Article:
Horton, M. orcid.org/0000-0002-6675-7335, Carlton, J. and Andrich, D. (2024) Re: the Upper Extremity Functional Scale for Prosthesis Users (UEFS-P): subscales for one and two-handed tasks. Disability and Rehabilitation. ISSN 0963-8288

https://doi.org/10.1080/09638288.2024.2433652

Letter to the Editor.

Mike Horton[1], Jill Carlton[2], David Andrich[3]

1. Director, Psychometric Laboratory for Health Sciences, University of Leeds, UK
2. Professor of Health Outcomes Research, SCHARR, University of Sheffield, UK
3. Emeritus Professor, The University of Western Australia, Australia

To the Editor,

We recently became aware of the article by Resnik et al., describing the development of the Upper Extremity Functional Scale for Prosthesis Users (UEFS-P): subscales for one and two-handed tasks. [1] We would firstly like to commend the authors on undertaking a thorough development and validation process for the scale, incorporating both qualitative and quantitative methodologies in order to inform the item content and the scale development process. The authors should also be commended on their usage of the Rasch model to assess the functioning of the response structure of the items, among other psychometric attributes. However, despite the strength of these commended elements, we believe there to be some discord between the qualitative and quantitative aspects with regard to the consideration of the item responses.

Whenever we try to create measurement in the social sciences this remains as much a qualitative as it is a quantitative undertaking, and this remains relevant during the application of any statistical measurement model, which is essentially an empty shell when considered without context. [2] By applying the Rasch model, the authors were able to identify that the original response structure of the scale was not functioning in the intended manner. The authors then addressed this lack of functionality by applying various post-hoc score-collapsing options to see which delivered the best fit statistics. However, when applying this psychometric methodology, we do not believe that the authors have fully considered the content and context of the response categories, leading to results that lack meaning, and measurement scales where the total score is uninterpretable.

Within the UEFS-P two-handed task subscale, when asked about completing a particular task, respondents are presented with the following response options: 1 ‘Did not do’; 2 ‘Could not do’; 3 ‘Very difficult’; 4 ‘Difficult’; 5 ‘Easy’; and 6 ‘Very easy’. When considering

the response options of a scale, the ordering requires that successive categories represent successively more (or less) of the underlying trait in question, and this is an *a priori* requirement that is independent of any statistical findings. [3] The lowest response category, 1, represents 'Did not do', but we do not believe that 'Did not do' represents a lower level of functionality than 'Could not do' - it just means that the task wasn't attempted within the given recall period. The response option 'Did not do' relates to the relevance of an item, and it essentially operates in the same way as 'not applicable', but whether a person *did* do something or *could* do something are not mutually exclusive. For example, if I were to consider the item 'Hold a nail to hammer', although I *would* find task this 'Very easy', I would respond 'Did not do', as I haven't had the need to complete this task in the last 4 weeks. By treating 'Did not do' in the way that it has been, this will corrupt the item response structures that the authors are trying to assess, as what is essentially a 'not applicable' response is being included within the active response hierarchy. With this in mind, we would recommend that 'Did not do' should be classified as missing data rather than forming part of the underlying scoring structure, and that the response 'Could not do' should represent the lowest response category within the hierarchy.

Acting on this recommendation will allow for the item set to be properly calibrated, and for the hierarchy of task difficulty to be established without the interference of the non-relevant items. It is acknowledged that this will lead to further issues that will need to be considered, but these issues relate to data quality rather than the methodological approach taken. This data quality issue still relates to the response category 'Did not do', which should perhaps be used as a separate indicator of validity, as this represents the relevance of an item to a target population. If a large proportion of the target population 'did not do' a task, then it is questionable as to whether it is an appropriate item to ask, if the aim is to create a scale that is relevant for all users with a total scale score that is comparable between users. Admittedly, this becomes more complicated when factoring in the use of a prosthesis, as we don't know whether respondents have avoided a task due to having a prosthesis, or whether the task is simply irrelevant for them. Nevertheless, in practical terms, a scale score cannot be confidently estimated if item-level missing data are high. [4] For the UEFS-P, the completion rate among all items (i.e. those who completed or attempted a task) ranged

from 10-66%, and for a standardised scale, even the maximum observed rate of 66% would be considered questionable.

With regard to standardisation, the original UEFS instructed to “Rate the difficulty of doing the activity or attempting to do the activity with your prosthesis in the past two weeks. If you did not do or attempt to do it, please rate how difficult you think it would be to do.” This approach is perhaps not perfect, due to the author-acknowledged challenge in estimating activity difficulty for those without recent experience of a task. However, these original scoring instructions represent an attempt to standardise the score, maximise the data completeness, and enable a direct comparison between respondents. When different people are essentially responding to different sets of items that are relevant to them, these will act as individualised measures. Although individualised measures are not necessarily a bad thing, they do present formidable scientific challenges [5], especially when it comes comparing individuals. Although individual scale scores could be compared within the framework of a Rasch analysis, no direct comparison can be made between their total raw scores.

Another important point that we wish to make also relates to the context of the scoring structure, and the rescoring process that followed. For the UEFS-P Two-handed subscale, the final scoring structure that was selected as the ‘best fitting’ is as follows:

1 = Did not do

2 = Could not do, Very ifficult, difficult, easy

3 = very easy

We have already discussed why ‘Did not do’ is inappropriately included as a response option, but alongside this, the combination of responses treated as a ‘2’ is extremely problematic. Here, someone that ‘cannot do’ a task is being scored *exactly the same* as someone who finds the task ‘easy’. Regardless of what the statistics on their own might say, this is illogical from a face validity perspective, and it therefore cannot be considered useful measurement.

Again, we would urge the authors to reconsider this post-hoc scoring structure, so that the separate response categories represent a distinct progression of the underlying trait that they are intended to reflect. We would also suggest that the response category structure is

discussed with some prosthesis user involvement to ensure that it makes sense from a user perspective. Ideally, any new response structure should also be re-tested using prospectively collected data, but this goes beyond the scope of what can be done with the pre-existing data that is currently available. However, it should be noted that collecting data with one set of categories and then combining them through a post-hoc rescoring process is not the same as collecting data with the newly defined response structure. [6]

With the UEFS-P One-handed subscale, the authors have combined both of the presented issues to create a dichotomous scale that relates to whether a particular task was attempted or not. We believe that this has actually resulted in a measure of task relevance, rather than measuring the function of the prosthesis users, as was intended.

For both the one-handed and two-handed subscales, we believe that it is likely that reclassifying the 'Did not do' category to missing data will have a knock-on effect on the functionality of the original response structure, and therefore the same level of response collapsing may not be necessary.

The intention of this letter is not to belittle or dishearten the authors in any way, but to provide a perspective that we hope will be useful in the continuation of their work. We believe that the authors can easily use their pre-existing data to re-assess the scale, and that taking our points into consideration will hopefully result in a final scale (or scales) that has both qualitative and quantitative value and can provide a functionally useful outcome to prosthesis users.

To elaborate briefly, and perhaps help the authors see our perspective, it appears to us that the authors have shifted from the paradigm of Rasch model application to a paradigm of statistical modelling. In a purely statistical modelling approach, the task is to find a model that fits the data, and that is seen as both necessary and sufficient for the task. Although the authors retained a Rasch model, they modified the data to fit the model and apparently considered the task complete. In a Rasch model application, this approach is chosen because it offers a criterion that a measuring instrument should meet, and as noted earlier, is chosen independently of the data. Although model fit is seen as necessary, it is not sufficient, and the data need to also meet the original substantive requirements to be a valid measuring

instrument. We suggest that the original substantive requirements seem to have been obscured with the search for model-fit.

Disclosure of Interest

None of the authors has any conflict of interest to declare.

Funding

No funding was received by any of the authors.

References:

1. Resnik, L., Borgia, M., Heinemann, A.W., Stevens, P., Clark, M.A., & Ni, P. (2023). The Upper Extremity Functional Scale for Prosthesis Users (UEFS-P): subscales for one and two-handed tasks, Disability and Rehabilitation, 45:22, 3768-3778, DOI: 10.1080/09638288.2022.2138572
2. Salzberger, T. (2023). How to Avoid a Parody of Measurement: Some Models are Wiser Than Others - A Commentary on the Pillars of Measurement Wisdom by George Engelhard, Jr. Journal of Applied Measurement, 23(3/4), 105-112.
3. Andrich, D. (2011). Rating scales and Rasch measurement. Expert Review of Pharmacoeconomics & Outcomes Research, 11(5), 571–585. https://doi.org/10.1586/erp.11.59
4. McHorney, C. A., Ware Jr, J. E., Lu, J. R., & Sherbourne, C. D. (1994). The MOS 36-ltem Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Medical care*, *32*(1), 40-66.
5. Tennant, A. (2007). Goal attainment scaling: Current methodological challenges. Disability and Rehabilitation, 29(20–21), 1583–1588. https://doi.org/10.1080/09638280701618828
6. Andrich, D. (1995). Models for measurement, precision, and the nondichotomization of graded responses. Psychometrika, 60, 7-26.