OPEN

The Modified Bristol Stool Form Scale: A Reliable and Valid Tool to Score Stool Consistency in Dutch (Non)Toilet-trained Toddlers

[*§]Carrie A.M. Wegh, [*]Gerben D.A. Hermes, [†]Margriet H.C. Schoterman, [‡]Elaine E. Vaughan, [*]Hauke Smidt, [*]Clara Belzer, and [§]Marc A. Benninga

ABSTRACT

Objective: The aim of the study was to assess whether the modified Bristol Stool Form Scale (m-BSFS) is reliable, valid and user-friendly to use by parents, grandparents, and day childcare employees to evaluate stool consistency in toilet and nontoilet-trained toddlers in the Netherlands.

Study design: Translation to Dutch and validity of the m-BSFS (scoring 32 general stool pictures) for 1 to 3 year old toddlers ($n = 89$) was evaluated by parents, grandparents, and day childcare employees. A subgroup of participants scored an additional 7 pictures of stools in a diaper to validate the m-BSFS for non-toilet-trained toddlers ($n = 16$). To determine inter-rater reliability, 2-way random effects single-rater intraclass correlation coefficient $(\text{ICC})_{\text{consistency}}$ was used. Intra-rater reliability was measured by Cohen kappa (κ) by rating the same pictures in random order twice, with at least 1 week between the first and second scoring.

Results: Inter- and intra-rater reliability of the m-BSFS were above recommended minimal standards of 0.61 for the 32 general stool pictures as well as for the 7 pictures of stools in a diaper. $\text{ICC}_{\text{consistency}}$ for the general stool pictures of the first and second ratings were 0.71 ($n = 89$) and 0.79 ($n = 77$), respectively, with a κ of 0.71 ($n = 77$). $\text{ICC}_{\text{consistency}}$ for the stools in diaper pictures of the first and second ratings were 0.93 ($n = 16$) and 0.93 ($n = 15$), respectively, with a κ of 0.77 ($n = 15$).

Conclusions: The m-BSFS is reliable, valid and user-friendly to use by Dutch-speaking parents, grandparents, and day childcare workers to evaluate stool consistency in both toilet- and nontoilet-trained toddlers in the Netherlands.

Key Words: Bristol Stool Form Scale, constipation, functional gastrointestinal disorders, modified Bristol Stool Form Scale for Children

An infographic is available for this article at: *http://links.lww.com/MPG/C426*.

(*JPGN* 2021;73: 210–216)

What Is Known

- Reliable and valid assessment of stool consistency is important for evaluating defecation patterns in toddlers and diagnosing gastrointestinal disorders.
- Reliability and validity has not been evaluated in current stool scales on defecation patterns in toddlers that may include stools from a diaper as well as from a toilet/potty.

What Is New

- The modified Bristol Stool Form Scale was reliable, valid, and user-friendly to use to evaluate stool consistency in both toilet- and nontoilet-trained toddlers in the Netherlands.
- To the best of our knowledge, this is the first stool scale that has been validated to score stool consistency in both toilet/potty as well as in diapers. This will be of value for monitoring bowel habits in young children in clinical research for food ingredients, medicines, or lifestyle changes that can impact this parameter at a critical age of toilet training.
- The reliability and validity of the modified Bristol Stool Form Scale has been assessed in caregivers.

Received August 18, 2020; accepted March 23, 2021.

From the [*]Laboratory of Microbiology, Wageningen University & Research, Wageningen, the [†]FrieslandCampina, Amersfoort, the [‡]Sensus (Royal Cosun), Roosendaal, and the [§]Department of Pediatric Gastroenterology and Nutrition, Emma's Children's Hospital, Amsterdam University Medical Center, Amsterdam, The Netherlands.

Address correspondence and reprint requests to Carrie A.M. Wegh, Stippeneng 4, 6708 WE, Wageningen, The Netherlands
(e-mail: carrie.wegh@wur.nl).

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*).

This study was supported by a grant from Topconsortium voor Kennis en Innovatie (TKI) (AF17067), FrieslandCampina, and Sensus.

Data statement: anonymized data of participant characteristics and all ratings can be requested at the corresponding author.

C.A.M.W. is supported by a grant from Topconsortium voor Kennis en Innovatie (TKI) (AF17067), FrieslandCampina and Sensus, G.H. is supported by the project MASTER (Microbiome Applications for Sustainable food systems through Technologies and Enterprise) funded by the European Union's H2020 research and innovation programme (No. 818368), M.H.C.S. is an employee of FrieslandCampina, E.E.V. is an employee of Sensus (Royal Cosun), M.A.B. has acted as a consultant for Shire, Sucampo, Astrazeneca, Norgine, Coloplast, Danone, FrieslandCampina, Sensus, and Novalac. For the remaining authors, there are no conflicts of interest.

DOI: 10.1097/MPG.0000000000003186

Alterations in stool frequency and stool consistency are associated with numerous organic and functional gastrointestinal (GI) disorders in children, such as irritable bowel syndrome (IBS), inflammatory bowel disease (IBD), and functional constipation (FC). These diseases may be so serious that they affect quality of life of the children and their parents (1–3). Medical professionals and scientists seek various approaches including dietary modifications and fiber-enriched foods/supplements to improve intestinal issues, such as constipation (4,5). Stool scales are often used to diagnose and evaluate GI disorders in children or toddlers. Several stool scales can be used depending on the age of the target group. Examples of these scales not only include the Bristol Stool Form Scale (BSFS) and the modified-BSFS (m-BSFS) but also scales that allow for scaling stools in diapers, such as the Amsterdam infant stool scale (AISS) and Brussels infant and toddler stool scale (BITSS) (6–9). These stool scales, however, are either general scales often used for defecation on a toilet or scales specifically developed to evaluate stools in a diaper. This contrasts with the fact that research on defecation patterns in toddlers may include stools from a diaper as well as from a toilet/potty. The m-BSFS is a pediatric 5-point stool form rating scale developed to score stools of toilet trained children, and validated to be scored by pediatric gastroenterologists and children >8 years of age. The evaluation of stools of toddlers to manage defecation problems is often done by caregivers, such as parents, grandparents, and day childcare employees. In addition, the m-BSFS is validated for the English language only, and in paper form, whereas currently, there is a shift towards more online-based methods for scoring in clinical research. Therefore, the aim of this study was to assess whether the m-BSFS in paper form and online is reliable, valid and user-friendly to use by Dutch-speaking parents, grandparents and day childcare employees to evaluate stool consistency in both toilet- and nontoilet-trained toddlers in the Netherlands (*http://links.lww.com/MPG/C373*).

MATERIALS AND METHODS

Translation and Cultural Adaptation of the Dutch Modified Bristol Stool Form Scale

Published guidelines were followed to translate and to culturally adapt the English version of the m-BSFS to Dutch (10–13). To achieve linguistic equivalence to the original m-BSFS, 2 native English-speaking forward translators with excellent knowledge of Dutch were asked to individually translate the m-BSFS into Dutch. Both translators were then asked to discuss their results with each other until consensus was reached. This Dutch m-BSFS was then presented to 2 native Dutch backward translators with excellent knowledge of English but with no previous knowledge of the English m-BSFS. The Dutch backward translators were asked to discuss their results until consensus was reached. The original English m-BSFS was then compared with the English backward translation. Subsequently, the Dutch version was given to 12 pediatric gastroenterology fellows to check if the scale's language matched the language used in practice, and if necessary, additional adaptations were made. The new Dutch m-BSFS was again sent to all 4 translators, who were asked whether they agreed with the new, culturally adapted translation.

Participants

The Medical Ethical Reviewing Committee of Wageningen University (METC-WU) reviewed the research file and concluded that this research does not fall within the remit of the Dutch 'Medical Research Involving Human Subjects Act'. Following International Conference on Harmonization—Good Clinical Practice (ICH-GCP) guidelines, however, informed consent was obtained. Participants were eligible if they were a parent of at least 1 child of 1 to 3 years old (12–36 months), day childcare employees working with toddlers of 1 to 3 years old, or grandparents of at least 1 grandchild of 1 to 3 years old. We sought to include approximately n = 100 participants, based on published recommendations, for the 32 general stool pictures in the ratio 3 : 1 : 1 for parents, grandparents, and day childcare employees, respectively (11). More parents were included as parents would be looking after their children most of the time in a real-life situation. In order to validate the 7 stools in diaper pictures and assess user-friendliness, we aimed to include approximately n = 20 participants. Flyers for the study were distributed over day childcare centers, public areas. such as public libraries and spread online via multiple media. Questionnaires were then sent by e-mail to those willing to participate.

Inter- and Intra-rater Reliability

The same 32 color pictures of stools as were used to initially evaluate and re-evaluate the m-BSFS by pediatric gastroenterologists and children of 3 to 18 years of age by Chumpitazi et al (9) and Lane et al (14) were obtained from the authors. These pictures will be referred to throughout as "general stool pictures." These pictures constituted focused, close-ups of entire stools in a toilet or potty but there were very few pictures of stools in diapers. To investigate whether it is possible to use only 1 scale for toddlers and avoid problems with comparisons between stool scales, and investigate if this scale can also be used for nontoilet-trained toddlers, 7 extra pictures were included in the validation. These additional pictures showed focused, close-ups of entire stools in diapers, as previously used by Huysentruyt et al (8) for their BITSS that are referred to as "stools in diaper pictures." Both general stool pictures and stools in diaper pictures depicted the full range of stool consistencies from type 1 to type 5 on the m-BSFS.

For the validation, interrater reliability was used as a measure for agreement between raters and intra-rater reliability as a measure for agreement within 1 person between the first and second time of scoring the pictures. To assess inter- and intra-rater reliability of the general stool pictures, participants were asked to complete a questionnaire that was built in the online platform Castor EDC (15). The questionnaires were sent to parents, grandparents, and day childcare employees as representatives for the people who most often take care of toddlers. Participants were asked to fill out the questionnaire twice with at least 1 week between the first and second scoring. The order of the pictures was different for both questionnaires to avoid bias.

For the stools in diaper pictures, inter- and intra-rater reliability was assessed in participants for 7 focused, close-up color images of bowels in diapers. Participants were asked to fill out the questionnaire twice with at least 1 week between the first and second scoring and again the order of the pictures was different for both questionnaires to avoid bias.

Paper Versus Online Use and User-friendliness in a Real-life Situation

For the scoring of the stools in diaper pictures, the participants were randomly assigned to scoring the stools in diapers either in a paper version or in an online version to investigate whether this would impact the $ICC_{consistency}$. Participants were also asked to fill out a 1-week diary, in which they scored all bowel movements of the child. Thus the parent, grandparent, or day childcare employee scored the bowel movements of the child(ren) for which they were present. Moreover, 3 statements were added to assess the user-friendliness, clarity of the instructions, and feasibility to use the m-BSFS on fresh stool samples in real-life situations. Each category for user-friendliness was scored on a 5-point Likert scale ranging from strongly agree to strongly disagree (16). Lastly, in case, participants scored the questionnaire as unclear, user-unfriendly or demanding, they were asked to elaborate on this as an open question.

Statistical Analysis

Statistical analyses were performed with R software, version 3.6.1 using the ‘‘irr’’ package version 0.84.1 and the ‘‘ICC.Sample.Size’’ package version 1.0 (17–19). A 2-way random effects single rater model intraclass correlation coefficient for consistency $(ICC)_{consistency}$ was used for inter-rater reliability with the ‘‘icc’ function, whereas Cohen κ (function ‘‘kappa2’’ was used for intra-rater reliability (11,20). As there were no comparable studies in terms of type of scale or type of people to score them to use for a priori power calculation, a posteriori analyses were conducted to check if the sample size used provided enough power to draw valid conclusions (function ‘‘calculateIccPower’’ from the ‘‘ICC.Sample.Size’’ package) (19). In addition, subject to item ratios were calculated for which many rules of thumb exist that range from a subject to item ratio of at least 2 : 1 to 20 : 1 (21).

RESULTS

Translation and Cultural Adaptation of the Dutch Modified Bristol Stool Form Scale

The original English m-BSFS and the final translated Dutch version of the m-BSFS as used in the validation study are presented in Figure 1.

Participants

In total, 93 participants completed the questionnaire on the general stool pictures, of whom 69% were parents, 20% grandparents, and 11% day childcare employees (Fig. 2). A total of 16 participants, constituting 69% parents, 19% grandparents, and 12% day childcare employees, completed the stools in diaper pictures and the user-friendliness questions in full. Three participants did not complete the questionnaire without giving reasons and 1 participant indicated that viewing and rating the pictures caused nausea. Participants reported to look after 1 to 2-year-old (52%) and 2 to 3-year-old toddlers (48%), 55% girls, and 45% boys, of whom 15% were completely toilet-trained, 7% only during the day, 1% only for urine, and 77% were nontoilet-trained.

Inter- and Intra-rater Reliability

Out of a total of 4505 ratings for both general stool pictures and stools in diaper pictures, 3505 (77.8%) were in agreement with the most commonly chosen rating, and 4272 (94.8%) were within 1 form type of the most commonly chosen rating for each stool picture. More specifically, for the general stool pictures, 3349 out of 4288 ratings (78.1%) were in agreement with the most commonly chosen rating, while this was the case for 156 out of 217 ratings (71.8%) for the stools in diaper pictures. For the general stool pictures, 4055 out of 4288 (94.6%) were within 1 rating from the most commonly chosen rating, compared with 217 out of 217 (100.0%) for the stools in diaper pictures.

The proportions of exact agreement of each individual picture are presented in Figure 3A and B for the general stool and stools in diaper pictures, respectively. Of the 32 general stool pictures, 3 pictures were most commonly scored as type 1, 8 as type 2, 8 as type 3, 8 as type 4, and 5 as type 5 (Fig. 3). For the 7 stools in diaper pictures, 1 was most commonly scored as type 1, 2 as type 2, 1 as type 3, 1 as type 4, and 2 as type 5 (Fig. 3B). Concerning the percentage of ratings that were in concordance with each other, the 3 weakest performing general stool pictures corresponded to m-BSFS type 4 (43%), type 5 (46%), and type 3 (47%). For the stools in diaper pictures, the 3 weakest performing pictures were m-BSFS type 4 (48%), type 3 (52%), and type 1 (52%).

The a posteriori power calculation shows that, for a power of 0.8 and an α of 0.05, sample sizes for both general stool pictures and the stools in a diaper pictures numbers were above the sample size

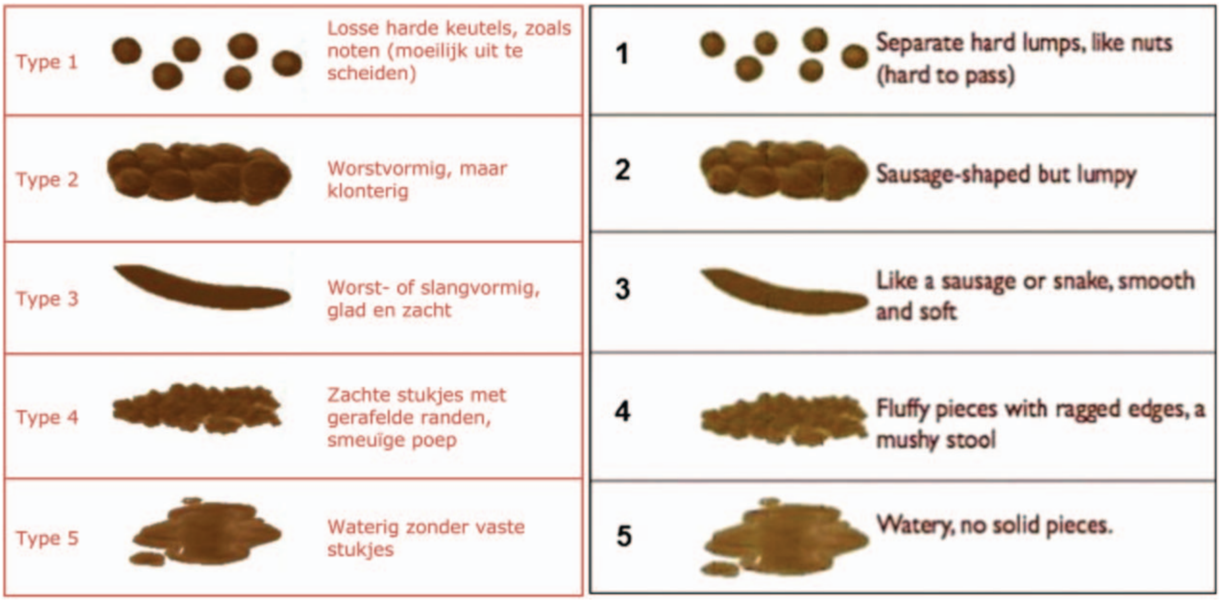

Type 1
Losse harde keutels, zoals noten (moeilijk uit te scheiden)
Type 2
Worstvormig, maar klonterig
Type 3
Worst- of slangvormig, glad en zacht
Type 4
Zachte stukjes met gerafelde randen, smeuïge poep
Type 5
Waterig zonder vaste stukjes
1
Separate hard lumps, like nuts (hard to pass)
2
Sausage-shaped but lumpy
3
Like a sausage or snake, smooth and soft
4
Fluffy pieces with ragged edges, a mushy stool
5
Watery, no solid pieces.

FIGURE 1. Left: Dutch version of the modified Bristol Stool Form Scale, right: the original English m-BSFS. Data from (13). m-BSFS = modified Bristol Stool Form Scale.

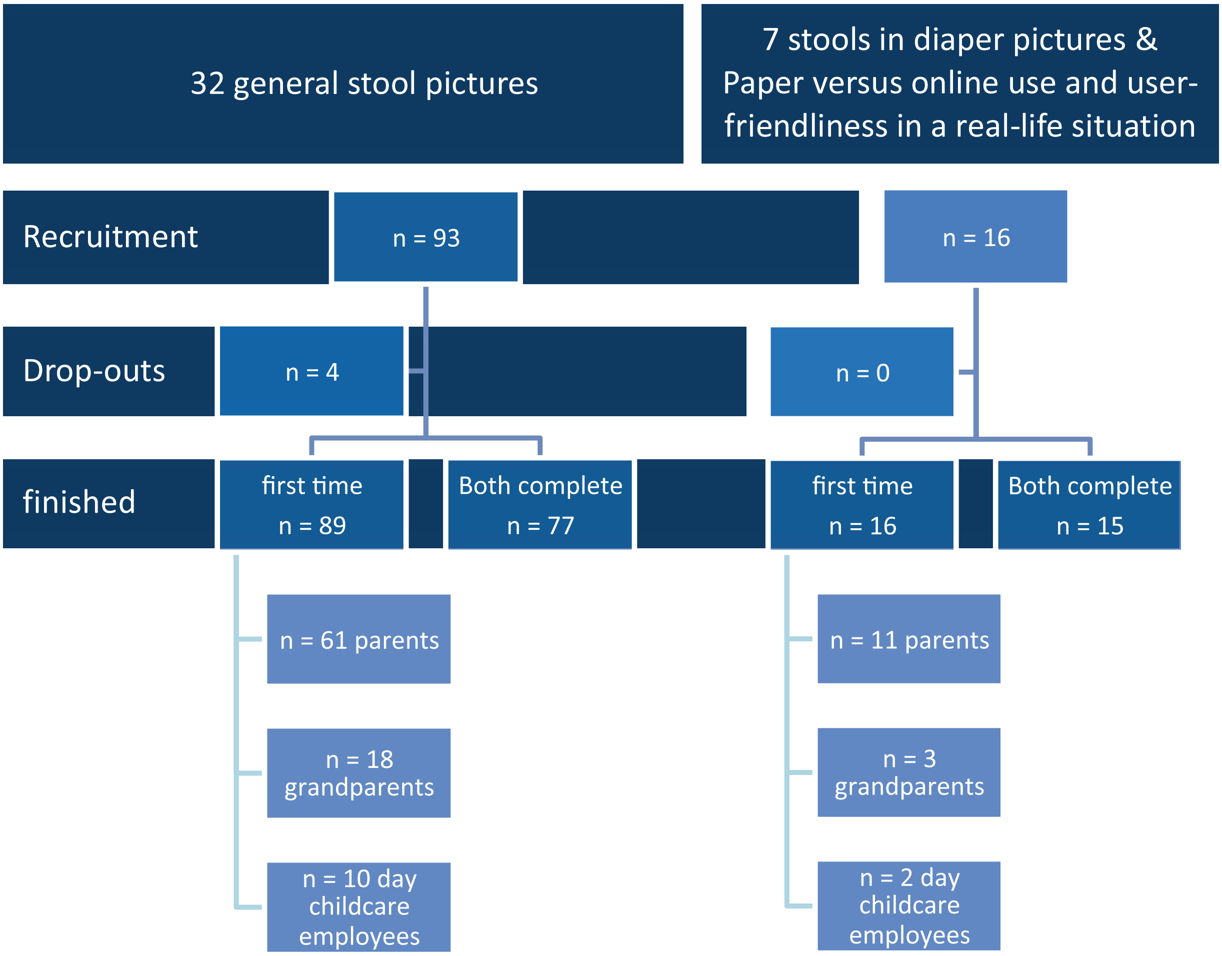


FIGURE 2. Flow chart of participants in the validation study.

as used in this study (Table 1). To further support this, the subject to item ratio was calculated. For the general stool pictures, this subject to item ratio was 2.8 for the first time and 2.4 for the second time and for the diaper-specific pictures this was 2.3 for the first time and 2.1 for the second time.

In conclusion, all inter- and intra-rater reliability scores were above the most commonly used thresholds of >0.61 or >0.7 (Table 1). Inter- and intra-rater reliabilities were found to be higher for the stools in diaper pictures, compared with the general stool pictures.

Paper Versus Online Use and User-friendliness in Real-life Situations

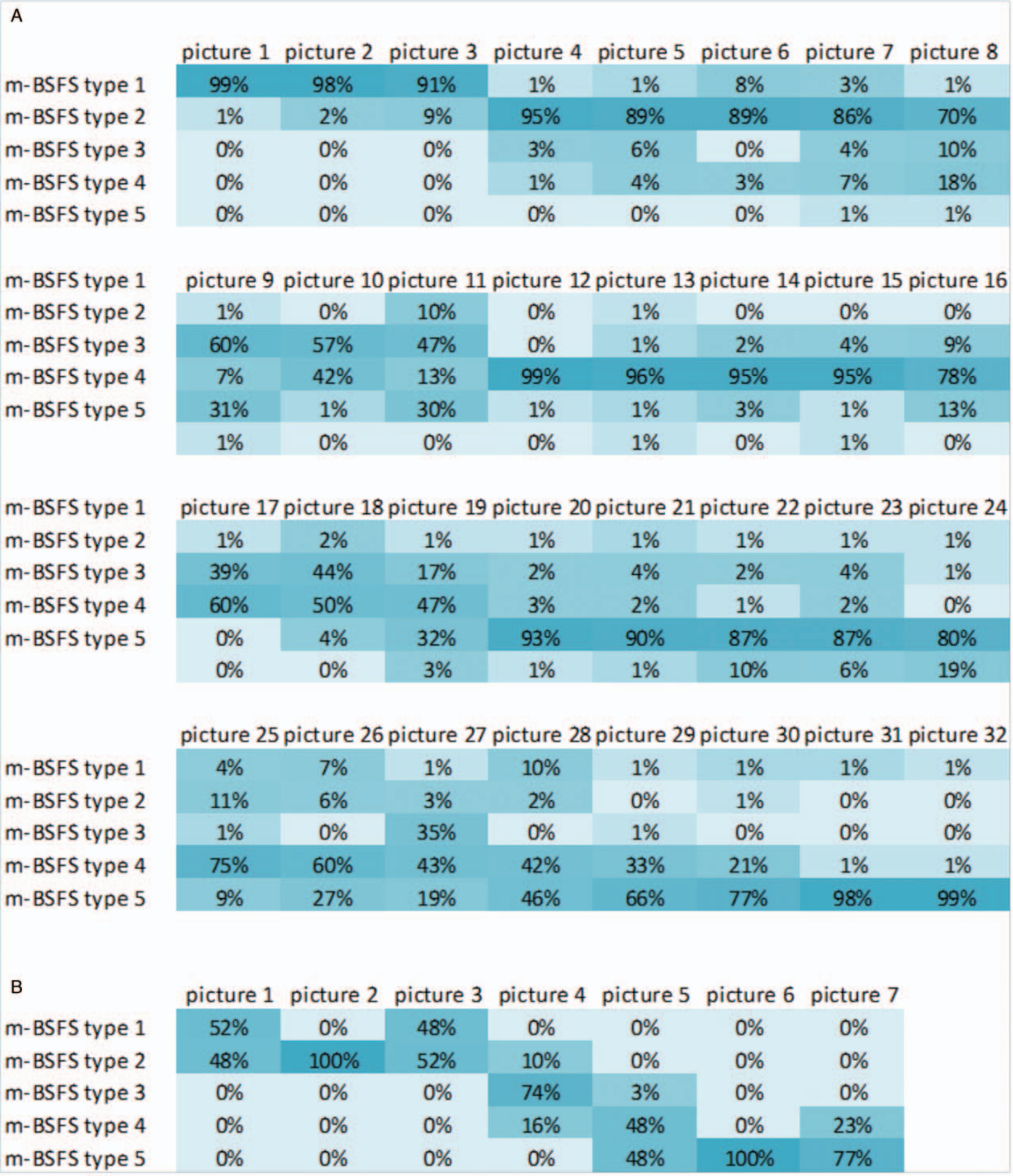
A comparison of the paper (n = 10) versus the online (n = 6) $ICC_{consistency}$ for the stools in diaper pictures revealed strong $ICC_{consistency}$ of 0.94 for both the online version (0.94) and paper version (0.93). To assess significance of the difference between these 2 ICCs, a 2-sided Fischer r-to-z transformation was used, resulting in a *P* value of 0.90, indicating no statistical significance between the ICC of the paper and the online version.

In terms of user-friendliness by means of the 5-point Likert scale, all 16 participants, experienced the m-BSFS as user friendly based on the responses for clarity of instructions, which were ‘‘neutral’’ for 10%, whereas 60% answered ‘‘agree’’ and 30% ‘‘strongly agree.’’ In response to the following translated statements, ‘‘I think it is demanding to use the m-BSFS. For example, it takes effort to remember to use the m-BSFS or it was demanding to compare fresh stools to the m-BSFS,’’ 20% of the participants answered with ‘‘strongly disagree,’’ 20% with ‘‘disagree’’ and 60% with ‘‘neutral.’’ In conclusion, both the paper and online version of the m-BSFS were largely considered user-friendly by the study participants.

DISCUSSION

The objective of this study was to assess whether the m-BSFS is reliable, valid, and user-friendly to use by parents, grandparents, and day childcare employees to evaluate stool consistency in both toilet- and nontoilet-trained toddlers in the Netherlands. Overall, the m-BSFS was successfully translated and culturally adapted to Dutch, and showed to have a high degree of inter-rater reliability, intra-rater reliability and user-friendliness, regardless of whether this was on paper or online. This scale is the first that can be used for rating stools both from a diaper as well as from a potty or toilet.

We showed that both inter- and intra-rater reliability were above thresholds of >0.61 or >0.7, as recommended by published guidelines (11,22–24). These findings are consistent with previous studies that aimed to validate the m-BSFS by either pediatric gastroenterologists or children that found inter- and intra-rater reliabilities ranging from 0.72 (from 8 to 10 years of age and up)



A

	picture 1	picture 2	picture 3	picture 4	picture 5	picture 6	picture 7	picture 8
m-BSFS type 1	99%	98%	91%	1%	1%	8%	3%	1%
m-BSFS type 2	1%	2%	9%	95%	89%	89%	86%	70%
m-BSFS type 3	0%	0%	0%	3%	6%	0%	4%	10%
m-BSFS type 4	0%	0%	0%	1%	4%	3%	7%	18%
m-BSFS type 5	0%	0%	0%	0%	0%	0%	1%	1%

m-BSFS type 1	picture 9	picture 10	picture 11	picture 12	picture 13	picture 14	picture 15	picture 16
m-BSFS type 2	1%	0%	10%	0%	1%	0%	0%	0%
m-BSFS type 3	60%	57%	47%	0%	1%	2%	4%	9%
m-BSFS type 4	7%	42%	13%	99%	96%	95%	95%	78%
m-BSFS type 5	31%	1%	30%	1%	1%	3%	1%	13%
	1%	0%	0%	0%	1%	0%	1%	0%

m-BSFS type 1	picture 17	picture 18	picture 19	picture 20	picture 21	picture 22	picture 23	picture 24
m-BSFS type 2	1%	2%	1%	1%	1%	1%	1%	1%
m-BSFS type 3	39%	44%	17%	2%	4%	2%	4%	1%
m-BSFS type 4	60%	50%	47%	3%	2%	1%	2%	0%
m-BSFS type 5	0%	4%	32%	93%	90%	87%	87%	80%
	0%	0%	3%	1%	1%	10%	6%	19%

	picture 25	picture 26	picture 27	picture 28	picture 29	picture 30	picture 31	picture 32
m-BSFS type 1	4%	7%	1%	10%	1%	1%	1%	1%
m-BSFS type 2	11%	6%	3%	2%	0%	1%	0%	0%
m-BSFS type 3	1%	0%	35%	0%	1%	0%	0%	0%
m-BSFS type 4	75%	60%	43%	42%	33%	21%	1%	1%
m-BSFS type 5	9%	27%	19%	46%	66%	77%	98%	99%

B

	picture 1	picture 2	picture 3	picture 4	picture 5	picture 6	picture 7
m-BSFS type 1	52%	0%	48%	0%	0%	0%	0%
m-BSFS type 2	48%	100%	52%	10%	0%	0%	0%
m-BSFS type 3	0%	0%	0%	74%	3%	0%	0%
m-BSFS type 4	0%	0%	0%	16%	48%	0%	23%
m-BSFS type 5	0%	0%	0%	0%	48%	100%	77%

FIGURE 3. Pictures are ordered to agreement on type. Participants received the pictures in random order. Colors of this figure are a gradient ranging from dark blue as highest values, light blue as lowest values. (A) Proportions of exact agreement for each of the 32 general stool pictures. (B) Proportions of exact agreement for each of the 7 stools in diaper pictures. m-BSFS = modified Bristol Stool Form Scale.

to 0.86 and 0.79 (from 8 to 10 years of age and up) to 0.87, respectively (1,14). Comparing our results to pediatric gastroenterologists' ratings on the same general stool pictures (in percentages) showed agreement with most commonly chosen ratings within 1 form type. ICCs were lower, 0.716 and 0.793 in our study compared with 0.85, but still well above recommended thresholds. Moreover, only 4 out of the 32 general stool pictures were scored differently by our participants compared with pediatric gastroenterologists, in which our participants scored 3 out of 4 pictures as a softer stool consistency and 1 out of 4 as a harder stool consistency (9). These differences could be explained by the difference in training and familiarity with stool patterns of gastroenterologists compared with our study participants. When comparing our results to those of children ages 3 to 18 years, our results concerning the ICC were

TABLE 1. Intraclass Correlation Coefficient$_{\text{consistency}}$ and Cohen κ for inter- and intra-rater reliabilities of the modified Bristol Stool Form Scale

	ICC/kappa	95% CI	Sample size boundary for power = 0.8
General stool pictures			
First time, n = 89	0.716	0.617 to 0.818, $P < 0.001$	57 to 64
Second time, n = 77	0.793	0.710 to 0.872, $P < 0.001$	20 to 22
Kappa, n = 77	0.706	$P < 0.001$	n.a.
Stools in diaper pictures			
First time, n = 16	0.925	0.828 to 0.984, $P < 0.001$	6
Second time, n = 15	0.934	0.847 to 0.986, $P < 0.001$	6
Kappa, n = 15	0.769	$P < 0.001$	n.a.

CI = confidence interval; ICC = intraclass correlation coefficient.

comparable to those obtained with children ages 8 to 10 years and up, who only used 10 out of 32 pictures for their final evaluation (14).

To the best of our knowledge, this is the first stool scale that has been validated for scoring stool consistency in stools in a potty/toilet and also for stools in a diaper. The added value of this validation is that data of potty-trained children can be directly compared with nonpotty-trained children and can also be used in a period when children are being potty-trained. Evaluation of the stools in diaper pictures resulted in a high ICC. Related to these findings concerning our 7 stools in diaper pictures and the findings by Lane et al (14) who also showed a high ICC in children ages 8 to 10 and up, with their 10 selected pictures, we noticed that the approach for calculating the ICC can lead to misleading results when comparing results for small numbers of items (pictures in this case) to those obtained with higher numbers of items. More specifically, the lower the number of items to be rated, the higher the ICC without having actual higher agreement (25). For example, when computing a random set of 7 items, n = 7, an ICC of 0.760 was calculated whereas 4 times the exact same 7 items, n = 28, gave an ICC of 0.738, without an actual difference in agreement percentages. Therefore, we recommend critically considering the type of ICC being used, how this is reported and checking if the number of items that were rated are comparable in order to directly compare ICCs to each other (20,25). Therefore, in order to compare ICCs one-on-one, it is recommended to use the same number of items. For this study, this does not only hold for comparing the data by Lane *et al* to our data but also for comparing the ICCs of the general stool pictures to the stools in diaper pictures. As the ICCs for our stools in diaper pictures are, however, very high, ranging from 0.925 to 0.934, we are confident that, even with the possible bias described above, our questionnaire is valid and well above the thresholds of 0.61 or 0.7 as the computations did not show a bigger difference than 0.06 (11,22–24). Altogether, we can conclude that we not only confirmed previous findings in a different target group, but with the extra stools in diaper pictures, we furthermore confirmed that the m-BSFS can be used for nontoilet-trained toddlers. To the best of our knowledge, this is the first stool scale that is validated for both stools in a toilet/potty and for stools in a diaper. The advantages are that results of older and younger children can be more easily compared, and research on stool consistency in toddlers, especially during potty training, can be done with only 1 stool scale instead of switching between validated scales. Moreover, despite the widespread use of stool scales in the research and management of GI diseases and functional GI disorders, this is one of the few stool scales that has been validated in a target age group.

Moreover, it has been suggested that stool form, as measured, for example, by the m-BSFS, is a proxy for colonic transit rate (6,26). In order to, however, confirm the validity of this statement, and use stool form as proxy for colonic transit rate, it becomes even more important to not only validate all different stool scales in the respective target group but also validate if this statement remains valid for other stool scales and different target groups.

In addition, most stool scales have previously been completed on paper while currently there is a shift towards more online-based methods (27). This shift towards online-based methods can have multiple reasons, of which the most obvious probably is the all-round presence of mobile devices. This comes with several challenges, such as data protection, validation of online tools and questionnaires, difficulties with reaching certain respondent groups, such as elderly and residents of remote areas and survey fraud. Online-questionnaires also come with many advantages, including automation in data input handling, quick inclusion and response of participants, real-time data collection, and anonymity that may lead to more honest responses as there is no social consequence to participation. Moreover, other important advantages include data validation and collection of all raw data in 1 database without losing or incorrectly entering data from paper files (27). To our best knowledge, this is the first study to compare a paper and online stool scale and we show that both questionnaires give comparable results in terms of ICC$_{\text{consistency}}$. In short, when the m-BSFS is used in practice, caregivers can use the paper or online version, or even mix and match during research or management of GI diseases or functional GI disorders to the method that suits best at that moment. Furthermore, in general, participants indicated that the m-BSFS is user friendly, with clear user instructions and not very demanding to use.

Strengths of this study includes that parents, grandparents, and day childcare employees were divided as proposed on forehand (3 : 1 : 1, respectively). Moreover, division of the toddlers in terms of gender and age ranges was close to 50 : 50, indicating a good representation of the population in which this stool scale might be used. Moreover, we did an a posteriori sample size calculation to check whether the number of participants was sufficient. By using our own data, that is, the actual ICCs found in this study, the power calculation is even more reliable as it is not an estimate based on comparable studies or study populations. We used the function ‘calculateIccPower’ from the ICC.Sample.Size package in R, which calculates a post hoc power from an ICC. This function demonstrates the additional power gained by increasing the number of subjects or the number of subjects needed to increase power. In addition, it determines the number of participants needed for a specific power. Consequently, we can conclude, based on the results from the ‘‘calculateIccPower’’ function, that the sample size boundary for a power of 0.8 would be 6 participants, as shown in Table 1. Our conclusion is, therefore, valid based on the a posteriori power calculation as well as confidence intervals of which even the lowest boundary is well above the most commonly used thresholds of 0.61 or 0.7. In addition, the subject to item ratios were between 2.8 and 2.1 in this study, which is according to

Anthoine et al (21) in line with the most commonly found number (92% of all studies are in the range of 2:1 to 20:1) in scale validation studies.

Considering limitations of this study, the 32 pictures generously provided by Chumpitazi et al are at least 10 years old and a few of which are relatively low in resolution and quality. The extra stools in diaper photographs helped to address this. This potentially leads to a lower $ICC_{consistency}$ or Cohen κ, not because of a true difficulty to discriminate the type of stool but because of the resolution and quality of the images. Possibly in line with this, we observed a higher $ICC_{consistency}$ and Cohen κ for the 7 stools in diaper samples. The latter pictures were more comparable to each other in terms of size, lighting and photographic composition but were also of higher resolution and quality, which might be a factor influencing $ICC_{consistency}$ and Cohen κ. Demographic data and socioeconomic status of the participants in our study are lacking, and it is, therefore, unknown if this sample is representative of the general population. We, however, do not expect that differences in demographics will significantly change the results of this study.

In conclusion, the modified m-BSFS, as paper or online version, is reliable, valid and user-friendly to use for Dutch-speaking parents, grandparents, and day childcare employees to evaluate defecation parameters in toddlers whether in diapers or toilet-trained. This validated m-BSFS is likely to prove useful in both clinical and research settings as a validated measure to record stool form from both diapers and toilet or potty. The m-BSFS can be used in clinical practice and clinical trials as tool for diagnosis, management, and evaluation of bowel patterns in healthy toddlers as well as in several disorders, such as functional constipation (FC), functional diarrhea (FD), or irritable bowel syndrome (IBS).

ACKNOWLEDGMENTS

The authors would like to thank all participants, Bruno P. Chumpitazi for generously providing us with the 32 general stool pictures, Maria Reumerman and Floriane Brunel for their assistance during this study and the translators Esther van Raamsdonk, Peter van Elswijk, James C. Dykstra, and Michael Rose.

REFERENCES

1. Hyams JS, Di Lorenzo C, Saps M, et al. Childhood functional gastrointestinal disorders: child/adolescent. *Gastroenterology* 2016;150: 1456–1468. e2.
2. Herzog D, Fournier N, Buehr P, et al., Swiss IBD Cohort Study Group. Age at disease onset of inflammatory bowel disease is associated with later extraintestinal manifestations and complications. *Euro J Gastroenterol Hepatol* 2018;30:598–607.
3. Drossman DA. Functional gastrointestinal disorders: history, pathophysiology, clinical features, and Rome IV. *Gastroenterology* 2016;150: 1262–1279. e2.
4. Wegh CA, Schoterman MH, Vaughan EE, et al. The effect of fiber and prebiotics on children's gastrointestinal disorders and microbiome. *Expert Rev Gastroenterol Hepatol* 2017;11:1031–45.
5. Firmansyah A, Chongviriyaphan N, Dillon DH, et al. Fructans in the first 1000 days of life and beyond, and for pregnancy. *Asia Pac J Clin Nutr* 2016;25:652.
6. Lewis SJ, Heaton KW. Stool Form Scale as a Useful Guide to Intestinal Transit Time. *Scand J Gastroenterol* 1997;32:920–4.
7. 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.
8. Huysentruyt K, Koppen I, Benninga M, et al., BITSS working group. The Brussels Infant and Toddler Stool Scale: a study on interobserver reliability. *J Pediatr Gastroenterol Hepatol Nutr* 2019;68:207–13.
9. Chumpitazi BP, Lane MM, Czyzewski DI, et al. Creation and initial evaluation of a Stool Form Scale for children. *J Pediatr* 2010;157:594–7.
10. Lohr KN. Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res* 2002;11:193–205.
11. Terwee CB, Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007;60:34–42.
12. Tsang S, Royse CF, Terkawi AS. Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine. *Saudi J Anaesth* 2017;11(Suppl 1):S80–9.
13. Sousa VD, Rojjanasrirat W. Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. *J Eval Clin Pract* 2011;17:268–74.
14. Lane MM, Czyzewski DI, Chumpitazi BP, et al. Reliability and validity of a modified Bristol Stool Form Scale for children. *J Pediatr* 2011;159:437.e1–41.e1.
15. Castor Electronic Data Capture, Ciwit BV, Amsterdam, The Netherlands. 2019.
16. Likert R. A technique for the measurement of attitudes. *Arch Psychol* 1932.
17. R Core Team. (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. *https://www.R-project.org/*. [Accessed 25 September 2019]
18. Matthias Gamer, Jim Lemon, Ian Fellows and Puspendra Singh. irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1. https://CRAN.R-project.org/package=irr. 2019. [Accessed 25 September 2019]
19. Alasdair Rathbone SS, Kumbhare D. ICC.Sample.Size: Calculation of Sample Size and Power for ICC. R package version 1.0. *https://CRAN.R-project.org/package=ICC.Sample.Size*. 2015. [Accessed 25 September 2019]
20. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;15:155–63.
21. Anthoine E, Moret L, Regnault A, et al. Sample size used to validate a scale: a review of publications on newly-developed patient reported outcomes measures. *Health Qual Life Outcomes* 2014;12:176.
22. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
23. Cicchetti DV, Sparrow SA. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *Am J Ment Defic* 1981;86:127–37.
24. Fleiss JL, Levin B, Paik MC. Statistical Methods for Rates and Proportions. New York: John Wiley & Sons; 2013.
25. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychol Rep* 1966;19:3–11.
26. Degen L, Phillips S. How well does stool form reflect colonic transit? *Gut* 1996;39:109–13.
27. Dewaele JM. Online Questionnaires. In: Phakiti A., De Costa P., Plonsky L., Starfield S. (eds) The Palgrave Handbook of Applied Linguistics Research Methodology. Palgrave Macmillan, London. 2018. https://doi.org/10.1057/978-1-137-59900-1_13.