

Defining Diarrhea: A Population-Based Validation Study of Caregiver-Reported Stool Consistency in the Amhara Region of Ethiopia

Kristen Aiemojy,^{1,2*} Solomon Aragie,³ Sintayehu Gebresillasie,³ Dionna M. Fry,¹ Adane Dagne,³ Dagnachew Hailu,³ Melsew Chanyalew,⁴ Zerihun Tadesse,³ Aisha Stewart,⁵ Kelly Callahan,⁵ Mathew Freeman,⁶ John Neuhaus,² Benjamin F. Arnold,⁷ and Jeremy D. Keenan¹

¹Francis I. Proctor Foundation, University of California, San Francisco, San Francisco, California; ²Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California; ³The Carter Center Ethiopia, Addis Ababa, Ethiopia; ⁴Amhara Regional Health Bureau, Bahir Dar, Ethiopia; ⁵The Carter Center, Atlanta, Georgia; ⁶Rollins School of Public Health, Emory University, Atlanta, Georgia; ⁷Division of Epidemiology, School of Public Health, University of California, Berkeley, Berkeley, California

Abstract. Diarrhea is a leading cause of death among children aged less than five years globally. Most studies of pediatric diarrhea rely on caregiver-reported stool consistency and frequency to define the disease. Research on the validity of caregiver-reported diarrhea is sparse. We collected stool samples from 2,398 children participating in two clinical trials in the Amhara region of Ethiopia. The consistency of each stool sample was graded by the child's caregiver and two trained laboratory technicians according to an illustrated stool consistency scale. We assessed the reliability of graded stool consistency among the technicians, and then compared the caregiver's grade with the technician's grade. We also tested if the illustrated stool consistency scale could improve the validity of caregiver's report. The weighted kappa measuring the agreement between the two laboratory technicians reached 0.90 after 500 stool samples were graded. The sensitivity of caregiver-reported loose or watery stool was 15.5% (95% confidence interval [CI]: 9.7, 24.2) and the specificity was 98.4% (95% CI: 97.1, 99.1). With the illustrated scale, the sensitivity was 68.5% (95% CI: 58.5, 77.1) and the specificity was 86.1% (95% CI: 79.3, 90.9). The results indicate that caregiver-reported stool consistency using the terms "loose or watery" does not accurately describe stool consistency as graded by trained laboratory technicians. Given the predominance of using caregiver-reported stool consistency to define diarrheal disease, the low sensitivity identified in this study suggests that the burden of diarrheal disease may be underestimated and intervention effects could be biased. The illustrated scale is a potential low-cost tool to improve the validity of caregiver-reported stool consistency.

INTRODUCTION

Diarrhea is a leading cause of childhood morbidity and mortality with an estimated 2 billion cases and 525,000 deaths annually.¹ In Ethiopia, diarrhea is the second leading cause of death among children aged less than five years, causing more deaths than human immunodeficiency virus, Tuberculosis, and malaria combined.^{2,3}

Most trials and epidemiologic studies of childhood diarrhea use caregiver reports of stool consistency and frequency to characterize disease status.⁴ The World Health Organization (WHO) definition of diarrhea is three or more loose or watery stools in a 24-hour period.⁵ In three recent systematic reviews of water, sanitation and hygiene (WASH) intervention trials to prevent pediatric diarrheal disease, all 22 studies used caregiver-reported symptoms to classify diarrhea.^{6–10} In a 2015 review of 55 studies of water quality interventions for reducing diarrheal disease, 36 used the WHO definition and 11 used another symptoms-based report.¹¹

Despite its widespread use, there is limited research on the validity of various definitions of diarrhea and the component items in those definitions (stool consistency and frequency).^{4,12–15} This knowledge gap is important, given misclassified caregiver-reported stool consistency could introduce measurement error and bias. True protective effects of interventions on diarrheal disease may be undetectable when measurement error is present. Underreporting of moderate and severe diarrhea may underestimate the disease burden and resulting cost-effectiveness of interventions designed to mitigate that burden.

Visual and descriptive stool consistency scales may standardize and improve the accuracy of reported stool consistency. A widely used stool form scale, The Bristol Stool Form Scale (BSFS), was developed in the late 1980s to measure gut transit time^{16,17} and later simplified to a five-level scale: the modified Bristol Stool Form Scale for children (mBSFS-C).¹⁸ The mBSFS-C could be used as a tool for eliciting self-reported or caregiver-reported diarrhea from epidemiological studies. However, to our knowledge, it has never been evaluated in a research setting in Africa.

Our objective was 3-fold: measure the inter-rater reliability of mBSFS-C among laboratory technicians, validate caregiver-reported stool consistency using laboratory technician-graded consistency as the reference standard, and determine if the mBSFS-C can improve the validity of caregiver-reported stool consistency.

METHODS

Study population. This study was conducted within two cluster-randomized trials studying WASH interventions in rural Ethiopia. In one, 14 communities in the East Gojjam zone were randomized to receive either a hand-dug well or no intervention^{19–21} (labeled Trial I in this report; clinicaltrials.gov NCT02373657). In the other one, 40 communities in the Wag Himra zone were randomized to either a comprehensive WASH package or no intervention (labeled Trial II in this report; clinicaltrials.gov NCT02754583). The primary outcome of both trials was ocular *Chlamydia*. The present study collected data at the final study visit of Trial I (April 2016) and the baseline visit of Trial II (January 2016).

In each study, we conducted a door-to-door population census approximately 3 weeks before the study visit to enumerate households and children eligible to participate in data

* Address correspondence to Kristen Aiemojy, Proctor Foundation, University of California, San Francisco, 513 Parnassus Avenue, MedSci S309, P.O. Box 0412, San Francisco, CA 94143. E-mail: kristen.aiemojy@ucsf.edu

collection. Based on sample size calculations for the primary outcome, in Trial I, all censused children aged 0–5 years were eligible to participate. In Trial II, a random sample of 40 children aged 0–5 years (up to their sixth birthday) and 40 children aged 6–9 years per community were eligible; if less than 40 children of a specific age group were censused, then all children in that age group were sampled. In each case, the sampling strategy was based on power calculations for the primary outcome of the trial.

Stool sample grading: reference standard. The mBSFS-C is a five-level adaptation of the original seven-level BSFS.^{18,22} It was five stool consistency categories with both cartoon depictions and descriptors: 1) hard lumps, 2) sausage-shaped but lumpy, 3) sausage-shaped and soft, 4) loose, and 5) watery (Figure 1). An Ethiopian clinician fluent in the Amharic and English languages translated the mBSFS-C into Amharic. The translations were adapted to cultural norms through consensus with three Amharic- or English-speaking clinicians. It was back-translated into English to check comparability with the original scale. The descriptive words “nuts” and “sausage” were not used because these are not part of a standard diet in rural Ethiopia.

Laboratory technicians attended a 2-day classroom training on stool sample collection and consistency grading using the mBSFS-C and a 1-day field training with real stool samples before data collection started. We emphasized the importance of masking the consistency grades during both classroom and field training.

Stool sample collection. Data were collected during the regularly scheduled study visits for the parent trials, in a centralized location in each community. Caregivers were instructed to take their children to a semiprivate outdoor place near the sample collection area and have their child defecate in a potty chair lined with a black plastic bag. For children unable to produce a stool within 2 hours, caregivers were asked to collect stool at home and bring it to a collection site the following day. All stool samples were graded in the field before they were set in a preservative and transported.

When the stool was returned to the field station, it was immediately inspected in the original collection container by two

medical laboratory technicians and the consistency of the sample was independently graded according to the mBSFS-C. The technicians were masked to each other’s grades. Masking was achieved by having the first technician silently enter their grade into custom-designed software on a smart-phone; once entered, this grade was immediately concealed in the application and impossible to change. The second technician then recorded their grade. The technicians were allowed to discuss their grades after both grades had been entered. The mBSFS-C diagram was available at the point of data collection in the data collection software and also as a laminated sheet.

Stool sample grading: caregiver report of “loose or watery” stool (index test 1). After the stool sample was graded by both laboratory technicians, the caregiver was asked: “Did your child have a loose or watery stool?” The wording of the question was designed to mimic stool consistency element of the standard definition of diarrhea: three or more loose or watery stools in a 24-hour period.⁵

Stool sample grading: caregiver mBSFS-C grade (index test 2). The caregiver was then shown the laminated color copy of the mBSFS-C with Amharic descriptions and asked to point to the consistency category most similar to their child’s stool.

Statistical methods. Agreement of reference standard. We assessed agreement in the five-level mBSFS-C stool consistency classification between the two graders using both an unweighted and quadratic-weighted kappa^{23,24} and used a bootstrap with 1,000 replicates to calculate bias-corrected 95% confidence intervals (CIs) with resampling by community.

We used a K-by-K confusion matrix to visualize absolute agreement and partial agreement for the mBSFS-C, where the first technician’s grades (in columns) are classified against the second technician’s grades (in rows).

We also compared the unweighted and weighted kappa by the number of samples graded, in increments of 100, to evaluate a change in kappa over time. We specified a kappa of 0.9 or greater between graders to use the first laboratory technician’s grade as the gold standard.






	Original mBSFS-C	Amharic	Back-translated to English
Type 1	 Separate hard lumps, like nuts (hard to pass)	በጣም የደረቀ ሰገራ (በጠጥ) ለመውጣት የሚስቸግር	Very dry stool (small and round like sheep feces), hard to pass
Type 2	 Sausage shapes but lumpy	የደረቀ ሰገራ (የተያያዘ በጠጥ ይመስላል)	Dry stool (a single mass of small round feces, like sheep feces formed together)
Type 3	 Like a sausage or snake, smooth and soft	ለለላ ለገና ደረቅ ያልሆነ የለብ ቅርፅ ያለው	Soft, not dry and its shape is like snake
Type 4	 Fluffy pieces with ragged edges, a mushy stool	በጣም ለለላ ቅርፅ የሌለው	Very soft and irregular shaped
Type 5	 Watery, no solid pieces	ቀጭን ተቅጣጥ	Watery stool

FIGURE 1. Modified Bristol Stool Form Scale for children (mBSFS-C), translated into Amharic. This figure appears in color at www.ajtmh.org.

Validity of caregiver report of “loose or watery” stool (index test 1). We compared the caregiver report of “loose or watery” stool consistency with the first laboratory technician grade (reference standard). We dichotomized the laboratory technician’s mBSFS-C grade with types 4 and 5 qualifying as “loose or watery” stool and types 1–3 as “not loose or watery.” We calculated the sensitivity and specificity using separate logit models with the dichotomous result of the index test as the dependent variable conditional on the reference standard being positive (sensitivity) or negative (specificity). We used a clustered sandwich estimator to adjust standard errors for clustering by community.^{25,26} Definitions of the validity measures are given in Table 1.

We stratified the models by caregiver type (mother or father). To test for an interaction between caregiver grade and type, we fit a logit model with the outcome of the index test as the dependent variable and the reference standard as an independent variable and an interaction term with caregiver type.^{27,28}

Validity and agreement of caregiver mBSFS-C grade (index test 2). We dichotomized the caregiver’s mBSFS-C into types 4 and 5 (loose or watery) and types 1–3 (not loose or watery) and compared this against the similarly dichotomized grade of the first laboratory technician (the reference standard). We calculated the sensitivity and specificity using the same method as for index test 1, with separate logit models for sensitivity and specificity and a clustered sandwich estimator to account for clustering by community.

To test if the mBSFS-C improved the sensitivity and specificity of caregiver-reported stool consistency, we used a logistic mixed-effects model with a random intercept for both community and child, to account for the paired comparison. We included an indicator variable for caregiver report with or without the mBSFS-C^{25,26} and used the Wald test for the coefficient of this indicator variable to evaluate statistical significance. The mixed-model yields estimates of median sensitivity and specificity conditional on the random effect; in this case, the child and the community.²⁶

We also evaluated agreement in five-level mBSFS-C stool consistency grade between caregivers and the laboratory technician using both a weighted and quadratic-weighted kappa and used a bootstrap with 1,000 replicates to calculate bias-corrected 95% CIs with resampling by community.

Analyses were run in Stata 15 (StataCorp, College Station, TX). Figures were generated in R Studio using R Version 3.4.1 (Foundation of Open Source Statistics, Boston, MA).

Ethics statement. Ethical committees at the University of California (San Francisco, CA); Emory University (Atlanta, GA); The Food, Medicine and Health Care Administration and Control Authority of Ethiopia; and the Ethiopian Ministry of Science

and Technology granted approval for this study. We obtained verbal informed consent in Amharic from all caregivers.

RESULTS

Characteristics of graders. Trial I used three medical laboratory technicians and one clinical nurse. Three worked at government clinics and one worked at a university clinic. The average number of years of experience was 5.5 years (standard deviation [SD] 3.9, range 1–10). Trial II used eight medical laboratory technicians as graders: seven from clinics operated by the ministry of health and one from a university clinic. On average, graders had 7.4 years of experience (SD 3.7, range 2–14). One medical laboratory technician was used in both studies (Table 2).

Characteristics of study population. A flow diagram of sampling and participation is shown in Figure 2. In Trial I, all 446 censused children were eligible to participate, 317 children presented for the study visit examination day, and 271 provided stool samples. The mean age of children with stool samples was 2.7 years, 48.3% (152/271) of the children were female, and 63% (170/271) of the caregivers were mothers. In Trial II, 2,400 children aged 0–9 years were randomly sampled, 2,362 children presented for the study visit examination day, and 2,127 children provided stool samples. Sixteen stool samples (0.75%) were collected the day after the study visit for children unable to produce a stool on the day of the study visit. Of children with stool samples, the mean age was 5.1 years, 51.2% (1,233/2,127) of the children were female, and 69.6% (1,678/2,127) of the caregivers were mothers (Table 3).

Agreement of reference standard. In Trial I, the two laboratory technicians agreed on 169/271 (62.6%) of the five-level mBSFS-C grades, with an unweighted kappa of 0.50 (95% CI 0.42, 0.57) and a quadratic-weighted kappa of 0.70 (0.61, 0.77). In Trial II, 1,870/2,127 (87.9%) of grades were in agreement, with an unweighted kappa of 0.84 (95% CI 0.82, 0.86) and a quadratic-weighted kappa of 0.92 (95% CI: 0.90, 0.93). See Table 4 for the K-by-K confusion matrix.

Kappa increased with the number of samples graded (Figure 3). When restricted to the first 271 samples (the size of Trial I), the unweighted kappa for Trial II [0.56; 95% CI 0.49, 0.62] was comparable to Trial I [0.50 (95% CI: 0.49, 0.53)] and the 95% CIs overlapped. In Trial II, the unweighted kappa surpassed 0.90 after 600 samples were assessed and the weighted kappa surpassed 0.90 after 500 samples were assessed by four field teams, approximately 125 samples per team. In Trial I, the weighted kappa did not reach 0.9.

To ensure robustness of our reference standard (laboratory technician-graded stool consistency) for the validity research questions in this study, we opted to permit a run-in period of

TABLE 1
Definition of validity measures

Caregiver assessment		First technician’s grade reference standard		
Index test 1	Index test 2	Loose/watery	Not loose/watery	Total
“Loose or watery” report	mBSFS-C grade	mBSFS-C: types 4 and 5	mBSFS-C: types 1–3	
Loose/watery	mBSFS-C: types 4 and 5	TP	FP	TP + FP
Not loose/watery	mBSFS-C: type 1–3	FN	TN	FN + TN
Total		TP + FN	FP + TN	

FN = false negative; FP = false positive; mBSFS-C = modified Bristol Stool Form Scale for children; TN = true negative; TP = true positive. Index test 1 = caregiver-reported loose or watery stool: “Did your child have a loose or watery stool?” Index test 2 = caregiver grade after visual inspection of the stool sample using the mBSFS-C (types 4 and 5 = loose or water; types 1–3 = not loose or watery). Reference standard = first laboratory technician-graded stool consistency according to the mBSFS-C (types 4 and 5 = loose or water; types 1–3 = not loose or watery). Sensitivity = TP/(TP + FN). Specificity = TN/(FP + TN).

TABLE 2
Characteristics of graders in Trial I and Trial II

	Trial I	Trial II
	N = 4	N = 8
Age in years, mean (SD)	27.5 (4.2)	28.9 (4.3)
Gender		
Female	1 (25%)	1 (13%)
Male	3 (75%)	7 (87%)
Profession		
Clinical nurse	1 (25%)	0
Medical laboratory technician	3 (75%)	8 (100%)
Years of experience, mean (SD)	5.5 (3.9)	7.4 (3.7)
Employer		
University	1 (25%)	1 (13%)
Ministry of health	3 (75%)	7 (87%)

SD = standard deviation. Numbers are *n* (%) unless otherwise indicated.

500 samples until the weighted kappa exceeded 0.90. Thus, the remaining validation research questions evaluate samples 501 through 2,127 in Trial II only (1,627 samples in total). Characteristics of the study population in the validation sample (allowing for the 500-sample run-in) are displayed in Table 3.

Validity of caregiver report of “loose or watery” stool (index test 1). Caregivers reported that 5.4% (87/1,627) of samples were “loose or watery” whereas laboratory technicians graded 26.7% (435/1,628) of samples as mBSFS-C types 4 and 5 (i.e., equivalent to “loose or watery”). The overall sensitivity of caregiver-reported “loose or watery” stool consistency was 15.6% (68/435; 95% CI: 9.7, 24.2) and the overall specificity was 98.4% (1,173/1,192; 95% CI: 97.1, 99.1). The sensitivity of the mother’s reported stool consistency was higher than that of the father’s report—16.8% (66/392; 95% CI: 10.4, 26.1) versus 3.0% (1/33; 95% CI: 0.4, 20.4); The specificity was 98.8% (941/952; 95% CI: 97.7, 99.4) for mothers and 95.4% (165/173; 95% CI: 87.8, 98.3) for fathers. The *P* value for the interaction term between caregiver type and grade was 0.004 (Table 5). A K-by-K confusion matrix with caregiver loose or water grade compared to technician’s 5-level mBSFS-C grade is presented in Supplemental Table 1.

Validity and agreement of caregiver mBSFS-C grade (index test 2). When caregivers used the mBSFS-C to grade

stool consistency, the overall sensitivity was 68.5% (298/435; 95% CI: 58.5, 77.1) and the specificity was 86.1% (1,026/1,192; 95% CI: 79.3, 90.9). The sensitivity was significantly higher when mothers used the mBSFS-C (index test 2) than when they did not (index test 1); *P* < 0.00001 in a mixed model with random intercepts for child and community.

Sensitivity improved for both mothers (271/392; 69.1%, 95% CI: 58.1, 78.4) and fathers (20/33; 60.6%, 95% CI: 46.1, 73.4); Specificity was 86.8% (826/952; 95% CI: 79.6, 91.7) for mothers and 88.4% (153/173; 95% CI: 82.2, 92.7) for fathers (Table 5). See Supplemental Figure 1 for a receiver operating characteristic space plot visualization of the change in sensitivity with and without the mBSFS-C by caregiver type.

The unweighted kappa using all five levels of the mBSFS-C between the caregiver’s grade and the laboratory technician reference standard was 0.35 (95% CI: 0.32–0.38); the quadratic-weighted kappa was 0.49 (0.44–0.53). The K-by-K confusion matrix is presented in Supplemental Table 2.

DISCUSSION

We documented the validity of caregiver-reported “loose or watery” stool consistency in two trials in the Amhara Region of Ethiopia using stool samples from 1,627 children. Our findings indicate that caregiver-reported “loose or watery” stool consistency has poor validity when compared with laboratory technician-graded stool “loose or watery” stool consistency. The low sensitivity is concerning given the terms “loose or watery” are key components of the widely used WHO definition of diarrhea: “three or more loose or watery stools in a 24-hour period.” The degree of misclassification suggests epidemiologic studies, randomized control trials, and global burden of disease estimates that rely on caregiver-reported “loose or watery” stool to define diarrheal disease in children may underestimate the prevalence of diarrhea and report potentially biased measures of association.

Symptom-based definitions of diarrhea are pervasive in epidemiology and clinical research, yet few studies have attempted to validate these measurements.^{4,12} The WHO definition of diarrhea is based on a 1991 longitudinal study of 512 children in Bangladesh investigating four definitions of

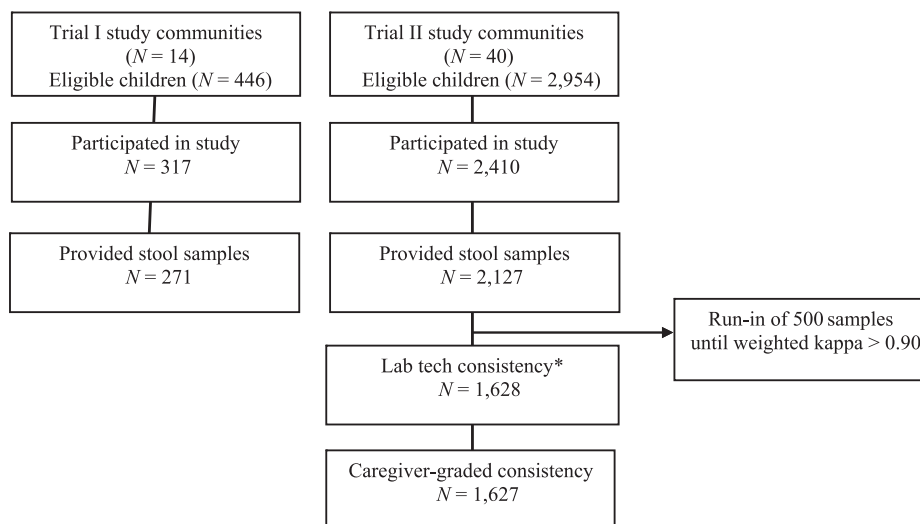


FIGURE 2. Selection and participation flow diagram.

TABLE 3
Characteristics of study populations

	Trial I		Trial II		Validation sample population* (Trial II only) N = 1,627
	Study communities (N = 14)		Study communities (N = 40)		
	Study population (N = 271)		Study population (N = 2,127)		
Age in years, mean (SD)	2.8 (1.9)		5.1 (2.7)		5.3 (2.7)
Female	152 (48.3%)		1,233 (51.2%)		834 (51.2)
Caregiver type					
Mother	170 (63%)		1,678 (69.6%)		1,344 (82.6)
Father	26 (8.3%)		296 (12.3%)		206 (12.7)
Aunt/uncle	14 (4.4%)		46 (1.9%)		26 (1.6)
Sibling	44 (14%)		18 (0.7%)		6 (0.4)
Self	13 (4.1%)		78 (3.2%)		42 (2.6)
Stool consistency type (according to reference standard)					
Type 1: pellets	14 (4.4%)		381 (15.8%)		290 (17.8)
Type 2: lumpy	89 (28.3%)		677 (28.1%)		491 (30.2)
Type 3: smooth	61 (19.4%)		563 (23.4%)		412 (25.3)
Type 4: loose	81 (25.7%)		407 (16.9%)		376 (22.5)
Type 5: watery	29 (9.2%)		100 (4.1%)		68 (4.18)

SD = standard deviation. Values are n (%) unless otherwise indicated.
* After 500-sample run-in period.

stool consistency and frequency against mothers' perception of diarrhea.²⁹ The definition "three or more loose or watery stools" was chosen because it had the highest sensitivity (77.8%) compared with mother's perception of diarrhea. The definition was not compared with direct observation of the child, stool sample, or a clinical diagnosis of diarrhea. A careful assessment of the WHO definition is likely warranted, given its pervasive use in epidemiology and clinical research.

We found that the mBSFS-C, an illustration of five stool consistency categories, had good agreement when used by local medical laboratory technicians to classify consistency of stool samples in a field-based research setting in Africa.

We specified a kappa of 0.9 or greater between laboratory technician graders to justify using the first technician's grade as the gold standard comparison. A kappa of 0.9 was reached after 500 samples were graded in the larger Trial I. However, in the smaller Trial II with only 271 samples in total, the highest weighted kappa reached was 0.83. Thus, all grades from the smaller trial were excluded from the validation study. The upward trajectory in kappa in the smaller Trial II mirrored that Trial I, signaling that a kappa of 0.9 may have been reached with more time and samples to grade.

The initial creation and assessment of the mBSFS-C in 2010 measured reliability between 14 physician graders using stool photographs with an overall intraclass correlation coefficient of 0.85 (95% CI 0.78, 0.91).¹⁸ Intraclass correlation coefficients are dependent on the variation of stool consistency within a study population and thus are not directly comparable

across study sites. Kappas have been reported for the original seven-level BSFS. A study published in 2016 was the first to assess reliability of the BSFS using real stool samples; when comparing patient and physician grades of stool consistency, 26% were in agreement with a weighted kappa of 0.67.³⁰ These results measure patient versus grader agreement rather than grader versus grader agreement and are comparable with the weighted kappa comparing caregivers grade and the laboratory technician's grade using the mBSFS-C.

Laboratory technicians may not be the best gold standard grader for diarrheal disease. Caregivers who are more familiar with their children's stool patterns may be better able to differentiate "diarrhea" from the normal stool pattern. Some iterations of the WHO definition of diarrhea include three or more loose or watery stools more than what is normal for the individual. Here, we focus our validation study on the specific words "loose or watery." For this narrower definition, the trained laboratory technicians are an appropriate reference standard. For validating the WHO or other definitions of diarrhea, other gold standards may be more appropriate such as a clinician's diagnosis or presence of a specific enteric pathogen.

The mBSFS-C improved the validity of caregiver-reported "loose or watery" stool consistency. The mBSFS-C is an option to improve the validity of caregiver-reported stool consistency when studies rely on caregiver-reported symptoms to define diarrheal disease. Reproducing the scale for use in epidemiologic studies and trials is easy to implement at a low

TABLE 4

K-by-K confusion matrix of caregiver-reported stool consistency using mBSFS-C vs. laboratory technician-graded stool consistency using mBSFS-C

Second laboratory technician	First laboratory technician					Total
	Type 1: pellets	Type 2: lumpy	Type 3: snake	Type 4: loose	Type 5: watery	
Type 1: pellets	261	8	2	1	1	273
Type 2: lumpy	21	460	20	3	0	504
Type 3: snake	8	20	376	16	2	422
Type 4: loose	0	1	12	341	4	358
Type 5: watery	0	2	2	6	61	71
Total	290	491	412	367	68	1,628

mBSFS-C = modified Bristol Stool Form Scale for children. After 500-sample run-in. Bold indicates the total number of grades that agree in each category.

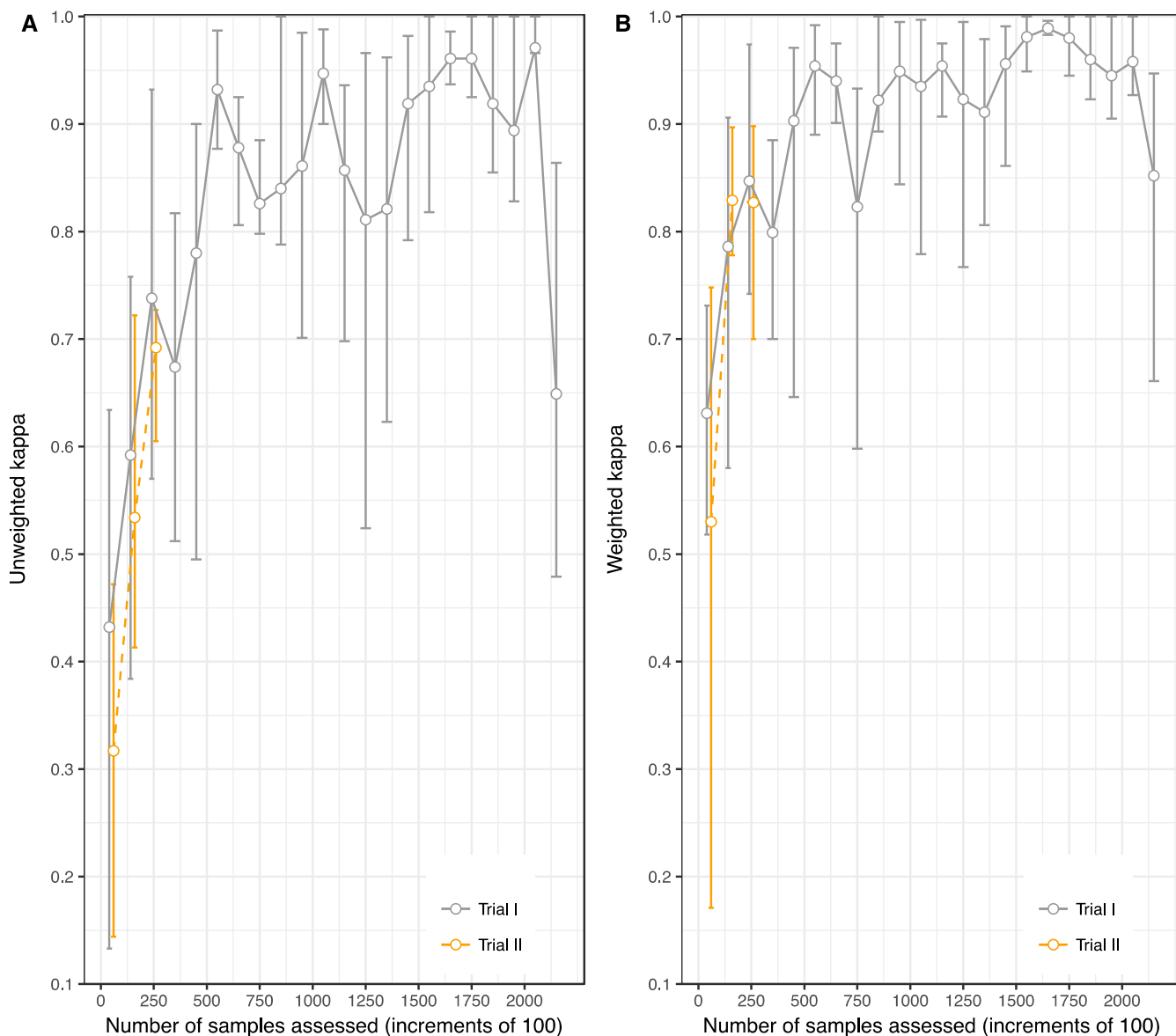


FIGURE 3. Both unweighted (A) and quadratic-weighted kappa (B) increase with the number of samples graded. This figure appears in color at www.ajtmh.org.

cost. Care must be taken to ensure that the stool-consistency descriptor translations are understood and culturally appropriate. In addition, caregivers should be informed that the cartoon pictures depict stool consistency and not merely appearance.

Although the mBSFS-C shows promise in improving caregiver-reported stool consistency, researchers must carefully consider the underlying construct that is of interest. Loose or watery stools can have both infectious and noninfectious causes. Moreover, enteric infections can be both symptomatic and asymptomatic. A pathogen-specific outcome is most likely more expensive than a symptoms-based outcome, but the specificity may be enough to outweigh the difference in costs. However, pathogen-specific outcomes may misclassify disease status because they include asymptomatic cases. Symptomatic outcomes may be contaminated with noninfectious symptomatic diarrhea cases, thus obscuring an effect of an intervention targeting

infection pathways. The most appropriate outcome for measuring diarrheal disease depends on a study's objectives, budget, and tolerance for misclassification.

We acknowledge several limitations of this study. First, we were unable to assess the accuracy of caregiver-reported stool frequency. Measurement error of reported stool frequency (number of bowel movements) is another threat to the validity of symptoms-based reporting of diarrhea and warrants further research. Second, we may have overestimated the sensitivity of caregiver-reported stool consistency by having caregivers report the consistency of a stool sample in front of them. When caregivers respond to traditional survey questions without actually observing their child's recent stools, we might expect more false negatives, and thus the true sensitivity of caregiver-reported stool consistency may be even lower than what we observed in this study. Third, we did not differentiate the stool of breastfed infants, which is typically looser than the stool of solid-fed infants. However, we

TABLE 5
Sensitivity and specificity caregiver's report of 'loose or watery' stool consistency and caregiver's mBSFS-C grade

	TP	TN	FP	FN	Sensitivity	Specificity
Index test 1: caregiver "loose or watery" grade						
Overall (N = 1,627)	68	1,173	19	367	15.6 [9.7, 24.2]*	98.4 [97.1, 99.1]†
Mothers (N = 1,347)	66	941	11	326	16.8 [10.4, 26.1]	98.8 [97.7, 99.4]
Fathers (N = 203)	1	165	8	32	3.0 [0.4, 20.4]	95.4 [87.8, 98.3]
Index test 2: caregiver mBSFS-C grade						
Overall (N = 1,627)	298	1,026	166	137	68.5 [58.5, 77.1]†	86.1 [79.3, 90.9]†
Mothers (N = 1,347)	271	826	126	121	69.1 [58.1, 78.4]	86.8 [79.6, 91.7]
Fathers (N = 203)	20	153	20	13	60.6 [46.1, 73.4]	88.4 [82.2, 92.7]

FN = false negative; FP = false positive; mBSFS-C = modified Bristol Stool Form Scale for children; TN = true negative; TP = true positive. Run-in period of 500 samples for reference standard (until weighted kappa > 0.9). Index test 1 = caregiver-reported loose or watery stool: "Did your child have a loose or watery stool?" Index test 2 = caregiver grade after visual inspection of the stool sample using the mBSFS-C (types 4 and 5 = loose or watery; types 1-3 = not loose or watery). Reference standard = first laboratory technician-graded stool consistency according to the mBSFS-C (types 4 and 5 = loose or watery; types 1-3 = not loose or watery). Sensitivity = TP/(TP + FN). Specificity = TN/(FP + TN). Standard errors account for clustering by village using a mixed-effects logistic regression with a random slope.

*Z = 10.25, $P < 0.001$; testing the difference in sensitivity with and without the mBSFS-C.
†Z = -8.83, $P < 0.001$; testing the difference in specificity with and without the mBSFS-C.

do not expect the validity of the caregiver's report to be affected as the reference standard grade would still classify breastfed stool as "loose." Fourth, this study was population-based and thus represented a normal spectrum of stool consistencies. A critically ill or hospitalized patient population may be more appropriate to validate definitions or diarrhea against clinical diagnoses. However, this study population is still relevant to the many public health intervention trials, epidemiologic studies, and surveys that use population-based samples and rely on caregiver report to define diarrheal disease. Finally, we did not assess intra-rater reliability of the mBSFS-C scale among laboratory technicians or caregivers. Stool samples were fixed in a liquid preservative immediately after collection, and thus the consistency could not be graded again.

Despite these limitations, our study had several strengths. We collected stool samples from more than 2,000 children from two distinct locations in Ethiopia. Each stool sample consistency was graded by a caregiver and two trained medical laboratory technicians according to an established stool consistency scale. We also evaluated the utility of a simple illustrative scale to improve the reporting of stool consistency. Our study was population based, representing a normal spectrum of stool consistencies and thus demonstrating the utility of the mBSFS-C for classifying stool consistency for population-based research.

Our findings are disconcerting for researchers and public health professionals who use caregiver-reported stool consistency to quantify diarrheal disease. We found that caregiver-reported "loose and watery" stool, a key component of the WHO definition of diarrhea, does not accurately reflect "loose or watery" stool as measured by the mBSFS-C. The degree of misclassification reported in this study would introduce substantial measurement error to studies quantifying diarrheal disease according to the reported stool consistency. Specifically, the adjusted prevalence can be estimated as the sum of the measured prevalence and specificity, minus the reciprocal of the sensitivity plus specificity minus one.³¹ It is not possible to directly calculate an adjusted prevalence of diarrhea for the present study because we have diagnostic accuracy estimates only of stool consistency and not frequency. However, as a thought exercise, if we assume that reported stool frequency is perfectly valid and that the error in reported frequency is independent of reported consistency, then the 13% 7-day prevalence of diarrhea observed in the

present study would be equivalent to an adjusted prevalence of 83%. This example illustrates the potential impact of the diagnostic accuracy of a screening test on the final prevalence estimate.

Given the global public health importance of diarrheal disease and the predominance of using caregiver-reported symptoms to identify cases, the low sensitivity identified in this study suggests that the burden of diarrheal disease may be underestimated and intervention effects could be biased. Researchers should take care when using caregiver-reported loose or watery stool to define diarrheal disease. If caregiver-reported stool consistency is the only option to measure pediatric diarrhea, a pictorial scale such as the mBSFS-C may improve the validity of caregiver report. These findings, if replicated in other settings, will have important implications for global estimates of diarrheal disease burden.

Received October 19, 2017. Accepted for publication January 2, 2018.

Published online February 26, 2018.

Note: Supplemental figure and tables appears at www.ajtmh.org.

Acknowledgments: We are thankful for the participation of communities, households, and individuals in this study. We are thankful for the collaboration with The Carter Center of Ethiopia and the Amhara Regional Health Bureau and are very appreciative of the time, effort, and commitment of all field teams, drivers, and laboratory technicians.

Financial support: This study was supported by the National Institute of Health (NEI U10 EY016214), (NICHD F31 HD088070-01A1 [to K. A.]), and (NIAID 1K01AI119180 [to B. F. A.]); That Man May See and The Sara & Evan Williams Foundation; and Research to Prevent Blindness.

Authors' addresses: Kristen Aiemjoy, Francis I. Proctor Foundation, University of California, San Francisco, San Francisco, CA, and Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, E-mail: kristen.aiemjoy@ucsf.edu. Solomon Aragie, Sintayehu Gebresillasie, Adane Dagne, Dagnachew Hailu, and Zerihun Tadesse, The Carter Center Ethiopia, Addis Ababa, Ethiopia, E-mails: solomon.aragie@cartercenter.org, sintayehugs@gmail.com, adane.dagne@cartercenter.org, dagnachew.hailu@cartercenter.org, and zerihun.tadesse@cartercenter.org. Dionna M. Fry and Jeremy D. Keenan, Francis I. Proctor Foundation, University of California, San Francisco, San Francisco, CA, E-mails: dionna.fry@ucsf.edu and jeremy.keenan@ucsf.edu. Melsew Chanyalew, Amhara Regional Health Bureau, Bahir Dar, Ethiopia, E-mail: yeshiwork97@yahoo.com. Aisha Stewart and Kelly Callahan, The Carter Center, Atlanta, GA, E-mails: aisha.stewart@cartercenter.org and kelly.callahan@cartercenter.org. Mathew Freeman, Rollins School of Public Health, Emory University, Atlanta, GA, E-mail: matthew.freeman@emory.edu. John Neuhaus, Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco,

CA, E-mail: John.Neuhaus@ucsf.edu. Benjamin F. Arnold, Division of Epidemiology, School of Public Health, University of California, Berkeley, Berkeley, CA, E-mail: benarnold@berkeley.edu.

REFERENCES

- World Health Organization, 2017. *Diarrhoeal Disease Fact Sheet*. Available at: <http://www.who.int/mediacentre/factsheets/fs330/en/>. Accessed July 13, 2017.
- World Health Organization, 2015. *Ethiopia: WHO Statistical Profile*. Available at: <http://www.who.int/gho/countries/eth.pdf?ua=1>. Accessed October 15, 2017.
- Institute for Health Metrics and Evaluation, 2017. *Global Burden of Disease*. Available at: <http://ghdx.healthdata.org/>. Accessed October 15, 2017.
- Johnston BC, Shamseer L, da Costa BR, Tsuyuki RT, Vohra S, 2010. Measurement issues in trials of pediatric acute diarrheal diseases: a systematic review. *Pediatrics* 126: e222–e231.
- World Health Organization. *Diarrhea (definition and sequelae)*. Geneva, Switzerland, WHO. Available at: <http://www.who.int/topics/diarrhoea/en/>. Accessed October 15, 2017.
- Clasen T, Schmidt WP, Rabie T, Roberts I, Cairncros S, 2007. Interventions to improve water quality for preventing diarrhoea: systematic review and meta-analysis. *BMJ* 334: 782.
- Cairncross S, Hunt C, Boisson S, Bostoen K, Curtis V, Fung IC, Schmidt WP, 2010. Water, sanitation and hygiene for the prevention of diarrhoea. *Int J Epidemiol* 39 (Suppl 1): i193–i205.
- Clasen TF, Bostoen K, Schmidt WP, Boisson S, Fung IC, Jenkins MW, Scott B, Sugden S, Cairncross S, 2010. Interventions to improve disposal of human excreta for preventing diarrhoea. *Cochrane Database Syst Rev* CD007180.
- Prüss-Ustün A et al., 2014. Burden of disease from inadequate water, sanitation and hygiene in low-and middle-income settings: a retrospective analysis of data from 145 countries. *Trop Med Int Health* 19: 894–905.
- Wolf J et al., 2014. Systematic review: assessing the impact of drinking water and sanitation on diarrhoeal disease in low-and middle-income settings: systematic review and meta-regression. *Trop Med Int Health* 19: 928–942.
- Clasen TF, Alexander KT, Sinclair D, Boisson S, Peletz R, Chang HH, Majorin F, Cairncross S, 2015. Interventions to improve water quality for preventing diarrhoea. *Cochrane Database Syst Rev* CD004794.
- Sinha I, Jones L, Smyth RL, Williamson PR, 2008. A systematic review of studies that aim to determine which outcomes to measure in clinical trials in children. *PLoS Med* 5: e96.
- Black RE, Brown KH, Becker S, Yunus M, 1982. Longitudinal studies of infectious diseases and physical growth of children in rural Bangladesh. I. Patterns of morbidity. *Am J Epidemiol* 115: 305–314.
- Koster FT, Palmer DL, Chakraborty J, Jackson T, Curlin GC, 1987. Cellular immune competence and diarrheal morbidity in malnourished Bangladeshi children: a prospective field study. *Am J Clin Nutr* 46: 115–120.
- Blum D, Feachem RG, 1983. Measuring the impact of water supply and sanitation investments on diarrhoeal diseases: problems of methodology. *Int J Epidemiol* 12: 357–365.
- O'donnell LJD, Virjee J, Heaton K, 1988. Pseudo-diarrhea in the irritable bowel syndrome—patients records of stool form reflect transit-time while stool frequency does not. *Gut* 29: A1455–A1455.
- Lewis SJ, Heaton KW, 1997. Stool form scale as a useful guide to intestinal transit time. *Scand J Gastroenterol* 32: 920–924.
- Chumpitazi BP, Lane MM, Czyzewski DI, Weidler EM, Swank PR, Shulman RJ, 2010. Creation and initial evaluation of a stool form scale for children. *J Pediatr* 157: 594–597.
- Aiemjoy K et al., 2016. 'If an eye is washed properly, it means it would see clearly': a mixed methods study of face washing knowledge, attitudes, and behaviors in rural Ethiopia. *PLoS Negl Trop Dis* 10: e0005099.
- Aiemjoy K et al., 2016. Is using a latrine “a strange thing to do”? A mixed-methods study of sanitation preference and behaviors in rural Ethiopia. *Am J Trop Med Hyg* 96: 65–73.
- Aiemjoy K, Gebresillasie S, Stoller NE, Shiferaw A, Tadesse Z, Chanyalew M, Aragie S, Callahan K, Keenan JD, 2018. Epidemiology of soil-transmitted helminth and intestinal protozoa infections in preschool-aged children in the Amhara region of Ethiopia. *Am J Trop Med Hyg* 96: 866–872.
- Lane MM, Czyzewski DI, Chumpitazi BP, Shulman RJ, 2011. Reliability and validity of a modified Bristol stool form scale for children. *J Pediatr* 159: 437–441.e1.
- Morton A, Dobson A, 1989. Assessing agreement. *Med J Aust* 150: 384–387.
- Cohen J, 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol Bull* 70: 213–220.
- Rogers W, 1994. Regression standard errors in clustered samples. *Stata Tech Bull* 3: 19–23.
- Genders TSS, Spronk S, Stijnen T, Steyerberg EW, Lesaffre E, Myriam Hunink MG, 2012. Methods for calculating sensitivity and specificity of clustered data: a tutorial. *Radiology* 265: 910–916.
- Coughlin SS, Trock B, Criqui MH, Pickle LW, Browner D, Tefft MC, 1992. The logistic modeling of sensitivity, specificity, and predictive value of a diagnostic test. *J Clin Epidemiol* 45: 1–7.
- Janssens ACJW, Deng Y, Borsboom GJJM, Eijkemans MJC, Habbema JDF, Steyerberg EW, 2005. A new logistic regression approach for the evaluation of diagnostic test results. *Med Decis Making* 25: 168–177.
- Baqui AH, Black RE, Yunus MD, Hoque ARA, Chowdhury HR, Sack RB, 1991. Methodological issues in diarrhoeal diseases epidemiology: definition of diarrhoeal episodes. *Int J Epidemiol* 20: 1057–1063.
- Blake MR, Raker JM, Whelan K, 2016. Validity and reliability of the Bristol stool form scale in healthy adults and patients with diarrhoea-predominant irritable bowel syndrome. *Aliment Pharmacol Ther* 44: 693–703.
- Rogan WJ, Gladen B, 1978. Estimating prevalence from the results of a screening test. *Am J Epidemiol* 107: 71–76.