point is lost in Mei and Grummer-Strawn (2007), who submit that WHO (1995) recommended "standard deviation ranges for data quality assessment" and claim to assess "whether these Z-score ranges still apply." I suggest they never did. Mei and Grummer-Strawn even concede that "the observed ranges of SD for all four indicators from our analysis were consistently wider than those recommended by WHO" (2007, 441). Yet these specific values were never given in WHO (1995) as the acceptable range for good quality surveys.

Citing WHO (1995), Mei and Grummer-Strawn assert that:

> On the basis of the 1978 WHO/National Center for Health Statistics (NCHS) growth reference, WHO has previously indicated that the SD of Z-scores of these indicators is reasonably constant across populations, irrespective of nutritional status, and thus can be used to assess the quality of anthropometric data. (Mei and Grummer-Strawn 2007, 441)

I think it is telling that they point to the 1995 technical report instead of pointing to John C. Waterlow et al., who were the actual developers of the WHO/National Center for Health Statistics (NCHS) growth reference[5] and who warned against universal principles: "Decisions of this kind have to be taken locally, and it is not possible to make international recommendations about them" (Waterlow et al. 1977, 491). Indeed, we need to make judgments backed up by logic, theory, and evidence, and not follow a binary decision rule that lacks contextual nuance. Waterlow et al. affirm that sub-populations are heterogenous, imploring us to make judgments on a case-by-case basis:

> Clearly, if there were differences dependent on different gene distributions, then the target for one population would not be the same as the target for another. … Because the reference population cannot be used as a universal target, the question of what is a realistic goal in any particular situation does become important. (Waterlow et al. 1977, 490)

The purpose of Waterlow et al. was to "present recommendations for the analysis and presentation of height and weight data" (1977, 489), *not* to present ways to exclude such data. All constraints that Waterlow et al. do propose are wholly directed at constructing a *reference* population. Whereas a *standard* represents a desirable target or norm, the sole aim of a *reference* is to be a common basis in order to group, analyze, and compare different populations (WHO 1995). Unfortunately, the distinction between references and standards was, and continues to be,

5. In 1971, as part of a long tradition for child growth references, the Maternal and Child Health Program, the United States Public Health Service, and the American Academy of Pediatrics concurred that more rigorous standards were needed for clinical characteristics of early childhood malnutrition. This decision was the impetus for the Health and Nutrition Examination Survey carried out by the Centers for Disease Control and Prevention's National Center for Health Statistics Task Force. First released in 1977, the National Center for Health Statistics Growth Curves were a combination of data from the National Center for Health Statistics' Health Examination Surveys, the Health and Nutrition Examination Survey, and the Fels Research Institute. Wanting in on the action, a WHO working group on nutritional surveillance made recommendations on the criteria for the anthropometric reference population and presented recommendations for the analysis of data from surveys involving nutrition and anthropometry, thus the "WHO/National Center for Health Statistics" growth reference.

indifferently heeded and left in unclarity.

The 1978 WHO/NCHS growth *reference* is distinct in its purpose and function from the 2006 WHO Multicentre Growth Reference Study (MGRS) growth *standards*. And neither can inform, through comparing standard deviations, whether or not any particular *sample* is of poor quality. But Mei and Grummer-Strawn assert that, "our analysis confirms the WHO assertion that the SD remains in a relatively small range for each indicator" (2007, 445). To do so, however, is to conflate *standards*, *references*, and *samples*.

In 1993, the Expert Committee on Physical Status, convened by WHO, concluded that previous *reference* growth charts had long been misconstrued as a *standard* for growth (de Onis and Habicht 1996). As a result, the WHO Multicentre Growth Reference Study was implemented between 1997 and 2003. The designers of the new Growth Reference were intentionally *prescriptive* rather than *descriptive* (Garza and de Onis 2004). They designed a growth chart for how children *should* grow rather than how children *actually* grow. In other words, it was purposely designed to produce an idealized *standard* rather than a baseline *reference*.

Even the initial sample data for the Multicentre Growth Reference Study did not have small and well-behaved standard deviations. To produce the growth standards, the sample was manipulated to fit specific distributional requirements (WHO 2006). And even though the MGRS sought out the healthiest, most ideal population to measure, 93 percent to 69 percent of the healthy populace were ineligible and did not conform to this ideal.[6] In other words, even in the healthiest

6. The Multicentre Growth Reference Study (July 1997–December 2003) consists of both cross-sectional and longitudinal surveys from six cities: Davis, California, USA; Muscat, Oman; Oslo, Norway; Pelotas, Brazil; in select affluent neighborhoods in Accra, Ghana; and South Delhi, India (WHO 2006). The distributions of children across the different survey countries for the longitudinal component are: 119 USA; 149 Oman; 148 Norway; 66 Brazil; 227 Ghana; and 173 India. The distributions of children across the different survey countries for the cross-sectional component are: 476 USA; 1,438 Oman; 1,385 Norway; 480 Brazil; 1,403 Ghana; and 1,487 India. Prior to constructing the standards, if a child was 3 SDs above the sample median or 3 standard deviations below the sample median they were excluded. For the cross-sectional sample the truncation procedure was even stricter. If a child was 2 SDs above the sample median or 2 SDs below the sample median they were excluded. Children were selected for inclusion based on: no known health or environmental constraints to growth, mothers willing to follow feeding recommendations, no maternal smoking before and after delivery, single term birth, and absence of significant morbidity. Of the 13,741 children screened for the longitudinal survey, less than 7 percent or 882 children (428 boys and 454 girls) were eligible and included in the final study. In addition, of the 21,520 children screened for the cross-sectional survey, less than 31 percent or 6,669 children (3,450 boys and 3,219 girls) were eligible, and included in the final study. In other words, 69 to 93 percent of the populace did not fit the standard. After selective sampling and exclusion, the sample was exceedingly skewed to the right (WHO 2006). To rectify the non-normality, the data were cleaved at the median, and then reflected to create two symmetrical distributions. Each mirrored distribution was used to derive standard deviation cut-off values (e.g., what is the severe wasting cutoff value where a WHZ score is less than 3 SDs from the median) for the respective upper and lower portions of the data.

and most ideal sub-populations, most children do not fit the growth standards, nor are they normally distributed with standard deviations close to one. The MGRS provided a growth standard intended for measuring benchmark distances from an idealized healthy child. It is not the only permittable distribution for a sample dataset nor is it relevant for measuring data quality.

Spurious theory and flawed logic

SD as QC may be believed by some to be loosely related to the seminal concepts of the eminent epidemiologist Geoffrey Rose, whose ideas transformed the strategy of preventive medicine. Central to Rose's analysis was an assumption that the width of the distribution of a variety of biological measures remains similar across different populations even as the mean of the distribution shifts: a mean-centric view of population (Rose 1992). He observed that most risk-factor distributions across populations appear to have uniform displacements, with risk changing the same amount at different parts of the risk-factor distributions. Rose's assumption implies that the mean of a distribution can be used as a proxy for a population's intrinsic traits.

But it is an untenable leap to go from Rose's notion that distributions of biological measures tend to have consistent spread, independent from the central tendency, to the misconception that any distribution of a biological measure that does not have a small and precise spread is invalid, inaccurate, and not insightful. Furthermore, Rose's conceptualization is anchored on the cohesiveness of populations, an assumption that may be violated by differential changes in the BMI distribution occurring globally within populations (Razak, Davey Smith, and Subramanian 2016).

Contrary to theoretical and observational expectations, some have claimed whole population distributions shift equally in the face of malnutrition stressors and that any data set which does not behave that way (i.e., any data set with z-score standard errors not equal to one) must be a low-quality survey (e.g., Blanton and Bilukha 2013; Bilukha et al. 2020; de Onis and Blössner 1997; Golden and Grellety 2002; Grellety and Golden 2016; Grellety and Golden 2018; Mei and Grummer-Strawn 2007). But the assertion remains unsubstantiated. If true, it would follow that whenever there was a famine (malnutrition stressor) anywhere in the world, you sitting at the breakfast table, drinking your coffee, oblivious to the famine, would also become slightly malnourished, too, to maintain a normally distributed population with a standard deviation of one. We all must move together to preserve the spread of the distribution, you see. Now, presumably, the SD as QC crowd would say that interpretation is preposterous, and that mean shifts in z-scores

do not occur for the entire planet but are only applicable to some smaller sub-population. Ah, then, by 'shifts in the population,' they don't really mean the Population. Okay, but they still have to contend with the problems of *sorites* and the fallacy of the transposed conditional (on that fallacy, see **here** or Appendix C).

If the effect is only valid for some sub-population then the boundaries of that sub-population must be defined, and the sub-population is by definition not representative of the whole population. So we should not be casual in talking of 'populations.' The meaningfulness of descriptive statistics depends on how meaningfully a population is defined in relation to the inherent intrinsic and extrinsic dynamic generative relationships by which they are constituted (Krieger 2012).

Prevalence and distributions of z-scores are therefore highly reliant on boundary definitions and cannot be extrapolated out of sample. Remember, too, that the 'reference population' used for judging a child's health is really a *standard* and by design a small sub-population of only the healthiest of healthy children. And, even still, those 'standard' children were not distributed standard normal with an SD of one (WHO 2006). There is no reason to believe that a healthy sub-population should behave the same way a malnourished sub-population does.

Standard deviation is merely the measure of dispersion for a set of values, unlike digit preference (heaping at 0 and 5), incompleteness (missing values), rounding errors (chop vs. nearest), data formatting (short, long, float, double), transposition and transcription errors (obvious typos), or procedural errors (e.g., a child measured lying down when they should have been standing), which are all direct quality control metrics of a specific error. For example, the standard deviation of WHZ only gauges the ratio of the weight-for-height sample standard deviation to that of the weight-for-height standard deviation of a reference population. The reference population (even if it is a *standard*) cannot signify anything qualitative about the sample data, nor should it. A reference population is merely a datum or a fixed point. It is a quantitative scale not a qualitative apparatus.

Measurement errors *might* generate inflated SD. Then again, they might not. Inflated SD does not necessarily imply measurement error (Biehl et al. 2013; Ulijaszek and Kerr 1999). The quality control maxim poses the prior "if the population is distributed normal, then the observed data will be distributed normal," and supposes wrongly "if data is observed, then the population it is drawn from is distributed normal." If *H*, then *O*, does not affirm if *O*, then *H*. It is the same as thinking if a person is hanged, then he will probably die; therefore, if observing a corpse, then one should conclude he was probably hanged (Ziliak and McCloskey 2008, 17).

Random errors lower precision by inflating confidence intervals. Random error is but one of many dozens of errors and seldom the biggest (Ziliak and

McCloskey 2008). It is systematic errors that we should be worried about. They cause bias. Especially when the costs of failure (i.e., child mortality) are high, the choice between low bias or low precision is not really a choice at all. If I can't be precisely right, I would rather be generally right than precisely wrong. More importantly, Ziliak and McCloskey note "sampling precision says nothing about the oomph of a variable or model" (2008, 25).

Systematic errors may even attenuate SD. A small spread in SD is not a necessary condition for a lack of systematic error, making SD a poor metric from which to judge quality. Suppose, for example, I performed an especially erroneous survey of child anthropometry in which instead of actually measuring different weights and heights, I just marked down the exact same value for every survey participant. Is my systematic measurement error captured by an inflated standard deviation? No. Obviously, this is an extreme and absurd example. But there exists a non-zero proportion of the total sample space in which systematic errors diminish rather than inflate standard deviation. Try to imagine the countless number of possible surveys with less extreme systematic error structures, all of which exhibit 'a standard deviation of approximately one.' If it is systematic errors that we are concerned with, SD signifies very little.

The obverse problem with SD as QC remains, too. Since Anscombe's quartet and the more recent Datasaurus Dozen, students of statistics have long known that different datasets with wildly varying graphical distributions can all have the exact same descriptive statistics, including standard deviation (Anscombe 1973; Matejka and Fitzmaurice 2017). Logic dictates SD is neither a necessary nor sufficient indicator of QC.

Informed dissent from the maxim

The debate surrounding standard deviation as a quality control metric is ongoing and unresolved. After two national nutrition surveys in Nigeria exhibited divergent estimates, both USAID and United Nations Children's Fund (UNICEF) staff in-country felt that substantial quality problems must exist in either one or both surveys (USAID 2016). In July 2015, the USAID Nutrition Division convened a technical meeting aimed at resolving the issues of accuracy and comparability of anthropometric data. Participants included representatives from USAID, CDC, UNICEF, WHO, the Pan American Health Organization, and external nutrition experts. The meeting report highlights that the importance of standard deviations for measuring data quality was a major point of contention. The report concludes that "there was no agreement on what is a reasonable standard deviation of z-scores to expect in heterogeneous populations" (USAID

2016, 17).

The meeting report features arguments for the SD as QC maxim given by an unspecified presenter from the CDC. In reference to the Demographic and Health Surveys, the CDC presenter asserted that high-quality anthropometric data will *always* be normally distributed with a standard deviation of approximately one regardless of population heterogeneity, and that a standard deviation greater than one *must* mean the data are of poor quality (USAID 2016, 16). One example they pointed to was the National Health and Nutrition Examination Survey in the United States with a (recent) stable trend of small standard deviations. Furthermore, they claimed the shape of the distribution does not change as a population becomes more malnourished, concluding there is no relationship between the mean z-score and standard deviation. In their estimation, this lack of relationship is sufficient to conclude standard deviation is a quality control metric.

The report suggests, however, that not all participants agreed with the SD as QC maxim. Some participants felt that standard deviations greater than one could reflect heterogeneity in the population. For the Demographic and Health Surveys in particular, they expressed concern regarding the emphasis on standard deviations of height-for-age, weight-for-age, and weight-for-height z-scores close to one as an indication of quality. The report details that other participants noted:

> In Kano state, Nigeria, for example, a majority of the within-cluster standard deviations were below 1, however, the average standard deviation in Kano state was more than 1. If the states are different, it is impossible for the standard deviation to be 1 in every state, and 1 for the country as a whole. (USAID 2016, 17)

Other researchers acknowledged that the Demographic and Health Surveys in particular did show the most variability in parameters such as standard deviation. But they also noted that the Demographic and Health Surveys Program has the largest number of surveys and covers the largest span of time; standard deviations may have changed with time as nutritional status of the populations changed or improved. One meeting facilitator affirmed that it is *not* true that the shape of the distribution does not change as nutritional status of the population changes. While others pointed out that in terms of the factors that influence anthropometric indicators (e.g., water, sanitation, and food security), the United States may be more homogeneous than other countries (e.g., India) (ibid., 16).

Given that standard deviations capture inherent population heterogeneity, there is no reason to assume that the standard deviation will be the same across all surveys. It is true that poor data quality could inflate the standard deviation of anthropometric measures, but given that anthropometric z-scores are biologic parameters, one would anticipate some population heterogeneity both within and

between countries, even in situations of high-quality data collection.

The Joint FAO/WHO Expert Committee on Nutrition (1971) noted that statistical evaluation cannot by itself distinguish between what is normal and abnormal in the biological sense. Even seminal author and pediatric expert Dr. Derrick Jelliffe (1966) emphasized the problems and difficulties of non-sampling errors, which cannot be detected with tests of sampling errors. And Jonathan Gorstein et al. (1994) noted that when the nature of a nutrition problem is unclear, it should be interpreted within the situational context.

Standard deviation is not indicative of quality control for some studies. There are researchers and journals confident enough in the quality of their findings even with standard deviations *not* approximately one. Yirgu Fekadu et al. (2015) found z-score standard deviations of 1.3 (weight-for-height), 1.33 (height-for-age), and 1.06 (weight-for-age) in Ethiopian children. Michel Garenne et al. (2009) found weight-for-height z-score standard deviations of 1.28 and 1.398 for Niakhar, Senegal, and Bwamanda, D. R. Congo, respectively. Afework Mulugeta et al. (2010) observed z-score standard deviations of 1.8 (height-for-age), 1.3 (weight-for-age), and 1.3 (weight-for-height) for children in northern Ethiopia.

In addition, Paul B. Spiegel et al. (2004) performed a meta-analytical quality assessment of anthropometric surveys with no mention of standard deviation. Daniel E. Roth et al. (2017) estimated that across 64 low- and middle-income countries, when mean height-for-age z-scores were zero, the standard deviation was 2.10 (95% CI 2.00 to 2.20), far above most QC thresholds. Examining mid-upper arm circumference (MUAC) for 852 cross-sectional nutritional surveys of children, Severine Frison et al. (2016) found that only 319, or 37.7 percent, follow a normal distribution.

In his survey of famines and economics, Martin Ravallion remarks on the unusual nature of malnourished communities: "I will say that a geographic area experiences famine when unusually high mortality risk is associated with an unusually severe threat to the food consumption of at least some people in the area" (1997, 1205). The phenomenon of malnutrition is by its very nature unusual, i.e., not normal. It would be bizarre to think that measures would behave the same in lean times as in abundance. In their appraisal of different anthropometric indices, André Briend et al. get to the heart of the matter when they observe "for most populations, little information is available on the amount of nutritional change one has to expect in a community and also on the standard deviations of some nutritional indices" (1989, 770).

Eschew the maxim

The SD as QC maxim is built on a history of shaky citations, corroborated with imprudent tests, substantiated by logical fallacies, and endorsed inconsistently by empiricists. It lacks archival, statistical, logical, theoretical, and practical merit. Of course, there are inaccurate surveys and samples that don't deserve our consideration, but other tests and conditions must be adopted.

Once the SD as QC maxim is abandoned, the therapeutic and ameliorative next step is more difficult. But good science is difficult. If it were easy, it would have already been done (Wasserstein, Schirm, and Lazar 2019). Good science embraces the explicable and ineffable (McCloskey 1994). Doing serious scientific inquiries calls for serious thinking about what makes a dataset 'good' or 'bad' and how its 'goodness' may impact the results. We need to consider the dozens of sources of *real* error, and reckon their effects on our results. As Ziliak and McCloskey put it, "After all, reconciling differences of effect, finding the common ground, is the point of statistics. … Most important is to minimize Error of the Third Kind, 'the error of undue inattention'" (2008, 246).

Appendix A
SD as QC in the literature

The practice of SD as QC is pervasive, almost to the point of being a norm or a given first principle of the field were citation and evidence are not required. And I believe that the SD as QC maxim is preventing more studies and surveys from being used and published. In Google Scholar, Mei and Grummer-Strawn (2007) are cited over 170 times, not to mention the over 8,950 articles citing WHO (1995) or the 760 citing de Onis and Blössner (1997). Clearly not all are relevant to the SD as QC discussion. To help illustrate the point I spent an afternoon tracking down articles that explicitly and openly abide by the SD as QC maxim in some form or another. Below are excerpts from a sample of 32 articles citing Mei and Grummer-Strawn (2007) where authors point to the SD as QC maxim. I have put some words in boldface for emphasis.

"Researchers also have analyzed ways in which use of the WHO standards might affect prevalences of wasting, stunting, and underweight worldwide, as well as **the distribution of z scores, a commonly used indicator of data quality** in international surveys" (Grummer-Strawn, Reinold, and Krebs 2010, 13).

"Accepted best practices for field-level quality control were followed. Systematic repeat data entries were done for all anthropometric data. Postanalysis **quality checks compared SDs of anthropometric data by site to WHO standards** and other studies for children <2 y of age" (Remans et al. 2011, 1636).

"There were another 5,010 children whose length-for-age z-scores (LAZs) were flagged in the DHS data files either as missing or as biologically implausible according to the WHO flags (Mei & Grummer-Strawn, 2007). **These children were excluded from the analysis**. We also removed 71 children whose mothers had a height of less than 130 cm, as these were considered to be implausible and likely due to measurement or recording errors" (Krasevec et al. 2017, 2).

"z score **SDs were within the valid range** accepted by the World Health Organization (WHO)" (Corvalán et al. 2009, 548).

"Summary statistics showed that **standard deviations** of the three indices Z score (weight for age, height for age and weight for height) were **between 0.92 and 1.03, indicating high quality** data" (El Mouzan et al. 2008, 339).

"The data were subjected to post-hoc methods of quality determination, and, if of suitable quality, included in the adequacy evaluation. … Accepted practices for field-level quality control were followed. However, systematic repeat measures, repeat sampling and inter-lab sampling were not available for quality control of the MICAH data. Therefore alternative, post-hoc methods were used for evaluating the quality of data collected. Some of these methods have been used previously, whereas others were developed for the purpose of this evaluation. … Comparison of magnitude of SDs of continuous variables to SDs in other, well-controlled studies… This method of **comparing SDs with reference populations has been recommended for anthropometrics**. We assume that common levels of variations will exist for other variables. … SDs of continuous variables in MICAH surveys in baseline (1996 or 1997), follow-up (2000) and endline (2004) compared with examples from the literature, for quality control purposes" (Berti et al. 2010, 613, 617, 618).

"In the analysis, plausibility of anthropometric Z scores were checked using the WHO protocol recommendations (2006), which provide **standard deviation cut points for anthropometric Z-scores as a data quality assessment tool**" (Abate and Belachew 2017, 6).

"Mei and colleagues previously reported a lack of a relationship between SD and mean HAZ across DHS surveys; however, they did not quantitatively

assess the change in SD with the age-related decline in mean HAZ, and **they interpreted their findings only as a justification for using SD as an indicator of anthropometric survey quality**" (Roth et al. 2017, e1255).

"Mei and Grummer-Strawn [2007] supported the use of **SD as a quality indicator** for anthropometric data" (Afifi et al. 2012, 2655).

"In our opinion reports from **surveys with an SD of more than 1.2 are unreliable**. … An analysis of DHS and MICS shows **elevated SD values with all of the mean SDs outside the acceptable range; none of mean SDs for any of the surveys was less than 1.0Z**. In agreement with the data from West Africa, the 5th and 95th centiles of the SDs of 51 recent DHS surveys were HAZ 1.35–1.95; WAZ 1.17–1.46, and WHZ 1.08–1.50. Mei & Grummer-Strawn conclude that they 'concur with **the WHO assertion that SD is in a relatively small range**'" (Grellety and Golden 2016, 19).

"Before turning to multivariate regressions, we relate our results to two indicators of measurement error used in previous work. The first step is to compare our December–January gap with the SD of HAZ. The SD of HAZ could reflect genuine dispersion related to health inequality but is **widely used as an indicator of survey errors** in both height and age (Assaf et al. 2015; Mei and Grummer-Strawn 2007)" (Larsen et al. 2019, 716–717).

"**Standard recommendations state that a standard deviation of greater than 1.3 for HAZ reflects excessive random variation** in either height measurements or age estimates. The standard deviation of HAZ in the three DHS greatly exceeds this threshold for data quality; however, this recommendation is based on the use of the old NCHS:CDC:WHO reference population. There is evidence that standard deviations for HAZ greater than 1.3 are common in DHS in other countries and may be normal when using the WHO Child Growth Standard (Mei & Grummer-Strawn 2007)" (Woodruff et al. 2017, 15).

"Many DHS surveys have **standard deviations greatly exceeding the quality criteria** defined by the World Health Organization. … Ranges are then used to describe the overall quality of the survey and arbitrary cut-offs are used to decide whether the data are acceptable or not" (Tuffrey and Hall 2016, 4–5, 14).

"We calculated z-score standard deviations (SD) and analyzed SD disaggregated by age (under and over two years of age) **to determine if the quality of measurements** differed by age. … We can consider z-score standard deviation to illustrate the importance of reaching consensus on interpretation and action. WHO

and the US CDC promote the use of normative ranges of SD to determine if survey quality is acceptable, but the ranges are based on surveys that have evidence of poor data quality. The most recent DHS data quality assessment showed that 30 of 52 countries had HAZ SD greater than 1.5, but only one country suppressed data because of poor quality. According to SMART **data quality is not acceptable if HAZ SD is above 1.2**, and a recent modeling study showed that SD of 1.5 can result in substantial overestimation of stunting prevalence. Meanwhile, the published normative range for HAZ SD that some organizations use to deem data quality acceptable is 1.35–1.95" (Conkle et al. 2017, 5, 10).

"Few studies have assessed the distribution of WFH. Two looked at the standard deviations of the WFH distributions. In 1977, **Waterlow et al. showed that the WFH distributions were skewed at the upper centiles**. Their analysis was performed on data from surveillance or surveys involving nutrition and anthropometry in young children up to the age of 10 years. In 2006, Mei et al. analysed data from 51 DHS surveys representing 34 developing Countries. They found a mean WFH and SD WFH (z-scores) of 0.06 and 1.40 respectively. The mean ranged from −0.91 to 0.83 and the SD range [*sic*] from 1.03 to 1.55. **They concluded that their analysis confirms the WHO assertion that the SD remains in a relatively small range (i.e. close to SD from a standard normal\ distribution)**, no matter the *Z*-score mean although the observed range of SD for was [*sic*] consistently wider" (Frison et al. 2016, 7).

"Summary statistics showed SDs of the 3 indices' *Z* score (weight for age, height for age, and weight for height) between 0.92 and 1.03, **indicating high-quality data**" (El Mouzan et al. 2009, 68).

"Previous research has demonstrated that Z-scores within a population are normally distributed with a SD of approximately 1.0; the shape of the distribution does not vary based on the nutritional status of the population, as measured by the mean Z-score. Based on the finding that SD remains in a relatively narrow range for each indicator regardless of mean Z-score, **WHO guidance recommends that the SD of Z-scores can be used as a data quality indicator** as well as a measure of variability. The introduction of random non-directional errors, such as those introduced when age is estimated rather than calculated or when teams are imprecise in measuring height or weight, can result in wider SD relative to the acceptable ranges outlined by WHO. … We therefore included **SD of the Z-scores to assess the degree to which data quality** in addition to variability impact DEFF in anthropometric surveys. … The SD of WHZ and WAZ were approximately 1.00, as expected in high-quality anthropometry surveys (WHZ

median = 1.03, WAZ median = 1.04)" (Hulland et al. 2016, 2–3, 10).

"Anthropometry **data quality indicators were extremely high (median SDs** for weight-for-length, length-for-age and weight-for-age z-scores 1.01, 0.98, and 1.03, respectively), likely due to extensive training, standardization, and monitoring efforts. … Anthropometry data quality indicators were monitored throughout the study. The medians of monthly standard deviations for weight-for-length, length-for-age, and weight-for-age z-scores were 1.01, 0.98, and 1.03, respectively; **close to the expected value of 1.0 for a reference distribution**. Standard deviations for z-scores varied month-to-month, but never reached the WHO thresholds for measurement error or incorrect age reporting" (Aceituno et al. 2017, 2, 8).

"The standard deviations reported in this study are much lower than the **suggested standard deviations** reported by Mei and Grummer-Strawn estimations in a cross-country analysis" (Sharma et al. 2020, 17).

"We also examined the quality of the 2009 data by **assessing the SD as a quality indicator** for anthropometric data (Mei and Grummer-Strawn 2007) and examining whether or not age heaping was evident. These assessments did not reveal any concerns" (Boylan et al. 2017, 2261).

"Based on the WHO Technical Report, the **SD for Weight-for-Height (WFH) should be between 0.8 and 1.2 Z-score units in all well-conducted surveys**. This has been confirmed empirically with well conducted surveys in both the developed world where large national surveys of heterogeneous populations have been conducted, for example the National Health and Nutrition Examination Survey (NHANES) from USA's National Centre for Health Statistics (NCHS) and the developing world. … The SD of organisation "t" differs significantly from the others (Student's t test < 0.0001), with 69% (53/77) of their surveys for WHZ having **an SD of more than 1.2 Z**. … For most anthropometric measurements the SD from single surveys should lie between 0.8 and 1.2, with about 80% between 0.9 and 1.1Z. For these reasons the **SD has been used as a useful measurement of data quality**" (Grellety and Golden 2018, 2, 3, 10).

"The median SD and range for HAZ were greater overall and across all surveys than for WHZ. The absolute difference in HAZ by MOB of age reporting should be close to 0 if there is no systematic error in age reporting, but was 0.25 (in $z$score units) overall and up to 0.90 in Timor-Leste in 2009. … HAZ SD and WHZ SD had the highest factor loadings in the data quality indices indicating that **SD is an important measure of anthropometric data quality**" (Perumal et al. 2020, 809S, 812S).

"Absent measurement error, distributions are expected to be approximately normal with a SD close to 1. … To exclude surveys with exceptionally poor anthropometry data quality or where data manipulation might be suspected, **we excluded from analysis surveys where the SD for WHZ, WAZ, HAZ, or BMIZ was outside of the following empirically defined cutoffs**: greater than 1.8 or lower than 0.8; or the SD for MUACZ greater than 1.8 and less than 0.7" (Bilukha et al. 2020, 2, 3).

"Anthropometric data collected during the 2008 to 2009 and 2014 Kenya surveys were reanalyzed to assess standard parameters of quality: standard deviation, skewness, and kurtosis of z-score values for 3 anthropometric indicators (weight for height, height for age, and weight for age)… The primary objective of the comparative analysis was to observe the quality of anthropometric variables. The **first metric of quality, standard deviation**, is presented in Table 3. … One key measure is SD of the continuous *z*-score distributions. As noted, previous research suggests that for a given population, Z-scores are normally distributed with an SD of approximately 1.0" (Leidman et al. 2018, 406, 412, 414).

"Careful interpretation is required, as the **standard deviations** for our anthropometric measurements are outside the World Health Organization range **for data quality assessment purposes**" (Bennett et al. 2020, 2038).

"Note that the standard deviations (SD) of WHZ and MUACZ in all rounds are near or even below 1.0, which gives us **confidence in the quality of the anthropometric data** (Grellety and Golden 2016b; Mei and Grummer-Strawn 2007). The average SD—across all four survey rounds—is 1.03 for WHZ and 0.95 for MUACZ" (Ecker et al. 2019, 10).

"Seventeen surveys had large standard deviations (SD) for HAZ, which could result in attenuated regression coefficients when HAZ was used as an explanatory variable in regression analyses. To avoid attenuation, HAZ values for each child were **adjusted to obtain a standard deviation for HAZ of 1.2** for each of these surveys by subtracting the survey mean for HAZ, dividing by the survey SD for HAZ, multiplying by 1.2, and then adding back the survey mean for HAZ" (Frongillo et al. 2017, 3038).

"The World Health Organization (WHO) has recommended the use of Z-score of these indicators to classify nutritional status, given the constancy of their values, independent of nutritional status, and can even be used as **indicators of the quality of anthropometric data**" (Martins et al. 2010, 1106).

"Z-score plausibility was determined using WHO cutoffs. We used the following WHO-defined **standard deviation (SD) ranges to assess the quality** of data (HAZ 1.1–1.3, WAZ 1.0–1.2, and WHZ 0.85–1.1)" (Gupta et al. 2020, 2–3).

"…as per WHO standards. Some individuals may have met >1 **exclusion criterion**" (Varghese and Stein 2019, 1208).

"Protocol used for obtaining data was an adaptation of that published by Lapham et al. and Mei et al." (Samiak and Emeto 2017, 2).

"Studies investigating the quality of the DHS data report the quality to be good (Mei Z and Grummer-Strawn LM., 2007, Mishra et al., 2006)" (Reda and Lindstrom 2014, 1160).

Appendix B
Z-score SD Proof

The aim here is to move away from the convoluted discussion of z-scores and standard deviations of z-scores to simply anthropometric index measurements and standard deviations of anthropometric index measurements. To make this simplification I will show that a z-score standard deviation is equivalent to the ratio of standard deviations of an anthropometric index to that of the reference population. The standard deviation of a given survey's anthropometric index is calculated as:

$$s_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2}$$

where:

- s_x: anthropometric index sample standard deviation
- N: is the number of children in the sample
- x_i: is a child's anthropometric index value (e.g., weight-for-height)
- $\overline{x}$: is the anthropometric index sample average given by: $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

A z-score tells you how many standard deviations away an individual data value falls from the mean. It is calculated as:

$$Z_i = \frac{\left(x_i - \mu\right)}{\sigma}$$

where:

- Z_i: is a child's z-score
- x_i: is a child's anthropometric index value (e.g., weight-for-height)
- μ: is the reference mean
- σ: is the reference standard deviation

A given survey's z-score standard deviation is calculated as:

$$s_Z = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(Z_i - \overline{Z}\right)^2}$$

where:

- s_Z: z-score sample standard deviation
- N: is the number of children in the sample
- Z_i: is a child's z-score
- $\overline{Z}$: sample average z-score given by $\overline{Z} = \frac{1}{N}\sum_{i=1}^{N} Z_i$

Thus, we are left with the question: Is the statement, *if an anthropometric survey has a z-score standard deviation greater than 1.3 it fails the test*, equivalent the statement, *if the sample standard deviation of an anthropometric index is 1.3 times that of the standard deviation of the reference population it fails the test*? Or in other words, is the ratio of the sample standard deviation of (weight-for-height) to the reference population standard deviation of (weight-for-height) equivalent to the standard deviation of (weight-for-height) z-scores.

Claim:

$$\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(Z_i - \overline{Z}\right)^2} = \frac{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2}}{\sigma}$$

Squaring both sides and reducing gives:

$$\sum_{i=1}^{N}\left(Z_i - \overline{Z}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2$$

Note x_i is a random variable and μ and σ are constants such that $Z_i = \frac{(x_i - \mu)}{\sigma} = \frac{-\mu}{\sigma} + \frac{1}{\sigma}x_i$ is a linear transformation of the form $Z_i = a + bx_i$.

If $Z_i = a + bx_i$ then,

$$E[Z_i] = E[a + bx_i] = a + bE[x_i] = a + b\bar{x}$$

and

$$Var[Z_i] = Var[a + bx_i] = b^2\sigma_x^2$$

where

$$\frac{1}{N}\sum_{i=1}^{N}\left(Z_i - \bar{Z}\right)^2 = \sigma_Z^2 = Var[Z_i]$$

and

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$$

giving

$$\frac{1}{N}\sum_{i=1}^{N}\left(Z_i - \bar{Z}\right)^2 = \sigma_Z^2 = b^2\sigma_x^2 = b^2\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$$

Note for our purposes $b = \frac{1}{\sigma}$ such that $b^2 = \frac{1}{\sigma^2}$ giving

$$\frac{1}{N}\sum_{i=1}^{N}\left(Z_i - \bar{Z}\right)^2 = \frac{1}{\sigma^2}\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$$

which reduces to

$$\sum_{i=1}^{N}\left(Z_i - \bar{Z}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$$

QED.

Appendix C
The fallacy

The fallacy of the transposed conditional, also known as confusion of the inverse or the statistical equivalent to the fallacy of affirming the consequent, is the jumbling of the probability of a set of data given a hypothesis, and the probability of a hypothesis given a set of data.

In statistical terms, the fallacy of the transposed conditional is corroborated through Thomas Bayes' (1763) theorem, given by:

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}$$

where A and B are two different outcomes or events (i.e., a hypothesis and a data set) and $\Pr(B) \neq 0$. Therefore, we can see $\Pr(A \mid B) = \Pr(B \mid A)$ holds true if and only if $\Pr(A) = \Pr(B)$ at the same time.

It is a fallacy if one claims to test the likelihood of a null hypothesis assuming the data are true, if what is actually tested is the likelihood of the data assuming the null hypothesis is true. It is incorrect to assume $\Pr(\text{Data} \mid H_0) = \Pr(H_0 \mid \text{Data})$.

In terms of rhetoric and logic, the fallacy of affirming the consequent is stated:

$$\frac{P \rightarrow Q,\ Q}{\therefore P}$$

where one takes the true statement $P \rightarrow Q$ and incorrectly concludes the converse $Q \rightarrow P$ to be true. In plain terms, the fallacy is demonstrated with the simple and absurd statement: All dogs are animals; therefore, all animals are dogs.

References

Abate, Kalkidan Hassen, and Tefera Belachew. 2017. Women's Autonomy and Men's Involvement in Child Care and Feeding as Predictors of Infant and Young Child Anthropometric Indices in Coffee Farming Households of Jimma Zone, South West of Ethiopia. *PloS One* 12(3): e0172885. **Link**

Aceituno, Anna M., Kaitlyn K. Stanhope, Paulina A. Rebolledo, Rachel M. Burke, Rita Revollo, Volga Iñiguez, Parminder S. Suchdev, and Juan S. Leon. 2017. Using a Monitoring and Evaluation Framework to Improve Study Efficiency and Quality During a Prospective Cohort Study in Infants Receiving Rotavirus Vaccination in El Alto, Bolivia: The Infant Nutrition, Inflammation, and Diarrheal

Illness (NIDI) Study. *BMC Public Health* 17: article 911. **Link**

Afifi, Hanan H., Mona S. Aglan, Moushira E. Zaki, Manal M. Thomas, and Angie M. S. Tosson. 2012. Growth Charts of Down Syndrome in Egypt: A Study of 434 Children 0–36 Months of Age. *American Journal of Medical Genetics Part A* 158(11): 2647–2655.

Allen, Courtney K., Trevor N. Croft, Thomas W. Pullum, and Sorrel M. L. Namaste. 2019. Evaluation of Indicators to Monitor Quality of Anthropometry Data During Fieldwork. *DHS Working Paper* 162. ICF International, Inc. (Rockville, Md.). **Link**

Anscombe, Francis J. 1973. Graphs in Statistical Analysis. *American Statistician* 27(1): 17–21.

Assaf, Shireen, Monica T. Kothari, and Thomas W. Pullum. 2015. An Assessment of the Quality of DHS Anthropometric Data, 2005–2014. *DHS Methodological Reports* 16. ICF International, Inc. (Rockville, Md.). **Link**

Bayes, Thomas. 1763. An Essay Towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London* 53: 370–418. **Link**

Bennett, Aisleen, Louisa Pollock, Khuzwayo C. Jere, Virginia E. Pitzer, Benjamin Lopman, Naor Bar-Zeev, Miren Iturriza-Gomara, and Nigel A. Cunliffe. 2020. Duration and Density of Fecal Rotavirus Shedding in Vaccinated Malawian Children With Rotavirus Gastroenteritis. *Journal of Infectious Diseases* 222(12): 2035–2040. **Link**

Berti, Peter R., Alison Mildon, Kendra Siekmans, Barbara Main, and Carolyn MacDonald. 2010. An Adequacy Evaluation of a 10-Year, Four-Country Nutrition and Health Programme. *International Journal of Epidemiology* 39(2): 613–629. **Link**

Biehl, Anna, Ragnhild Hovengen, Haakon E. Meyer, Jøran Hjelmesæth, Jørgen Meisfjord, Else-Karin Grøholt, Mathieu Roelants, and Bjørn Heine Strand. 2013. Impact of Instrument Error on the Estimated Prevalence of Overweight and Obesity in Population-Based Surveys. *BMC Public Health* 13: article 146. **Link**

Bilukha, Oleg, Alexia Couture, Kelly McCain, and Eva Leidman. 2020. Comparison of Anthropometric Data Quality in Children Aged 6–23 and 24–59 Months: Lessons From Population-Representative Surveys From Humanitarian Settings. *BMC Nutrition* 6: article 60. **Link**

Black, Robert E., Cesar G. Victora, Susan P. Walker, Zulfiqar A. Bhutta, Parul Christian, Mercedes de Onis, Majid Ezzati, Sally Grantham-McGregor, Joanne Katz, Reynaldo Martorell, Ricardo Uauy, and the Maternal and Child Nutrition Study Group. 2013. Maternal and Child Undernutrition and Overweight in Low-Income and Middle-Income Countries. *The Lancet* 382(9890): P427–P451. **Link**

Blanton, Curtis J., and Oleg O. Bilukha. 2013. The Probit Approach in Estimating the Prevalence of Wasting: Revisiting Bias and Precision. *Emerging Themes in Epidemiology* 10: article 8. **Link**

Boylan, Sinead, Seema Mihrshahi, Jimmy Chun Yu Louie, Anna Rangan, Norsal Salleh, Ilham Md Ali, Roseyati Dato Paduka, and Timothy Gill. 2017. Prevalence and Risk of Moderate Stunting Among a Sample of Children Aged 0–24 Months in Brunei. *Maternal and Child Health Journal* 21(12): 2256–2266.

Briend, André, Kh. Z. Hasan, K. M. A. Aziz, Bilqis A. Hoque, and F. J. Henry. 1989.

Measuring Change in Nutritional Status: A Comparison of Different Anthropometric Indices and the Sample Sizes Required. *European Journal of Clinical Nutrition* 43(11): 769–778.

Castro Bedriñana, Jorge, and Doris Chirinos Peinado. 2014. Z-Score Anthropometric Indicators Derived from NCHS-1977, CDC-2000 and WHO-2006 in Children Under 5 Years in Central Area of Peru. *Universal Journal of Public Health* (Horizon Research Publishing, San Jose, Calif.) 2(2): 73–81. **Link**

Conkle, Joel, Usha Ramakrishnan, Rafael Flores-Ayala, Parminder S. Suchdev, and Reynaldo Martorell. 2017. Improving the Quality of Child Anthropometry: Manual Anthropometry in the Body Imaging for Nutritional Assessment Study (BINA). *PloS One* 12(12): e0189332. **Link**

Corvalán, Camila, Ricardo Uauy, Aryeh D. Stein, Juliana Kain, and Reynaldo Martorell. 2009. Effect of Growth on Cardiometabolic Status at 4 y of Age. *American Journal of Clinical Nutrition* 90(3): 547–555. **Link**

Crowe, Sonya, Andrew Seal, Carlos Grijalva-Eternod, and Marko Kerac. 2014. Effect of Nutrition Survey 'Cleaning Criteria' on Estimates of Malnutrition Prevalence and Disease Burden: Secondary Data Analysis. *PeerJ* 2: e380. **Link**

de Onis, Mercedes, and Monika Blössner. 1997. *WHO Global Database on Child Growth and Malnutrition*. WHO/NUT/97.4. Geneva: World Health Organization. **Link**

de Onis, Mercedes, and Jean-Pierre Habicht. 1996. Anthropometric Reference Data for International Use: Recommendations from a World Health Organization Expert Committee. *American Journal of Clinical Nutrition* 64(4): 650–658. **Link**

Ecker, Olivier, Jean-François Maystadt, and Zhe Guo. 2019. Can Unconditional Cash Transfers Mitigate the Impact of Civil Conflict on Acute Child Malnutrition in Yemen?: Evidence From the National Social Protection Monitoring Survey. *Middle East and North Africa Regional Program Working Paper* 17. International Food Policy Research Institute (Washington, D.C.). **Link**

El Mouzan, Mohammad I., Abdullah S. Al Herbish, Abdullah A. Al Salloum, Peter J. Foster, Ahmad A. Al Omar, Mansour M. Qurachi, and Tatjana Kecojevic. 2008. Comparison of the 2005 Growth Charts for Saudi Children and Adolescents to the 2000 CDC Growth Charts. *Annals of Saudi Medicine* 28(5): 334–340. **Link**

El Mouzan, Mohammad I., Peter J. Foster, Abdullah S. Al Herbish, Abdullah A. Al Salloum, Ahmad A. Al Omar, Mansour M. Qurachi, and Tatjana Kecojevic. 2009. The Implications of Using the World Health Organization Child Growth Standards in Saudi Arabia. *Nutrition Today* 44(2): 62–70.

Fekadu, Yirgu, Addisalem Mesfin, Demewoz Haile, and Barbara J. Stoecker. 2015. Factors Associated with Nutritional Status of Infants and Young Children in Somali Region, Ethiopia: A Cross-Sectional Study. *BMC Public Health* 15: article 846. **Link**

Frison, Severine, Francesco Checchi, Marko Kerac, and Jennifer Nicholas. 2016. Is Middle-Upper Arm Circumference "Normally" Distributed? Secondary Data Analysis of 852 Nutrition Surveys. *Emerging Themes in Epidemiology* 13: article 7. **Link**

Frongillo, Edward A., Shibani Kulkarni, Sulochana Basnet, and Filipa de Castro. 2017. Family Care Behaviors and Early Childhood Development in Low- and Middle-Income Countries. *Journal of Child and Family Studies* 26(11): 3036–3044.

Garenne, Michel, Douladel Willie, Bernard Maire, Olivier Fontaine, Roger Eeckels, André Briend, and Jan Van den Broeck. 2009. Incidence and Duration of Severe Wasting in Two African Populations. *Public Health Nutrition* 12(11): 1974–1982. **Link**

Garza, Cutberto, and Mercedes de Onis. 2004. Rationale for Developing a New International Growth Reference. *Food and Nutrition Bulletin* 25: S5–S14. **Link**

Golden, Michael H. 2009. Proposed Recommended Nutrient Densities for Moderately Malnourished Children. *Food and Nutrition Bulletin* 30(3): S267–S342. **Link**

Golden, Michael H., and Yvonne Grellety. 2002. Population Nutritional Status During Famine. Presented at the Standardized Monitoring and Assessment of Relief and Transitions (SMART) Workshop, University of Aberdeen (Aberdeen, UK). **Link**

Gorstein, Jonathan, Kevin Sullivan, Ray Yip, Mercedes de Onis, Frederick Trowbridge, Peter Fajans, and Graeme Clugston. 1994. Issues in the Assessment of Nutritional Status Using Anthropometry. *Bulletin of the World Health Organization* 72(2): 273–283. **Link**

Grellety, Emmanuel, and Michael H. Golden. 2016. The Effect of Random Error on Diagnostic Accuracy Illustrated with the Anthropometric Diagnosis of Malnutrition. *PloS One* 11(12): e0168585. **Link**

Grellety, Emmanuel, and Michael H. Golden. 2018. Change in Quality of Malnutrition Surveys Between 1986 and 2015. *Emerging Themes in Epidemiology* 15: article 8. **Link**

Grummer-Strawn, Laurence M., Chris Reinold, and Nancy F. Krebs. 2010. Use of World Health Organization and CDC Growth Charts for Children Aged 0–59 Months in the United States. *Morbidity and Mortality Weekly Report* 59(RR-9): 1–15. **Link**

Gupta, Priya M., Emily Wieck, Joel Conkle, Kristina A. Betters, Anthony Cooley, Selena Yamasaki, Natasha Laibhen-Parkes, and Parminder S. Suchdev. 2020. Improving Assessment of Child Growth in a Pediatric Hospital Setting. *BMC Pediatrics* 20: article 419. **Link**

Hulland, Erin N., Curtis J. Blanton, Eva Z. Leidman, and Oleg O. Bilukha. 2016. Parameters Associated With Design Effect of Child Anthropometry Indicators in Small-Scale Field Surveys. *Emerging Themes in Epidemiology* 13: article 13. **Link**

Jacob, Amita, Leah Thomas, Kezia Stephen, Sam Marconi, Joseph Noel, K. S. Jacob, and Jasmin Prasad. 2016. Nutritional Status and Intellectual Development in Children: A Community-Based Study from Rural Southern India. *National Medical Journal of India* 29(2): 82–84. **Link**

Jelliffe, Derrick B. 1966. *The Assessment of the Nutritional Status of the Community (with Special Reference to Field Surveys in Developing Regions of the World)*. Geneva: World Health Organization. **Link**

Joint FAO/WHO Expert Committee on Nutrition. 1971. *Joint FAO/WHO Expert Committee on Nutrition, Eighth Report: Food Fortification, Protein-Calorie Malnutrition*. Geneva: World Health Organization. **Link**

Kaiser Family Foundation (KFF). 2016. U.S. Funding for International Nutrition Programs. April 25. Henry J. Kaiser Family Foundation (San Francisco). **Link**

Krasevec, Julia, Xiaoyi An, Richard Kumapley, France Bégin, and Edward A. Frongillo. 2017. Diet Quality and Risk of Stunting Among Infants and Young

Children in Low- and Middle-Income Countries. *Maternal & Child Nutrition* 13(Supp. 2): e12430. **Link**

Krieger, Nancy. 2012. Who and What Is a "Population"? Historical Debates, Current Controversies, and Implications for Understanding "Population Health" and Rectifying Health Inequities. *Milbank Quarterly* 90(4): 634–681. **Link**

Kwena, Arthur M., Dianne J. Terlouw, Sake J. de Vlas, Penelope A. Phillips-Howard, William A. Hawley, Jennifer F. Friedman, John M. Vulule, Bernard L. Nahlen, Robert W. Sauerwein, and Feiko O. ter Kuile. 2003. Prevalence and Severity of Malnutrition in Pre-School Children in a Rural Area of Western Kenya. *American Journal of Tropical Medicine and Hygiene* 68: 94–99. **Link**

Larsen, Anna Folke, Derek Headey, and William A. Masters. 2019. Misreporting Month of Birth: Diagnosis and Implications for Research on Nutrition and Early Childhood in Developing Countries. *Demography* 56(2): 707–728. **Link**

Leidman, Eva, Louise Masese Mwirigi, Lucy Maina-Gathigi, Anna Wamae, Andrew Amina Imbwaga, and Oleg O. Bilukha. 2018. Assessment of Anthropometric Data Following Investments to Ensure Quality: Kenya Demographic Health Surveys Case Study, 2008 to 2009 and 2014. *Food and Nutrition Bulletin* 39(3): 406–419. **Link**

Martins, Rita C. B., José E. Corrente, Ana Paula Vioto, Emilia Alonso Balthazar, and Maria Rita Marques de Oliveira. 2010. Body Mass Index as Indicator of the Nutritional Status of School Children from the Public Schools of Piracicaba-SP from 2003 to 2006. In *Proceedings of the 9th Brazilian Conference on Dynamics, Control and Their Applications*, 1106–1112. São Carlos, Brazil: Brazilian Society of Applied and Computational Mathematics. **Link**

Matejka, Justin, and George Fitzmaurice. 2017. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. In *CHI'17: Proceedings of the 2017 ACM SIGCHI Conference on Human Factors in Computing Systems*, 1290–1294. New York: Association for Computing Machinery. **Link**

McCloskey, Deirdre N. 1994. Economics: Art or Science or Who Cares? *Eastern Economic Journal* 20(1): 117–120.

McCloskey, Deirdre N., and Stephen T. Ziliak. 1996. The Standard Error of Regressions. *Journal of Economic Literature* 34(1): 97–114.

Mei, Zuguo, and Laurence M. Grummer-Strawn. 2007. Standard Deviation of Anthropometric Z-Scores as a Data Quality Assessment Tool Using the 2006 WHO Growth Standards: A Cross Country Analysis. *Bulletin of the World Health Organization* 85: 441–448. **Link**

Mulugeta, Afework, Fitsum Hagos, Gideon Kruseman, Vincent Linderhof, Barbara Stoecker, Zenebe Abraha, Mekonen Yohannes, and Girmay G. Samuel. 2010. Child Malnutrition in Tigray, Northern Ethiopia. *East African Medical Journal* 87(6): 248–254. **Link**

Perumal, Nandita, Sorrel Namaste, Huma Qamar, Ashley Aimone, Diego G. Bassani, and Daniel E. Roth. 2020. Anthropometric Data Quality Assessment in Multisurvey Studies of Child Growth. *American Journal of Clinical Nutrition* 112(Supp. 2): 806S–815S. **Link**

Quételet, Adolphe. 1832. Recherches Sur Le Poids De L'homme Aux Différens Âges [Research on the Weight of Man at Different Ages]. *Nouveaux mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles* 7: 1–83.

Quételet, Adolphe. 1835. *Sur L'homme Et Le Développement De Ses Facultés* [*On Man and the Development of His Faculties*]. Paris: Bachelier.

Ravallion, Martin. 1997. Famines and Economics. *Journal of Economic Literature* 35(3): 1205–1242.

Razak, Fahad, George Davey Smith, and Sankaran Venkata Subramanian. 2016. The Idea of Uniform Change: Is It Time to Revisit a Central Tenet of Rose's "Strategy of Preventive Medicine"? *American Journal of Clinical Nutrition* 104(6): 1497–1507. **Link**

Reda, Ayalu A., and David Lindstrom. 2014. Recent Trends in the Timing of First Sex and Marriage Among Young Women in Ethiopia. *African Population Studies* 28(2 Supp.): 1157–1170. **Link**

Remans, Roseline, Paul M. Pronyk, Jessica C. Fanzo, Jiehua Chen, Cheryl A. Palm, Bennett Nemser, Maria Muniz, Alex Radunsky, Alem Hadera Abay, Mouctar Coulibaly, Joseph Mensah-Homiah, Margaret Wagah, Xiaoyi An, Christine Mwaura, Eva Quintana, Marie-Andree Somers, Pedro A. Sanchez, Sonia E. Sachs, John W. McArthur, and Jeffrey D. Sachs. 2011. Multisector Intervention to Accelerate Reductions in Child Stunting: An Observational Study from 9 Sub-Saharan African Countries. *American Journal of Clinical Nutrition* 94(6): 1632–1642. **Link**

Rose, Geoffrey. 1992. *The Strategy of Preventive Medicine*. New York: Oxford University Press.

Roth, Daniel E., Aditi Krishna, Michael Leung, Joy Shi, Diego G. Bassani, and Aluisio J. D. Barros. 2017. Early Childhood Linear Growth Faltering in Low-Income and Middle-Income Countries as a Whole-Population Condition: Analysis of 179 Demographic and Health Surveys from 64 Countries (1993–2015). *Lancet Global Health* 5(12): e1249–e1257. **Link**

Samiak, Louis, and Theophilus I. Emeto. 2017. Vaccination and Nutritional Status of Children in Karawari, East Sepik Province, Papua New Guinea. *PloS One* 12(11): e0187796. **Link**

Sen, Amartya. 1984. *Resources, Values, and Development*. Oxford: Basil Blackwell.

Sharma, Nikita, Madhu Gupta, Arun Kumar Aggarwal, and Mutyalamma Gorle. 2020. Effectiveness of a Culturally Appropriate Nutrition Educational Intervention Delivered Through Health Services to Improve Growth and Complementary Feeding of Infants: A Quasi-Experimental Study from Chandigarh, India. *PloS One* 15(3): e0229755. **Link**

Shekar, Meera, Jakub Kakietek, Julia Dayton Eberwein, and Dylan Walters. 2017. *An Investment Framework for Nutrition: Reaching the Global Targets for Stunting, Anemia, Breastfeeding, and Wasting*. Washington, D.C.: World Bank. **Link**

Siegel, Jacob S., and David A. Swanson. 2004. *The Methods and Materials of Demography*. London: Elsevier Academic Press.

Spiegel, Paul B., Peter Salama, Susan Maloney, and Albertien van der Veen. 2004. Quality of Malnutrition Assessment Surveys Conducted During Famine in Ethiopia. *Journal of the American Medical Association* 292(5): 613–618. **Link**

Standardized Monitoring & Assessment of Relief & Transitions (SMART). 2006. *Measuring Mortality, Nutritional Status, and Food Security in Crisis Situations: SMART Methodology*. Toronto: Action Against Hunger Canada. **Link**

Subramanian, Shankar, and Angus Deaton. 1996. The Demand for Food and Calories. *Journal of Political Economy* 104(1): 133–162.

Tanner, James M. 1952. The Assessment of Growth and Development in Children. *Archives of Disease in Childhood* 27: 10–33. **Link**

Tanner, James M. 1981. *A History of the Study of Human Growth*. Cambridge, UK: Cambridge University Press.

Tuffrey, Veronica, and Andrew Hall. 2016. Methods of Nutrition Surveillance in Low-Income Countries. *Emerging Themes in Epidemiology* 13: article 4. **Link**

Ulijaszek, Stanley J., and Deborah A. Kerr. 1999. Anthropometric Measurement Error and the Assessment of Nutritional Status. *British Journal of Nutrition* 82(3): 165–177. **Link**

UNICEF, WHO, and World Bank. 2018. Levels and Trends in Child Malnutrition: Key Findings of the 2018 Edition. World Health Organization (Geneva). **Link**

USAID. 2016. Anthropometric Data in Population-Based Surveys: Meeting Report, Washington, DC, July 14–15, 2015. Food and Nutrition Technical Assistance III Project (Washington, D.C.). **Link**

Varghese, Jithin Sam, and Aryeh D. Stein. 2019. Malnutrition Among Women and Children in India: Limited Evidence of Clustering of Underweight, Anemia, Overweight, and Stunting Within Individuals and Households at Both State and District Levels. *American Journal of Clinical Nutrition* 109(4): 1207–1215. **Link**

Wasserstein, Ronald L., Allen L. Schirm, and Nicole A. Lazar. 2019. Moving to a World Beyond "$p < 0.05$." *American Statistician* 73(Supp. 1): 1–19. **Link**

Waterlow, J. C., R. Buzina, W. Keller, J. M. Lane, M. Z. Nichaman, and J. M. Tanner. 1977. The Presentation and Use of Height and Weight Data for Comparing the Nutritional Status of Groups of Children under the Age of 10 Years. *Bulletin of the World Health Organization* 55(4): 489–498. **Link**

WHO Working Group. 1986. Use and Interpretation of Anthropometric Indicators of Nutritional Status. *Bulletin of the World Health Organization* 64(6): 929–941. **Link**

WHO Expert Committee on Physical Status. 1995. *Physical Status: The Use of and Interpretation of Anthropometry*. Geneva: World Health Organization.

WHO. 2006. *WHO Child Growth Standards: Length/Height-for-Age, Weight-for-Age, Weight-for-Length, Weight-for-Height and Body Mass Index-for-Age: Methods and Development*. Geneva: World Health Organization. **Link**

WHO and UNICEF. 2019. *Recommendations for Data Collection, Analysis and Reporting on Anthropometric Indicators in Children Under 5 Years Old*. Geneva: World Health Organization. **Link**

Wijaya-Erhardt, Maria. 2019. Nutritional Status of Indonesian Children in Low-Income Households with Fathers That Smoke. *Osong Public Health and Research Perspectives* (Korea Disease Control and Prevention Agency, Cheongju) 10(2): 64–71. **Link**

Woodruff, Bradley A., James P. Wirth, Adam Bailes, Joan Matji, Arnold Timmer, and Fabian Rohner. 2017. Determinants of Stunting Reduction in Ethiopia 2000–2011.

Maternal & Child Nutrition 13(2): e12307. Link

Yip, Ray. 1993. Expanded Usage of Anthropometry Z-Scores for Assessing Population Nutritional Status and Data Quality. Abstract. In *Abstracts Book No. 1, 15th International Congress of Nutrition (Adelaide)*, 279. Adelaide: International Union of Nutritional Sciences.

Ziliak, Stephen T., and Deirdre N. McCloskey. 2008. *The Cult of Statistical Significance: How the Standard Error Is Costing Jobs, Justice, and Lives*. Ann Arbor: University of Michigan Press.

About the Author

Austin Sandler is a Ph.D. candidate in the Department of Geographical Sciences at the University of Maryland and Consultant at The World Bank Group. He holds an M.S. in Applied Economics from the University of Minnesota and an M.S. in Agricultural and Applied Economics from the University of Wyoming. His teaching includes principles of economics and human geography, mapping and geographic information science, and economic geography. His research covers economic development and growth, food security and childhood malnutrition, spatial economics and econometrics, and history of scientific thought. His email address is sandlera@terpmail.umd.edu.

Go to archive of Economics in Practice section
Go to March 2021 issue