## PEER REVIEW HISTORY



## ARTICLE DETAILS

| | |
|---|---|
| **TITLE (PROVISIONAL)** | Psychometric evaluation of the modified Kessler seven-item version (K7) for measuring psychological distress using Rasch analysis. A cross-sectional study in a rural district of Bangladesh. |
| **AUTHORS** | Uddin, Mohammed; Islam, Fakir |

## VERSION 1 - REVIEW

| | |
|---|---|
| **REVIEWER** | Kaj Sparle Christensen<br>Institute of Public Health<br>Aarhus University<br>Denmark |
| **REVIEW RETURNED** | 19-Oct-2019 |

| | |
|---|---|
| **GENERAL COMMENTS** | Authors have previously validated the K10 using Rasch analysis and suggest a modified version including 7 items and four response categories, the K7.<br><br>The idea with this paper is to field-test the construct validity of the K7 using Rasch analysis, as applied in Rumm2030.<br><br>First of all I would like to commend the authors for bringing the research in this field an important step further.<br><br>Overall the paper needs a thorough revision of the English writing to be acceptable for publication.<br><br>Abstract: In the results section authors claim that results showed good overall fit (to the Rasch model, I assume). However, the fit statistics revealed a p-value of 0.0245. Many people would consider a p-value less than 5% an indication of misfit. Please explain.<br><br>Background section: Only the K10 and K7 is mentioned. What about the K6? Wouldn't it would be fair to include this abbreviation of the K10 in the background section?<br><br>Material and Methods: Page 9, line 207: 'For a scale to be unidimensional, under 5% of the t-tests ought to be significant,..'. What t-tests do you refer to? Please specify.<br><br>Results:<br>Page 11, line 239: please elaborate on the 'good overall fit' with a p-value of 0.0245. |

| | |
|---|---|
| | Figure 2: please attach a title to the figure and sort items in location order as well.<br>Page 12, line 248: please elaborate on 'no misfit of items were identified' - according to Table 3, item 5 reveals a p-value of 0.011?<br>Please include and comment on an item-person distribution map.<br>You may consider moving Table 5 to the extra material section.<br><br>Discussion: I suggest that the discussion should compare the findings with studies on K6 and K10.<br><br>Conclusion: authors may recommend the K7 to be used in rural Bangladesh, I don't think the study can do so. |

| | |
|---|---|
| **REVIEWER** | Ze Lu (Steve)<br>Clinical Research Lab,Roth\|McFarlane Hand and Upper Limb Centre, St. Joseph's Health Centre<br>Canada |
| **REVIEW RETURNED** | 05-Dec-2019 |

| | |
|---|---|
| **GENERAL COMMENTS** | Line 31<br>During the recruitment, your study group did use a strategy to control the percentage of males and females. That was not a strict random sampling.<br><br>Line 128<br>Social demographic information was listed in the result section.<br>I am wondering whether or not you need to introduce the national population here.<br><br>Line 148<br>Why do you need to interview the participants for your data collection? Is K7 a patient-reported outcome measure? Please be precise about the word here. If you just teach subject to finish K7 using CommCare, that is not an interview.<br>Different methods to administrate the questionnaire may influence the study result.<br><br>Line 198<br>The item-trait interaction result provided by RUM2030 doesnot include the Bonferroni correction for this overall item fit statistic. Please add another reference to support your correction instead of the general introduction of Bonferroni. Is there any other published Rasch paper that performed this correction for the overall item fit?<br><br>Line 204<br>Change 'not ready' to 'not able'.<br><br>Line 212-213<br>Please elaborate on the PSI. It has its own definition and interpretation.<br><br>Line 224-226<br>Please delete 'Although the results are being published in peer-reviewed journals', this is confusing.<br><br>Line239-240 |

| | |
|---|---|
| | Your model fit is based on the Bonferroni correction. Please check my feedback on line 198<br><br>Line 242<br>Please rephrase the words 'no problem'.<br><br>Line 249-250<br>The interpretation of PSI is different from Cronbach's alpha. Please explain your result 0.85 in detail.<br><br>Line 344<br>Why the single occasion is the limitation of your study? Do you mean cross-sectional data?<br>Do you recommend the future researchers should include the longitudinal data? There is a problem when you include the longitudinal data into Rasch analysis since you may introduce the within-subject variation into your analysis besides the original between-subject variation.<br><br>Other recommendations<br>Please also include the person-item distribution in your result. |

| | |
|---|---|
| **REVIEWER** | Thach Tran<br>Monash University<br>Australia |
| **REVIEW RETURNED** | 10-Dec-2019 |

| | |
|---|---|
| **GENERAL COMMENTS** | This manuscript reported a Rasch analysis of the K7 in rural Bangladesh. I have two major concerns:<br>1. The authors have published similar data from a previous study (Uddin at al. BMJ Open, 2018. 8(6): e022967). Therefore, this manuscript adds very little to the knowledge. They aimed to 'evaluate the appropriateness of the modified K7 scale … in CLINICAL settings'. However, the sample was a community sample that was similar to the previous study. If they used a sample from clinical settings, that would add something new and achieve what they aimed to.<br>2. This manuscript is really difficult to read for most of the readers of the BMJ Open. It would be more suitable to be published in a psychometrics journal. Rasch analysis, many technical terms, and results were not explained adequately.<br><br>Detail comments:<br><br>Abstract<br>1. 'Validation' is a broad term, should be more specific on what kind of validation.<br>2. 'DIF adjustment' should be spelled out and explained.<br>3. What does it mean by 'different cohort'?.<br>4. 'SPSS25 … for analyses' can be removed from the Abstract.<br>5. 'The K7 showed adequate internal consistency, reliability…'. Internal consistency is a kind of reliability.<br>6. 'Results showed good overall fit, as indicated by a non-significant item-trait interaction (($\chi 2$ = 44.54, df = 28, p =0.0245)'. P-value < 0.05 showing it is significant?<br>7. 'FURTHER validation of K7 in DIFFERENT population'. 'FURTHER' and 'DIFFERENT' are confusing to readers. |

| | |
|---|---|
| | 8. The author concluded that 'the tool is psychometrically robust and suitable for routine measure of psychological distress and thus provides an effective screening instrument among the rural Bangladeshi population'. It should be toned down the conclusion because the Rash analysis provides no evidence on 'routine measure of psychological distress' or 'effective screening instrument'.<br><br>Background<br>9. There was no concrete rationale for this study to conduct stated.<br><br>Methods<br>10. The sample size was calculated based on 'item difficulty' but this study did not report item difficulty.<br>11. The authors stated that 'large sample sizes can result in type 1 errors'. However, 320 would not be much different to 300. Please specify how 300 participants included in the analysis were drawn from 320 participants?<br>12. DIF analyses were conducted by several sociodemographic characteristics. Please have a statement about the power for those analyses. DIF statistical significance is influenced by the sizes of sub-groups.<br>13. What are the criteria of DIF categories?<br>14. Please specify how sociodemographic characteristics were collected.<br>15. How age, education, socio-economic conditions were categorised in the DIF analyses?<br><br>Results<br>16. Table 3 – p-value of item 5 < 0.05. Please interpret it.<br>17. Table 4 – Please spell out MS, F, DF, Prob.<br>Discussion<br>18. Paragraph 3: 'It has commonly been assumed that, deduction of items from the scale would reduce at least some redundancy [48-51]. However, our examination recognised that Cronbach's alpha for the K7 scale (0.89) was comparable to the earlier validated K7 Cronbach's alpha (0.88)'. I cannot understand this statement because both are the same K7 scale. |

**VERSION 1 – AUTHOR RESPONSE**

Reviewer: 1
Reviewer Name: Kaj Sparle Christensen
Institution and Country:
Institute of Public Health
Aarhus University
Denmark
Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below
Authors have previously validated the K10 using Rasch analysis and suggest a modified version including 7 items and four response categories, the K7.

The idea with this paper is to field-test the construct validity of the K7 using Rasch analysis, as applied in Rumm2030.

First of all I would like to commend the authors for bringing the research in this field an important step further.

Overall the paper needs a thorough revision of the English writing to be acceptable for publication.

Thank you for the comments, the manuscript has now been edited by an English language editor.

Abstract: In the results section authors claim that results showed good overall fit (to the Rasch model, I assume). However, the fit statistics revealed a p-value of 0.0245. Many people would consider a p-value less than 5% an indication of misfit. Please explain.

Thanks for your comments, “To take into account of multiple testing Bonferroni corrections were applied to adjust the chi-square p-value (Bland & Altman, 1995). The p-values are compared with the Bonferroni corrected p-value 0.007 (0.05/7=0.007). The p-value 0.007 means the significance at level 0.05 because the number of items is seven (0.05/7=0.007). Therefore, any p-values greater than 0.007 would considered to be non-significant. Many people would consider it significant however with Bonferroni correction which was frequently used in Rasch analysis it was non-significant p-value.

Background section: Only the K10 and K7 is mentioned. What about the K6? Wouldn't it would be fair to include this abbreviation of the K10 in the background section?

Thanks for your comments. K6 has now been incorporated in the background as well as in the discussion sections. (line 99-114 page 5 and 3 line 328-343 on pages 16-17)

Material and Methods: Page 9, line 207: 'For a scale to be unidimensional, under 5% of the t-tests ought to be significant,..'. What t-tests do you refer to? Please specify.

Thank you for the comments. We have restructured the sentence for a clearer understanding. “The independent t-test has been used to determine the difference between the two subsets identified from the principal component analysis (PCA) of residuals”. (line 230-34 page 10)

Results:
Page 11, line 239: please elaborate on the 'good overall fit' with a p-value of 0.0245.

Thanks for your comments. we have mentioned earlier, the p-value 0.007 means the significance at level 0.05 because the number of items is seven (0.05/7=0.007). Therefore, any p-values greater than 0.007 would considered to be non-significant. (line 269-70 page 12)

Figure 2: please attach a title to the figure and sort items in location order as well.

Thank you for your suggestion. This has now been included in the manuscript and figure 2 has been sorted as location order.
.
Page 12, line 248: please elaborate on 'no misfit of items were identified' - according to Table 3, item 5 reveals a p-value of 0.011?

Thanks for your comments. Comparing the p-value 0.011 after Bonferroni correction (0.05/7=0.007), the item 5 p-value is greater than 0.007. So, it was considered non-significant Chi-square p-value according to Bonferroni adjustment. (line 281-82 page 13)

Please include and comment on an item-person distribution map.

Thanks for your suggestion. This has been included in the manuscript and titles as ‘figure 5’ and interpretation was included in the result section of the manuscript. (line 305-309 pages 14-15)

You may consider moving Table 5 to the extra material section.

Thank you for the suggestion. Table 5 has been moved to the appendix as recommended. (line 299 page 14)

Discussion: I suggest that the discussion should compare the findings with studies on K6 and K10.

Thanks for your suggestions, we have now compared our findings with the K6 and K10 findings. (lines 328-343 page 16)

Conclusion: authors may recommend the K7 to be used in rural Bangladesh, I don't think the study can do so.

Thanks for your comments. This is the first study to validate the K7 version in rural Bangladesh. Uddin et al. first validated the K10 in a rural district (Narial) in Bangladesh using a Rasch analysis and proposed some modifications to fulfils the Rasch properties. An adjustment has reduced three items from K10 and formed a 7-item K7 to be used in rural Bangladesh. Further validation was required in different rural populations to check whether the newly developed K7 was valid or not. Therefore, the K7 was tested in a different community. The study was not validated in clinical settings because first we needed to validate its cross-cultural validation in rural Bangladesh before validating in a clinical setting. We have now completed the construct validity of the K7 scale, following all the assumptions of Rasch theory. The author believes that we can recommend the K7 for use in rural Bangladesh.

Reviewer: 2
Reviewer Name: Ze Lu (Steve)
Institution and Country:
Clinical Research Lab,Roth|McFarlane Hand and Upper Limb Centre, St. Joseph’s Health Centre Canada
Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below

Line 31 During the recruitment, your study group did use a strategy to control the percentage of males and females. That was not a strict random sampling.

Thanks for your comments, yes that was not strictly a random sampling technique. However, to maintain an approximately equal number of male and female participants, one female was interviewed immediately after a male participant and in the same way one adult was interviewed immediately after an elderly participant. We agree that an adult could appear immediately after an adult, but we have managed the same procedure, male after female and adult after elderly.

Line 128 Social demographic information was listed in the result section.
I am wondering whether or not you need to introduce the national population here.

Thanks for your comments. We have added more information of national population data (line 142-45 pages 6-7)

Line 148 Why do you need to interview the participants for your data collection? Is K7 a patient-reported outcome measure? Please be precise about the word here. If you just teach subject to finish K7 using CommCare, that is not an interview. Different methods to administrate the questionnaire may influence the study result.

Thanks for your comments. In Bangladesh the majority of the surveys needed the interview-administered method to collect data directly from respondents due to the literacy level is not very high among rural population, which is currently 60%. Interview-administered method assisted the respondents to understand the meaning of the items. It was not a patient-reported outcome. We have used tablets or mobile phone devices instead of a paper format. In our understating introduce of tablet or mobile phone devices to collect data would not influence the study results.

Line 198 The item-trait interaction result provided by RUM2030 does not include the Bonferroni correction for this overall item fit statistic. Please add another reference to support your correction instead of the general introduction of Bonferroni. Is there any other published Rasch paper that performed this correction for the overall item fit?

Thanks for your comments. Yes, many published papers used Bonferroni correction for overall as well as for individual item fits. We have mentioned some as referenced below:
(Chachamovich, Fleck, & Power, 2010; Chachamovich, Fleck, Trentini, & Power, 2008; Pallant & Tennant, 2007; Pomeroy, Tennant, & Young, 2013; Van de Winckel, Feys, Lincoln, & De Weerdt, 2007; Verheyden & Kersten, 2010)

Line 204 Change ‘not ready’ to ‘not able’.

Thanks for your comments, this has been changed to “not able”. (line 228 page 10)

Line 212-213 Please elaborate on the PSI. It has its own definition and interpretation.

Thanks for your comments, PSI definition has been elaborated as follows “The PSI of the Rasch analysis consists of indices developed as an approximation of the proportion of the true or error-free variance. This applies throughout the distribution of person estimates relative to the sum of this variance and error variance in these estimates. With Rasch measurement, instead of reliability indices, the person separation index is used.” (line 243-248 page 11).

Line 224-226 Please delete ‘Although the results are being published in peer-reviewed journals’, this is confusing.

Thanks for your comments. We have deleted the sentence from the manuscript. (line 257 page 11)

Line239-240 Your model fit is based on the Bonferroni correction. Please check my feedback on line 198

Thanks for the comments. This has been already addressed and reference included. (line 222-23-24 and line 242 on page 10-11)

Line 242 Please rephrase the words ‘no problem’.

Thanks for your comments. “No problem” was rephrased to “suggesting that the respondents have no difficulty differentiating between the response’s choices with the 4-point liker-type scale used in the K7 scale”. (line 272-73 page 12)

Line 249-250The interpretation of PSI is different from Cronbach's alpha. Please explain your result 0.85 in detail.

Thanks for your comments. Cronbach's alpha has been explained differently "The value of the Cronbach's alpha (0.89) of the K7 scale demonstrates good internal consistency" (line 284-85 page 13)

Line 344 Why the single occasion is the limitation of your study? Do you mean cross-sectional data? Do you recommend the future researchers should include the longitudinal data? There is a problem when you include the longitudinal data into Rasch analysis since you may introduce the within-subject variation into your analysis besides the original between-subject variation.

Thanks for your very good comments. No, we did not recommend using a longitudinal data. We would like to say single occasion data is a barrier for generalisation of the scale. This is possible data collection from many occasions could demonstrate some different conclusions. Therefore, single occasion is considered to be a limitation.

Other recommendations
Please also include the person-item distribution in your result.

Thanks for your comments. This has been included in the result section as figure 5 and commented. (line 305-309 pages 14-15).

Reviewer: 3
Reviewer Name:Thach Tran
Institution and Country:
Monash University
Australia
Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below
This manuscript reported a Rasch analysis of the K7 in rural Bangladesh. I have two major concerns:
1.The authors have published similar data from a previous study (Uddin at al. BMJ Open, 2018. 8(6): e022967). Therefore, this manuscript adds very little to the knowledge. They aimed to 'evaluate the appropriateness of the modified K7 scale … in CLINICAL settings'. However, the sample was a community sample that was similar to the previous study. If they used a sample from clinical settings, that would add something new and achieve what they aimed to.

Thank you for the comments. This is the first study to validate the K7 version in rural Bangladesh. Uddin et al. first validated the K10 in a rural district Narial in Bangladesh using a Rasch analysis and proposed some modifications to fulfils the Rasch properties. An adjustment has reduced three items from K10 and formed a 7-item K7 to be used in rural Bangladesh. Further validation was required in different rural populations to check whether the newly developed K7 was valid or not. Therefore, the K7 was tested in a different community. The study was not validated in clinical settings because first we need to validate its cross-cultural validation in rural Bangladesh before validating in a clinical setting.

2.This manuscript is really difficult to read for most of the readers of the BMJ Open. It would be more suitable to be published in a psychometrics journal. Rasch analysis, many technical terms, and results were not explained adequately.

Thanks for your comments, the “Psychometric evaluation of an interview-administered version of the Kessler 10-item questionnaire (K10) for measuring psychological distress in rural Bangladesh” which was published in BMJ open and has reached 9 citations. We have now incorporated a more detailed explanation of the technical term and included an explanation in the result section which will be more readable for the researcher.

Detail comments:

3. Abstract
‘Validation’ is a broad term, should be more specific on what kind of validation.

Thanks for your comments. This is to field-test the construct validity of the K7 using Rasch analysis.

‘DIF adjustment’ should be spelled out and explained.

Thanks for your comments, this has been spelled out. (line 35 page 2)

What does it mean by ‘different cohort’?.

Thanks for your comments. We have validated earlier in Narial and now we have collected the data form another location in Narial which was not included in the earlier data collection of 2017.

’‘ SPSS25 … for analyses’ can be removed from the Abstract.

Thanks for your comments, this has been removed from abstract. (line 37 page 2)

‘The K7 showed adequate internal consistency, reliability…’. Internal consistency is a kind of reliability.

Thank you for the comment, internal consistency has been removed. (line 42 page 2)

‘Results showed good overall fit, as indicated by a non-significant item-trait interaction ((χ2 = 44.54, df = 28, p =0.0245)’. P-value < 0.05 showing it is significant?

Thanks for your comments. “To take into account of multiple testing Bonferroni corrections were applied to adjust the chi-square p-value (Bland & Altman, 1995). The p-values are compared with the Bonferroni corrected p-value 0.007 (0.05/7=0.007). The p-value 0.007 means the significance at level 0.05 because the number of items is seven (0.05/7=0.007). Therefore, any p-value greater than 0.007 would considered to be non-significant (Chachamovich et al., 2010; Chachamovich et al., 2008; Pallant & Tennant, 2007; Pomeroy et al., 2013; Van de Winckel et al., 2007; Verheyden & Kersten, 2010). (line 39 page 2)

‘FURTHER validation of K7 in DIFFERENT population’. ‘FURTHER’ and ‘DIFFERENT’ are confusing to readers.

Thanks for your comments, the term ‘further’ and ‘different’ population both were removed from the manuscript. (line 46 page 2)

The author concluded that ‘the tool is psychometrically robust and suitable for routine measure of psychological distress and thus provides an effective screening instrument among the rural Bangladeshi population’. It should be toned down the conclusion because the Rash analysis provides no evidence on ‘routine measure of psychological distress’ or ‘effective screening instrument’.

Thanks for your comments. We have now toned down our conclusion as recommended “the tool is suitable for measuring psychological distress among the rural Bangladeshi population”. (line 46 page 2)

Background

There was no concrete rationale for this study to conduct stated.

Thanks for your comments. We have now rewritten the background section incorporating the K6 and other studies. (line 86-114 pages 4-5)

Methods

The sample size was calculated based on ‘item difficulty’ but this study did not report item difficulty.

Thanks for your comment. The concern with fit statistics in the Rasch analysis is that the greater the sample size, the higher the likelihood of finding the probability of detecting deviations from the Rasch model (D. Andrich, 2010; Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008). Nevertheless, there are no clear guidelines for sample size when implementing the Rasch Measurement Theory (David Andrich & Marais, 2019). Thus, we used the sample size of 300, which is more appropriate (Smith et al., 2008). We put this as a limitation of the study (line 380-84 page 17).

The authors stated that ‘large sample sizes can result in type 1 errors’. However, 320 would not be much different to 300. Please specify how 300 participants included in the analysis were drawn from 320 participants?

Thanks for your comments. In previous response we have mentioned the sample size concern in Rasch analysis. However, we have restructured the methods section as “Interviewers used a mobile data collection platform CommCare on their android phone to collect data from the respondents. To mitigate the effect of selection bias, 320 respondents were used with an equal proportion of adults and older adults, further partitioned into gender. This study excluded 20 participants randomly as 300 participants were deemed sufficient for the Rasch Measurement Theory”. (line 166-70 page 7-8)

DIF analyses were conducted by several sociodemographic characteristics. Please have a statement about the power for those analyses. DIF statistical significance is influenced by the sizes of sub-groups.

Thanks for your comments. We agree that the small number of participants in the subgroup makes our claims weaker and that the models are suitable for different levels of education, age, gender and socioeconomic conditions. However, we have broadly classified education into two groups (no education 135 vs have some education 165), gender (male 150 vs female 150), age (adult 150 vs elderly 150), SES (low 221 vs high 79) to check DIF. There was no significant DIF observed. Although there may still be an association between these factors and psychological distress in the literature, the lack of DIF means in our research that the domains function in the same way regarding their DIF categories. Furthermore, in table 1, we have reported the socio-economic variables in detail with their group categories. However, when we did DIF analysis we applied only the broad categories of the variables.

What are the criteria of DIF categories?

Thanks for your comments. We have addressed this in earlier comments.

Please specify how sociodemographic characteristics were collected.

Thanks for your comments. We have collected data through an interview administer method with survey questionnaire.

How age, education, socio-economic conditions were categorised in the DIF analyses?

Thanks for your comments, all the socio-economic variable we have categorised into two broad categories as we mentioned in earlier response.

Results

Table 3 – p-value of item 5 < 0.05. Please interpret it.

Thanks for your comments. Comparing the item 5 p-value 0.011 which is <0.05 and seemed to be significant. However, after Bonferroni correction for seven items, the adjusted p-value would be 0.007 (0.05/7=0.007) which is equivalent to 0.05 significance level. So, any p-value >0.007 was considered non-significant Chi-square p-value.

Table 4 – Please spell out MS, F, DF, Prob.

Thanks for your comments. All abbreviations have been spelled out and included in a table legend (line 291 and 302-03 on page 13-14).

Discussion
Paragraph 3: 'It has commonly been assumed that, deduction of items from the scale would reduce at least some redundancy [48-51]. However, our examination recognised that Cronbach's alpha for the K7 scale (0.89) was comparable to the earlier validated K7 Cronbach's alpha (0.88)'. I cannot understand this statement because both are the same K7 scale.

Thanks for your comments. We have deleted the sentence and restructured the whole paragraph (line 336-343 on page 16).

Reference:

Andrich, D. (2010). Understanding the Response Structure and Process in the Polytomous Rasch Model. Handbook of Polytomous Item Response Theory Models, 123-152. Retrieved from ://WOS:000289781300007
Andrich, D., & Marais, I. (2019). A course in Rasch measurement theory: Measuring in the educational, social and health sciences: Springer.
Chachamovich, E., Fleck, M. P., & Power, M. (2010). Is Geriatric Depression Scale-15 a suitable instrument for measuring depression in Brazil? Results of a Rasch analysis. Psychology Health & Medicine, 15(5), 596-606. doi:10.1080/13548506.2010.487108
Chachamovich, E., Fleck, M. P., Trentini, C., & Power, M. (2008). Brazilian WHOQOL-OLD Module version: a Rasch analysis of a new instrument. Revista De Saude Publica, 42(2). Retrieved from ://WOS:000255682000017

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). British Journal of Clinical Psychology, 46, 1-18. Retrieved from ://WOS:000244295300001
Pomeroy, I. M., Tennant, A., & Young, C. A. (2013). Rasch Analysis of the Whoqol-Bref in Post Polio Syndrome. Journal of Rehabilitation Medicine, 45(9), 873-880. doi:10.2340/16501977-1186
Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. Bmc Medical Research Methodology, 8, 33. doi:10.1186/1471-2288-8-33
Van de Winckel, A., Feys, H., Lincoln, N., & De Weerdt, W. (2007). Assessment of arm function in stroke patients: Rivermead Motor Assessment arm section revised with Rasch analysis. Clinical Rehabilitation, 21(5), 471-478. doi:10.1177/0269215507071783
Verheyden, G., & Kersten, P. (2010). Investigating the internal validity of the Trunk Impairment Scale (TIS) using Rasch analysis: the TIS 2.0. Disability and rehabilitation, 32(25), 2127-2137. doi:10.3109/09638288.2010.483038

## VERSION 2 – REVIEW

| | |
|---|---|
| **REVIEWER** | Kaj Sparle Christensen<br>Institute of Public Health<br>Aarhus University<br>Denmark |
| **REVIEW RETURNED** | 13-Jan-2020 |

| | |
|---|---|
| **GENERAL COMMENTS** | Authors have responded adequately to my concerns. |

| | |
|---|---|
| **REVIEWER** | Ze Lu(Steve)<br>Clinical Research Lab, Roth\|McFarlane Hand and Upper Limb Centre, St. Joseph's Health Centre, Canada |
| **REVIEW RETURNED** | 15-Jan-2020 |

| | |
|---|---|
| **GENERAL COMMENTS** | All my concerns were addressed in the revision. |