PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Can manual ability be measured with a generic ABILHAND scale? A cross-sectional study conducted on six diagnostic groups
AUTHORS	Arnould, Carlyne; Vandervelde, Laure; Batcho, Charles; Penta, Massimo; Thonnard, Jean-Louis

VERSION 1 - REVIEW

REVIEWER	Merkies, Ingemar Spaarne Hospital, Department of Neurology
REVIEW RETURNED	13-Aug-2012

GENERAL COMMENTS	This is a very elegant and meticulously described study by a well-experienced group of researchers in the field of measurements. The study demonstrates the difficulty of manual activities being diagnosis dependent, thus emphasizing the importance of implement disease-specific assessments rather than general measures, and therefore serves as a good example of modern assessment methodology using Rasch analyses. In the introduction the difficulties and assumptions made by those using generic measures are well presented, a matter that is seen throughout medical world, not only in neurology and rehabilitation, but (unfortunately) also in other disciplines like orthopaedics, surgery, internal medicine etc. Also, the sample size as a factor that might influence the results is highlighted in the discussion, a matter frequently neglected by researchers. For the purposes of the current study a large cohort was used of patients with chronic illnesses or illnesses leading to chronic activity and participation restrictions. The study is very well described and the results are systematically presented with a thorough back-up of literature findings. I have only a couple of minor questions / suggestions to this paper: 1. The authors have succeeded in constructed a unidimensional scale for the 5 (randomly selected?) groups of illnesses of which some items were split based on the DIF found. Also, the authors have already been involved in creating various disease specific AbilHand measures. My question is: would not it have been more convenient to search for equivalent items in the various already constructed disease-specific AbilHand OMs, subsequently aiming to construct ‘one ruler’ through anchoring (calibrating one measure to the other)-approach? Since we have a quite large amount of other chronic illness (e.g., within neurology Parkinson, MS, epilepsy, lumbar disc herniation etc - or from other disciplines) this approach would be more feasible to incorporate data of future AbilHand other chronic conditions. Perhaps that the authors could comment on this

	in the discussion. 2. Some DIF parameters (symm ⇓◊ asymm) were used, while others were not (e.g. duration of illness or age / gender); could the author comment on this. 3. Were the selected manual ability activities based on the WHO ICF categorization + literature search/reference tracing? 4. Page 10 – last paragraph: “Based upon PCA results,…….and items that did not share a common location between the diagnostic groups.” I have some difficulty understanding the last part of the sentence in relation to splitting the items based on DIF. Please clarify. 5. Regarding Table 3 showing “strong correlations” between generic and disease specific measures: what is the meaning of these correlations found? IF the generic and the disease specific are highly correlated, are not we measuring the same construct with both? Again, I complement the group for the wonderful paper.

REVIEWER	Ching-yi Wu, Professor, Department of Occupational Therapy and Graduate Institute of Behavioral Sciences, Chang Gung University, Taiwan
REVIEW RETURNED	18-Aug-2012

GENERAL COMMENTS	This study attempted to establish a generic ability scale unbiased by diagnosis across various populations. The authors combined DIF tests, PCA, and manual activities categorization about the test nature to develop this generic scale. The findings suggested that both generic and disease-specific scales has its own clinical value ad provides unique information corresponding to specific objectives. There are some comments as follows. This study aimed to develop a generic ability scale and claimed that it allows examination of cost-effectiveness of health interventions across a variety of diseases and disorders while the disease-specific scales should be preferred in daily clinical settings for treatment, planning and follow-up. However, in the Abstract, the conclusion exclusively emphasized the importance of implementing disease-specifi assessments. It seems a little contradictory and inconsistent about the conclusion derived from the obtained results It follows it seems confusing that what is the point and major conclusions based on the results. It is not quite clear either that how the authors explain the advantage about the generic scale. I suggested reorganize the last three paragraphs in the Discussion section to clearly point out the major conclusion of this study Would the author suggest the 11 common items to be constructed to a generic scale or the 52 items? If 52 items are employed across various diagnostic groups, how to explain or apply the results given that the 41 items are category-specific . Patients with five types of disorders with large sample sizes were recruited, which is the strength of this study. However, some of the disorders might be a heterogeneous group. For example, stroke

	patients may cover a variety of different levels of motor, perceptual, or cognitive deficits. Does the sample of one specific disorder used in this study sufficiently represent the disorder? In other words, what are the clinical characteristics for each diagnostic group? The results showed that the generic scale is globally less accurate than the disease-specific scales. The results also showed that manual ability measures of generic and disease-specific scales were highly correlated. The results seem contradictory intuitively. Please elaborate these findings and their implications.

REVIEWER	Stefan Cano Associate Professor of Psychometrics Clinical Neurology Research Group Peninsula College of Medicine and Dentistry Room N13 (ITTC Building 1) Tamar Science Park Davy Road Plymouth PL6 8BX In relation to this manuscript I have no competing interests.
REVIEW RETURNED	23-Aug-2012

THE STUDY	I was pleased to have received this manuscript as I have interest in both the clinical and research areas presented. I have also been aware and very much admire the work that this group has produced for over the last 20 years. Thus, reviewing this manuscript has been difficult for me, and in particular, trying to think of the most productive way of presenting my views back to the authors. I first want to say that I think it's great to see a manuscript of this type submitted to this journal. I was also pleased to see the authors tackle such a complex topic. My overall feeling is that, in order to get the message across in an appropriate way: 1) the text needs to be simplified, modified or better explained; 2) the methods and reasoning behind the study needs to be better argued and signposted; and 3) the main messages and implications for future research better explained. I hope the below comments are helpful: 1. The crux of my issue with the manuscript is this: In his seminal 1960 book, in which is included the origin and development of the simple logistic model (Rasch model), Georg Rasch outlines the cornerstones of a linking design study in an educational setting. I am sure the authors are more than aware of this fact. When I read the current manuscript I assumed that this was, in effect, the same type of research I was reading. But after reading the manuscript several times I struggled to pull out the key messages or follow some of the logical steps. 2. The justification for why the authors embarked on this study is missing for me, and this then has an impact on everything else in the manuscript. Without a clear basis, it almost feels like the authors carried out this study to show that a generic ABILHAND “can’t be done”. Is this just an empirical reaction to Simone et al, 2011? If this is the purpose, perhaps this can be set up a little better.

	3. The whole piece needs to be better signposted and, given the general readership of the journal, some of the terms (eg item difficulty calibrations, DIF, bias, invariance) really need to be explained, perhaps in an appendix glossary, or similar. I think this will help guide readers through the sections. A flow diagram may also help. The Methods need to be more clearly labelled in subsections relating to the steps involved, and the Results really need to directly mirror these. I found the latter particularly difficult to follow. The Discussion section includes the reporting of results, which also made this difficult to read and reconcile with the rest of the manuscript. I also struggled with some of the language. For example: “Data analysis The RUMM2020R Rasch analysis computer program analysed all responses. Manual ability was the only personal attribute theorized to account for the probability of choosing a given response. This requirement, called “unidimensionality”, has been tested using fit statistic indices as described elsewhere.[16, 17] Unidimensionality also requires that patients with identical ability, but different diagnoses, have the same probability of succeeding any particular item. Consequently, the invariance of item difficulty across patient diagnostic groups must be controlled. To investigate the invariance of item difficulty hierarchy, a twoway ANOVA was computed on the standardized residuals. [16, 17] Significant diagnostic main effects represented group differences in item difficulty hierarchy.” 4. If the authors have used RUMM 2020, and therefore followed a "RUMM paradigm" approach, I struggled to understand the role of the PCA analysis (which is outside of the paradigm). And I also struggled with getting to the bottom of (ie implications of, against what criteria) what phrases such as the following were trying to communicate: “The difficulty of most manual activities was diagnosis-dependent and depends on the specificity of the underlying disease since the vast majority (85%) of the difficulty variations observed in manual activities across diagnostic groups was explained by 1) the symmetric or asymmetric nature of the disorder (57% of the variance) and 2) the proximal or distal nature of the disorder (28% of the variance).” 5. Coming back to my first comment, why then did the authors not approach this with a link design, employing anchoring and then examine DIF by clinical sub-group? This could have been achieved iteratively with different disease groups, with further examinations examining and attempting to account for DIF. Thus, for me, the steps carried out need to be made clearer. 6. The sample sizes of the respective clinical groups makes me wonder what would happen if they were larger. I think some of the sub-sample sizes are so small that perhaps the language of the manuscript really should be toned down to reflect this fact. 7. I also wonder about trying to link across adult and children populations, and whether the thing to do would be to try and initially explore doing this in the two separate groups first. Aren't there also other problems with trying to do this? I think the authors need to better set up the justification for this approach.

	8. I think there needs to be Appendices that include the various versions of the ABILHAND or some sort of schema that helps the reader get a better feel for s/he is reading about. 9. Finally I was left confused by some of the key points in manuscript, and once again struggled to pick out the main message of the manuscript. Thus: "The strong correlations (R ≥ 0.94) observed between the generic scale and each of the disease-specific ABILHAND scales supported the assertion that these instruments successfully measure manual ability" "In our study, the generic scale was globally less accurate than the disease-specific scales which often included a greater number of disease-relevant activities." "Nevertheless, using 11 linked items unbiased by diagnosis, we successfully constructed a unidimensional scale common to six diagnostic groups by separating asymmetric from symmetric disorders. This new generic scale that allows the manual ability of patients with different diagnoses to be compared may be used to examine cost-effectiveness of health interventions across a variety of diseases and disorders." "Our findings emphasize the importance of implementing disease-specific assessments (with ABILHAND) and highlight the risk of using generic scales without prior investigation of the diagnosis-invariance of item difficulties."

VERSION 1 – AUTHOR RESPONSE

Reviewer 1: INGEMAR S.J. MERKIES

We address here the comments of the first reviewer.

This is a very elegant and meticulously described study by a well-experienced group of researchers in the field of measurements. The study demonstrates the difficulty of manual activities being diagnosis dependent, thus emphasizing the importance of implement disease-specific assessments rather than general measures, and therefore serves as a good example of modern assessment methodology using Rasch analyses.

In the introduction the difficulties and assumptions made by those using generic measures are well presented, a matter that is seen throughout medical world, not only in neurology and rehabilitation, but (unfortunately) also in other disciplines like orthopaedics, surgery, internal medicine etc. Also, the sample size as a factor that might influence the results is highlighted in the discussion, a matter frequently neglected by researchers.

For the purposes of the current study a large cohort was used of patients with chronic illnesses or illnesses leading to chronic activity and participation restrictions. The study is very well described and the results are systematically presented with a thorough back-up of literature findings.

I have only a couple of minor questions / suggestions to this paper:

1. The authors have succeeded in constructed a unidimensional scale for the 5 (randomly selected?) groups of illnesses of which some items were split based on the DIF found. Also, the authors have already been involved in creating various disease specific AbilHand measures. My question is: would not it have been more convenient to search for equivalent items in the various already constructed disease-specific AbilHand OMs, subsequently aiming to construct 'one ruler' through anchoring (calibrating one measure to the other)-approach? Since we have a quite large amount of other chronic illness (e.g., within neurology Parkinson, MS, epilepsy, lumbar disc herniation etc - or from other disciplines) this approach would be more feasible to incorporate data of future AbilHand other chronic conditions. Perhaps that the authors could comment on this in the discussion.

Although the anchoring approach seems easier at first glance, this method also requires that the selected items fit a unidimensional construct. This involves the following steps (Equating/linking with anchors. *RMT* 2004,**18**(3):993; Ingebo G. Linking tests with the Rasch model. *RMT* 1997,**11**(1):549; Wright B. Anchoring & standard-errors. *RMT* 1993,**6**(4):259): i) selecting 'equivalent' items in the various disease specific scales , ii) verifying that the so called 'equivalent' items do not present significant DIF between various diseases, iii) anchoring the items common to the various diseases and iv) verifying that the additional, disease-specific items, fit with the construct defined by the anchored items. Practically speaking, this can prove as complex as the co-calibration and adjustment for DIF approach followed in our study. So, the anchoring approach requires that 'equivalent' items free of DIF can be found in the various disease-specific scales, which was not the case in our dataset.

In addition, many previous studies have applied the co-calibration approach in the field of rehabilitation (Lundgren-Nilsson A, Tennant A, Grimby G, et al. 2006; Dallmeijer AJ, de Groot V, Roorda LD, et al. 2007; Wann-Hansson C, Klevsgard R, Hagell P 2008; Simone A, Rota V, Tesio L, et al. 2011). The justification for the selected equating method has been added in the analysis process section (page 11, last paragraph).

2. Some DIF parameters (symm asymm) were used, while others were not (e.g. duration of illness or age / gender); could the author comment on this.

The scope of this study was to investigate the DIF across diagnoses. In this respect the nature of the disorder (e.g. symmetric vs. asymmetric, proximal vs. digital) was shown to be the most important sources of DIF (up to 85% of the item difficulty hierarchy variation). Other demographic clinical aspects were also taken into account but show minor effects (e.g. age). Extensive DIF analyses were performed for each disease-specific scale in previous studies, showing the invariance of the different disease-specific ABILHAND scales and indicating that the heterogeneity of the disorder has no impact on the item difficulty hierarchy. This was not the scope of the present study.

3. Were the selected manual ability activities based on the WHO ICF categorization + literature search/reference tracing?

The 83 items was not based on the WHO ICF categorization. All items came from existing scales and were selected as part of the development of disease-specific ABILHAND scales. The items were selected by researchers, disease-experts or patients to extend the range of activities explored by each ABILHAND questionnaire. This information has been added in the Methods (Manual ability measure section; pages 9-10).

4. Page 10 – last paragraph: "Based upon PCA results,…….and items that did not share a common location between the diagnostic groups." I have some difficulty understanding the last part of the sentence in relation to splitting the items based on DIF. Please clarify.

This has been clarified in the Methods (Analysis process section; page 12, lines 8-13).

5. Regarding Table 3 showing "strong correlations" between generic and disease specific measures: what is the meaning of these correlations found? IF the generic and the disease specific are highly correlated, are not we measuring the same construct with both?

Indeed, the strong correlations ($R \geq 0.94$) observed between the generic scale and each of the disease-specific ABILHAND scales indicate that they measure the same construct, namely, manual ability. However, disease-specific scales enable more accurate measures (i.e., patient estimates have lower standard errors) than the generic scale. This has been clarified in the Discussion (page 20, lines 6-12).

Again, I complement the group for the wonderful paper.

Reviewer 2**: Ching-yi Wu**

We address here the comments of the second reviewer.

This study attempted to establish a generic ability scale unbiased by diagnosis across various populations. The authors combined DIF tests, PCA, and manual activities categorization about the test nature to develop this generic scale. The findings suggested that both generic and disease-specific scales has its own clinical value ad provides unique information corresponding to specific objectives.

There are some comments as follows.
This study aimed to develop a generic ability scale and claimed that it allows examination of cost-effectiveness of health interventions across a variety of diseases and disorders while the disease-specific scales should be preferred in daily clinical settings for treatment, planning and follow-up. However, in the Abstract, the conclusion exclusively emphasized the importance of implementing disease-specifi assessments. It seems a little contradictory and inconsistent about the conclusion derived from the obtained results

The conclusion of the abstract has been modified by focusing on the major conclusion of the paper, namely, that it is dangerous to use generic scales without prior investigation of item invariance across diagnostic groups as most of the item difficulties were disease-dependent.

It follows it seems confusing that what is the point and major conclusions based on the results. It is not quite clear either that how the authors explain the advantage about the generic scale. I suggested reorganize the last three paragraphs in the Discussion section to clearly

point out the major conclusion of this study

The last three paragraphs of the discussion (pages 20-21) were revised to bring out the major conclusion of the study, namely, that it is dangerous to use generic scales without prior investigation of item invariance across diagnostic groups as most of the item difficulties were disease-dependent.

Would the author suggest the 11 common items to be constructed to a generic scale or the 52 items? If 52 items are employed across various diagnostic groups, how to explain or apply the results given that the 41 items are category-specific.

The « generic » manual ability scale includes 52 items: 11 items share a common location between disorders and 41 items have a location that is specific to asymmetric vs. symmetric disorders. The 11 "common" items cannot be used alone as a generic scale since they are few in number and they are not responded by all diagnostic groups (one item was responded by all diagnoses, 3 items by 4 diagnoses, 3 items by 3 diagnoses, and 4 items by only 2 diagnoses). A scale including only the 11common items provides a reliability that is quite low (R = 0.80) and would lack sensitivity. We therefore consider that 52 items are needed to constitute the generic scale, using the 11 common items as a link in the calibration of the 41 items with disorder-specific difficulties. A possible means of applying disorder specific items is by implementing the scale calibration on a web site that allows clinicians to exploit the research results without the need for Rasch software. This strategy has already been applied with all ABILHAND scales and others, see www.rehab-scales.org.

Patients with five types of disorders with large sample sizes were recruited, which is the strength of this study. However, some of the disorders might be a heterogeneous group. For example, stroke patients may cover a variety of different levels of motor, perceptual, or cognitive deficits. Does the sample of one specific disorder used in this study sufficiently represent the disorder? In other words, what are the clinical characteristics for each diagnostic group?

The clinical characteristics of the sample have been described in previous studies as well as the inclusion criteria for each study. A summary has been added in Table 1 for clarification (Methods section; Subjects subsection; page 9). Nevertheless, our previous studies have shown, through DIF tests, the invariance of the different disease-specific ABILHAND scales indicating that the heterogeneity of the disorders has no impact on the item difficulty hierarchy.

The results showed that the generic scale is globally less accurate than the disease-specific scales. The results also showed that manual ability measures of generic and disease-specific scales were highly correlated. The results seem contradictory intuitively. Please elaborate these findings and their implications.

The fact that 1) manual ability measures of generic and disease-specific scales are highly correlated, and 2) the generic scale is globally less accurate than the disease-specific scales are not contradictory. The high correlations ($R \geq 0.94$) indicate that both generic and disease-specific scales measure the same construct, namely manual ability. However, disease-specific scales which often included a greater number of disease-relevant activities enable more accurate measures (i.e., patient estimates have lower standard errors) than the generic scale. This is most likely due to the fact that

disease-specific scales have been constructed to maximize their person separation reliability and therefore also their accuracy. This has been clarified in the Discussion (page 20, lines 6-12).

Reviewer 3: **Stefan Cano**

We address here the comments of the third reviewer.

In relation to this manuscript I have no competing interests.

I was pleased to have received this manuscript as I have interest in both the clinical and research areas presented. I have also been aware and very much admire the work that this group has produced for over the last 20 years. Thus, reviewing this manuscript has been difficult for me, and in particular, trying to think of the most productive way of presenting my views back to the authors.

I first want to say that I think it's great to see a manuscript of this type submitted to this journal. I was also pleased to see the authors tackle such a complex topic.

My overall feeling is that, in order to get the message across in an appropriate way: 1) the text needs to be simplified, modified or better explained; 2) the methods and reasoning behind the study needs to be better argued and signposted; and 3) the main messages and implications for future research better explained. I hope the below comments are helpful:

The crux of my issue with the manuscript is this: In his seminal 1960 book, in which is included the origin and development of the simple logistic model (Rasch model), Georg Rasch outlines the cornerstones of a linking design study in an educational setting. I am sure the authors are more than aware of this fact. When I read the current manuscript I assumed that this was, in effect, the same type of research I was reading. But after reading the manuscript several times I struggled to pull out the key messages or follow some of the logical steps.

1. The justification for why the authors embarked on this study is missing for me, and this then has an impact on everything else in the manuscript. Without a clear basis, it almost feels like the authors carried out this study to show that a generic ABILHAND "can't be done". Is this just an empirical reaction to Simone et al, 2011? If this is the purpose, perhaps this can be set up a little better.

The study was not designed in reaction to Simone et al (2011) but was performed to explore the applicability of a generic manual ability scale unbiased by diagnosis across various populations. Setting out this objective, we also intended to improve our understanding of the nature of manual ability and especially its interaction with diagnosis. When using a generic scale, it is implicitly assumed that the item difficulties are invariant across diagnoses. Our finding that most of the manual item difficulties were disease-dependent emphasizes the danger of using generic scales without prior investigation of item invariance across diagnostic groups. Nevertheless, a generic manual ability scale could be developed by adjusting and accounting for activities perceived differently in various disorders. The justifications for the study and the key message have been clarified in the Introduction and Discussion.

2. The whole piece needs to be better signposted and, given the general readership of the journal, some of the terms (eg item difficulty calibrations, DIF, bias, invariance) really need to be explained, perhaps in an appendix glossary, or similar. I think this will help guide readers through the sections. A flow diagram may also help. The Methods need to be more clearly labelled in subsections relating to the steps involved, and the Results really need to directly mirror these. I found the latter particularly difficult to follow. The Discussion section includes the reporting of results, which also made this difficult to read and reconcile with the rest of the manuscript. I also struggled with some of the language. For example: "Data analysis: The RUMM2020R Rasch analysis computer program analysed all responses. Manual ability was the only personal attribute theorized to account for the probability of choosing a given response. This requirement, called "unidimensionality", has been tested using fit statistic indices as described elsewhere.[16, 17] Unidimensionality also requires that patients with identical ability, but different diagnoses, have the same probability of succeeding any particular item. Consequently, the invariance of item difficulty across patient diagnostic groups must be controlled. To investigate the invariance of item difficulty hierarchy, a twoway ANOVA was computed on the standardized residuals. [16, 17] Significant diagnostic main effects represented group differences in item difficulty hierarchy."

Although RUMM2020 has been used to perform the Rasch analyses, the DIF approach was slightly modified in order to address the specific purpose of this study. An original approach was developed and is described step-by-step in the analysis process flow (Analysis process section; pages 11-12). A figure has also been added (Figure 1) in order to clarify the analysis process. The whole manuscript has been revised to clarify the language and technical aspects of Rasch analysis. More details have been provided in the Methods section. Note that the manuscript has also been revised for the editing by Dr. Ann Power Smith of the Write Science Right editing company.

3. If the authors have used RUMM 2020, and therefore followed a "RUMM paradigm" approach, I struggled to understand the role of the PCA analysis (which is outside of the paradigm). And I also struggled with getting to the bottom of (ie implications of, against what criteria) what phrases such as the following were trying to communicate:
"The difficulty of most manual activities was diagnosis-dependent and depends on the specificity of the underlying disease since the vast majority (85%) of the difficulty variations observed in manual activities across diagnostic groups was explained by 1) the symmetric or asymmetric nature of the disorder (57% of the variance) and 2) the proximal or distal nature of the disorder (28% of the variance)."

As indicated in the response to the previous comment, the RUMM paradigm was not strictly followed for the analysis of DIF. The original method proposed is described in the analysis process flow (Analysis process section; pages 11-12) and in Figure1. Essentially, the DIF analysis proposed by RUMM investigates the CI residuals with a 2-factor ANOVA, allowing significant effects to be identified. The effects are either the first factor (i.e. diagnosis) or the second factor (i.e. CI or ability) or their interaction. In this analysis, significant DIF is observed when one or more diagnostic groups do share the same item characteristic curve as the other diagnoses. Nevertheless, this analysis does not allow the identification of the item characteristics that interact with diagnosis. For instance, a unimanual activity may be easier, relative to the other items, for a patient with an asymmetric disorder because the less affected limb can be used without difficulty. This type of interaction is what we were looking for in order to improve the current understanding of the nature of manual ability and its interaction with diagnosis. By contrast, the PCA was performed to identify the potential factors

explaining the variations in item difficulty hierarchy observed across the diagnostic groups. The PCA was performed on the differences between 1) the item difficulty specific to each diagnostic group and 2) the average item difficulty for all diagnoses as these differences reflect disease-specific patterns of item difficulty. So, the PCA is used to understand the nature of the DIF and thus departs from the analysis package proposed by RUMM. The intent and use of the PCA has been clarified in the Methods (Analysis process section; page 12, lines 8-13). The results of the PCA indicate that the first factor is the symmetric or asymmetric nature of the disorder (explaining 57% of the variation in item difficulty between disorders) and the second factor is the proximal or distal nature of the disorder (explaining 28% of the variance). Together, these factors explain 85% of the variance of the difference in item difficulty hierarchy between disorders, and this is the main result of the PCA. We hope this clarifies the key message.

4. Coming back to my first comment, why then did the authors not approach this with a link design, employing anchoring and then examine DIF by clinical sub-group? This could have been achieved iteratively with different disease groups, with further examinations examining and attempting to account for DIF. Thus, for me, the steps carried out need to be made clearer.

A link design is feasible provided that the linking items do not show DIF between all diagnostic groups, which is not the case in our dataset. Another approach, called 'co-calibration'or 'concurrent equating' was followed instead. In the co-calibration approach, all items are merged together as one scale with empty spaces for missing values. Note that many previous studies have also applied the co-calibration approach in the field of rehabilitation (Lundgren-Nilsson A, Tennant A, Grimby G, et al. 2006; Dallmeijer AJ, de Groot V, Roorda LD, et al. 2007; Wann-Hansson C, Klevsgard R, Hagell P 2008; Simone A, Rota V, Tesio L, et al. 2011). The justification for the selected equating method has been added in the analysis process section (page 11, last paragraph) and the analysis steps have been clarified in the same paragraph (page 12) and in the additional Figure 1.

5. The sample sizes of the respective clinical groups makes me wonder what would happen if they were larger. I think some of the sub-sample sizes are so small that perhaps the language of the manuscript really should be toned down to reflect this fact.

Generally, the higher the sample size, the smaller a significant change can be observed. Therefore, if larger samples would be used, the observed DIF would be even more significant if the same tendency is maintained in a larger sample. Concerning the particular sample size of our study, i.e. at least 100 patients per diagnostic group (ranging from 103 patients for CS to 156 for SSc), according to Scott, Fayers, Aaronson, et al. (2009), a uniform DIF of 1 logit, i.e. the approximate amplitude of uniform DIF observed for the items split between symmetric and asymmetric disorders (see Figure 2), in a test containing 10 items or more answered by at least 100 subjects can be detected at a significance level of 0.05 with a power of 95% or more. This indicates that the power of the DIF observed in our study is more than adequate considering the study setup (i.e. test length, sample size and significance level). In addition, according to Linacre (2009), a sample size between 64 and 144 subjects (from the best to the poor targeting on the items) is enough to obtain with a 95% confidence interval that the "true" item difficulty is within 0.5 logit of its reported estimate [i.e. the 95% confidence interval is of 1 logit]. In our study, the obtained standard errors (SE) on items estimates on the generic ABILHAND scale range from 0.09 to 0.56 logits, average 0.20 logits and indicate that the corresponding confidence interval (approx. 4*SE) for the items' difficulty estimates range from 0.36 to 2.24 logits and average 0.80 logits. This indicates that the errors on item estimates observed in our study match the predictions of

Linacre (2009) regarding sample size and targeting. This has been clarified in the Discussion (page 17, second paragraph, lines 7-13; page 20, lines 3-6).

6. I also wonder about trying to link across adult and children populations, and whether the thing to do would be to try and initially explore doing this in the two separate groups first. Aren't there also other problems with trying to do this? I think the authors need to better set up the justification for this approach.

The link between children and adults was initially established in NMD patients (Vandervelde et al. 2007). Essentially, this was done to allow the manual ability of patients to be followed from childhood to adulthood. As far as manual ability is concerned, the main difference between children and adults consists in the selection of manual activities: some activities are common to both children and adults, while some activities are not meaningful to either group and are therefore coded as missing for this group. Nevertheless, missing values are treated without problem in Rasch analysis provided that linking items are available for both children and adults and a single manual ability construct was successfully achieved in NMD children and adults (Vandervelde et al. 2007). In the present study, CP children were also included and adult patients have not been selected on their age. Neuromuscular children and adults have also been considered as two separate groups to avoid a methodological bias due to non-equivalent number of subjects in each group. This also allowed the investigation of the difficulty of manual activities with age. Results show that age influences at least partially the perceived difficulty of manual activities but to a lesser extent than the symmetric/asymmetric and proximal/digital nature of the disorders. A discussion of these results has been added in the manuscript (page 19, from line 10).

7. I think there needs to be Appendices that include the various versions of the ABILHAND or some sort of schema that helps the reader get a better feel for s/he is

reading about.

A supplementary file including the original set of items answered by at least two diagnostic groups has been added.

8. Finally I was left confused by some of the key points in manuscript, and once again struggled to pick out the main message of the manuscript. Thus:
The different key points pointed by the reviewer have been clarified. More emphasis has been put in the main message of the manuscript, namely 1) that most item difficulties were disease-dependent, which emphasizes the danger of using generic scales without prior investigation of item invariance across diagnostic groups and 2) that it is possible to develop a generic manual ability scale by adjusting and accounting for activities perceived differently in various disorders.

“The strong correlations (R ≥ 0.94) observed between the generic scale and each of the disease-specific ABILHAND scales supported the assertion that these instruments successfully measure manual ability” “In our study, the generic scale was globally less accurate than the disease-specific scales which often included a greater number of disease-relevant activities.”

The strong correlations (R ≥ 0.94) observed between the generic scale and each of the disease-specific ABILHAND scales indicate that they measure the same construct, namely, manual ability. However, disease-specific scales enable more accurate measures (i.e., patient estimates have lower standard errors) than the generic scale. This has been clarified in the Discussion (page 20, lines 6-12).

"Nevertheless, using 11 linked items unbiased by diagnosis, we successfully constructed a unidimensional scale common to six diagnostic groups by separating asymmetric from symmetric disorders. This new generic scale that allows the manual ability of patients with different diagnoses to be compared may be used to examine cost-effectiveness of health interventions across a variety of diseases and disorders."

The sentence was clarified to indicate that the items that presented difficulties specific to asymmetric and to symmetric disorders were separated (i.e., split): "Nevertheless, using 11 linked items unbiased by diagnoses, we successfully constructed, from a metric point of view, a unidimensional scale common to six diagnostic groups by separating items with difficulties specific to asymmetric and to symmetric disorders" (Discussion; page 20, lines 1-3). This has been emphasised in the abstract and in the discussion.

"Our findings emphasize the importance of implementing disease-specific assessments (with ABILHAND) and highlight the risk of using generic scales without prior investigation of the diagnosis-invariance of item difficulties."

The abstract has been clarified by focusing on the major conclusion of the paper, namely, that it is dangerous to use generic scales without prior investigation of item invariance across diagnostic groups as most of the item difficulties were disease-dependent.

VERSION 2 – REVIEW

REVIEWER	Ching-yi Wu Professor Chang Gung University Taiwan
REVIEW RETURNED	01-Oct-2012

GENERAL COMMENTS	I have no further questions based upon authors' responses.

REVIEWER	Stefan Cano Associate Professor of Psychometrics Peninsula College of Medicine and Dentistry UK
REVIEW RETURNED	24-Sep-2012

GENERAL COMMENTS	I would like to thank the authors for the responses to my queries.