Disability and the Global South, 2025
Vol.12, No. 1, 2532-2550

OPEN ACCESS
ISSN 2050-7364
www.dgsjournal.org

Reliability and Feasibility of the Child Functioning Module– Teacher Version: Findings from a pilot test conducted in Somalia

Henok Tesfay Zeratsion[a*] and Marius Rohdin Karlsen[b]

[a,b] *Save the Children Norway. Corresponding Author- Email: Henok.zeratsion@reddbarna.no*

The Child Functioning Module– Teacher Version was developed by the Washington Group and UNICEF as a tool for collection of disability data about children aged 5 to 17 years in a school setting. Teachers serve as proxy respondents to questions about functional difficulties their students may have. This paper studied reliability and administrative feasibility of the tool. Two teachers independently rated the functional difficulty levels of each of their 328 (39% girls) primary school learners in Garowe district, Somalia. The study found that the percentage agreement between the ratings given to the learners' functional difficulties by the two teachers was more than 75% in 9 of the 12 functional domains. Logistic regression analysis found that the rating disagreement was likely to significantly decrease across all 12 domains when the rating was done in non-IDP schools compared the rating done in IDP schools (odds ratios: 0.02-0.25). Learners' age and educational level were significant predictors for rating disagreement only in a few domains. Both teacher raters agreed in 5% of the learners as having disability in the same functional domain. Cohen's Kappa analysis found a fair agreement (Kappa value 0.21-0.40) in most of the domains. The findings suggest acceptable reliability and feasibility of the tool.

Keywords: disability; functional difficulty; child functioning module – teacher version

Introduction

Article 24 of the United Nations Convention of the Rights of Persons with Disabilities aims not only at preventing the exclusion of persons with disabilities from education, but it also promotes the provision of reasonable accommodation and required support for persons with disabilities (UN, 2006). In line with this, Goal of the Sustainable Development Goals, focuses on enabling education systems across member countries to provide equal access to education to vulnerable persons, including children with disabilities, by 2030 (UN, 2015). However, there has been a scarcity of internationally comparable and reliable disability data that enable an analysis of the link between disability and various development goals (Loeb et al., 2008; Madans et al., 2017). Addressing this data scarcity was identified as one of the priorities in the United Nations 2030 Agenda for Sustainable Development where the aim was stated as enhancing capacity-building support for developing countries to generate high-quality, timely and reliable data on disability by 2020 (UN, 2015). In response to the global need for enhanced

capacity of disability data collection, the Washington Group on Disability Statistics (Washington Group) and United Nations Children's Fund (UNICEF) have developed standardized tools for collection of comparable disability data from populations living in a variety of cultural and economic contexts (Madans et al., 2017). These tools, known as the Washington Group questions (Washington Group, n.d) and the Child Functioning Module question sets (UNICEF, 2022), are intended to be used for collection of data from a community setting. Several previous studies have recommended these sets of questions for disability inclusive monitoring of progress on the Sustainable Development Goals and national level surveys (Schneider et al., 2009; Madans et al., 2017; Abualghaib et al., 2019; Fotso et al., 2019; Mitra et al., 2022).

While the strengths of these question sets as a tool for collection of comparable disability data from a community setting were well documented in a previous study (see Mactaggart et al., 2016), a major gap remains the lack of a verified set of questions that can be used to collect disability data about learners in a school setting, where caregivers of learners are normally not available to serve as proxy respondents to questions about functional difficulty that their children may have. The Child Functioning Module–Teacher Version (CFM-TV) was recently developed by the Washington Group and UNICEF to address this gap. Teachers are required to serve as proxy respondents to the questions in the CFM-TV questionnaire for identification of the type and level of functional difficulty of their learners.

The CFM-TV is believed to be relevant for addressing issues of non-comparable student disability data that is evident in Somalia and other low- and middle-income countries. In Somalia's context, the proportion of children with disabilities enrolled in school remains unclear due to several reasons including the lack of a standardized disability data collection tool that can provide reliable data to the education management information system; lack of clarity about the definition of children with disability; and negative teacher and community attitudes towards children with disability (UNESCO, 2022). The issue of significantly different disability prevalence reported in surveys using different disability measurement tools, was also found in a study from Cameroon (Fotso et al., 2019). As teachers are considered the foundation for inclusive education, the need to provide them with relevant skills and supportive attitudes is critical for effective disability inclusive education services (Edusei et al., 2015). Due to cultural and religious beliefs, disability is considered a disadvantage and curse in many countries, hence exposing children with disabilities to violence and abuse (Nyangweso, 2021), isolation and even hiding as in the case of Somalia (UNESCO, 2022). The disadvantages that persons with disabilities face is largely the result of the social context, when this society is responsible for accommodating the barriers that persons with disabilities face (Aas, 2020).

The United Nations Convention on the Rights of Persons with Disabilities takes into account the extent to which the social context accommodates the barriers to participation in its definition of persons with disabilities. According to this Convention, '[p]ersons with

disabilities include those who have long-term physical, mental, intellectual or sensory impairments which in interaction with various barriers may hinder their full and effective participation in society on an equal basis with others' (UN, 2006: 3). Based on this definition of disability, the Washington Group and Child Functioning Module set of questions have the qualities of a standardized disability measure that can provide reliable and comparable disability data on different levels and type of functional difficulties as the questions are designed to be non-stigmatizing, that can be interpreted in the same way across socio-cultural contexts (Madans et al., 2017). As it uses questions selected from these question sets, the CFM-TV is also expected to fulfill these qualities. However, as stated in an informational meeting of the Washington Group (2021), evidence on reliability and implementation of CFM-TV has been limited. In order to address this lack of evidence, a pilot test of the CFM-TV has recently been done by some organizations that documented their findings in different study reports (Brus et al., 2019; School-to-School International, 2023; de Kadt, 2023). The current study was conducted based on the pilot test of the CFM-TV that was done by Save the Children International country office in Somalia. Data was collected in 2022 from a sample of schools that were supported with a five-year (2019-2023) Save the Children program funded by the Norwegian Agency for Development Cooperation (Norad).

The objectives of the current study are to assess the reliability of the CFM-TV using interrater reliability analysis; assess feasibility of administering CFM-TV by teachers; and document lessons learned from the pilot test. In light of these objectives, the current study addresses the reliability of the CFM-TV as a tool for collection of comparable disability data from a school setting. In this effort, it starts by comparing the disability prevalence reported by the two different teacher raters who independently rated each student in the sample for their functional difficulties using the CFM-TV questionnaire. The functional domains included in the CFM-TV questionnaire that was used for the pilot study were seeing, hearing, mobility, communication, learning, remembering, concentrating/attention, coping with change, controlling behavior, relationships (making friends), and affect (anxiety and depression). The definition of the functional domains is available in the Module on Child Functioning – Manual for Interviewers (UNICEF, 2018). Data on functional difficulty in each of these domains is expected to provide a reliable estimate of disability prevalence among 5-17-year-old children. Based on a survey from year 2011, the World Health Organization documented that 5% of children under 15 years of age live with a moderate or severe disability (WHO, 2015), although a more recent estimate of disability prevalence of 10% among 0-17 year-old children was reported based on global survey done in 2021 (UNICEF, 2021). After providing disability prevalence disaggregated by relevant student and school characteristics, the article presents the results from the analysis of agreement between the ratings of the teacher raters, by functional domain. This is followed by analysis of whether student and school characteristics were predictors for the observed rating disagreements. Finally, the statistical results and findings are discussed in light of relevant literature and feedback from teacher raters to provide conclusions and relevant recommendations for effective implementation of the CFM-TV.

Methods

A CFM-TV set of questions was used to collect data on learners' health-related difficulties in 12 functional domains. First, a team of staff members of Save the Children International country office in Somalia was trained by Save the Children Norway on Administering the CFM-TV. The country office staff had already experience administering Child Functioning Module set of questions, and Washington Group Short Set of questions that were designed for collection of disability data from a community setting. Then, using the training materials and pilot test guidelines developed by Save the Children, the country office team provided a two-days training for teachers on Administering the CFM-TV. The teachers who attended the training were selected by head teachers. At the end of the training, only teachers who were voluntarily willing to participate in the pilot test were assigned to a pilot-test class. Two teachers, a teacher who had responsibility for the class and another teacher who taught the class at least three days per week, were assigned to a class to independently do the rating of functional difficulty levels on each learner in the class. They did not receive any financial or in-kind incentives for their participation. Three months after the school year began in September, each teacher started to use their observation of and interaction with the learners to understand the functional difficulties of each learner in the class. Each of the teachers had the observation log integrated into their attendance sheet and observation notes on learners' functional difficulties were continuously entered in the observation log. After the observation and interaction period of two months was ended, each of the two teachers used two weeks to independently rate the functional difficulty level of each learner by completing the CFM-TV questionnaire for each learner. The CFM-TV was translated from English to Somali language before the training was provided to the teachers. The Module on Child Functioning – Manual for Interviewers (UNICEF, 2018) and other relevant technical guidance documents were used by country office staff to provide technical support to the teachers whenever needed.

Study design and sampling

The study used both quantitative and qualitative data. The quantitative disability data was collected from a total of 328 (61% boys, 39% girls) primary school learners who were sampled from four schools that were supported by the Norad funded Save the Children program 2019-2023. The four schools were located in the Garowe District, Nugal Region. To ensure a good representation of various age groups in the primary education level, it was decided to include learners from grade 3, 4, 5 and 7 into the sample. The average age of the sample of students was 12 years.

In the interest of ensuring efficiency in conducting the pilot test, a convenience sampling technique was used to select the four schools for the pilot test. As the four schools were located in Garowe district, they were in close proximity to the Save the Children International country office that was also located in Garowe, making it easier to provide training, technical support

and continuous monitoring of the pilot test processes. There was normally only one class of students for each grade in the schools. The availability of a teacher who taught in a given class, and who voluntarily decided to participate in the pilot test, was one of the factors considered for deciding which class of students to include in the pilot-test.

Table 1: Description of the study sample and percent of learners rated as children with disability by each of the two teacher raters, disaggregated by student and school characteristics.

Student and school characteristics	**Disaggregation levels**	**Percentage of total sample (N=328)**	**Number and % of students rated by a teacher as children with disability**			
			Teacher 1 rater		**Teacher 2 rater**	
			#	**%**	**#**	**%**
Gender	All learners	100%	67	20%	73	22%
	Male	61%	46	23%	46	23%
	Female	39%	21	16%	27	21%
Age (years)	Childhood (7-10)	22%	6	8%	14	19%
	Early adolescence (11-14)	64%	51	24%	47	23%
	Middle adolescence (15-17)	14%	10	22%	12	27%
Grade	3	39%	21	16%	22	17%
	4	9%	8	27%	13	43%
	5	30%	20	21%	13	13%
	7	22%	18	25%	25	35%
School type	School for IDPs	65%	62	29%	60	28%
	School for non-IDPs	35%	5	4%	13	11%

Note: *IDPs: internally displaced persons*

Teachers who participated in rating of functional difficulties and focus group discussion

In total, 11 classes of learners were selected from the four schools. Average class size was 30 learners. Two different teachers who teach in the same class were assigned to independently rate the level of functional difficulty of learners in the class; thus, a total of 22 teachers (36% female) participated in the pilot test. The teachers taught 3-6 days a week (4.4 days on average) in the assigned class and used two months for observation and interaction with the learners in order to understand the level of functional difficulty each child in the class might have had in each of the 12 functional domains. The teachers were instructed to do the rating of a particular learner's functional difficulty independently. In addition, they were instructed not to discuss a

child's functional difficulty status with each other and with the concerned child. The two different teachers assigned to a class are randomly labelled as Teacher 1 and Teacher 2 in the study.

Focus group discussion

Eight (50% female) of the 22 teachers who administered the CFM-TV during the pilot test, participated in a focus group discussion (FGD) immediately after the teachers completed data collection using the CFM-TV questionnaire. The FGD participants were represented from all four pilot test schools based on their availability and willingness to participate. The purpose of the FGD was to document the teacher raters' challenges, innovative solutions and lessons learned from administering the CFM-TV. All of the FGD participants were physically present in the discussion venue in Garowe, and a monitoring and evaluation staff from Save the Children International Somalia country office facilitated the discussion using the Somali version of the eight FGD questions that were originally written in English by the researchers. The discussion facilitator was supported by a note taker from the country office. The data was analysed to identify viewpoints that were repeated several times or that gained consensus among discussion participants. Attention was also paid to statements that contradicted the viewpoint of the majority in the group.

Data quality

The training on administering the CFM-TV questionnaire that was provided to teacher raters, and the translation of the questionnaire into Somali language were a few of the activities carried out to improve the quality of collected data. In addition to this, consistency between disability prevalence obtained from the current study and prevalence documented in previous reports was considered as indicator of data quality. The descriptive data analysis results shown in Table 1 above indicate that the disability prevalence increased with an increase in age. Findings about disability prevalence that increases with age was also reported in a previous study (Fotso et al., 2019). It is also worthwhile noting that disability prevalence (28-29%) among learners from internally displaced communities (IDPs) was higher compared to the prevalence (4-11%) among their peers who did not experience displacement due to a humanitarian situation. The European Civil Protection and Humanitarian Aid Operations (2023) have also stated that disability prevalence increases in communities that are in a humanitarian crisis.

Data analysis

Ten of the 12 health-related functional difficulty questions in the CFM-TV questionnaire had four response/rating options: "no difficulty", "some difficulty", "a lot of difficulty", and "cannot do at all". A Learner rated as having "a lot of difficulty", or "cannot do at all" in a specific domain was identified as having "disability" in that domain. For questions on anxiety

and depression, five response/rating options were provided regarding how often the learner exhibited the difficulty. The response options were "daily", "weekly", "monthly", "a few times a year", and "never". The rating "daily" was the cut-off point for defining disability in terms of anxiety or depression. Using these cut-off points that are recommended by the Washington Group and UNICEF for defining "disability" (Washington Group, 2020; UNICEF, 2017), each of the 328 learners was finally identified as a child with disability or a child without disability in each of the functional domains based on the rating done independently by Teacher 1 and Teacher 2.

An interrater reliability analysis was conducted to assess the level of agreement between the ratings that the two different teacher raters gave to the functional difficulty levels they observed in their learners. Analysis of percentage agreement (percentage disagreement) between the ratings of Teacher 1 and Teacher 2 was done based on data collected on all 328 learners. For this purpose, the data on functional difficulty ratings were categorized into three levels: "no difficulty", "some difficulty", and "a lot of difficulty". For the four response questions in the CFM-TV questionnaire, "a lot of difficulty" and "cannot do at all" responses were categorized into "a lot of difficulty". For questions on anxiety and depression that had five response options, the responses "weekly", "monthly", and "a few times a year" were categorized as "some difficulty", while "daily" and "never" were categorized as "a lot of difficulty" and "no difficulty", respectively. A binary logistic regression model was fitted per functional domain to investigate which student, teacher and school characteristics were related with the observed disagreement between the ratings given by the two teachers. Cohen's Kappa analysis was also done to investigate further the level of agreement between the teacher raters when their ratings were categorized into "no", "some" or "a lot" of difficulties. SPSS version 25 was used to do data analysis.

Ethics

Adhering to common practices of doing research in Somalia, the study received ethical approval from the Ministry of Education, Garowe District Office, through request for ethical approval submitted by the Save the Children International country office in Somalia in good time before the pilot test started. Adequate provisions were done to anonymize both learners and teachers who participated in the research. Informed consent for data collection was received from school head teachers in all pilot test schools. The authors have taken necessary actions to adhere to the General Data Protection Regulations of the European Union to protect personal and sensitive data of research participants.

Results and findings

When the students in the sample were categorized into those who had "no", "some", or "a lot"

of difficulties based on the two teachers' independent ratings of functional difficulties per domain, the teachers' percentage agreement was 75% or more in 9 of the 12 domains (see Table 2). The percentage agreement values indicate that about 69%-88% of the functional difficulty data collected by teachers using the CFM-TV tool were correct data. Cohen's Kappa analysis was done to assess agreement between the teacher raters controlling for agreement by chance. There was slight agreement (Kappa value 0.0-0.20) in four domains, fair agreement (Kappa value 0.21-0.40) in seven domains, and moderate agreement (Kappa value 0.41-0.60) in one domain.

Table 2: Percent agreement (Kappa values), and percent disagreement between the two teachers' rating of learners' functional difficulty as "no", "some", or "a lot" of difficulty, per domain (N=328)

Functional domain	**% Agreement (*Kappa values*)**	**% Disagreement**			
		Total	**"No" vs "Some"**	**"Some" vs "A lot"**	**"No" vs "A lot"**
Seeing	83.5% (*0.21*)	16.5%	9.5%	0.9%	6.1%
Hearing	88.4% (*0.44*)	11.6%	6.4%	1.2%	4.0%
Walking	88.1% (*0.22*)	11.9%	5.5%	1.8%	4.6%
Communication	81.1% (*0.21*)	18.9%	13.1%	1.5%	4.3%
Learning	69.2% (*0.14*)	30.8%	21.0%	3.4%	6.4%
Remembering	75.0% (*0.23*)	25.0%	19.0%	1.0%	5.0%
Concentrating	78.1% (*0.16*)	21.9%	13.4%	1.5%	7.0%
Accepting change	76.8% (*0.28*)	23.2%	13.7%	4.3%	5.2%
Controlling behaviour	83.8% (*0.11*)	16.2%	11.9%	0.6%	3.7%
Making friends	78.7% (*0.14*)	21.3%	14.9%	1.8%	4.6%
Anxiety	71.9% (*0.36*)	28.1%	16.8%	5.8%	5.5%
Depression	73.5% (*0.33*)	26.5%	18.0%	2.7%	5.8%

The disability prevalence independently reported by Teacher 1 and Teacher 2 was 20% and 22%, respectively (Table 1). Additional analysis has shown that both teacher raters agreed that 5% (n=15) of the learners (N=328) had disability in the same functional domain.

In addition to the analysis of percentage agreement, it is worthwhile investigating the reliability of the CFM-TV in more depth by analysing the disagreement between the teachers' ratings. A *disagreement* is defined to mean that one of the two teachers rated a functional difficulty observed in a specific learner as "no" difficulty, while the other teacher rated it as "some" difficulty or "a lot" of difficulty. Likewise, it is a disagreement when one of the teachers gave a rating of "some" difficulty while the other teacher gave it a rating of "a lot" of difficulty. The way the functional difficulty ratings initially provided by teachers are re-categorized into only

three categories of ratings – "no difficulty", "some difficulty", and "a lot of difficulty" – is explained in the section on *data analysis* above.

The total percentage disagreement per domain was in the range 11.6%-30.8%. The total percentage disagreement is the sum of the percentage disagreements between two types of ratings that are presented in the last three columns of Table 2. With the exception of the learning, anxiety and depression domains, the total percentage disagreement is not higher than 25% per domain. There was lower total percentage disagreement (11.6%-16.5%) in functional domains such as walking, hearing and seeing where the functional difficulty is easier to observe than internalizing problems such as depression and anxiety. Difficulties of controlling behaviour can also be easily observable to a larger extent when exhibited as externalising behaviours; thus, among domains with lower percentage disagreement. The highest percentage disagreement was observed when one of the teachers rated a child's functioning as "no difficulty" while the other teacher rated it as "some difficulty", across all the domains.

Table 3: Odds ratio from a binary logistic regression on predictors of disagreement between the ratings of the two different teacher raters by functional domain (confidence interval), N=328.

Functional domain	**Learner's age**	**Learner's sex (female)**	**Learner's educ. level**	**Teacher raters' sex (same sex)**	**School type (none-IDP school)**
Seeing	0.91 (0.75-1.10)	1.10 (0.58-2.08)	1.36 (1.05-1.75)*	0.45 (0.23-0.88)*	0.07 (0.03-0.22)***
Hearing	0.75 (0.59-0.94)*	0.90 (0.42-1.91)	1.74 (1.28-2.37)***	0.54 (0.25-1.18)	0.02 (0.01-0.15)***
Walking	0.80 (0.64-1.02)	0.55 (0.25-1.19)	1.66 (1.22-2.25)**	0.44 (0.20-0.94)*	0.06 (0.02-0.23)***
Communication	0.92 (0.78-1.09)	1.22 (0.67-2.23)	1.14 (0.90-1.45)	1.12 (0.58-2.17)	0.05 (0.01-0.16)***
Learning	0.96 (0.83-1.11)	0.80 (0.47-1.39)	1.09 (0.89-1.34)	1.57 (0.86-2.86)	0.25 (0.13-0.45)***
Remembering	0.98 (0.84-1.14)	1.34 (0.77-2.34)	1.04 (0.84-1.29)	1.44 (0.78-2.66)	0.24 (0.12-0.46)***
Concentrating	0.88	1.01	1.28	1.37	0.12

	(0.75-1.04)	(0.56-1.82)	(1.02-1.61)*	(0.72-2.62)	(0.05-0.27)***
Accepting change	0.83 (0.70-0.98)*	0.67 (0.37-1.21)	1.36 (1.07-1.72)*	1.09 (0.58-2.07)	0.06 (0.02-0.14)***
Control. behavior	0.88 (0.73-1.05)	1.10 (0.58-2.08)	1.07 (0.82-1.38)	0.76 (0.39-1.50)	0.16 (0.07-0.38)***
Making friends	0.91 (0.77-1.07)	1.20 (0.67-2.17)	1.21 (0.96-1.52)	1.37 (0.71-2.63)	0.09 (0.03-0.21)***
Anxiety	1.08 (0.88-1.19)	1.09 (0.63-1.89)	1.12 (0.91-1.39	0.47 (0.26-0.85)*	0.15 (0.08-0.31)***
Depression	1.05 (0.91-1.22)	1.18 (0.68-2.05)	0.78 (0.64-1.00)	0.64 (0.34-1.16)	0.23 (0.14-0.45)***

Note: statistical significance level at p-value < 0.05, p-value < 0,01**, p-value < 0.001***, NS: not significant*

A binary logistic regression model was fitted per functional domain to examine whether or not learner characteristics, teacher characteristics or school characteristics were associated with the disagreements observed between the ratings given by the two teacher raters. The categorical dependent variable was "agreement status" between teacher raters, where 'agreed ratings' was coded 0, and 'disagreed ratings' was coded 1. The predictor variables used in the model were learner's age, sex (ref.=male), and education level; teacher raters' sameness of sex (ref.=different sex); and school type (ref.=IDP school).

School type was the only predicator variable that was significantly associated with the rating disagreements of the teacher raters in each of the 12 functional domains (odds ratio: 0.02-0.25; p-value < 0.001). Taking the inverted value of the odds ratio 0.25, the results indicate that teachers who did the rating in IDP schools were at least 4 times more likely to exhibit disagreement in their ratings compared to teachers who did the rating on learners enrolled in non-IDP schools (confidence interval: 2.2-7.7). Learners' education level was also significantly associated with the rating disagreement in seeing, hearing, walking, concentrating and accepting change domains (odds ratios: 1.28-1.74). This indicates that for ratings done among learners in every higher education level, the teacher raters were about 1.28 to 1.74 times more likely to disagree in their ratings of functional difficulties in these five domains. Learner's age was found to be a significant predictor of rating disagreements only in hearing (odds ratio: 0.75) and accepting change (odds ratio: 0.83) domains. Inverse relationship was found between

age of learners and rating disagreement between teacher raters. The sameness of sex of teacher raters was significant predictor only in seeing domain (odds ratio: 0.44), walking domain (odds ratio: 0.45) and anxiety domain (odds ratio: 0.47) indicating that the disagreement between the teacher raters was likely to decrease by about 53-56% when the sex of both teacher raters was the same rather than when the two raters had different sex.

Discussion

The results and findings presented in the section above raise interesting discussion points. Agreement by two independent teacher raters on whether the observed child had "no", "some" or "a lot" of functional difficulty in a given functional domain was used to analyse percentage agreement between the raters. "A lot" of difficulty category refers to the ratings initially given by the teacher raters as "a lot" or "cannot do at all" level of difficulties; thus, children with disability. The current study found a disability prevalence of 5% based on the number of students in the sample about whom the two teachers, who did the rating independently, agreed on existence of the same type of disability in a student. This disability prevalence of 5% is very similar to the global average of disability prevalence that was previously documented by World Health Organisation (WHO, 2015), although a more recent estimate of disability prevalence of 10% among 0-17 years-old children was reported based on a global survey that was done in 2021 (UNICEF, 2021). Even when the Washington Group questions were used in a correct manner, countries tended to report prevalence rates usually ranging from 6% to 12% among people of all age groups (Mont, 2019). This indicates that the 5% prevalence found in the 7-15 year-old learners in the current study is not only a reasonably reliable prevalence estimate, but also a higher and more realistic estimate compared to prevalence rates reported in previous studies that used Child Functioning Module for disability data collection from a community setting in Mexico, Samoa, and Serbia (Cappa et al., 2018). It is also worth noting that the CFM-TV identified higher proportion of children with disabilities compared to the disability prevalence of 3.4% - 3.9% that was estimated using binary questions in the Somalia Health and Demographic Survey conducted in year 2020 (Federal Government of Somalia, 2020).

The disability prevalence among children enrolled in school has been unclear in Somalia (UNESCO, 2022) due to lack of standardized disability measurement instruments. Previous studies from Zambia (Loeb et al., 2008) and Cameroon (Fotso et al., 2019) have noted the effect of different disability measurement instruments on disability prevalence estimation. The difference in the words used in survey questions to identify persons with disabilities does not produce comparable disability data (Schneider et al., 2009; Fotso et al., 2019). According to Madans et al. (2017), questions similar to 'Do you have a disability?' make respondents understand the concept of disability differently based on their general life experiences, cultural contexts and association between disability and stigma. Asking someone directly whether or not they have a disability can be very stigmatizing in cultural and religious contexts where disability is considered as a disadvantage or a curse (Nyangweso, 2021). As in the case of

Somalia where children with disabilities are exposed to isolation and hiding from others (UNESCO, 2022), the parents are less likely to report the correct disability status of their children if they find the question stigmatizing. The questions, answer options, and words in the Child Functioning Module and Washington Group question sets are formulated in a way that is non-stigmatizing and that are understood as having the same meaning in different social, cultural and economic contexts (Loeb et al., 2008; Madans et al., 2017). For this reason, the less stigmatizing and easy to understand questions in the CFM-TV help to identify the majority of children with disability who have been hidden and invisible in education management information systems partly due to negative attitudes and cultural beliefs. Thus, children with disability can be as much visible, counted and included in social and political decisions as their peers without disability (UNICEF, 2021).

With regards to the interrater reliability of the CFM-TV, the percentage disagreement per functional domain was only 25% or less in 9 of the 12 domains (Table 2). The 25% disagreement indicates that the two different teacher raters agreed in at least 75% of their ratings in 9 of the 12 domains. Literature on interrater analysis, states that the percentage values of absolute agreement in the range of 75%-90% demonstrate acceptable level of agreement (Graham, 2012); thus, an important indicator of a good reliability of the CFM-TV. This high percentage agreement in most of the functional domains can be considered a reliable measure of interrater reliability because the teacher raters had received needed training on administering the CFM-TV and they used two months to observe the health-related functional difficulties their learners might have had. If raters are well trained, little guessing is likely to exist, and the level of percentage agreement can be better relied on to determine interrater reliability (McHugh, 2012). The percentage agreement level reported in the current study shows that multiple teacher raters are likely to consistently rate learners' functional difficulties in similar ways in about 75% of the times.

Further investigation of interrater reliability was done using Cohen's Kappa analysis. There was a fair level of agreement (Kappa value 0.21-0.40) in most of the domains, while there was a slight agreement (Kappa value 0.0-0.20) in four domains and a moderate agreement (Kappa value 0.41-0.60) in one domain (Table 2). It may seem that there is discrepancy between the high level of percentage agreement discussed above and the Cohen's kappa results. Clarification for such seemingly discrepant findings were provided in previous studies (Viera and Garrett, 2005; McHugh, 2012). According to these studies, when the expected prevalence of the outcome of interest is low, small kappa values may not necessarily reflect low agreement. In the current study, Cohen's kappa analysis was done per functional domain where normally a low disability prevalence is expected in a specific domain. In addition to this, a contextual reality characterised by a large number of out-of-school children with disabilities in Somalia (Federal Government of Somalia, 2020) led to identification of only a substantially smaller proportion of children with disabilities in the school setting, compared to the actual proportion of school aged children with disabilities who live in the communities. Due to these reasons, the

low Kappa values may not necessarily indicate low agreement between the teacher raters. In the current study, more weight can be given to the interrater agreement provided in terms of percentage agreement.

Taking note of the total percentage disagreement per domain presented in Table 2, a logistic regression model was fitted per functional domain to identify the predictors of the observed disagreement (Table 3). The regression results showed that *school type* was a statistically significant predictor of rating disagreement in each of the 12 functional domains. Two teachers who independently rated learners that were enrolled in non-IDP schools were significantly less likely to disagree in their ratings compared to two teacher raters who independently did the ratings on learners enrolled in IDP schools. This implies that it is more challenging for teachers to accurately determine the type and/or the level of functional difficulties of learners in a humanitarian context compared to learners in a non-humanitarian situation. A more rigorous study is needed to investigate the humanitarian factors that make rating of functional difficulties challenging. Furthermore, in five of the functional domains, there was a higher likelihood of rating disagreement for each additional educational level of learners. On the contrary, a lower likelihood of rating disagreement for each additional year in the age of learners was found in the hearing domain and accepting change domain. This implies that the older the children the easier for the teacher proxy respondents to more accurately rate the functional difficulty of the learners. A lower likelihood of rating disagreement was also found when two teacher raters that had the same sex, did the rating as compared to when the two teacher raters had different sex. In countries such as Somalia where the population strictly adhere to gender-sensitive interaction norms between individuals of different sex might have provided similar level of interaction for the two teacher raters of the same sex with learners that had the same sex as the raters. A few of the male teachers who participated in the focus group discussion have stated that it was not easy for them to get closer to a female student to engage in more interaction with her with an effort to understand the type or level of a suspected functional difficulty. Whether or not gender roles influence rating accuracy in societies that have conservative social and religious norms is a topic that needs more research. To sum up, the findings from the logistic regression imply that the functionality and reliability of the CFM-TV is not influenced largely by the characteristics of the learners and the teacher raters.

A previous study has documented that rater training on understanding and administering a data collection tool is one of the most important factors that improve agreement between raters (McHugh, 2012). A correctly designed training can improve common understanding of terms and address such issues as personal beliefs, bias, and difference in interpreting the same observation. The training needs to include sufficient time for discussion on the question-by-question specifications that are available in UNICEF's Module on Child Functioning – Manual for Interviewers (UNICEF, 2018). The focus group discussion of teachers has revealed that the questions about depression and anxiety were somehow challenging to understand. This could be the reason why anxiety and depression domains were among those with highest percentage

disagreement (Table 2). The two months allocated for observation were meant to help the teacher raters have more or less the same level of familiarity as the caregivers had about functional difficulties of their children. However, some teachers stated that it was challenging to determine the specific type and/or level of functional difficulty with confidence as the two months were not sufficient for effective observation. For example, there was uncertainty regarding whether to record an observed functional difficulty as a concentration difficulty or as a hearing difficulty. The fact that the highest percentage disagreement per domain was between "no difficulty" and "some difficulty" (Table 2) indicates that most of the rating disagreement was related to the challenge the teachers had regarding whether to give the rate "no difficulty" or "some difficulty". Such a challenge was also documented in a previous study which found that severe impairments were reported relatively evenly across the functional difficulty levels "some difficulty", "a lot of difficulty" and "cannot do at all" and most of the moderate impairments were reported as "some difficulty" (Sprunt et al., 2019). Additional data analysis has shown that the two teacher raters agreed in 9% of the learners (N=328) as having disability regardless of agreement in the type of disability they identified in a specific learner. This 9% compared to the 5% of the learners that were identified by both teacher raters as having the same type of disability, highlights that the raters had some level of challenge to correctly classify the type of disability they observed. While training of teacher raters is key to improved technical skills needed for collection of reliable disability disaggregated data, the importance of the training needs to be understood from the broader objectives of disability inclusive education. As stated by Edusei et al. (2015), teachers are the foundation for inclusive education and need to be continuously supported to develop relevant technical competency and attitude to deliver disability inclusive education service. Enabling teachers to understand the concept of disability from a social model (Aas, 2020), can help them to identify barriers to disability inclusive education, and to take a leader role in accommodating the health-related functional difficulties of their learners both at school level and at community level.

It is important to understand the findings and related discussions in light of the strengths and limitations of the CFM-TV questionnaire and implementation of the pilot test. The non-stigmatizing and mostly easy-to-understand questions of CFM-TV was mentioned by the teachers as the most important strength of the tool. This enabled the teachers to use the tool without major challenges after receiving a few days training tailored for the purpose of the pilot test. A translation of the CFM-TV set of questions into local language following the translation guidelines from the Washington Group, a voluntary and committed participation of teachers in the pilot test, and availability of technical support to teacher raters by Save the Children International country office in Somalia were very important for the success of the pilot study. The unintended benefits reported by teachers improved awareness about disability inclusive education, understanding of the concept of disability from a human rights perspective, and ability to practically identify learners with health-related functional difficulty.

Limitations of the study

There was shortage of time and capacity to translate the UNICEF's Module on Child Functioning – Manual for Interviewers into Somali, which could have provided sufficient clarification to teacher raters about the terms used in the CFM-TV questionnaire. The lack of practical training activities regarding how to more accurately identify the type or level of the observed functional difficulty was noted as an area that needed improvement. Some teachers had to use the time previously allocated for teaching activities in order to administer the CFM-TV; thus, less time left for teaching activities. On average, 10 minutes were used to complete a CFM-TV questionnaire per learner after the observation period of two months ended. The focus group discussion data also had some limitations. The researchers were not able to travel to Somalia to facilitate and observe the focus group discussion. Due to this, data about group dynamics and interaction, and how group members influenced or changed each other's viewpoints was not captured well.

Conclusion and recommendations

Conclusion

The disability prevalence of 5% found in the study based on agreement by both teacher raters on existence of disability in same functional domain, and the percentage agreement greater than 75% in most of the functional domains, indicate good reliability of CFM-TV as a tool for collection of disability disaggregated data from a school setting. The tool enables one to correctly identify more children with disability compared to the use of binary questions that ask whether or not a child has disability. The use of the Child Functioning Module – Teacher Version is an important step towards filling the gap that the education information system in many countries has due to lack of comparable and reliable disability data about students enrolled in schools. Data collected using the CFM-TV tool is expected to substantially contribute to disability inclusive learning and child development objectives.

Recommendations

In order to collect comparable and reliable disability data, sufficient teacher training is critical. Training topics include disability inclusive education, the concept of disability, administering the CFM-TV set of questions, understanding of how to do observation of health-related functional difficulties, and a practical exercise on rating the type and level of functional difficulties.

Teachers who are selected to administer the CFM-TV in a class should have good familiarity with the class. Teachers who have the role of a "classroom contact teacher" can be prioritized over subject teachers, provided that they also teach in the class for at least three days a week.

Subject teachers who teach the class for at least three days a week, can also be selected to administer the CFM-TV tool in the class if the class contact teacher actually teaches less than three days per week in the class.

A minimum of three months is needed for observation and interaction, followed by two weeks for completing the CFM-TV questionnaire. The CFM-TV questionnaire, the UNICEF Child Functioning Module - Manual for Interviewers, and any relevant technical guidance document on administering the CFM-TV should be translated adhering to the translation protocols from the Washington Group. Continuous advocacy work and joint-planning needs to be done with the ministry of education and its sub-national offices to officially integrate the CFM-TV into the education management information system.

Further study is needed to rigorously investigate factors that might influence the reliability of the CFM-TV as a tool for disability data collection from school settings characterized by a humanitarian or emergency situation.

Acknowledgements

We would like to thank Save the Children Somalia country office in general, and Yasin Hersi Hassan, MEAL Manager in the country office in particular, for making the pilot test of the CFM-TV possible by training teacher raters on administering the CFM-TV, leading the translation of the questionnaire and related documents into Somali, and for sharing good quality data for the study. We are also thankful to Nikolai Holm (PhD), Director of Evidence and Learning, and Christina W. Fægri (PhD), Senior Advisor MEAL, Save the Children Norway, for editing the article and for providing input that helped to improve the study report. We thank the Norwegian Agency for Development Cooperation for supporting the five-year development program that enabled the collection of disability disaggregated data that could be used for this study.

Funding information

There was no financial resource that was specifically allocated to conduct this study. However, the salary of the authors and salary of the Save the Children International Somalia country office staff who worked on the pilot test was paid from the grant received from the Norwegian Agency for Development Cooperation (Norad). The grant's name was Leaving No Child Behind: Norad Framework Agreement 2019-2023, grant number QZA-18/0373.

References

Aas, S. (2020). Disability, Society, and Personal Transformation. *Journal of Moral Philosophy*, 18(1), 49-74.

Abualghaib, O., Groce, N. et al. (2019). Making Visible the Invisible: Why Disability-Disaggregated Data is Vital to "Leave No-One Behind". *Sustainability*, 11(11), 3091.

Brus, A., Deleu, M. et al. (2019). Testing a Teacher Version of the UNICEF/Washington Group Child Functioning Module (CFM-TV) in Senegal: A Humanity and Inclusion Publication RS/FP/25. Available at: https://www.washingtongroup-disability.com/fileadmin/uploads/wg/TestingTeacherVersionWG_CFM_Senegal_FP_RS_25b.pdf

Cappa, C., Mont, D. et al. (2018). The development and testing of a module on child functioning for identifying children with disabilities on surveys. III: Field testing. *Disability and Health Journal*, 11(4), 510–518.

de Kadt, J., Kaindaneh, S. et al. (2023). Collecting Disability Data in Schools in Sierra Leone. Inclusive Education Initiative blog, 20th November. Available at: https://www.inclusive-education-initiative.org/blog/collecting-disability-data-schools-sierra-leone

Edusei, A.K., Mprah, W.K. et al. (2015). Attitude of teacher trainees towards children with disabilities in the Northern Region of Ghana. *Journal of Disability Studies*, 1(2), 55-60.

European Civil Protection and Humanitarian Aid Operations (2023). Disability Inclusion. Brussels: European Commission. Available at: https://civil-protection-humanitarian-aid.ec.europa.eu/system/files/2023-03/fst%20Disability%20Inclusion%20EN.pdf

Federal Government of Somalia (2020). A rapid assessment of the status of children with disabilities in Somalia. Somalia: Ministry of Women and Human Rights Development. Available at: https://resourcecentre.savethechildren.net/pdf/rapid-assessment-children-with-disabilites-in-somalia_report_fa_digital-1-1_1.pdf/

Federal Government of Somalia (2020). The Somali Health and Demographic Survey 2020. Somalia: Directorate of National Statistics. Somalia: The Directorate of National Statistics. Available at: https://somalia.unfpa.org/sites/default/files/pub-pdf/FINAL%20SHDS%20Report%202020_V7_0.pdf

Fotso, A.S., Duthe, G. and Odimegwu, C. (2019). A comparative analysis of disability measures in Cameroonian surveys. *Population Health Metrics*, 17(1):16. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC6896774/

Graham, M., Milanowski, A. and Miller, J. (2012). Measuring and promoting inter-rater agreement of teacher and principal performance ratings. Center for Educator Compensation Reform. Available at: https://www.researchgate.net/publication/265562724_

Loeb, M. E., Eide, A. H., Mont, D. (2008). Approaching the measurement of disability prevalence: The case of Zambia. *Alter*, 2, 32-43.

Mactaggart, I., Kuper, H. et al. (2016). Measuring Disability in Population Based Surveys: The Interrelationship Between Clinical Impairments and Reported Functional Limitations in Cameroon and India. *PLoS ONE Journal*, 11(10).

Madans, J., Loeb, M. et al. (2017). Measuring Disability and Inclusion in relation to the 2030 Agenda on Sustainable Development. *Disability and the Global South*, 4(1), 1164-1179.

McHugh, M. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3), 276–282.

Mitra, S., Chen, W. et al. (2022). Invisible or Mainstream? Disability in Surveys and Censuses in Low- and Middle-Income Countries. *Social Indicators Research*, 163, 219-249. Available at: https://link.springer.com/article/10.1007/s11205-022-02879-9

Modeer, U. and Viera, J. (2023). Rethinking Disability Inclusion for the SDGs. A UNDP blog, 19th June. Available at: https://www.undp.org/blog/re-thinking-disability-inclusion-sdgs

Mont D. (2019). Differences in reported prevalence rates: Is something wrong if I don't get 15%? A Washington Group blog, 22nd August. Available at: https://www.washingtongroup-disability.com/wg-blog/differences-in-reported-disability-prevalence-rates-is-something-wrong-if-i-dont-get-15-120/

Nyangweso, M. (2021). Disability in Africa: A Cultural/Religious Perspective. In T. Falola and N. Hamel (eds.). *Disability in Africa: Inclusion, Care, and the Ethics of Humanity* (pp. 115-136). US: Boydell and Brewer.

Schneider, M., Dasappa, P. et al. (2009). Measuring disability in censuses: The case of South Africa. *Alter*, 3, 245-265.

School-to-School International (2023). Final Study Report on the Validity of the Child Functioning Module- Teacher Version. Available at: https://pdf.usaid.gov/pdf_docs/PA021754.pdf

Sprunt, B., McPake, B. et al. (2019). The UNICEF/Washington Group Child Functioning Module–Accuracy, Inter-Rater Reliability and Cut-Off Level for Disability Disaggregation of Fiji's Education Management Information System. *International Journal of Environmental and Research and Public Health*, 16, 806. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6427525/pdf/ijerph-16-00806.pdf

UNESCO (2022). Education sector analysis: Assessing opportunities for rebuilding the country through education; Federal Government of Somalia, IIEP-UNESCO Dakar. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000380838

UNICEF (2017). Module on Child Functioning: Tabulation plans, narrative and syntaxes. Available at: https://data.unicef.org/resources/module-child-functioning-tabulation-plan-narrative/

UNICEF (2018). Module on Child Functioning: Manual for Interviewers. New York: UNICEF. Available at: https://data.unicef.org/resources/module-on-child-functioning-manual-for-interviewers/

UNICEF (2021). Seen, Counted, Included: Using data to shed light on the well-being of children with disabilities. New York: UNICEF. Available at: https://data.unicef.org/resources/children-with-disabilities-report-2021/

UNICEF (2022) Module on Child Functioning: Questionnaires. Available at: https://data.unicef.org/resources/module-child-functioning/

UN (2006). United Nations Convention on the Rights of Persons with Disabilities. New York: UN. Available at: https://www.un.org/disabilities/documents/convention/convoptprot-e.pdf

UN (2015). Transforming Our World: The 2030 Agenda for Sustainable Development. New York: UN. Available at: https://documents.un.org/doc/undoc/gen/n15/291/89/pdf/n1529189.pdf

Viera, A. J. and Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic, *Family Medicine*, 37(5), 360-363.

Washington Group (2020). Analytic Guidelines: Creating Disability Identifiers Using the Washington Group Short Set on Functioning (WG-SS) SPSS Syntax. Available at: https://www.washingtongroup-disability.com/fileadmin/uploads/wg/Documents/WG_Document__5A_-_Analytic_Guidelines_for_the_WG-SS__SPSS_.pdf

Washington Group (2021). Informational Meeting on Experiences with the Child Functioning Module-Teacher Version. 9th September. Available at: https://www.washingtongroup-disability.com/events/informational-meeting-on-experiences-with-the-child-functioning-module-teacher-version-539/

Washington Group (n.d.) Washington Group on Disability Statistics: Question sets. Available at: https://www.washingtongroup-disability.com/question-sets/

WHO (2015). Global Disability Action Plan 2014-2021: Better health for all people with disability. Geneva: WHO. Available at: https://iris.who.int/bitstream/handle/10665/199544/9789241509619_eng.pdf?sequence=1