ORIGINAL ARTICLE: GASTROENTEROLOGY

CME

The Brussels Infant and Toddler Stool Scale: A Study on Interobserver Reliability

[*]Koen Huysentruyt, [†]Ilan Koppen, [†]Marc Benninga, [‡]Tom Cattaert, [‡]Jiqiu Cheng, [*]Charlotte De Geyter, [§]Christophe Faure, [||]Frédéric Gottrand, [¶]Badriul Hegar, [#]Iva Hojsak, [**]Mohamad Miqdady, [††]Seksit Osatakul, [‡‡]Carmen Ribes-Koninckx, [§§]Silvia Salvatore, [||||]Miguel Saps, [¶¶]Raanan Shamir, [##]Annamaria Staiano, [***]Hania Szajewska, [†††]Mario Vieira, and [*]Yvan Vandenplas, and the BITSS working group

ABSTRACT

Objectives: The Bristol Stool Form Scale (BSFS) is inadequate for non-toilet trained children. The Brussels Infant and Toddler Stool Scale (BITSS) was developed, consisting of 7 photographs of diapers containing stools of infants and toddlers. We aimed to evaluate interobserver reliability of stool consistency assessment among parents, nurses, and medical doctors (MDs) using the BITSS.

Methods: In this multicenter cross-sectional study (2016–2017), BITSS photographs were rated according to the BSFS. The reliability of the BITSS was evaluated using the overall proportion of perfect agreement and the linearly weighted κ statistic.

Results: A total of 2462 observers participated: 1181 parents (48.0%), 624 nurses (25.3%), and 657 MDs (26.7%). The best-performing BITSS photographs corresponded with BSFS type 7 (87.5%) and type 4 (87.6%), followed by the BITSS photographs representing BSFS type 6 (75.0%), BSFS type 5 (68.0%), BSFS type 1 (64.8%), and BSFS type 3 (64.6%). The weakest performing BITSS photograph corresponded with BSFS type 2 (49.7%). The overall weighted κ-value was 0.72 (95% CI 0.59–0.85; good agreement). Based on these results, photographs were categorized per stool group as hard (BSFS type 1–3), formed (BSFS type 4), loose (BSFS types 5 and 6), or watery (BSFS type 7) stools. According to this new categorization system, correct allocation for each photograph ranged from 83 to 96% (average: 90%). The overall proportion of correct allocations was 72.8%.

Conclusions: BITSS showed good agreement with BSFS. Using the newly categorized BITSS photographs, the BITSS is reliable for the assessment of stools of non-toilet trained children in clinical practice and research. A multilanguage translated version of the BITSS can be downloaded at *https://bitss-stoolscale.com/*.

Key Words: Brussels Infant and Toddler Stool Scale, infants, stool scale, toddlers

(*JPGN* 2019;68: 207–213)

What Is Known

- Reliable assessment of stool consistency is important for evaluating children's defecation pattern and diagnosing gastrointestinal disorders.
- The reliability of the Bristol Stool Form Scale, developed for adults, has been debated for young children who are non-toilet trained and wear diapers.

What Is New

- The Brussels Infants and Toddlers Stool Scale was validated as a reliable instrument to assess stools of non-toilet trained children via assessment of interobserver reliability among 2462 study participants, including parents nurses and medical doctors.

In pediatric gastroenterology, reliable assessment of stool consistency is of key importance in the evaluation of a child's defecation pattern and for diagnosing gastrointestinal disorders such as functional constipation, diarrhea and irritable bowel syndrome (1,2). Moreover, stool consistency is commonly considered as an outcome measure in clinical trials for functional constipation and irritable bowel syndrome (3–5). In clinical practice, eliciting an adequate description of a child's stool pattern can, however, be difficult and unreliable. Therefore, visual stool scales are

Received April 17, 2018; accepted June 25, 2018.

From the *KidZ Health Castle, UZ Brussel, Vrije Universiteit Brussel, Brussels, Belgium, the †Department of Pediatric Gastroenterology and Nutrition, Emma Children's Hospital/Academic Medical Center, Amsterdam, the Netherlands, the ‡DICE, Brussels, Belgium, the §Department of Pediatric Gastroenterology, Sainte-Justine Hospital, Montreal, Quebec, Canada, the ||Department of Pediatric Hepatology, Gastroenterology and Nutrition, CHU Lille, University Lille, France, the ¶Department of Child Health University of Indonesia, Jakarta, Indonesia, the #Children's Hospital Zagreb, University of Zagreb School of Medicine, University J.J. Strossmayer School of Medicine Osijek, Osijek, Croatia, the **Department of Pediatric Gastroenterology, Hepatology and Nutrition, Sheikh Khalifa Medical City, Abu Dhabi, United Arab Emirates, the ††Department of Pediatrics, Faculty of Medicine, Prince of Songkla University, Hat Yai, Songkhla, Thailand, the ‡‡Department of Pediatric Gastroenterology, Hepatology and Nutrition, La Fe University Hospital, Valencia, Spain, the §§Department of Pediatrics, University of Insubria, Ospedale ''F. Del Ponte,'' Varese, Italy, the ||||Division of Pediatric Gastroenterology, Hepatology and Nutrition, University of Miami, Miller School of Medicine, Miami, FL, the ¶¶Schneider Children's Medical Centre of Israel, Sackler Faculty of Medicine, Tel-Aviv University, Tel-Aviv, Israel, the ##Department of Translational Medical Science, Section of Pediatrics, University of Naples Federico II, Naples, Italy, the ***Department of Pediatrics, The Medical University of Warsaw, Warsaw, Poland, and the †††Department of Pediatrics, Pontifical University of Paraná and Center for Pediatric Gastroenterology—Hospital Pequeno Príncipe, Curitiba, Brazil.

K.H. and I.K. are co–first authors.

Address correspondence and reprint requests to Yvan Vandenplas, PhD, MD, Prof Dr, Department of pediatrics, UZ Brussel, Laarbeeklaan 101, 1090 Brussels, Belgium (e-mail: yvan.vandenplas@uzbrussel.be); Koen Huysentruyt, PhD, MD, Universitair Ziekenhuis Brussel, Vrije Universiteit Brussel (VUB), Brussels, Belgium (e-mail: koen.huysentruyt@uzbrussel.be).

This article has been developed as a Journal CME Activity by NASPGHAN. Visit *http://www.naspghan.org/content/59/en/Continuing-Medical-Education-CME* to view instructions, documentation, and the complete necessary steps to receive CME credit for reading this article.

Copyright © ESPGHAN and NASPGHAN. All rights reserved.

commonly used as an aid in the assessment of stool consistency. The most commonly used visual stool scale is the Bristol Stool Form Scale (BSFS), which was developed to be used in adults (6). This stool scale is frequently used in children with defecation disorders as well (7,8). The BSFS consists of 7 descriptions of different stool forms, ranging from hard stools to watery stools, which are typically accompanied by drawings (6). Although the BSFS is frequently applied in children, its reliability in young children who are not toilet trained and wear diapers has been debated (8–10). A recent study among parents of infants and toddlers showed only fair agreement ($\kappa = 0.335$) between the BSFS and verbal parental report of stool consistency (8).

In 2009, the Amsterdam Infant Stool Scale (AISS) was developed and validated, providing an assessment tool for stools of infants under 1 year of age (9). The AISS enables evaluation of stool consistency, volume and color. Although the AISS has been reported to be more suitable for use in infants than the BSFS (10), it is not universally used in clinical practice or pediatric research, probably due to its complexity.

Recently, our working group developed the Brussels Infant and Toddler Stool Scale (BITSS), a visual stool form scale adapted to infants and toddlers wearing diapers (11). A detailed description of the development of the BITSS was published previously (11). The BITSS consists of 7 color photographs of diapers containing stools of infants and toddlers who are not toilet trained (Fig. 1). These photographs were selected through multiple voting sessions based on their resemblance to the original BSFS according to a group of nurses and medical doctors (MDs). The objective of the current study was to evaluate the interobserver reliability of stool consistency assessment among parents, nurses and MDs using the BITSS.

METHODS

Participant Recruitment

Eighteen centers were invited to participate in this study based on their experience with research and their geographical location, representing a variety of countries in Europe, Asia, and the Americas. Each participating center was instructed to include a minimum of 50 parents, 25 nurses and 25 MDs. Methods for participant selection were at the discretion of the principal investigator at each center. No specific selection criteria were applied as to the type of pediatric patients of whom parents were invited to participate. Nurses and MDs were included at varying pediatric departments.

Interobserver Reliability

Each observer was shown all 7 BITSS photographs and was asked to match these with the BSFS, which was accompanied by its descriptors in the local language. In most countries, the BSFS had already been translated into the local language before this study; if this was not the case, the local investigator provided a translation of the descriptors. In all but 1 of the countries, observers were instructed to match each photograph with only 1 item on the BSFS. A different approach was used in the Netherlands, where observers were free to match multiple photographs with the same item on the BSFS. Observers performed the ratings independently in writing and were not allowed to communicate during the assessment. The BITSS photographs are depicted in Figure 1.

Statistical Analysis

For the individual photographs of the BITSS, the proportion of exact agreement and the mode of the BSFS type chosen for each photograph were determined. Comparisons were made using χ^2 analysis between parents, nurses, and MDs and between responders from different countries and continents. Overall, the performance of the BITSS scale was tested via the proportions of exact agreements and the proportions of agreements in which the ratings deviated with maximum 1 for the reference BSFS stool type. The BITSS photographs were categorized a priori per stool group as hard stools (BSFS types 1 and 2), normal formed stools (BSFS types 3 and 4), normal loose stools (BSFS type 5), or watery stools (BSFS types 6 and 7) (11). The performance of this a priori grouped BITSS scale was tested via the proportions of exact agreements. Lastly, a new

Supplemental digital content is available for this article. Direct URL citations appear in the printed text, and links to the digital files are provided in the HTML text of this article on the journal's Web site (*www.jpgn.org*).

BITSS working group: A. Chogle: Division of Pediatric Gastroenterology, Hepatology and Nutrition, Children's Hospital of Orange County, Orange County, CA; C.A. Velasco-Benítez: Department of Pediatrics, Universidad de Granada, España; C.R. Ramirez: Hospital Maria Inmaculada, Florencia, Colombia; E. Játiva Mariño: Universidad de La Frontera, Chile; E. Privat: Department of Pediatrics, Centre Hospitalier Régional Universitaire de Lille, Lille, France; J.M. Łukasik: Department of Paediatrics, The Medical University of Warsaw, Warsaw, Poland; M. Mejía-Castro: Hospital Infantil de Nicaragua Manuel de Jesus Rivera, Managua, Nicaragua; R. Chanís: Servicio de Gastroenterología Pediátrica y Endoscopia Digestiva, Hospital del Niño Dr José Renán Esquivel, Ciudad de Panamá, Panamá; R. Zablah, Hospital Nacional de Niños Benjamin Bloom, San Salvador, El Salvador.

www.clinicaltrials.gov registration number: NCT02913950

Koen Huysentruyt and Ilan Koppen contributed equally to the writing of the article.

The authors have no financial relationships relevant to this article to disclose.

M.A.B. is a scientific consultant for Shire, Sucampo, Astrazeneca, Norgine, Zeria, Coloplast, Danone, Friesland Campina, Sensus, United Pharmaceuticals. C.F. has participated as clinical investigator, and consultant for Sucampo. I.H. has participated as a speaker for Medias Adria and GM Pharma. M.S.M. has participated as a clinical investigator, and/or advisory board member, and/or consultant, and/or speaker for Abbott Nutrition, Danone, Nestle, Nutricia, Merck, and Wyeth. C.R.-K. has participated as a clinical investigator, and/or advisory board member, and/or consultant, and/or speaker for Shire, Nestle Health Science, Nutricia, and Mead Johnson Nutrition. S.S. has participated as a consultant and/or speaker for Deca, IMS Health, Danone, Menarini, Nestlé Health Science. M.S. is a scientific consultant for Sucampo, Forest, Quintiles, Ardelyx, QOL Medical, IMHealth Science, Nutricia. R.S. has participated as a consultant and/or speaker for Abbott, Danone, Enzymotec, Nestle Nutrition Institute, NG Solutions, and Nutrinia. A.S. is clinical investigator for Aboca, was clinical investigator for Nestlé, is advisory board member for Sucampo, and consultant for Aboca and D.M.G. Italy, and speaker for Angelini, Danone, Miltè, Valeas. H.S. has participated as a clinical investigator, and/or advisory board member, and/or consultant, and/or speaker for companies manufacturing infant formulas, that is, Arla, Danone, HiPP, Nestle, Nestle Nutrition Institute, Nutricia, and Mead Johnson. M.C.V. has participated as a consultant and/or speaker for Danone, Nestlé Nutrition Institute, and Aché Laboratories. Y.V. has participated as a clinical investigator, and/or advisory board member, and/or consultant, and/or speaker for Abbott Nutrition, Aspen, Biocodex, Danone, Nestle Health Science, Nestle Nutrition Institute, Nutricia, Mead Johnson Nutrition, Merck, Phacobel, Rontis, United Pharmaceuticals, Wyeth, and Yakult. The remaining authors report no conflicts of interest.

Copyright © 2018 by European Society for Pediatric Gastroenterology, Hepatology, and Nutrition and North American Society for Pediatric Gastroenterology, Hepatology, and Nutrition

DOI: 10.1097/MPG.0000000000002153

Copyright © ESPGHAN and NASPGHAN. All rights reserved.

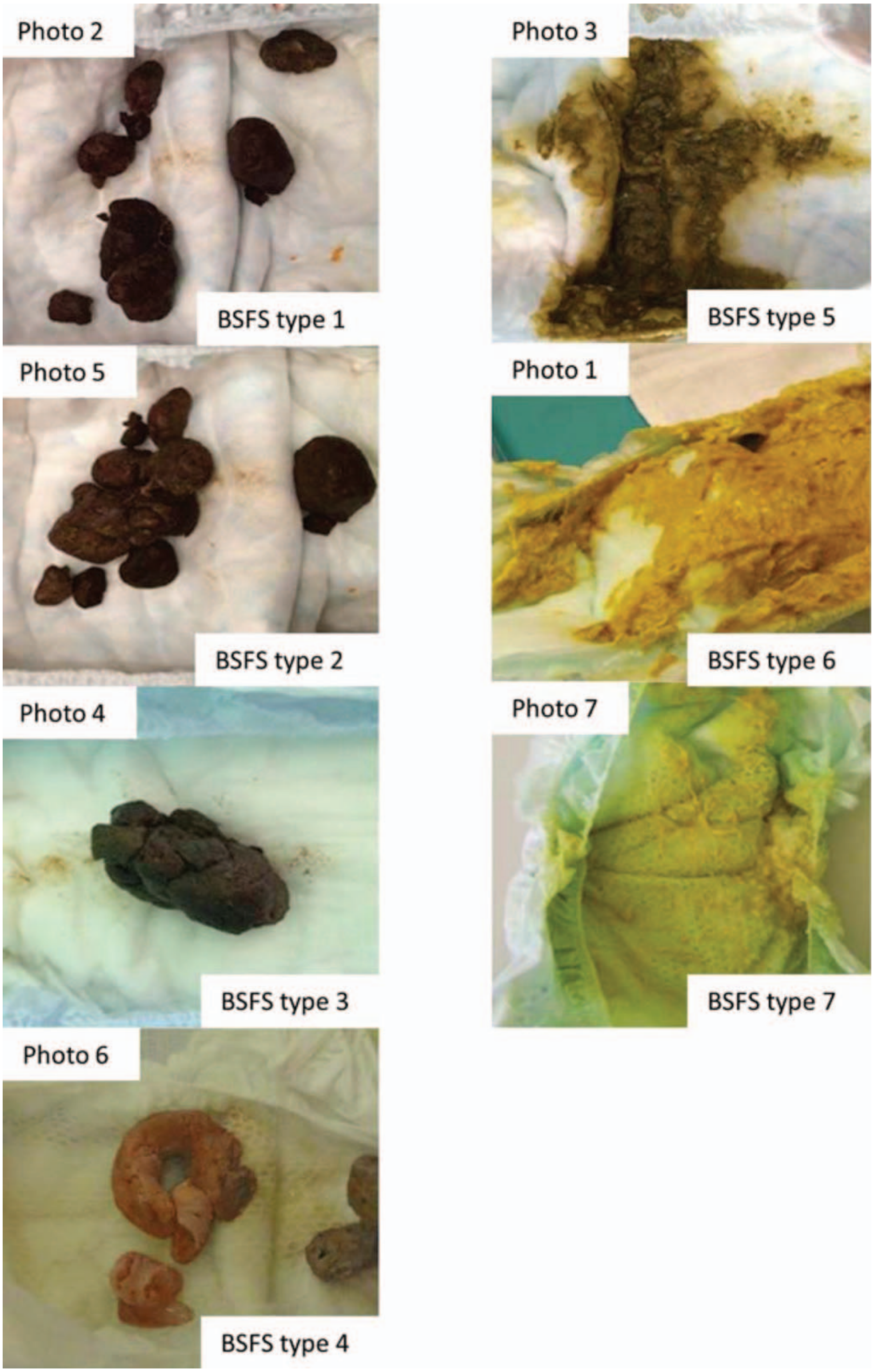

FIGURE 1. Brussels Infants and Toddlers Stool Scale photographs with their matching Bristol Stool Form Scale classification.

Copyright © ESPGHAN and NASPGHAN. All rights reserved.

grouped BITSS scale was proposed based on our results to maximize its performance.

Going beyond χ^2 analysis comparing different groups, multiple logistic regression models were also built, considering continent, country, observer group, and continent by observer group interaction as potential predictors.

The reliability of the BITSS was further evaluated using the overall proportion of perfect agreement and calculation of the linearly weighted κ statistic, which is used to measure the degree of agreement in classification accounting for that which would be expected by chance. Based on the value of κ, agreement was categorized as poor ($\kappa \le 0.2$), fair ($0.21 \le \kappa \le 0.40$), moderate ($0.41 \le \kappa \le 0.60$), good ($0.61 \le \kappa \le 0.80$) or excellent ($0.81 \le \kappa \le 1.00$) (12). A 'moderate' κ value (>0.41 to 0.60) was considered to be indicative of an acceptable level of agreement.

The results from the Netherlands were analyzed separately because of the different study design, in which observers were free to match multiple photographs with the same item on the BSFS instead of ranking the 7 BITSS photographs according to the 7 BSFS descriptions. This provided additional information on the standalone performance of each individual photograph, as there was no longer a "ranking effect."

Analyses were performed using SPSS v24.0 (IBM, Armonk, NY) and R version 3.4.3 (R Core Team, http://www.R-project.org); a P value <0.05 was considered significant. Data were considered categorical throughout the analyses. We hypothesized that there would be an acceptable level of reliability of the BITSS, both among health care workers and parents, regardless of the different geographical regions (11).

RESULTS

Demographic Data

A total of 2462 participants (including those from the Netherlands) performed the ratings: 1181 parents (48.0%), 624 nurses (25.3%), and 657 MDs (26.7%). Demographic data of the study participants are presented in Supplementary file 1 (Supplemental Digital Content 1, *http://links.lww.com/MPG/B487*). Responders originated from the following continents: Asia (n = 210, 8.5%), Europe (n = 1370, 55.6%), and the Americas (n = 882, 35.8%). The distribution of the observer groups was significantly different between different countries ($P < 0.001$) and continents ($P = 0.008$).

Performance of the Individual Photographs

The proportions of exact agreement for each individual photograph are presented in Figure 2. The best-performing BITSS photographs corresponded with BSFS type 7 (87.5%) and BSFS type 4 (87.6%), followed by the BITSS photographs representing BSFS type 6 (75.0%), BSFS type 5 (68.0%), BSFS type 1 (64.8%), and BSFS type 3 (64.6%). The weakest performing BITSS photograph corresponded with BSFS type 2 (49.7%). The multiple logistic regression model including terms for continent and observer group showed an effect of both factors for each photo (Supplementary file 2, Supplemental Digital Content 2, *http://links.lww.com/MPG/B488*).

Overall, the mode for each BITSS photograph corresponded with the respective reference BSFS stool types. The modes were consistent over the different observer groups and across the different continents, except for BITSS photograph 5, for which the mode was BSFS type 1 instead of type 2 in Asia (Supplementary file 3, Supplemental Digital Content 3, *http://links.lww.com/MPG/B489*). When looking at countries separately, the mode deviated from the reference BSFS type for BITSS photograph 2 (BSFS type 2 instead of BSFS type 1) in observers from United Arab Emirates, Mexico, and USA and for BITSS photograph 5 (BSFS type 1 instead of BSFS type 2) in observers from United Arab Emirates, Mexico, USA, Spain, Uruguay, Canada, and the Netherlands. The mode for BITSS photograph 4 was BSFS type 2 instead of type 3 in Spain, Canada, and the Netherlands, and the mode for BITSS photograph 3 was BSFS type 6 instead of type 5 in the Netherlands and Canada (Supplementary file 3, Supplemental Digital Content 3, *http://links.lww.com/MPG/B489*).

The proportions of correct allocations per observer group are presented in Table 1. Health care professionals (nurses and MDs) had significantly more correct allocations for each BITSS photograph to the correct reference BSFS type than parents, except for BITSS photograph 2 (BSFS type 1; $P = 0.086$). MDs also consistently scored significantly better than nurses, except in the case of BITSS photograph 5 (BSFS type 2, $P = 0.818$) and BITTS photograph 2 (BSFS type 1, $P = 0.852$).

The results from the Netherlands (Supplementary file 4A, Supplemental Digital Content 4, *http://links.lww.com/MPG/B490*) confirmed the good performance of BITSS photographs 1 and 7 representing BSFS stool types 6 and 7, respectively. This, however, was not the case for BITSS photograph 3, which represented BSFS

	Photo 2 (BSFS 1) %	Photo 5 (BSFS 2) %	Photo 4 (BSFS 3) %	Photo 6 (BSFS 4) %	Photo 3 (BSFS 5) %	Photo 1 (BSFS 6) %	Photo 7 (BSFS 7) %
BSFS 1	64.8	27.1	6.6	0.5	0.6	0.3	0.1
BSFS 2	21.9	49.7	25.0	1.4	1.5	0.5	0.0
BSFS 3	9.3	16.5	64.6	3.5	4.7	0.8	0.6
BSFS 4	0.9	2.0	1.5	87.6	6.8	0.8	0.5
BSFS 5	2.8	4.0	1.2	5.8	68.0	14.1	4.1
BSFS 6	0.2	0.6	1.0	1.0	15.1	75.0	7.1
BSFS 7	0.3	0.1	0.1	0.2	3.2	8.6	87.5

FIGURE 2. Proportions of exact agreement for each individual brussels infant and toddler stool scale photograph. BSFS = Bristol Stool Form Scale. Green cells: allocation matches the reference BSFS type for corresponding photograph. Orange cells: allocation deviates by 1 level from the reference BSFS type for corresponding photograph. Red cells: allocation deviates by more than 1 level from the reference BSFS type for corresponding photograph.

Copyright © ESPGHAN and NASPGHAN. All rights reserved.

TABLE 1. Correct allocations per observer group

BITSS photograph (Reference BSFS)	Total, n (%)	Parents, n (%)	Medical doctor, n (%)	Nurse, n (%)	P
Photo 2 (BSFS 1)	1523 (64.8)	706 (63.0)	421 (66.6)	396 (66.1)	0.225
Photo 5 (BSFS 2)	1170 (49.7)	498 (44.4)	343 (54.3)	329 (54.9)	<0.001
Photo 4 (BSFS 3)	1519 (64.6)	627 (55.9)	477 (75.5)	415 (69.3)	<0.001
Photo 6 (BSFS 4)	2060 (87.6)	923 (82.3)	594 (94.0)	543 (90.7)	<0.001
Photo 3 (BSFS 5)	1600 (68.0)	698 (62.3)	483 (76.4)	419 (69.9)	<0.001
Photo 1 (BSFS 6)	1765 (75.0)	794 (70.8)	520 (82.3)	451 (75.3)	<0.001
Photo 7 (BSFS 7)	2059 (87.5)	933 (83.2)	592 (93.7)	534 (89.1)	<0.001

BITSS = Brussels Infant and Toddler Stool Scale; BSFS = Bristol Stool Form Scale.

type 5; this photo was allocated by only 10% to BSFS type 5 (66% allocated it to BSFS type 6). BITSS photograph 5 (allocated by 25% to BSFS type 2) and BITSS photograph 4 (allocated by 25% to BSFS type 3) also performed poorly, while the results for BITSS photograph 2 (allocated by 65% to BSFS type 1) were more in line with the general results.

Performance of the Brussels Infant and Toddler Stool Scale

The BITSS photographs matched perfectly with the reference BSFS stool types for 819 (34.8%) observers. Significantly more MDs (278, 44.0%) than nurses (230, 38.6%) or parents (308, 27.7%) scored a perfect match ($P < 0.001$). The proportion of perfect matches was also significantly different across continents (Europe 44.5%, Americas 25.3%, and Asia 16.7%, $P < 0.001$). The multiple logistic regression model including terms for continent and observer group showed that both effects persisted when adjusted for the other (Supplementary file 2, Supplemental Digital Content 2, *http://links.lww.com/MPG/B488*). No evidence was found for a country effect, nor for interaction between continent and observer group (no effect modification).

The BITSS photographs showed maximum 1 class deviation with respect to the reference BSFS stool types for 1647 (70.0%) observers. Significantly more MDs (80.2%) than nurses (73.8%) or parents (62.2%) showed maximum 1 class deviations ($P < 0.001$). The proportion showing maximum 1 class deviations was also significantly different across continents (Europe 76.9%, Americas 64.6%, and Asia 51.4%, $P < 0.001$). Again, the regression model showed that both effects persisted when adjusted for the other (Supplementary file 2, Supplemental Digital Content 2, *http://links.lww.com/MPG/B488*).

The overall linearly weighted κ-value for the BITSS was 0.72 (95% CI 0.59 to 0.85), corresponding to good agreement (11). Similar values were found in the Netherlands (linearly weighted κ = 0.71 [95% CI 0.52 to 0.90]). The linearly weighted κ-values varied across the different observer groups, but all could be classified as good agreement (MDs κ = 0.80 [95% CI 0.68 to 0.91]; nurses: κ = 0.74 [95% CI 0.61 to 0.87]; and parents: κ = 0.67 [95% CI 0.53 to 0.81]). The same was true for the different continents: Europe: κ = 0.65 (95% CI 0.41 to 0.89); Asia: κ = 0.78 (95% CI 0.66 to 0.90); and Americas: κ = 0.66 (95% CI 0.54 to 0.78).

Performance of the Grouped Brussels Infants and Toddlers Stool Scale

The photographs were categorized a priori into the following groups: hard stools (BSFS types 1 and 2), normal formed stools (BSFS types 3 and 4), normal loose stools (BSFS type 5) and watery stools (BSFS type 6 and 7) (11). The application of this grouping to our data increased the overall number of correct classifications considerably to 49.1%; however, BITSS photograph 3 (BSFS 5, allocated as loose by 68.0%), BITSS photograph 4 (BSFS 3, allocated as normal by 66.1%), and BITSS photograph 5 (BSFS 2, allocated as hard stools by 76.9%) still stood out as weaker

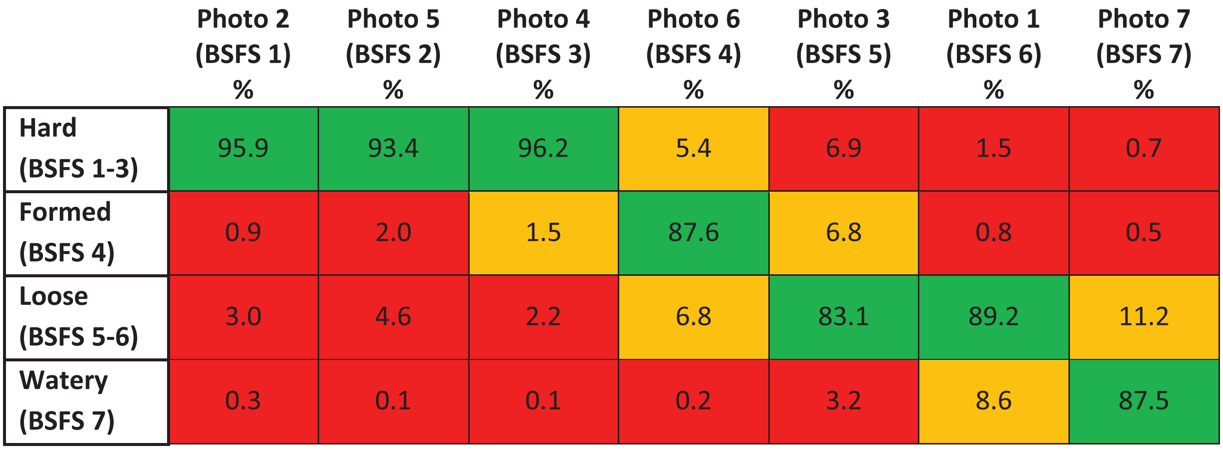

	Photo 2 (BSFS 1) %	Photo 5 (BSFS 2) %	Photo 4 (BSFS 3) %	Photo 6 (BSFS 4) %	Photo 3 (BSFS 5) %	Photo 1 (BSFS 6) %	Photo 7 (BSFS 7) %
Hard (BSFS 1-3)	🟩 95.9	🟩 93.4	🟩 96.2	🟧 5.4	🟥 6.9	🟥 1.5	🟥 0.7
Formed (BSFS 4)	🟥 0.9	🟥 2.0	🟧 1.5	🟩 87.6	🟧 6.8	🟥 0.8	🟥 0.5
Loose (BSFS 5-6)	🟥 3.0	🟥 4.6	🟥 2.2	🟧 6.8	🟩 83.1	🟩 89.2	🟧 11.2
Watery (BSFS 7)	🟥 0.3	🟥 0.1	🟥 0.1	🟥 0.2	🟥 3.2	🟧 8.6	🟩 87.5

FIGURE 3. Proportions of exact agreement for each individual brussels infant and toddler stool scale photograph according to new categorization system. BSFS = Bristol Stool Form Scale. Green cells: allocation matches the reference BSFS type for corresponding photograph. Orange cells: allocation deviates by max 1 level from the reference BSFS type for corresponding photograph. Red cells: allocation deviates by more than 1 level from the reference BSFS type for corresponding photograph.

Copyright © ESPGHAN and NASPGHAN. All rights reserved.

performing BITSS photographs (Supplementary file 5, Supplemental Digital Content 5, *http://links.lww.com/MPG/B491*). This was especially true when observers were free to match multiple photographs with the same BSFS type. The correct allocation of BITSS photographs 3 to 5 in the Netherlands was only 10.0%, 27.3%, and 70%, respectively. The number of correct allocations of this priori grouped BITSS differed significantly among observer groups (MDs 386, 61.1%; nurses 325, 54.3%; parents 443, 39.5%, $P < 0.001$) and continents (Europe 726, 57.6%; Asia 74, 35.2%; Americas 354, 40.1%, $P < 0.001$). Once more, the regression model showed that both effects persisted when adjusted for the other (Supplementary file 2, Supplemental Digital Content 2, *http://links.lww.com/MPG/B488*).

As a consequence of the findings, a new categorization system was created to enhance the performance of the grouped BITSS scale (Fig. 3). For this new categorization system, photographs representing BSFS type 1 to 3 were categorized as hard stools, the photograph representing BSFS type 4 was considered formed stool, the photographs representing BSFS types 5 and 6 were considered loose stools and the photograph representing BSFS 7 was considered watery stool. The BITSS photographs were correctly matched with the new categorization for 1713 (72.8%) of the observers. For each photo correct allocation into 1 of the 4 categories ranged from 83% to 96%, with an average of 90%. The number of correct allocations of this new grouped BITSS differed significantly among observer groups (MDs 539, 85.3%; nurses 448, 74.8%; parents 726, 64.8%, $P < 0.001$) and continents (Europe 1,006, 79.8%; Asia 132, 62.9%; Americas 575, 65.2%, $P < 0.001$). Once more, the regression model showed that both effects persisted when adjusted for the other (Supplementary file 2, Supplemental Digital Content 2, *http://links.lww.com/MPG/B488*). This scale also performed well when using the Dutch approach (Supplementary file 4B, Supplemental Digital Content 4, *http://links.lww.com/MPG/B490*): all BITSS photographs had over 80% correct allocations.

DISCUSSION

This study describes the interobserver reliability of stool consistency assessment using the BITSS, developed for infants and toddlers wearing diapers, among parents, nurses, and MDs from 18 countries. Overall, the BITSS showed good agreement with the BSFS. Photos representing BSFS types 4 (normal stools) and 7 (watery stools) performed very well. Photos, however, representing BSFS types 1 to 3 were easily mixed up among each other. Overall, the results indicate that agreement between the BITSS photographs and the BSFS varied strongly between pictures. When grouping different photographs together into clinically relevant groups, performance of the BITSS improved substantially. Finally, we proposed a new categorization of the BITSS based on our results. This new categorization system may be clinically more appropriate for the assessment of stools in diapers of young children, showing 80% agreement among health care professionals and 65% agreement among parents.

Previous studies utilizing the AISS have shown that assessment of stool consistency of young children who are not toilet trained is difficult and results in interobserver disagreement (9,10). A direct comparison with these results is, however, difficult, as different approaches were used in these studies. In the study by Bekkali et al, 2 observers (MD and medical student) rated pictures of stools according to the AISS and a 78% agreement rate was reported (9). Certain factors need to be considered when interpreting these results. First, these observers were also involved in the development of the AISS, which may explain why their agreement was so high. Furthermore, a learning effect can be expected when 2 observers are asked to use a scale >500 times, whereas observers were shown each BITSS photo only once in our approach. The exact agreement among MDs for the BITSS photographs grouped a priori into 4 stool consistency groups (hard stools, normal formed stools, loose normal stools, and watery stools) in our study was 61%; when a new categorization system was used, adapted according to the observed results, the percentage of perfect correct matches among MDs increased to 85%, a level of agreement that was never previously reached. Ghanma et al compared ratings of fresh stools by 2 nurses using the AISS and reported an exact agreement of 65% for stool consistency and similar results when using the BSS (69%) (10). The exact agreement among nurses for the BITSS photographs in our study was 54% using the a priori categorization system but it reached 75% when the new categorization system was used.

In the current study, only 35% of observers matched all 7 pictures of the BITSS perfectly with the BSFS reference types, with better results for medical professionals compared to parents. When the BITSS photographs were grouped together into 4 stool consistency groups (hard stools, normal formed stools, normal loose stools, and watery stools) that are used for the BSFS, the percentage of correct matches was 49%. When the new categorization system was used, the percentage of perfect correct matches, however, increased to 73%. This new categorization system grouped the BITSS photographs representing BSFS types 1 to 3 together as hard stools. Originally, BSFS type 3 is considered as a variation of normal stools in adults. Stool consistency, however, becomes harder with increasing age in infants, reflecting changes in diet and the maturation of the gut. Therefore, stools with the appearance of BSFS type 3 may be considered too hard for young children by parents, nurses, and MDs. In a future study, it would be interesting to use the BITSS for the assessment of fresh stools and obtain information about the observer's interpretation of the stool consistency as well. A prospective longitudinal study is also needed to investigate if the BITSS could serve as a reliable instrument to measure slow or accelerated transit in non-toilet trained children.

The BSFS is commonly used in children of all ages; however, it was not specifically designed for use in the pediatric population, and certainly not for use in non-toilet trained children. This aspect reflects the ambiguity of our research: while there is a consensus that the BSFS is not adapted for non-toilet trained children, the BSFS is used so frequently that we decided to compare the findings of the new approach (BITSS photographs) to the BSFS, considering it as the ''gold standard.'' The main problem with using the BSFS in the assessment of stools in diapers seems to be that stools may look different in diapers as compared to stools in toilets. Especially the form of soft stools is altered when it is pressed together between the buttocks and is spread out in the diaper. Also, the duration that the stools have been in the diaper will change the appearance. A previous attempt to overcome this problem resulted in the development of the AISS, which was developed specifically for infants under 1 year of age and included approximately 90% of prematurely born infants (9). The AISS is not commonly used in daily practice or research, probably due to its complexity and potentially because it is difficult to compare the results with other stool form scales. With the development of the BITSS, an alternative visual stool form scale has, however, been provided for non-toilet trained children. The BITSS can be used in clinic to help parents describe their child's stool consistency in order to gain insights into defecation patterns and potential gastrointestinal disorders. Moreover, the BITSS can also be considered for research purposes; for instance, to assess the effects of laxative treatment in a clinical trial for childhood constipation in children who are not toilet trained. We do advise a training session for possible investigators, as our results showed an influence of

Copyright © ESPGHAN and NASPGHAN. All rights reserved.

geographical location and observer group on the performance of both the original and the newly categorized BITSS scale.

Strengths of this study include the large sample size and the fact that this study was performed in several countries around the world. However, some limitations need to be taken into account when interpreting our results. The majority of participants were instructed to match each BITSS photograph with only 1 BSFS type, which may have resulted in bias. This bias, however, was not present in the results from the Netherlands, providing almost identical results to those with the newly proposed categorization system, although ideally every participant would have assessed the BITSS photographs using both instructions to assess the true effect of this bias. Moreover, no specific selection criteria were applied during the recruitment of responders and selection bias may have occurred. For example, the fact that all included health care professionals were active in pediatric departments could have led to a bias in the results, and it is unsure if nurses or MDs who are only sporadically involved into pediatric health care would perform the same. Demographic data of the parents are lacking, and it is unknown if this sample is representative of the general population.

In conclusion, our results show that after grouping BITSS photographs together into groups, this visual stool form scale is likely to prove useful in the assessment of stool consistency of non-toilet trained children both in clinical practice and for research purposes. The proposed new categorization should now be validated using a large number of fresh stools, rated by a limited number of observers.

Table of Contents Summary

The BITSS scale was validated as a reliable instrument to assess stools of non-toilet trained children via assessment of interobserver reliability among 2462 study participants.

REFERENCES

1. Benninga MA, Faure C, Hyman PE, et al. Childhood functional gastrointestinal disorders: neonate/toddler. *Gastroenterology* 2016;150:1443–55.
2. Hyams JS, Di Lorenzo C, Saps M, et al. Functional disorders: children and adolescents. *Gastroenterology* 2016;150:1456–68.
3. Kuizenga-Wessel S, Benninga MA, Tabbers MM. Reporting outcome measures of functional constipation in children from 0 to 4 years of age. *J Pediatr Gastroenterol Nutr* 2015;60:446–56.
4. Kuizenga-Wessel S, Heckert SL, Tros W, et al. Reporting on outcome measures of functional constipation in children—a systematic review. *J Pediatr Gastroenterol Nutr* 2016;62:840–6.
5. Saps M, van Tilburg MA, Lavigne JV, et al. Recommendations for pharmacological clinical trials in children with irritable bowel syndrome: the Rome foundation pediatric subcommittee on clinical trials. *Neurogastroenterol Motil* 2016;28:1619–31.
6. Lewis SJ, Heaton KW. Stool form scale as a useful guide to intestinal transit time. *Scand J Gastroenterol* 1997;32:920–4.
7. Vriesman MH, Velasco-Benítez CA, Ramirez CR, et al. Assessing children's report of stool consistency: the agreement between the pediatric Rome III Questionnaire and the Bristol Stool Scale. *J Pediatr* 2017;190:69–73.
8. Koppen IJN, Velasco-Benitez CA, Benninga MA, et al. 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.e1–8.e1.
9. Bekkali N, Hamers SL, Reitsma JB, et al. Infant stool form scale: development and results. *J Pediatr* 2009;154:521–6.
10. Ghanma A, Puttemans K, Deneyer M, et al. Amsterdam infant stool scale is more useful for assessing children who have not been toilet trained than Bristol stool scale. *Acta Paediatr* 2014;103:e91–2.
11. Vandenplas Y, Szajewska H, Benninga M, et al. Development of the Brussels Infant and Toddler Stool Scale ("BITSS"): protocol of the study. *BMJ Open* 2017;7:e014620.
12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.

Copyright © ESPGHAN and NASPGHAN. All rights reserved.