



## Generalisability of the Barthel Index and the Functional Independence Measure: robustness of disability measures to Differential Item Functioning

Antonio Caronni & Stefano Scarano

**To cite this article:** Antonio Caronni & Stefano Scarano (02 Sep 2024): Generalisability of the Barthel Index and the Functional Independence Measure: robustness of disability measures to Differential Item Functioning, *Disability and Rehabilitation*, DOI: [10.1080/09638288.2024.2391554](https://doi.org/10.1080/09638288.2024.2391554)

**To link to this article:** <https://doi.org/10.1080/09638288.2024.2391554>



© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 02 Sep 2024.



[Submit your article to this journal](#)





[View related articles](#)



[View Crossmark data](#)

# Generalisability of the Barthel Index and the Functional Independence Measure: robustness of disability measures to Differential Item Functioning

Antonio Caronni<sup>a,b</sup>  and Stefano Scarano<sup>a,b</sup> 

<sup>a</sup>Department of Neurorehabilitation Sciences, IRCCS Istituto Auxologico Italiano, Milan, Italy; <sup>b</sup>Department of Biomedical Sciences for Health, University of Milan, Milan, Italy

## ABSTRACT

**Purpose:** Differential Item Functioning (DIF), an item malfunctioning, causes Differential Test Functioning (DTF), thus biasing questionnaire measures. The current study evaluates the relationship between DIF and DTF for the Barthel Index and the Functional Independence Measure, likely the most used disability measures. The aim is to understand under which conditions DIF can be ignored as its DTF is negligible.

**Methods:** A simulation study was run. Disability measures were obtained for the Barthel Index and FIM motor domain using Rasch analysis with previously published item calibrations. Several DIF scenarios have been assessed. DTF was tolerable if  $\leq 0.50$  logits.

**Results:** Simulations showed that the larger the DIF, the larger the DTF and that, keeping the overall DIF constant, the total number of items with DIF does not affect DTF. DIF of the items with the lowest or highest calibrations is the most dangerous. The DIF of central items should be so massive to matter in DTF terms that it is unlikely to happen in practice. The FIM robustness to DIF is better than that of the Barthel Index.

**Conclusions:** The FIM and the Barthel Index show remarkable robustness to DIF. Thanks to this feature, sample invariant, generalisable disability measures are available.

## ARTICLE HISTORY

Received 16 April 2024

Revised 16 June 2024

Accepted 8 August 2024

## KEYWORDS

Disability measurement; item bias; psychometrics; Rasch analysis; rehabilitation

## > IMPLICATIONS FOR REHABILITATION

- The Barthel Index and the Functional Independence Measure are two assessment procedures for evaluating disability.
- It is demonstrated that the Functional Independence Measure and the Barthel Index are remarkably resistant to Differential Item Functioning.
- As a result, these questionnaires yield very generalisable disability measures.

## Introduction

Disability, understood as the need for assistance from a person to complete activities of daily living, is one of the substantial variables in medicine and rehabilitation medicine in particular. For example, it is paramount to set the patient's prognosis in terms of disability, and disability is an outcome of therapies in the clinic and trials (e.g., [1,2]).

Several instruments are available for measuring disability, and, being a latent variable [3], they consist of questionnaires usually filled out by clinicians about a person's independence.

The Barthel Index [4], dating back to the sixties, is among the first questionnaires developed for assessing disability, and it is still one of the most used. From this, the Modified Barthel Index [5] was derived, primarily consisting of the Barthel Index with a more articulated categories structure. Another instrument that originated from the Barthel Index is the Functional Independence Measure (FIM) [6], a younger but well-established disability measure.



Regarding the items' content, the Barthel indices and the FIM share nearly identical items [7], which primarily assess the need for assistance in completing basic activities of daily living, such as walking and eating.


Like any questionnaire, the Barthel Index, the Modified Barthel, and the FIM return total scores, which are ordinal measures per the classification by Stevens [8]. However, in strict metrological terms, these scores *are not* measures [3].

The most convincing argument that ordinal scores are not measures is what could be called the "3–2≠2–1" paradox. In the case of the Barthel Index, there is no reason to assume that a patient whose total score increases from 5 to 10 has improved their disability of the same quantity as when the total score goes from 10 to 15. This is simply because ordinal scores lack a measurement unit.

However, thanks to the Rasch analysis [3,9–11], a psychometric technique, measures can be extracted from questionnaires. As long as a set of axioms is complied with, the Rasch analysis provides a score-to-measure conversion to turn the questionnaire's total ordinal scores into interval measures, with "logits" as the measurement unit.

In the Rasch analysis framework, like in mathematics and physics, measures must be unidimensional, i.e., measures should reflect the quantity of a single variable. A thermometer reading only reflects temperature, so the Barthel Index and FIM scores should only reflect the person's disability level.

**CONTACT** Antonio Caronni  [a.caronni@auxologico.it](mailto:a.caronni@auxologico.it)  Department of Neurorehabilitation Sciences, IRCCS Istituto Auxologico Italiano, Ospedale San Luca, Capitanio, Via Giuseppe Mercalli, 28, 20122 Milano, Italy

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/09638288.2024.2391554>.

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

The dimensionality of questionnaire measures can be assessed in different ways. Among these, there is testing if the Differential Item Functioning (DIF) deteriorates the questionnaire's items.

Regarding the FIM and the Barthel Index, given two persons with the same disability level but belonging to different (e.g., diagnostic) groups (e.g., a stroke patient and a multiple sclerosis patient), an item has DIF if the scores of these two persons on this item differ. In this condition, an additional variable (e.g., the disease type) affects the item's score independently from disability. The unidimensionality requirement is thus violated, given that the item's (and therefore the questionnaire's) scores depend on both the disability and this additional, undesired variable.

The multidimensionality flagged by the DIF causes a generalisability problem: measures are not independent anymore of what is being measured [12].

While it is clear that DIF introduces a measurement error, for practical purposes, it matters even more to understand if this artefact is harmful, i.e., if DIF severely distorts persons' measures. If the DIF-related bias is negligible, the questionnaire's measures remain generalisable despite DIF.

Hence, the research question underlying this work is: how much DIF is too much?

The current study aims to assess the measurement bias, referred to as Differential Test Functioning (DTF), caused by the DIF affecting one or more items of the Barthel Index, Modified Barthel Index and the FIM. In particular, the study evaluates the extent to which DIF can be ignored since the DTF it causes is negligible.

## Methods

The current study uses Rasch analysis simulations to assess the consequences of DIF on disability measures. The FIM motor domain, the Barthel Index and the Modified Barthel Index were chosen as questionnaires of physical disability.

In the first set of simulations, items and threshold calibrations of the FIM motor domain from our previous study [11] have been used. These calibrations have been obtained from a Rasch analysis (rating scale model) of the FIM scale collected on discharge from inpatient rehabilitation.

Instead, a more general approach was preferred for the two Barthel indices. A systematic Pubmed search was run, and the items' calibrations were eventually retained from 12 papers (see Note 1 in [Supplementary Materials 1](#)).

### The FIM scale and the Barthel indices: measuring disability

The FIM scale [6] consists of motor and cognitive domains, measuring physical and cognitive disability, respectively. Since the focus here is on physical disability (i.e., the need for assistance in completing activities of the type of eating and walking), only the FIM motor domain will be considered.

The FIM motor domain consists of 13 items, labelled from A to K ([Table 1](#)), each scored from 1 to 7. The total score ranges from 13 to 91.

The Barthel Index [4] consists of 10 items scored on two, three or four categories. In what is probably the most used version, each item is scored 0, 5, 10 or 15, but a version in which the items' categories are scored from 0 to 2 is also used [13], which is the version investigated here. The total score of the Barthel Index ranges from 0 to 100 or 0 to 20.

The Modified Barthel Index [5] represents a Barthel Index refinement developed to increase its responsiveness to small

**Table 1.** Ranking on the item map of the items of the FIM motor domain and Barthel Index.

Position	FIM motor	Barthel Index
1	A, Eating	Feeding
2	H, Bowel	Bowel
3	G, Bladder	Transfer
4	I, Bed Chair	Bladder
5	B, Grooming	Toilet
6	J, Toilet	Dressing
7	D, Dress Up	Grooming
8	L, Walk	Walking
9	C, Bathing	Bathing
10	F, Toileting	Stairs
11	E, Dress Low	
12	M, Stairs	
13	K, Tub	

Position: item position on the item map; this corresponds to the item calibration rank. FIM motor: FIM motor domain. A keyword indicates the content of each item. As customary, a capital letter is also used for the FIM motor domain. Items are ordered according to their calibration from the Rasch analysis. The item occupying the lowest position on the item map (i.e., on the first rank) indicates the easiest activity to complete independently. That with the highest calibration (rank 13 for the FIM motor domain and 10 for the Barthel Index) indicates the most difficult one. The items' positions of the FIM motor domain are taken from [10,11]; those of the Barthel Index are taken from [48]. The items of the modified Barthel Index have the same labels and ranking as the Barthel Index.

changes in independence. This questionnaire is made of 10 items, each in five categories. The numerals labelling these categories are arranged so that the questionnaire's total score ranges from 0 to 100. Here, the categories are re-labelled so that the total score ranges from 0 to 40.

For all three questionnaires, the higher the score, the greater the autonomy (i.e., the less the disability).

### Differential Item Functioning and differential test functioning: the overall idea

#### Differential Item Functioning and the split-item procedure

In the simulations run here, two participant groups were considered (Group A and Group B), and items with DIF were set as more difficult in Group B than A. At the same time, it was also set that the two participant groups had the same mean ability.

If an item is more difficult for Group B, Group B participants will be more likely to fail this item than Group A participants. The questionnaire total score will thus be lower in Group B than in Group A. If the same score-to-measure conversion is used for both groups to turn the questionnaire's scores into measures, measures of Group B will be lower than those of Group A. Since it has been set *a priori* that the participants from the two groups measure the same (i.e., the two groups have the same mean ability), an actual measurement artefact is introduced.

The "split-item procedure" [14–16] is often applied to correct this measurement artefact caused by DIF. Consider a test where one item (say item 2) is corrupted by DIF, making it more difficult for Group B than for Group A. With the split-item procedure, two virtual items are derived from this item: "Item 2 – Group A" and "Item 2 – Group B". In the former (i.e., the virtual item "Item 2 – Group A"), values for respondents from Group B will be considered as missing, while in the latter (i.e., "Item 2 – Group B") values for Group A will be considered as missing.

The Rasch analysis is then run on the new item set (including "Item 2 – Group A" and "Item 2 – Group B" as well as all the questionnaire's items, but the original item 2), items' difficulties are obtained, and eventually used to arrange two score-to-measure conversions, one for each of the two groups.

These group-specific score-to-measure transformations return participant measures unbiased from DIF. So, while using the total score (and the measures from a single score-to-measure transformation) for comparing a person from Group A with one from Group B would be unfair, measures from the group-specific score-to-measure transformations allow fair comparisons between groups.

One thousand participants were simulated for each DIF group, so each simulation run in the current study had a sample size of 2000 people. Regarding the sample size, the one used here is the same as that of a previous study, which has inspired the current one, also investigating DIF with the Rasch analysis [17]. A sample of 2000 participants can be considered large in the Rasch analysis since 250 participants are "appropriate for most purposes," even in the case of a high-stakes test [18].

#### *Estimating the differential test functioning from DIF*

As previously done [15,19], the artefact caused by DIF on the whole questionnaire measures, i.e., the Differential Test Functioning (DTF), is quantified by comparing the score-to-measure transformations of the two DIF groups.

Given the above, using a single score-to-measure conversion to measure both Group A and Group B would introduce a measurement artefact. However, is the DIF so large to make this measurement artefact substantial? This would be the case if the difference in the actual measures of Group A and Group B is so large that it cannot be ignored. In that case, group-specific score-to-measure conversions would be necessary.

The difference between Group B and Group A in the measures from the score-to-measure transformation can be calculated for each of the questionnaire's total score values, and a difference  $> 0.5$  logits is considered substantial, as customary in the Rasch analysis [20].

In sum, when there is DIF, people scoring the same can have different measures. This difference in measures can be quantified thanks to the groups' specific score-to-measure transformations from the split-item procedure. If the difference between measures from the two groups, given the same total score, is  $> 0.5$  logit, the DTF produced by DIF flags a non-negligible measurement artefact.

R 4.2.3 and Winsteps 5.4.3.0 (batch mode) were used to run the simulations. The above analysis steps are detailed in Notes 2 and 3 in [Supplementary Materials 1](#).

## Results

### *Differential test functioning of the FIM motor domain: the case of a single item with DIF*

[Figure 1](#) shows the relationship between the DIF size, the position of a single item with DIF on the item map (i.e., its calibration rank) and DTF.

There is a clear positive relationship between the size of DIF and the DTF it causes: regardless of the rank of the item with DIF, the larger the DIF, the larger the DTF.

In addition, the DTF size depends on the position of the item with DIF along the item map, with DTF being more substantial when the item with DIF lies at the extremes of the item map. For example, a 1.50 logit DIF of item A causes a maximum DTF of almost 0.50 logits. In comparison, the maximum DTF caused by a 1.50 logit DIF of items B and J is less than 0.25 logits.

Moreover, the DTF is not constant for different questionnaire total scores, and the DIF of items located at the extremes of the

item map (i.e., the easiest and the most difficult activities to complete independently) causes a DTF that is maximum for extreme questionnaire scores. For example, the DIF of item A (i.e., the easiest item of the FIM motor) causes some malfunctioning for low questionnaire total scores, while the DTF tapers as the total score increases. On the contrary, item K (i.e., the most difficult item of the FIM motor) causes the largest DTF for high total scores but no malfunctioning for low ones.

### *Differential test functioning of the FIM motor domain: the case of two items with DIF*

[Figure 2](#) reports the DTF for DIF affecting two items of the FIM motor domain and occupying the item map's lower, middle or higher third.

As before, the larger the overall DIF (i.e., the sum of the DIF size of the two items), the larger the DTF. Second, the maximum DTF varies with the position of the items with DIF on the item map and again, the DTF is larger when the items affected by DIF lie on the map extremes. Third, the DTF is not constant for different questionnaire total scores.

If the two- and the one-item-with-DIF cases are compared, the DTF originating in the case of DIF affecting two items is systematically larger. For example, item A with a DIF of 1.5 logit causes a maximum DTF that is smaller than 0.50 logits ([Figure 1](#)), while a 1.5 logit DIF affecting both items A and H causes a maximum DTF of about 0.75 logits.

[Table S2.1](#) in [Supplementary Materials 2](#) reports the maximum DTF for all the possible combinations of two items with DIF of the FIM motor domain.

Any two items with DIF up to 0.75 logits cause no substantial DTF. For DIF up to 1 logit, the DTF is  $> 0.50$  logit in the case (also depicted in [Figure 2](#)) in which the items with DIF occupy the first and the second (i.e., items A and H) or the first and the third position (i.e., items A and G) on the item map. All the remaining two-item combinations are safe in DTF terms. Most item combinations caused no harmful DTF when DIF was 1.25 and 1.5 logits (73 and 57 out of 78 combinations, respectively).

These simulations in which two items have DIF also highlight that DTF increases as the items with DIF are close on the items map (e.g., items ranking one and two on the items map) and is reduced when these items are apart (e.g., items ranking one and four).

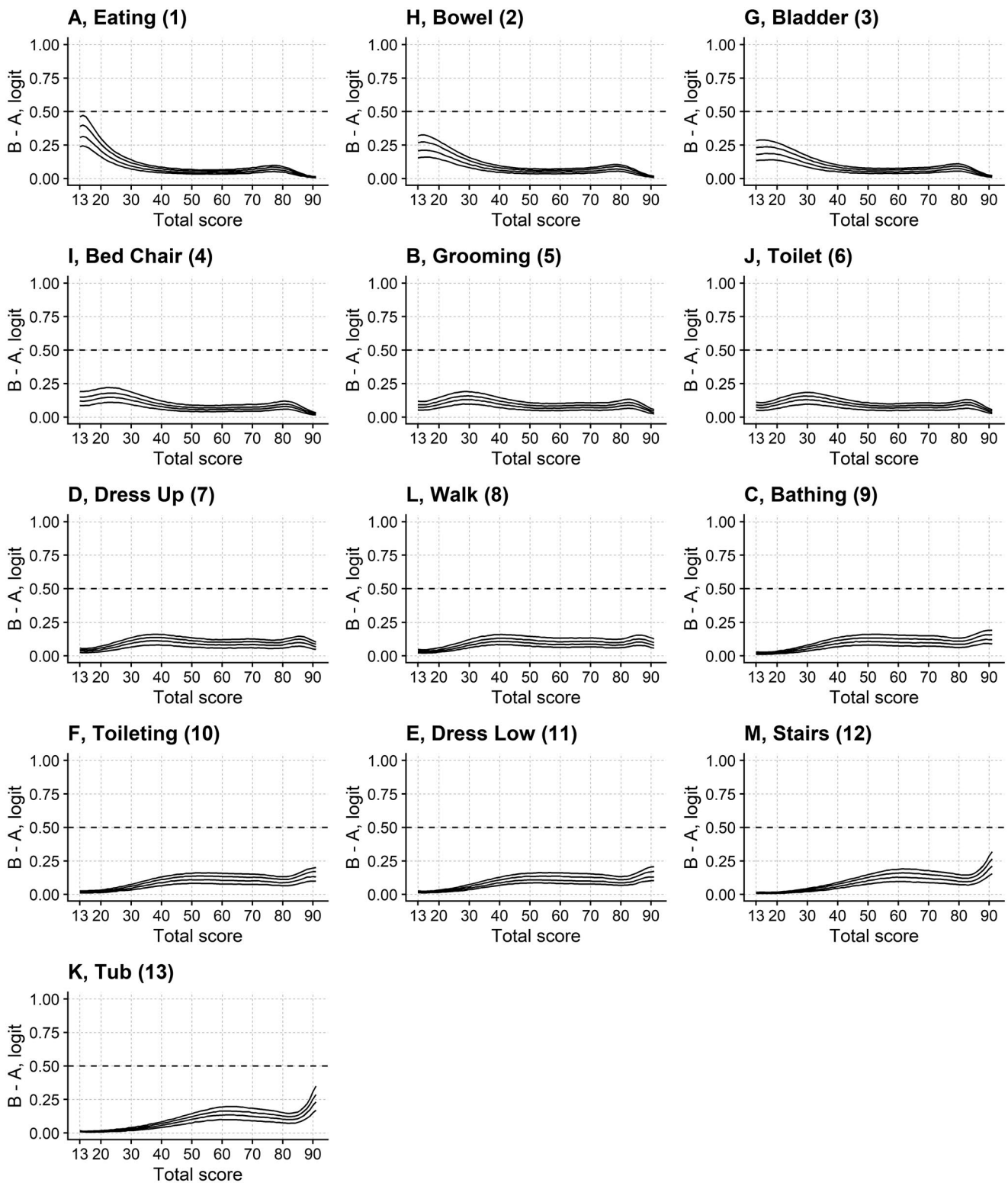
Further simulations with three items with DIF are reported in [Supplementary Materials 3](#).

### *Additive effects of the Differential Item Functioning on the differential test functioning*

The analyses reported in the previous paragraph show that multiple items' DIF somehow add to the DTF. This feature has been explored further in this paragraph.

What matters in DTF terms is the total amount of DIF (i.e., the overall DIF) rather than the number of items affected by DIF. The DTF seems unaffected by the total number of items with DIF as long as the overall DIF remains constant. Indeed, in each graph of [Figure 3](#), showing the relationship between the DTF and the number of items with DIF given an overall DIF, the different curves are substantially superimposed.

In addition, the figure stresses the remarkable robustness of the FIM motor domain to DIF. When DIF affects the items in the central part of the items map, an overall DIF of up to 4 logits causes a maximum DTF  $< 0.5$  logits.



**Figure 1.** Differential Test functioning of the FIM motor domain: one item with DIF. Items are indicated by a letter and a keyword, while the number between brackets reports the item's rank on the item map. Line plots: median of 20 simulations. Within each plot, moving top down, curves show the Differential Item Functioning (DTF) caused by a Differential Item Functioning (DIF) of size 1.5, 1.25, 1.0 and 0.75 logits. Total score: total ordinal score of the FIM motor domain. B - A: difference (in logit) between the measures from the score-to-measure conversion in group B and those from group A. The horizontal dashed line marks 0.5 logits, set in the current study as the maximum tolerable DTF.

**Differential test functioning of the Barthel and modified Barthel Index**

The findings from the FIM motor domain simulations are confirmed in the Barthel Index simulations.

Again, the larger the DIF, the larger the DTF (Figure 4). The DIF is more hazardous when it affects the items located at the item map's extremes. The DTF caused by DIF is not constant for different total score values and is most prominent for the total score extremes when DIF affects an item on the map's extreme (e.g., item Feeding).

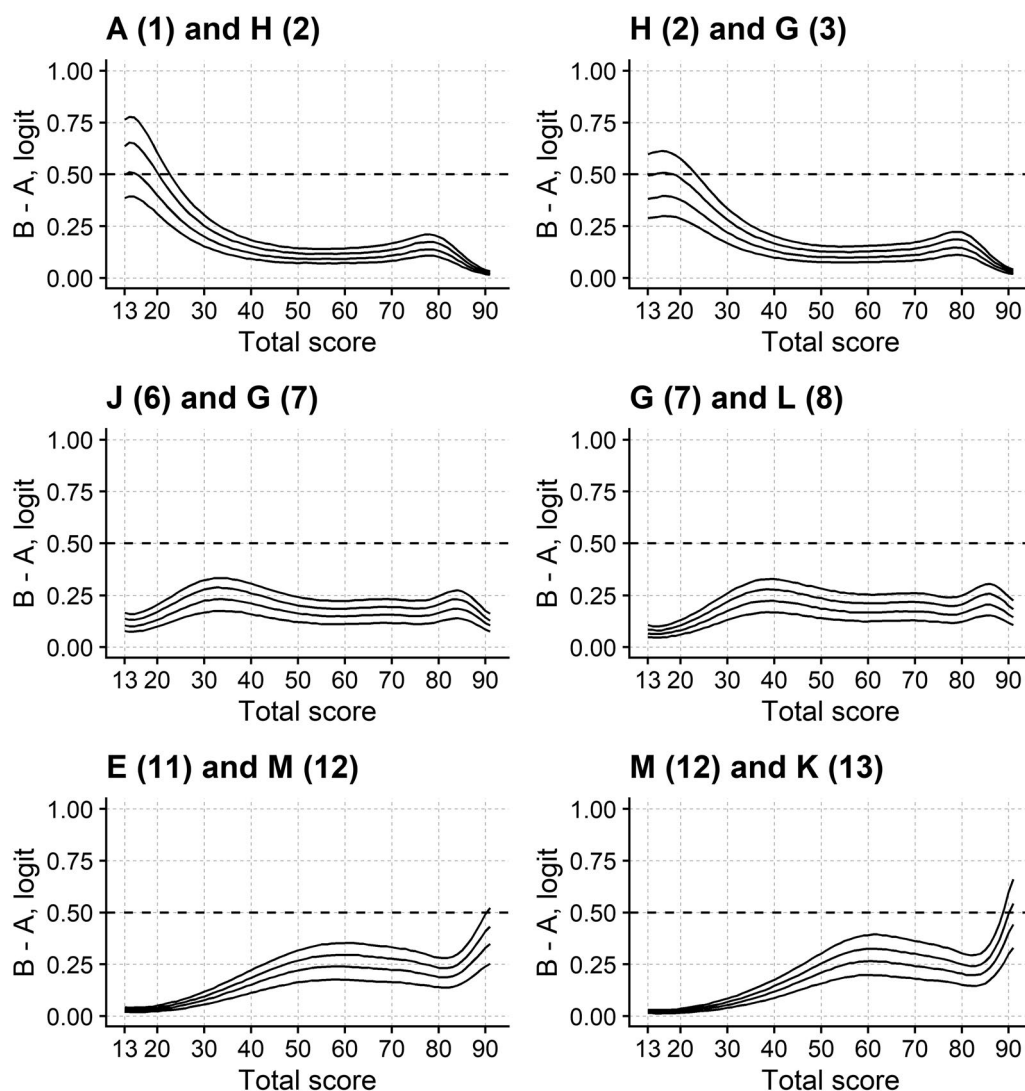


Figure 2. Differential Test functioning of the FIM motor: two items with DIF. Same abbreviations as Figure 1.

DIF is the most dangerous when it affects more than one item, and its effects on DTF are the largest when the items with DIF are contiguous on the item map (Supplementary Materials 2 and 3).

The Barthel Index seems less robust to DIF than the FIM motor domain. For example, a DIF of size  $\geq 1$  logit affecting the item with calibration rank one causes a negligible DTF of the FIM motor domain but a substantial DTF of the Barthel Index.

The simulations of the Modified Barthel Index returned results comparable to those relative to the Barthel Index (see Supplementary Materials). The similarity between the Barthel and Modified Barthel Index findings suggests that the total number of categories (and thresholds) does not substantially affect the DTF when the item map and threshold ranges are the same.

## Discussion

This study used simulations to assess the measurement artefact, here called DTF, caused by the DIF of one or more items of the three main questionnaires used for assessing physical disability: the Barthel Index, the Modified Barthel Index and the FIM motor domain.

The study's key findings are hereby summarised.

First, disability measures from the FIM motor domain and the Barthel indices show remarkable robustness to DIF. One item with

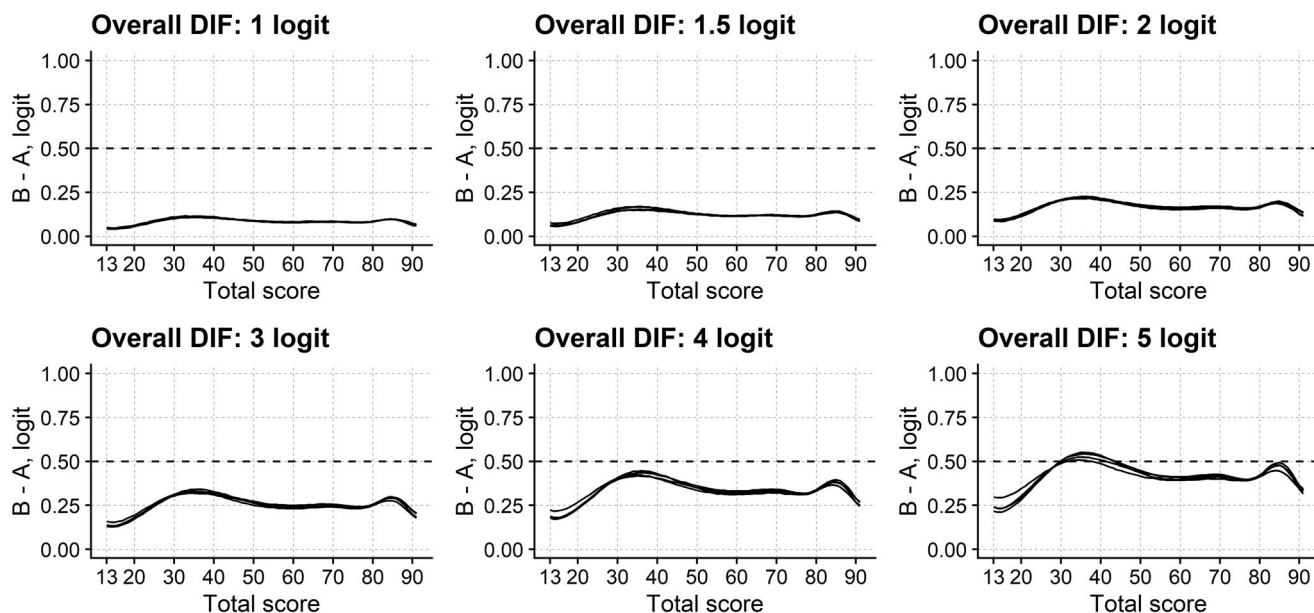
DIF up to 1 logit does not harm the questionnaire's measures. Except for the case in which DIF affects the items of the Barthel indices with the lowest or highest calibrations, even the DIF of a single item up to 1.5 logits causes no substantial DTF.

Second, the disability measures from the FIM motor domain are more robust to DIF than those from the Barthel questionnaires.

Third, most troubles with DIF come when it affects the items located at the extremes of the item map, while DTF is negligible even when a huge DIF affects the items occupying the map's central position.

The findings reported here align with those from previous studies. It has already been noticed that unless DIF is large and mainly in one direction (i.e., the type of DIF tested here), the impact of DIF on measures can be small [20]. For example, biases of 1.0 logits are unlikely to have much impact on a test instrument and can be set as an upper limit in the item calibration error [21] (see Supplementary Materials 1 for a broader discussion on the DIF-DTF relationship).

One point is worth stressing about the seemingly inconsequential nature of DIF. The fact that the DIF of some items can be ignored when using a questionnaire takes advantage of a statistical/mathematical feature of the questionnaire's measures. Questionnaire measures are mathematically robust to some



**Figure 3.** Differential Test functioning caused by the same overall DIF split into a different number of items. Each graph shows the differential Test functioning (DTF) caused by the Differential Item Functioning (DIF) of up to four items of the FIM motor scale. Overall DIF of 1.0, 1.5, 2.0, 3.0, 4.0 and 5.0 logits is divided between two, three, four and five items. For example, for an overall DIF of 3 logits and four items with DIF, each of these four items suffers a DIF of 0.75 logits. Only items with a Central calibration are considered in these simulations, i.e., items B, J, D, L, and C, ranking 5 to 9 on the item map. For the two items with DIF condition, DIF affected items J and L. Items J, D, and L had DIF in the three items simulations, and items B, J, L, and C had DIF in the four items simulations. Note that the four curves are substantially superimposed and that the maximum DTF changes with increasing levels of overall DIF but not as the number of items with DIF increases. Abbreviations as in Figure 3.

amount of DIF. However, can we trust a questionnaire with, say, half of the items showing DIF in the same direction for the same group of respondents? What is the measured variable when different groups understand half of the items differently?

DIF poses a content validity problem [22], not just a construct validity one [15,22]. Deciding about DIF urges one to reason about the measured variable.

#### ***DIF of the Barthel indices and the FIM scale: previous reports and solutions from the literature***

DIF has been reported for both the Barthel and Modified Barthel Index.

For example, regarding the Barthel Index, most items showed DIF related to the country, a finding leading the Authors to conclude that "Barthel Index scores should not be compared between cultures" [23], which, in strict statistical terms, is a correct conclusion.

For the Modified Barthel Index, DIF was found for the clinical phenotype in a study which recruited stroke patients, with six items showing DIF related to the severity of the upper limb paresis [24].

The FIM scale has been extensively investigated with the Rasch analysis, and, in this framework, that some FIM items are affected by DIF is relatively common [25,26].

In a study explicitly assessing the DIF of the FIM items [27], difficulties of about half of the items of the FIM motor domain differed between diagnostic groups, i.e., showed DIF for the neurological diagnosis of the patient disability.

In detail, calibrations of seven items differed in stroke and multiple sclerosis, two in multiple sclerosis and traumatic brain injury patients, and the calibration of three items differed in stroke and traumatic brain injury. Only four items of the FIM motor domain were found to have no DIF for the patient diagnosis.

The authors of this study [27] rightly stressed that no DIF should be present if the FIM has to be compared between patient

groups or when pooling data from different conditions. On this line of reasoning, they suggested adjusting the FIM items calibration to solve for DIF.

Similar to our study, the split-item procedure was applied, and item calibrations specific to three diagnostic groups, i.e., stroke, multiple sclerosis, and traumatic brain injury, were provided.

However, it is paramount to note that, entirely agreeing with the findings reported here, even if some differences were found between the DIF-adjusted and the unadjusted measures, the agreement between the two sets of measures was high, and differences were negligible. According to the Authors, "adjusting for DIF seems to have only minor impact on the person abilities" [27].

In another study recruiting neurological patients [28], most FIM items were affected by DIF for diagnosis with a different calibration in spinal cord injury patients than in other neurological patients. The number of items with DIF was so large (9 out of 13) that it was not feasible to amend this DIF by resorting to the split item procedure.

The authors assessed the clinical meaning of this statistical DIF by evaluating its impact on measures. Score-to-measures curves for the different diagnoses were compared, and the difference between curves (i.e., the DTF) was judged as causing no harm in clinical terms.

Showing that DIF is not substantially harmful provides us with generalisable measures. This idea of generalisability is pretty in line with the use of the term "generalisability" in other branches of statistics, where "to generalise" means to extend findings to other settings and samples [29]. In this context, items' scores are generalisable if a specific score indicates a specific disability level, which is the same in this and other samples.

The need for generic questionnaire measures has already been pointed out [30].

Disease-specific questionnaires, e.g., [12,31], have been contrasted with generic ones (e.g., invariant across diseases or

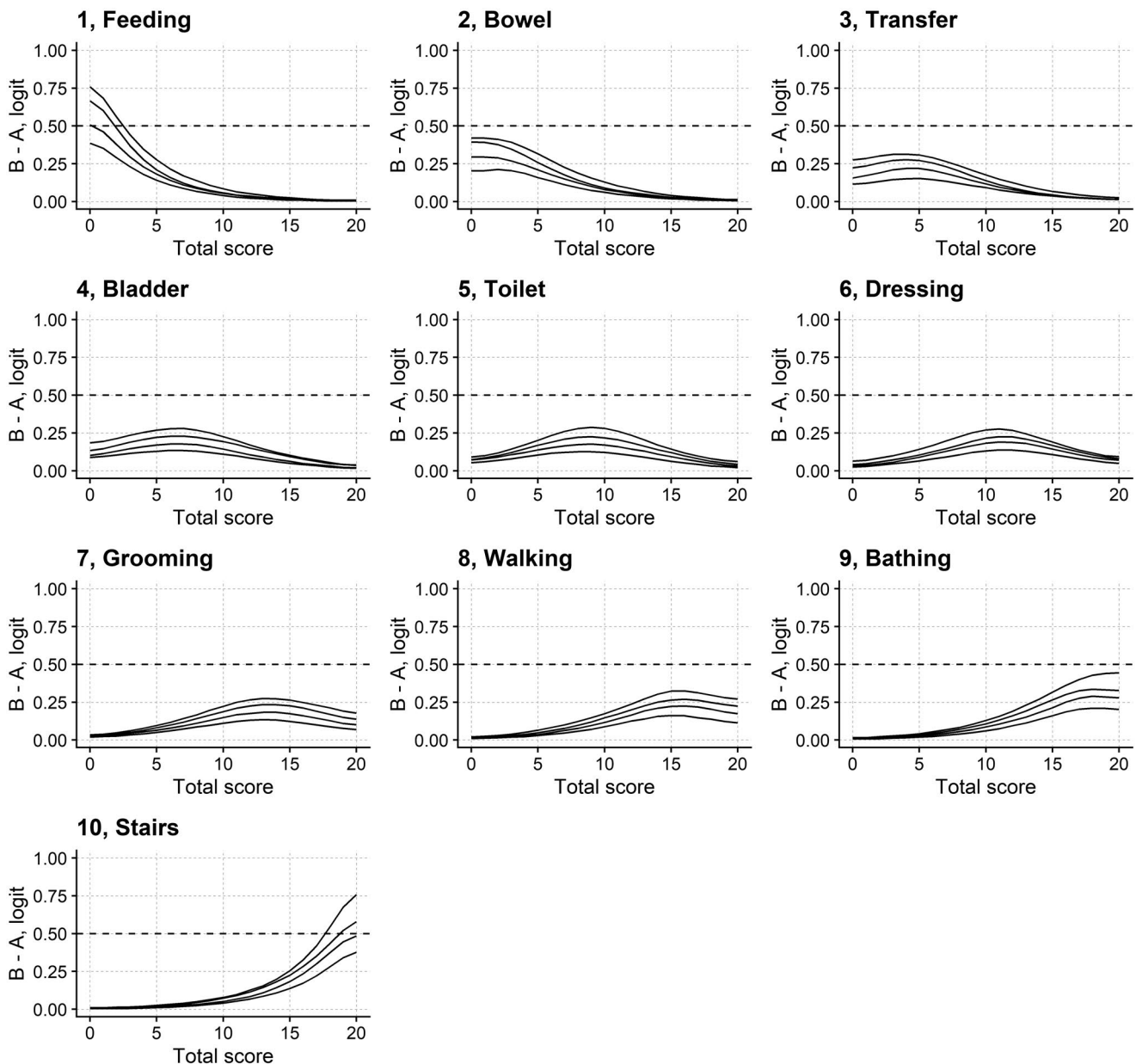


Figure 4. Differential Test functioning of the Barthel Index: one item with DIF. Same abbreviations as Figure 1. Items are labelled with a keyword. The number indicates the item rank on the item map.

cultures), e.g., [32], and it has been noted that while disease-specific measures provide well-tailored patient measures, this specificity pays the price of no generalisability. Comparing the variable of interest in different groups becomes harder.

It has also been stressed that generic measures are necessary in medical contexts, such as rehabilitation medicine, where a condition (e.g., disability) is treated regardless of the underlying disease [30].

The importance of generalisable measures is further developed in [Supplementary Materials 1](#). The procedures used when DIF is found are also treated and compared to those proposed here.

#### ***Really, but who cares about DIF, and who cares about DIF of scales for disability measurement?***

In this excursus: (i) case studies about what DIF is, how DIF originates and what problems it causes are provided, and (ii) how

the findings reported here can be used from a practical standpoint in the clinic and research is explained.

There are two issues with DIF. First, the measurement artefact caused by DIF poses a fairness problem. A test would be easier for some respondents and more challenging for others regardless of their ability level. Second, DIF causes a validity problem [33,34], urging us to consider which variable the questionnaire measures (which variables drive the respondents' scores to an item?).

Depending on the case, DIF can cause more of the first or the second issue.

The current study shows that the measurement artefact caused by DIF, and thus the DIF-related fairness issue, is minor in most cases unless DIF is massive (substantial in size and simultaneously affecting several items all in the same direction). Instead of bothering about statistics, an evaluation of the variable(s) the questionnaire measures – that is, the validity of the questionnaire – should be done in the first place when DIF is discovered.



Let us first deal with the issue of fairness. That will be clear with the following (toy) example from educational measurement. It is worth remembering that many DIF studies and investigations are conducted in the educational field.

A typical instance of DIF is when knowledge of the matter under evaluation (say knowledge of the European literature of the first half of the twentieth century) and high language proficiency (which is not strictly related to the knowledge of literature, i.e., the variable to be measured with the test) are both needed to answer a question correctly [35].

The test below, consisting of four open-ended questions, scored 1 if answered correctly and 0 elsewhere, is administered to assess knowledge of literary works:

1. Who is the protagonist in "Steppenwolf"?
2. Who is the protagonist in "Der Zauberberg"?
3. Who is the protagonist in "Ulysses"?
4. Who is the protagonist in "The Metamorphosis"?

Consider now a student answering questions 1, 3 and 4 correctly and leaving unanswered question 2. Another student answers correctly to all questions. The total score of the second student on the test is higher than that of the first, so it would be concluded that the second is better in European literature.

The book title mentioned in question 2 is translated into English ("The Magic Mountain"), the question is re-administered to the first student, and the student correctly answers "Hans Castorp".

So, do we still conclude that the second student is more proficient in literature than the first? We would not say so if, for example, it were known that the second student, but not the first, is bilingual in English and German. Instead, we would probably consider the second question unfair towards students who do not understand German.

Question 2 has different difficulties, i.e., a different chance of being passed, for students belonging to different groups (with and without knowledge of German) despite the students' same ability level, i.e., despite the same level of knowledge of literary works (they both score 1 on question 2, eventually).

Regarding medical questionnaires and scales, an example comparable to the one reported above is an oral test for assessing semantic memory administered to a patient with expressive aphasia. This person would probably score poorly on this test, but we would not conclude that they have a memory impairment. Instead, we would conclude that it is unfair to administer such a test in case of aphasia.

Now, let's examine the DIF-caused fairness issue for disability questionnaires.

A certain level of motor and cognitive abilities is needed to complete the tasks that comprise the Barthel Index and the FIM disability scales. If these motor or cognitive competencies are too low, one or more tasks would become too difficult for the patient, who would need help from another person to complete these tasks. i.e., they have an activity limitation and suffer some disability.

To provide fair disability measures, it is paramount that the tasks included in the disability scales retain the same difficulty for participants with the same ability, even if these are different in some other respects, i.e., even if participants belong to different groups. If this is not the case, based on what has been discussed, DIF is present.

Studies found DIF for the Barthel Index and the FIM scale because of the patient's diagnosis. For example, the calibration of item E, "Dressing – lower," of the FIM motor domain has been

reported to be higher in multiple sclerosis patients than in stroke patients [27].

The overall difficulty of the task considered by item E is partly due to how difficult it is to mobilise the lower limbs. This difficulty increases with the severity of the lower limb's paresis and spasticity. Since it often causes spinal cord involvement, these impairments could be bilateral and thus more severe in multiple sclerosis than stroke. If this is the case, "mobilising the lower limbs," an activity necessary for "Dressing – lower," would thus be more challenging in multiple sclerosis than stroke.

Given a stroke and a multiple sclerosis patient who both score low on item E, since the stroke patient is failing an easier task, the actual disability of the stroke patient could be worse than that of the multiple sclerosis patient. The stroke patient could be more disabled, actually, and the difference in the difficulty of item E should be considered for a fair disability comparison between the two patients and diseases.

The second DIF-related issue, which is how DIF and DIF remedies may threaten a questionnaire's validity, is considered in the following example.

The five-item Barthel index [36] is a short form of the Barthel index consisting, as the name implies, of five items only: Transfers, Bathing, Toilet use, Stairs, and Mobility. This pocketable version of the Barthel Index considers disability primarily as a gross motor function limitation, i.e., a limitation with transferring and locomotion.

Consider now a stroke patient with hemiparesis and a paraparetic patient because of a dorsal spinal cord injury. Both are dependent on bathing, need some help with toilet use, need major help with transfers, are wheelchair independent for mobility and are unable to manage stairs. The total score on the scale is 3 for both; therefore, according to the five-item Barthel Index, their disability level is the same.

Now, go on administering the other items of the Barthel Index. The Feeding item is administered with the spinal cord injury patient independent on this task, thus scoring 2 out of 2, and the stroke patient needing help cutting and requiring a modified diet, thus scoring 1.

For the five-item Barthel Index, the disability level of the patients would be the same (i.e., their ability is the same). So, they are expected to score the same on the feeding item, which is not the case. Since it is easier for spinal cord injury than stroke, the feeding item is thus affected by DIF for the patient's diagnosis.

This case of DIF for the feeding item of the Barthel Index is substantially different from the case of DIF of "Dressing – lower" (item E) of the FIM motor domain.

Disability is the need for assistance in basic daily activities, including self-care and sphincter control, not only transferring and locomoting. Probably, any clinician and researcher working on disability would consider a person needing assistance for eating as a person who has some disability.

On these bases, the DIF of the feeding item points out more of a content validity problem of the original questionnaire: is the variable of interest entirely sounded out?

As shown here, mathematical amendments to DIF, such as the split-item procedure, are available. Another extreme solution to DIF is to drop the item with DIF from the questionnaire. It has been noted that these remedies to DIF are not free of consequence since both could deteriorate the content validity of the measure [34,37].

Removing the feeding item since it is affected by DIF would avoid considering a crucial facet of disability, impoverishing the questionnaire content validity. The split-item procedure would reduce the disability measure of the stroke patient (which would

make this patient less disabled than they are), but we said this is against any clinical common sense.

Here, DIF is stressed as a content validity threat, understood as "the degree to which the content of an instrument is an adequate reflection of the construct to be measured" [34]. In addition, DIF also threatens construct validity.

Measurements are assumed to be unidimensional. In the case of DIF of item E of the FIM motor domain described above, the score of participants to this item is affected by two variables: the disability severity and the severity of the upper motor neuron syndrome (i.e., the paresis and spasticity severity). Note some amount of orthogonality between them: a paraplegic person can suffer a lower limb motor impairment of extreme severity, but their disability could be low (e.g., they are competent in getting autonomously around the city with a wheelchair, which is with a different form of locomotion). Two variables (dimensions) thus drive item E scores, and the unidimensionality assumption about the construct is violated.

It is worth stressing that no multidimensionality issue is apparent in the case of DIF of item feeding for the Barthel Index short form. This item just grasps a different facet of the disability [38].

After this overview, it is evident that DIF can cause (i) a measurement artefact and, thus, a measurement bias and a fairness problem, (ii) a content and (iii) a construct validity issue.

So, who should care about DIF? DIF potentially matters anytime a patient's measure is at stake.

This study can be used by psychometricians when developing and assessing questionnaires. We can say that it makes life easier for the researcher in the psychometrics lab.

Assessing DIF is a mandatory step in the psychometric assessment of a questionnaire, with guidelines also dealing with it [39,40]. When DIF is found, and some amount of DIF is always found due to the large number of tests commonly run in a DIF analysis and being defined in probabilistic terms, the psychometrician has to assess the effects of DIF or amend DIF by applying the split-item procedure.

Both assessing the DIF consequence, for example, by comparing the score-to-measure conversions or amending the DIF with the split-item procedure, are articulated and cumbersome. In addition, as detailed above, amending DIF is not consequence-free regarding content validity. Strictly speaking, this solution could solve a statistical malfunction at the cost of causing a clinical malfunction.

Our study encourages not bothering too much from a strict statistical, metrological point of view in front of a questionnaire structurally similar to the FIM and Barthel indices when the DIF of one or more of their items is within the boundaries reported here. In psychometric terms, the test can be considered valid despite DIF.

For the clinical researcher, choosing the instrument with no harmful DIF among different instruments measuring the same variable means choosing a robust measurement instrument, eventually increasing the chance of reaching the correct conclusion about treatment effectiveness [41].

Time-related DIF can be a severe problem for a clinical study [15,32]. This type of DIF could happen, for example, when a participant learns some tasks included in the test, and thus, the test becomes easier to complete. For example, the respondent could learn the three words "apple, table and penny" of the Mini-Mental State Examination [42] because of a practice effect from repeated test administrations rather than genuinely improving their memory and ability to register and recall [43].

If there is DIF because of the passing of time, a patient's difference between two subsequent time points, such as before

treatments and at follow-up, could be a measurement artefact. Patient scores worsening in a longitudinal study could also be an artefact instead of indicating the disease's progression.

In multicentre international studies, DIF for cultures or languages (e.g., [16]) makes comparing the variable of interest in participants' samples from different countries unsuitable.

DIF also matters to clinicians, even if they are not directly involved in research. For example, in everyday clinical practice in some health systems (e.g., [44]), treatment access requires a specific disability level, as indicated by the score of a disability questionnaire. If disability scores are unfair towards some patients, they could be excluded from treatments when entitled to them. For an example of significant interest in everyday clinical practice in rehabilitation, a study is suggested in which the DIF caused by assistive devices in balance measurement has been evaluated [15].

In more general, but not less critical, terms, since some amount of DIF is invariably found (e.g., [28]), the clinician and the clinical researcher would wonder if fair measures of the paramount (latent) variables of the physical and rehabilitation medicine (e.g., disability, pain) are possible. Showing that in some circumstances, DIF is too small to matter provides clinicians with fair, unbiased measures, which is a comforting solution to this issue.

### **Boundaries in the applicability of the current findings and ideas for future research**

While several parameters have been varied in the simulations presented here, others have not. Therefore, future research is needed to understand the effects of these parameters on the DTF.

For example, the person's mean measure was set to 0 logits for both groups, i.e., the persons were perfectly centred on the item map. Of course, this is not necessarily the case with real data, where poor questionnaire targeting can be found, e.g., [45].

Moreover, the participants' mean measures can differ in the two groups. However, another simulation study [46] showed that differences in group measures had negligible effects on measurement bias.

All the participants' measures and items simulated here perfectly fit the Rasch model, which is not necessarily true with real data. For the FIM motor domain, some misfits, for example, for the bladder and bowel management items, are typical [10,25,47]. Misfits of the Barthel Index items have also been reported [48].

In the case of two or three items with DIF, only the case of DIF affecting all these items to the same extent and in the same direction has been considered. Evaluating DIF with opposite signs and amplitudes for different items seems of practical importance (see [Supplementary Materials 1](#) for a discussion on DIF cancellation [49]).

Another intrinsic feature of the current work that could be considered a limitation is that the DIF analysis reported here is exclusively based on the Rasch analysis measurement model. However, Classical Test Theory and Item Response Theory are also available for questionnaire evaluation, and each has suitable DIF assessment tools.

As detailed in the Methods section, in the current study, items with DIF had a *different calibration* in the two groups of respondents (i.e., were set more difficult in Group B than A). On the contrary, the two participant groups were set with the *same mean ability*, fixed to 0 logits.

This way of simulating DIF is entirely in line with the DIF analysis by Linacre [20], according to which persons' abilities from the primary analysis (here, their disability level) are anchored (i.e.,

fixed), and item difficulties are estimated unanchored, i.e., free to differ in the two participants' groups.

In computational terms, this "anchor abilities method" is consistent with the Mantel-Haenszel statistics, commonly used for DIF assessment in the Classical Test Theory and the "anchor theta method" of the Item Response Theory [20].

According to the Mantel-Haenszel procedure [50], respondents are classified based on their ability level, which is determined by the questionnaire's total score. Since it works on total raw scores, the Mantel-Haenszel procedure does not allow for missing items.

Next, for each ability class and item, the number of respondents passing the item and the number of respondents failing it is compared in the two DIF groups (here, these would be Group A and Group B).

In Item Response Theory, DIF also looks at the chance of correctly answering or endorsing an item conditioned on the latent trait called "theta," in Item Response Theory jargon.

Several methods are available for DIF assessment in the Item Response Theory. Among these, a procedure based on logistic regression, which aligns with the Mantel-Haenszel procedure, is often used.

The probability of affirming/passing/endorsing an item is the response variable of a logistic regression model with the participants' questionnaire's total score, membership to one of the DIF groups, and the interaction between the total score and group membership as the predictors.

A significant group membership would indicate that the probability of endorsing an item is different in the groups of the DIF analysis, regardless of any difference in the overall ability, as indicated by the questionnaire total score.

Even if computational differences can be found, the similarities between the DIF assessment in the Classical Test Theory, Rasch analysis and the Item Response Theory are remarkable. Moreover, it is worth noting that, mathematically speaking, the dichotomous Rasch model, i.e., the original model by Georg Rasch [9], is the one-parameter model of the Item Response Theory [51]. Thus, DIF assessment techniques for the Item Response Theory could be applied to the Rasch measurement framework.

As part of the future development of current research, the following studies could assess the DTF caused by DIF in the context of different psychometric theories. For example, it could be that participants' metrics extracted from questionnaires' total scores with models from the Item Response Theory could show different robustness to DIF compared to the measures from the Rasch analysis.

As a last note, it also emphasised that this analysis has only looked at the uniform DIF case. The case of the non-uniform DIF remains to be developed in full.

## Conclusions

It is shown here that disability measures from the FIM motor domain and the Barthel indices are remarkably robust to the DIF, i.e., the malfunctioning of their items.

Regarding measurement error, the bias (i.e., the unfairness) caused even by a considerable DIF is often small enough to be ignored for practical purposes.

As long as the amplitude of the bias caused by DIF is known, some error is tolerated, and DIF does not threaten the questionnaire's validity, the robustness of questionnaire measures to DIF makes it possible to benefit from generalisable measures.

## Acknowledgements

We thank the Reviewers for offering us the chance to explain and treat in greater detail some crucial aspects of the analysis of the Differential Item Functioning. The section "Really, but who cares about DIF, and who cares about DIF of scales for disability measurement?" sparked from a fruitful discussion with them.

## Ethical approval

Not applicable: the study presents the results of statistical simulations using parameters from previously published studies accessible from PubMed. On these bases, no ethical approval for the current study was deemed necessary. No participant has been primitively recruited here.

## Disclosure statement

Stefano Scarano is a member of the Editorial Board of *Disability and Rehabilitation*. The authors declare that they have no competing interests.

## Funding

BIBLIOSAN provided APC funding.

## ORCID

Antonio Caronni  <http://orcid.org/0000-0003-3051-1031>

Stefano Scarano  <http://orcid.org/0000-0002-9217-4117>

## Data availability statement

The simulations about the FIM motor domain have used item calibrations tabulated in a former study [11] whose data have been uploaded on Zenodo [47]. For the Barthel Index and the Modified Barthel Index, a systematic search on PubMed was run, the results of which are detailed in full in [Supplementary Material 1](#), Note 1.

## References

- [1] Bracard S, Ducrocq X, Mas JL, et al. Mechanical thrombectomy after intravenous alteplase versus alteplase alone after stroke (THRACE): a randomised controlled trial. *Lancet Neurol*. 2016;15(11):1138–1147. doi: [10.1016/S1474-4422\(16\)30177-6](https://doi.org/10.1016/S1474-4422(16)30177-6).
- [2] Kwakkel G, Wagenaar RC, Twisk JW, et al. Intensity of leg and arm training after primary middle-cerebral-artery stroke: a randomised trial. *Lancet*. 1999;354(9174):191–196. doi: [10.1016/S0140-6736\(98\)09477-X](https://doi.org/10.1016/S0140-6736(98)09477-X).
- [3] Tesio L, Scarano S, Hassan S, et al. Why questionnaire scores are not measures: a question-raising article. *Am J Phys Med Rehabil*. 2023;102(1):75–82. doi: [10.1097/PHM.0000000000002028](https://doi.org/10.1097/PHM.0000000000002028).
- [4] Mahoney FI, Barthel DW. Functional evaluation: the Barthel Index. *Md State Med J*. 1965;14:61–65.
- [5] Shah S, Vanclay F, Cooper B. Improving the sensitivity of the Barthel Index for stroke rehabilitation. *J Clin Epidemiol*. 1989;42(8):703–709. doi: [10.1016/0895-4356\(89\)90065-6](https://doi.org/10.1016/0895-4356(89)90065-6).
- [6] Granger CV, Hamilton BB, Linacre JM, et al. Performance profiles of the functional independence measure. *Am J Phys Med Rehabil*. 1993;72(2):84–89. doi: [10.1097/0002060-199304000-00005](https://doi.org/10.1097/0002060-199304000-00005).

- [7] Prodinge B, O'Connor RJ, Stucki G, et al. Establishing score equivalence of the Functional Independence Measure motor scale and the Barthel Index, utilising the International Classification of Functioning, Disability and Health and Rasch measurement theory. *J Rehabil Med.* 2017;49(5):416–422. May 16 doi: [10.2340/16501977-2225](https://doi.org/10.2340/16501977-2225).
- [8] Stevens SS. On the theory of scales of measurement. *Science.* 1946;103(2684):677–680. doi: [10.1126/science.103.2684.677](https://doi.org/10.1126/science.103.2684.677).
- [9] Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research; 1960.
- [10] Tesio L, Caronni A, Kumbhare D, et al. Interpreting results from Rasch analysis 1. The “most likely” measures coming from the model. *Disabil Rehabil.* 2023;5:1–13.
- [11] Tesio L, Caronni A, Simone A, et al. Interpreting results from Rasch analysis 2. Advanced model applications and the data-model fit assessment. *Disabil Rehabil.* 2023;46(3):604–617. doi: [10.1080/09638288.2023.2169772](https://doi.org/10.1080/09638288.2023.2169772).
- [12] Caronni A, Zaina F, Negrini S. Improving the measurement of health-related quality of life in adolescent with idiopathic scoliosis: the SRS-7, a Rasch-developed short form of the SRS-22 questionnaire. *Res Dev Disabil.* 2014;35(4):784–799. doi: [10.1016/j.ridd.2014.01.020](https://doi.org/10.1016/j.ridd.2014.01.020).
- [13] Wade DT, Hewer RL. Functional abilities after stroke: measurement, natural history and prognosis. *J Neurol Neurosurg Psychiatry.* 1987;50(2):177–182. doi: [10.1136/jnnp.50.2.177](https://doi.org/10.1136/jnnp.50.2.177).
- [14] Tennant A, Penta M, Tesio L, et al. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. *Med Care.* 2004;42(1 Suppl):I37–48. doi: [10.1097/01.mlr.0000103529.63132.77](https://doi.org/10.1097/01.mlr.0000103529.63132.77).
- [15] Caronni A, Picardi M, Scarano S, et al. Differential item functioning of the Mini-BESTest balance measure: a Rasch analysis study. *Int J Environ Res Public Health.* 2023;20(6):5166. doi: [10.3390/ijerph20065166](https://doi.org/10.3390/ijerph20065166).
- [16] Negrini S, Zaina F, Buyukaslan A, et al. Cross-cultural validation of the Italian Spine Youth Quality of Life questionnaire: the ISYQOL international. *Eur J Phys Rehabil Med.* 2023;59(3):364–376. doi: [10.23736/S1973-9087.23.07586-X](https://doi.org/10.23736/S1973-9087.23.07586-X).
- [17] Hagquist C, Andrich D. Determinants of artificial DIF—a study based on simulated polytomous data. *Psychol Test Assess Model.* 2015;57:342.
- [18] Linacre JM. Sample size and item calibration (or person measure) stability. *Rasch Meas Trans.* 1994; 7:4:328.
- [19] Vaganian L, Boecker M, Bussmann S, et al. Psychometric evaluation of the Positive Mental Health (PMH) scale using item response theory. *BMC Psychiatry.* 2022;22(1):512. doi: [10.1186/s12888-022-04162-0](https://doi.org/10.1186/s12888-022-04162-0).
- [20] Linacre JM. DIF - DPF - bias - interactions concepts. *Winsteps Help for Rasch Analysis* [Internet]. Available from: <https://www.winsteps.com/winman/difconcepts.htm>.
- [21] Wright BD, Douglas GA. Best test design and self-tailored testing. Chicago: Statistical Laboratory, Department of Education, University of Chicago; 1975.
- [22] de Vet HCW, Terwee CB, Mokkink LB, et al. Validity. In: *Measurement in medicine: a practical guide. Practical guides to biostatistics and epidemiology.* Cambridge: Cambridge University Press; 2011. p. 150–201.
- [23] Yi Y, Ding L, Wen H, et al. Is Barthel Index suitable for assessing activities of daily living in patients with dementia? *Front Psychiatry.* 2020;11:282. doi: [10.3389/fpsy.2020.00282](https://doi.org/10.3389/fpsy.2020.00282).
- [24] Yang H, Chen Y, Wang J, et al. Activities of daily living measurement after ischemic stroke: rasch analysis of the modified Barthel Index. *Medicine.* 2021;100(9):e24926. doi: [10.1097/MD.00000000000024926](https://doi.org/10.1097/MD.00000000000024926).
- [25] Granger CV. Comments to differential item functioning of the functional independence measure in high performing neurological patients. *J Rehabil Med.* 2006;38(6):391–392; author reply 393. doi: [10.1080/06501970600799351](https://doi.org/10.1080/06501970600799351).
- [26] Granger CV, Linn RT. Biologic patterns of disability. *J Outcome Meas.* 2000;4(2):595–615.
- [27] Dallmeijer AJ, Dekker J, Roorda LD, et al. Differential item functioning of the Functional Independence Measure in higher performing neurological patients. *J Rehabil Med.* 2005;37(6):346–352. doi: [10.1080/16501970510038284](https://doi.org/10.1080/16501970510038284).
- [28] Lundgren-Nilsson Å, Tennant A, Grimby G, et al. Cross-diagnostic validity in a generic instrument: an example from the Functional Independence measure in Scandinavia. *Health Qual Life Outcomes.* 2006;4(1):55. doi: [10.1186/1477-7525-4-55](https://doi.org/10.1186/1477-7525-4-55).
- [29] Kukull WA, Ganguli M. Generalizability. *Neurology.* 2012; 78(23):1886–1891. June 5 doi: [10.1212/WNL.0b013e318258f812](https://doi.org/10.1212/WNL.0b013e318258f812).
- [30] Simone A, Rota V, Tesio L, et al. Generic ABILHAND questionnaire can measure manual ability across a variety of motor impairments. *Int J Rehabil Res.* 2011;34(2):131–140. doi: [10.1097/MRR.0b013e328343d4d3](https://doi.org/10.1097/MRR.0b013e328343d4d3).
- [31] Caronni A, Sciumè L, Donzelli S, et al. ISYQOL: a Rasch-consistent questionnaire for measuring health-related quality of life in adolescents with spinal deformities. *Spine J.* 2017;17(9):1364–1372. doi: [10.1016/j.spinee.2017.05.022](https://doi.org/10.1016/j.spinee.2017.05.022).
- [32] Caronni A, Picardi M, Redaelli V, et al. The Falls Efficacy Scale International is a valid measure to assess the concern about falling and its changes induced by treatments. *Clin Rehabil.* 2022;36(4):558–570. April doi: [10.1177/02692155211062110](https://doi.org/10.1177/02692155211062110).
- [33] Borsboom D, Mellenbergh GJ, van Heerden J. The concept of validity. *Psychol Rev.* 2004;111(4):1061–1071. October doi: [10.1037/0033-295X.111.4.1061](https://doi.org/10.1037/0033-295X.111.4.1061).
- [34] Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol.* 2010;63(7):737–745. July doi: [10.1016/j.jclinepi.2010.02.006](https://doi.org/10.1016/j.jclinepi.2010.02.006).
- [35] Martinková P, Drabinová A, Liaw Y-L, et al. Checking equity: why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE Life Sci Educ.* 2017;16(2):rm2. doi: [10.1187/cbe.16-10-0307](https://doi.org/10.1187/cbe.16-10-0307).
- [36] Hobart JC, Thompson AJ. The five item Barthel index. *J Neurol Neurosurg Psychiatry.* 2001;71(2):225–230. doi: [10.1136/jnnp.71.2.225](https://doi.org/10.1136/jnnp.71.2.225).
- [37] Hagquist C, Andrich D. Recent advances in analysis of differential item functioning in health research using the Rasch model. *Health Qual Life Outcomes.* 2017;15(1):181. doi: [10.1186/s12955-017-0755-0](https://doi.org/10.1186/s12955-017-0755-0).
- [38] Tesio L. Items and variables, thinner and thicker variables: gradients, not dichotomies. *Rasch Meas Trans.* 2014;28:1477–1479.
- [39] Mallinson T, Kozlowski AJ, Johnston MV, et al. Rasch Reporting Guideline for Rehabilitation Research (RULER): the RULER statement. *Arch Phys Med Rehabil.* 2022;103(7):1477–1486. doi: [10.1016/j.apmr.2022.03.013](https://doi.org/10.1016/j.apmr.2022.03.013).
- [40] Van de Winckel A, Kozlowski AJ, Johnston MV, et al. Reporting guideline for RULER: Rasch reporting guideline for rehabilitation research: explanation and elaboration. *Arch Phys Med Rehabil.* 2022;103(7):1487–1498. doi: [10.1016/j.apmr.2022.03.019](https://doi.org/10.1016/j.apmr.2022.03.019).
- [41] Hobart JC, Cano SJ, Zajicek JP, et al. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *Lancet Neurol.* 2007;6(12):1094–1105. doi: [10.1016/S1474-4422\(07\)70290-9](https://doi.org/10.1016/S1474-4422(07)70290-9).

- [42] Folstein MF, Folstein SE, McHugh PR. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res.* 1975;12(3):189–198. doi: [10.1016/0022-3956\(75\)90026-6](https://doi.org/10.1016/0022-3956(75)90026-6).
- [43] Benedict RH, Zgaljardic DJ. Practice effects during repeated administrations of memory tests with and without alternate forms. *J Clin Exp Neuropsychol.* 1998;20(3):339–352. doi: [10.1076/jcen.20.3.339.822](https://doi.org/10.1076/jcen.20.3.339.822).
- [44] Castiglia SF, Galeoto G, Lauta A, et al. The culturally adapted Italian version of the Barthel Index (IcaBI): assessment of structural validity, inter-rater reliability and responsiveness to clinically relevant improvements in patients admitted to inpatient rehabilitation centers. *Funct Neurol.* 2017;22(4):221–228. doi: [10.11138/fneur/2017.32.4.221](https://doi.org/10.11138/fneur/2017.32.4.221).
- [45] Caronni A, Ramella M, Arcuri P, et al. The Rasch analysis shows poor construct validity and low reliability of the Quebec User Evaluation of Satisfaction with Assistive Technology 2.0 (QUEST 2.0) Questionnaire. *Int J Environ Res Public Health.* 2023;20(2):1036. doi: [10.3390/ijerph20021036](https://doi.org/10.3390/ijerph20021036).
- [46] Rouquette A, Hardouin J-B, Vanhaesebrouck A, et al. Differential Item Functioning (DIF) in composite health measurement scale: recommendations for characterizing DIF with meaningful consequences within the Rasch model framework. *PLoS One.* 2019;14(4):e0215073. doi: [10.1371/journal.pone.0215073](https://doi.org/10.1371/journal.pone.0215073).
- [47] Tesio L, Caronni A, Simone A, et al. Interpreting results from rasch analysis 2. Advanced model applications and the data-model fit assessment.2023. Available from: doi: [10.5281/zenodo.7550135](https://doi.org/10.5281/zenodo.7550135).
- [48] de Morton NA, Keating JL, Davidson M. Rasch analysis of the barthel index in the assessment of hospitalized older patients after admission for an acute medical condition. *Arch Phys Med Rehabil.* 2008;89(4):641–647. doi: [10.1016/j.apmr.2007.10.021](https://doi.org/10.1016/j.apmr.2007.10.021).
- [49] Wyse AE. DIF cancellation in the Rasch model. *J Appl Meas.* 2013;14:118–128.
- [50] Holland PW, Thayer DT. Differential item functioning and the ManTel-Haenszel procedure. *ETS Res Rep Ser.* 1986;1986(2):i–24. doi: [10.1002/j.2330-8516.1986.tb00186.x](https://doi.org/10.1002/j.2330-8516.1986.tb00186.x).
- [51] Kean J, Brodke DS, Biber J, et al. An introduction to Item Response Theory and Rasch Analysis of the Eating Assessment Tool (EAT-10). *Brain Impair.* 2018;19(Spec Iss 1):91–102. doi: [10.1017/Brlmp.2017.31](https://doi.org/10.1017/Brlmp.2017.31).