Pediatr Gastroenterol Hepatol Nutr. 2021 Jan;24(1):38-44
https://doi.org/10.5223/pghn.2021.24.1.38
pISSN 2234-8646·eISSN 2234-8840

Pediatric Gastroenterology, Hepatology & Nutrition
pghn

Original Article

Real Time Versus Photographic Assessment of Stool Consistency Using the Brussels Infant and Toddler Stool Scale: Are They Telling Us the Same?

Berthold Albert Aman,[1] **Elvira Ingrid Levy**,[1] **Benjamine Hofman**,[2] **Yvan Vandenplas**,[1] **and Koen Huysentruyt** [1]

[1]KidZ Health Castle, UZ Brussel, Brussels, Vrije Universiteit Brussel, Brussels, Belgium
[2]Day Care Centre, Vrije Universiteit Brussel, Brussels, Belgium

Received: Apr 21, 2020
1st Revised: Jun 29, 2020
2nd Revised: Aug 18, 2020
Accepted: Sep 10, 2020

Correspondence to
Yvan Vandenplas
KidZ Health Castle, UZ Brussel, Vrije Universiteit Brussel, Laarbeeklaan 101, 1090 Brussels, Belgium.
E-mail: yvan.vandenplas@uzbrussel.be

ORCID iDs
Berthold Albert Aman
https://orcid.org/0000-0003-2447-8554
Elvira Ingrid Levy
https://orcid.org/0000-0002-4397-7095
Benjamine Hofman
https://orcid.org/0000-0003-3579-3656
Yvan Vandenplas
https://orcid.org/0000-0002-1862-8651
Koen Huysentruyt
https://orcid.org/0000-0002-6402-1762

Conflict of Interest
The authors have no financial conflicts of interest.

ABSTRACT

Purpose: Digital communication is becoming increasingly important in clinical practice and research. The finding that stool consistency can be evaluated similarly using either "in vivo" or photographic material by health care professionals will decrease subjective interpretation by parents. The primary outcome of this study was the reliability of stool consistency scoring using the Brussels Infant and Toddler Stool Scale (BITSS) between fresh stools and their photos; the secondary outcome was the inter-rater reliability based on the fresh stools.
Methods: Fresh stool samples from healthy children were collected in a day care center. These stools, and one month later the corresponding photos presented in a random order, were presented to 14 observers. Reliabilities were analyzed using absolute agreements and weighted and unweighted Cohen's κ.
Results: In total, 202 samples were rated 576 times. Absolute agreement between photographic and real time assessment ranged between 71.1% and 83.3% among observers. This corresponded with substantial agreement (unweighted κ=0.70 [95% CI, 0.61–0.78]; weighted κ=0.86 [95% CI, 0.78–0.88]). The inter-observer agreement showed similar percentages of absolute agreement (81.4–82.0%) and κ-values corresponding with fair-to-moderate agreement.
Conclusion: Our findings suggest that the assessment of fresh stool consistency can also reliably be done on photographic material when using the BITSS. This opens opportunities in scientific surroundings and in our daily life communication with parents and caretakers.

Keywords: Brussels Infant and Toddler Stool Scale; Photographic assessment; Fecal consistency; Feces; Observer variation

INTRODUCTION

The Brussels Infant and Toddler Stool Scale (BITSS) was developed because of the need for a stool consistency scoring system for non-toilet trained children [1]. In a large validation study involving 18 countries, the original scale was regrouped to a four-point scale consisting of a total of seven colored photographs [2]. **Fig. 1** shows the photographs used in the BITSS, arranged according to the four-point scale; this figure is also accessible at https://bitss-stoolscale.com. The reliable and reproducible assessment of stool consistency is

Pediatric Gastroenterology, Hepatology & Nutrition pghn

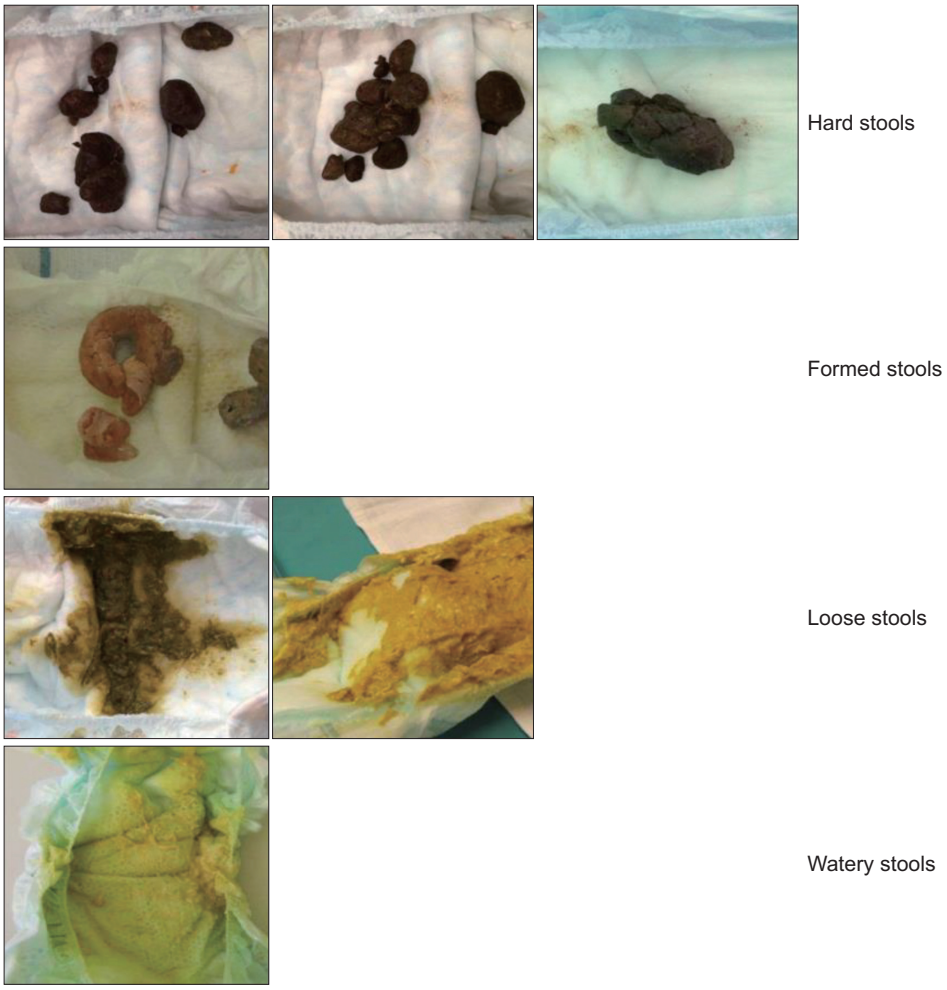

Fig. 1. Brussels Infant and Toddler Stool Scale.

paramount to differentiate pathological from normal stools in order to gain insight into the defecation patterns of children and to monitor response to treatment both for clinical and research purposes. Three other scoring systems for children were developed previously: the Amsterdam Infant Stool Scale (AISS) [3], the Stool Scale for Diapered Infants [4], and the Parental Report of Stool Consistency [5]. The BITTS is, however, the only tool so far that has been studied in a large population and is currently being used to develop an automated machine learning tool for stool consistency scoring [6].

In modern medicine, communication between parents and doctors occurs digitally along with photographical material. Apart from conventional emails, the use of mobile applications such as WhatsApp, Messenger, and Siilo plays an important role in communication [7,8]. The COVID-19 pandemic has accelerated this evolution through the use of teleconsultations. Furthermore, there is a research interest if the number of study visits can be reduced without compromising the amount of reliably collected data on stool consistency. This could reduce the burden on parents and children who participate in clinical trials. This is why we believe a stool consistency assessment scoring system with similar performance as a photographic material will be of interest. The only study similar to this study involving non-toilet trained children that we are aware of is a recent Polish study that used the AISS to compare the

https://doi.org/10.5223/pghn.2021.24.1.38

performance of photographic materials with an "in vivo" stool assessment system [9]. However, the aforementioned study did not provide a clear answer to our research question since the in vivo ratings and the photographic ratings were performed by different people.

Therefore, the primary goal of this study was to investigate the variability between real time and photographic assessment of stool consistency using the BITSS. The secondary aim was to explore the inter-rater agreement based on the real time assessment. Given the simplicity of our stool scale and the results of our previous study [2], we hypothesize that there will be a significant agreement (classically defined as a κ of 0.6–0.8 [10]) between the in vivo samples and the photographs.

MATERIALS AND METHODS

We performed a monocentric comparative study in Brussels, Belgium. Diapers containing fresh stool samples were rated and photographed by staff members at the baby units of a single children's day care center. The stool samples were first presented as fresh samples to the observers. Each stool sample was rated in vivo and immediately thereafter photographed using a NIKON COOLPIX A100, equipped with a CCD-sensor of 20.1 megapixels. When taking photos, the camera was used in the auto mode and neither pre-sets nor filters were used. Pictures were taken in daylight. No flashlight was used. One month later, the photos were presented on a computer screen to the same observers in a random order. When multiple staff members rated the same fresh stool or digital photograph, they were blinded from each other's rating. The staff received a brief explanation about the BITSS scale, but no formal training session was held.

A total of 49 babies were enrolled in the study. Their ages ranged from 1 to 16 months. To be included in the study babies needed to be healthy. Infants with an acute illness (vomiting, fever, and acute diarrhea) were excluded. Those with a chronic intestinal illness were also excluded. In total, 202 fresh stool samples were rated using the BITSS scale. One observer rated all stools, the others (n=13) rated between 6 and 48 stool samples.

The design and protocol of this study have been approved by the Ethics Committee (2018/365). The trial was registered (NCT02913950). Parents were informed orally and in writing. They had the option to withdraw from the study at any time.

Statistics

A power calculation was performed using the "kappaSize" package for R (v 3.6.2; R Development Core Team, 2013; http://www.r-project.org). The κ0 (the null hypothesis level, i.e. the level that is considered unacceptably low) was set at 0.4, κ1 (the value we considered as minimally acceptable) was set at 0.6, α was set at 0.05, and β was set at 0.8. For the input of the probability distribution input, a grid of all possible combinations of proportions of stools in each of the four classes was created (using an increment of 5% at each iteration). Of these possible combinations, 802 were retained. These were the combinations in which the percentages of all four consistency classes summed up to 100%. The power calculation using the previously mentioned arguments was run 802 times. The worst case scenario was used to determine the required sample size needed for the intra-rater reliability between fresh stools and their photographs. Two outliers were found, each requiring 250 diapers. The underlying distribution for these two combinations were, however, disregarded as they seemed highly

https://doi.org/10.5223/pghn.2021.24.1.38

unlikely to occur in our healthy study population (one scenario required 85% of the samples to be loose and all other categories 5%; the other required 85% of the samples to be watery and all other categories 5%). This approach yielded a "worst case scenario" of 197 diapers needed for our primary outcome.

The primary analyses for the intra-rater variability were performed on the observer meeting the required amount of ratings to achieve adequate power for these analysis. To examine the robustness of these findings, we replicated these analyses twice: first, using the data analyzed by the two other observers who performed over 40 ratings, and second, on all ratings taken together, where the data were treated as if the ratings were performed by the same observer. Descriptive statistics, including the number of absolute agreements were performed. Furthermore, a Cohen's κ was calculated (both weighted using squared weights as unweighted) to evaluate the agreement in ratings between fresh samples and photographs. In accordance with the most commonly used guidelines for the interpretation of the Cohen's kappa, a κ of 0–0.2 was considered to indicate a slight agreement, a κ of 0.2–0.4 was considered to indicate a fair agreement, a κ of 0.4–0.6 was considered to indicate a moderate agreement, a κ of 0.6–0.8 was considered to indicate a substantial agreement, and a κ of 0.8–1 was considered to indicate an almost perfect agreement [10]. Confidence intervals (CIs) were calculated using bootstrapping. The difference between weighted and unweighted kappa values is that in a weighted kappa, the higher the degree of misclassification, the more this will negatively affect the kappa value (i.e. misclassifying a loose stool as hard is worse than misclassifying it as formed); in an unweighted kappa, every misclassification is treated similarly. For the secondary goal of our study, which was to determine the inter-rater reliability of the BITSS for fresh stool samples, we applied a similar strategy. However, we limited ourselves to only observers who have performed over 40 ratings. The observer who rated every stool served as the reference observer. IBM SPSS Statistics for Windows, Version 26.0 (IBM Co., Armonk, NY, USA) and R v3.6.2 were used for statistical analyses.

RESULTS

The stools of 49 infants were analyzed. In total, 50 babies were part of the baby units in the day care center. One child was excluded due to acute illness (fever). The median age of these infants was eight months, ranging from four to16 months. We collected, in total, 576 in vivo and 576 photo ratings for the 202 samples: 52 (25.7%) of the stools were rated by two observers, 135 (66.8%) were rated by three observers, and 15 (7.6%) were rated by four or more observers. Overall, the ratings consisted of 22 (3.8%) hard stools, 304 (52.8%) formed stools, 196 (34.0%) loose stools, and 54 (9.4%) watery stools. For the subanalysis of the results of the observers who rated over 40 stools, 202 stool samples were used. These stool samples were scored 295 times by a total of three observers; 112 (55.4%) of the stools were rated once, 87 (43.1%) of the stools were rated two times, and 3 (1.5%) of the stools were rated three times.

As presented in **Table 1**, a high overall absolute agreement between photographic and real time assessment was found (77.4%). It ranged from 71.1% to 83.3% among all observers. Based on the unweighted kappa, a substantial agreement between fresh stool samples and their photos was found for the observer who assessed all the stool samples. This result was in agreement with our primary hypothesis. The results that were obtained based on the weighted kappa were even better, demonstrating substantial agreement. Similar values

https://doi.org/10.5223/pghn.2021.24.1.38

Table 1. Agreement between fresh stool samples and their corresponding photos

Observer	N	Absolute agreement (%)	Unweighted kappa (95% CI)	Weighted kappa (95% CI)
Observer 1	202	82.2	0.70 (0.61–0.78)	0.86 (0.78–0.88)
Observer 2	48	83.3	0.73 (0.53–0.90)	0.79 (0.73–0.95)
Observer 3	45	71.1	0.47 (0.23–0.70)	0.66 (0.49–0.82)
All observers	567	77.4	0.63 (0.58–0.69)	0.79 (0.77–0.83)

CI: confidence interval.

Table 2. The inter-observer agreement based on fresh stool sample assessments

Observer	N	Absolute agreement (%)	Unweighted kappa (95% CI)	Weighted kappa (95% CI)
Observer 1 vs. 2	50	82.0	0.69 (0.50–0.85)	0.80 (0.64–0.92)
Observer 1 vs. 3	43	81.4	0.64 (0.41–0.85)	0.80 (0.58–0.94)

CI: confidence interval.

were found when all the ratings were analyzed together or when the other two observers performed more than 40 ratings (albeit with wider CIs due to a lower number of ratings). When a disagreement occurred, photographic assessment was more commonly classified as softer (70%) than harder (30%) compared to real time assessment.

The results of the inter-observer agreement based on fresh stool sample assessments are presented in **Table 2**. Although there is a substantially low number of ratings, similar percentages of absolute agreement were noted similar to the agreement between fresh stool samples and photos. The same is true for the weighted and unweighted kappa values, although wide CIs were noted due to a low number of ratings.

DISCUSSION

This study emphasizes the possibility of using photographic material in daily medical practice. We showed that there is a substantial agreement between the assessment of stool consistency based on fresh stool samples and their photos when using the BITSS. This study also presents promising results regarding the inter-observer reliability based on fresh stool samples, which was not studied previously.

Contrary to the study by Wojtyniak et al. [9], our study involved multiple observers, while their study involved only the parents and two medical doctors. Their study was flawed in the sense that their results were affected by inter-rater variability and variability due to the use of "in vivo" assessment vs. photographic material [9].

Our study showed that the BITSS performed well with multiple observers and all the observations were consistent. Furthermore, our results show that it has a good performance on photos as well. The strengths of our study include the fact that it was sufficiently powered to enable us to investigate the variability between photographic and fresh stool assessments. The interval of one month and the randomization of the order of the presented photos eliminated the possibility of bias, which could artificially inflate our results. A possible limitation of our study is the fact that we performed our study on samples from healthy children. Therefore, the pathological samples (hard and watery stools) are underrepresented. The disagreements tended to be more frequent in the extremes of the scale (**Supplementary Table 1**), although the numbers of hard and watery stools are too low for further inferences. The distribution in our medical practice is probably different because the pathological samples are more frequently an issue. Using a group comprised of children less than 16

https://doi.org/10.5223/pghn.2021.24.1.38

months old, we overcame this problem, since, we believe, they still had the widest variance in stool consistency [11]. Additionally, only non-medically trained healthcare workers were involved in this study. The international validation study reported a difference in the reliability of the BITSS among nurses, medical doctors, and parents [2]. We suspect that the level of staff members of a day care center will be somewhere between that of nurses and parents, as they are not medically trained, but do have a lot of experience dealing with infants and diapers. In a way, these results probably come close to a "worst case scenario" for the reliability of the BITSS, which makes its use even more promising. Furthermore, a digital camera was used to take the photos, whereas most parents will probably use a smartphone to do so. To overcome this problem, we used the camera only in the auto mode, without any pre-sets, filters, or flashlight.

Further study is needed to investigate whether these results can be reproduced in cohorts involving children with gastro-intestinal disorders and parents. Larger numbers of children are needed to further investigate the inter-rater reliability when using the BITSS to rate the consistency of fresh stools.

Based on our findings, the assessment of photographic images could replace at least some of the live assessment of stool consistency in clinical practice or research. This finding is even more interesting with the recent developments due to the COVID-19 pandemic. During this period and in the period hereafter, teleconsultations will become increasingly more important. The BITSS is a good tool that can be used during these teleconsultations.

In conclusion, we found a substantial agreement between stool consistency assessments based on fresh stool samples and their photos when using the BITSS for healthy infants. We also found promising results regarding inter-observer reliability, although this needs to be replicated in a larger group. Overall, these findings suggest that the BITSS is a good clinical tool that can be used to evaluate photographic material, both in scientific surroundings as in our daily life communication with parents and caretakers. Furthermore, it suggests that the BITSS can be used in daily practice by both professional health care workers and people without a medical background. This study adds value especially during the COVID-19 pandemic and the era post COVID-19, as these periods are expected to encourage teleconsultation.

ACKNOWLEDGEMENTS

We would like to express our gratitude to the staff of the daycare center of the Vrije Universiteit Brussel who participated in this study.

SUPPLEMENTARY MATERIAL

Supplementary Table 1

Agreement between fresh stool samples and their corresponding photos

Click here to view

REFERENCES

1. Vandenplas Y, Szajewska H, Benninga M, Di Lorenzo C, Dupont C, Faure C, et al. Development of the Brussels Infant and Toddler Stool Scale ('BITSS'): protocol of the study. BMJ Open 2017;7:e014620.
PUBMED | CROSSREF
2. Huysentruyt K, Koppen I, Benninga M, Cattaert T, Cheng J, De Geyter C, et al. The Brussels Infant and Toddler Stool Scale: a study on interobserver reliability. J Pediatr Gastroenterol Nutr 2019;68:207-13.
PUBMED | CROSSREF
3. Bekkali N, Hamers SL, Reitsma JB, Van Toledo L, Benninga MA. Infant stool form scale: development and results. J Pediatr 2009;154:521-6.e1.
PUBMED | CROSSREF
4. Gustin J, Gibb R, Kenneally D, Kutay B, Waimin Siu S, Roe D. Characterizing exclusively breastfed infant stool via a novel infant stool scale. JPEN J Parenter Enteral Nutr 2018;42 Suppl 1:S5-11.
PUBMED | CROSSREF
5. Koppen IJN, Velasco-Benitez CA, Benninga MA, Di Lorenzo C, Saps M. Using the Bristol Stool Scale and parental report of stool consistency as part of the Rome III criteria for functional constipation in infants and toddlers. J Pediatr 2016;177:44-8.e1.
PUBMED | CROSSREF
6. Ludwig T, Wong J, Oukid I, Huysentruyt K, Roy P, Foussat AC. Automated stool consistency scoring for non-toilet trained children by machine learning algorithms [abstract]. J Paediatr Child Health 2019;2019:26-7.
7. Krynski L, Goldfarb G, Maglio I. Technology-mediated communication with patients: WhatsApp Messenger, e-mail, patient portals. A challenge for pediatricians in the digital era. Arch Argent Pediatr 2018;116:e554-9.
PUBMED
8. Masterman M, Cronin RM, Davis SE, Shenson JA, Jackson GP. Adoption of secure messaging in a patient portal across pediatric specialties. AMIA Annu Symp Proc 2017;2016:1930-9.
PUBMED
9. Wojtyniak K, Horvath A, Dziechciarz P. In vivo assessment by parents and a physician using the Amsterdam Infant Stool Scale provided better inter-rater agreement than photographic evaluation. Acta Paediatr 2018;107:529-31.
PUBMED | CROSSREF
10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
PUBMED | CROSSREF
11. Tham EB, Nathan R, Davidson GP, Moore DJ. Bowel habits of healthy Australian children aged 0-2 years. J Paediatr Child Health 1996;32:504-7.
PUBMED | CROSSREF