NursingOpen
Open Access

Comparing questionnaires across cultures: using Mokken scaling to compare the Italian and English versions of the MOLES Index

Journal:	*Nursing Open*
Manuscript ID	NOP-2018-Jan-0019.R1
Wiley - Manuscript type:	Research Article
Search Terms:	Cancer, Public Health Nursing, Mokken Scaling

Abstract:

Aim The aims of this study were: i) to translate the MOLES index from English to Italian and to compare the two versions using non-parametric item response theory.

Design A online survey was used to gather data.

Methods Forward and back-translation was used to prepare the Italian version of the MOLES which was then analysed using the non-parametric item response theory of Mokken scaling.

Results Mokken scales were found in both the English and the Italian versions of the MOLES index. However, the two scales—while the total scale score was not significantly different—showed different properties, and Mokken scaling selected different items from each scale.

Comparing questionnaires across cultures: using Mokken scaling to compare the Italian and English versions of the MOLES Index

ABSTRACT

Aim The aims of this study were: i) to translate the MOLES index from English to Italian and to compare the two versions using non-parametric item response theory.

Design A online survey was used to gather data.

Methods Forward and back-translation was used to prepare the Italian version of the MOLES which was then analysed using the non-parametric item response theory of Mokken scaling.

Results Mokken scales were found in both the English and the Italian versions of the MOLES index. However, the two scales—while the total scale score was not significantly different—showed different properties, and Mokken scaling selected different items from each scale.

INTRODUCTION

A range of methods exist to study the dimensional properties of questionnaires. Such dimensions are known as 'latent' traits as, essentially, they are hidden within the items of questionnaires and may not be obvious without specific multivariate analysis or, when they are purported to exist, require specific multivariate analysis to demonstrate this. A simple example of a commonly used questionnaire that has demonstrable dimensions is the HADS (Hospital Anxiety and Depression Scale). The HADS is comprised of 14 items and seven of these purportedly measure depression as distinct from the seven items that purportedly measure anxiety. Indeed, the two-dimensional nature of the HADS can be demonstrated by appropriate multivariate analysis which, in the case of the HADS is factor analysis.

A range of multivariate techniques exists to study the dimensional properties of questionnaires and these fall under two broad umbrellas: classical test theory (CTT) and item response theory (IRT) and these will, briefly, be considered. Classical test theory is, essentially, based on correlation—a measure of the common variance between two or more variables. Therefore, multivariate statistical techniques such as Cronbach's alpha, principal components analysis and factor analysis—both exploratory and confirmatory—fall under this umbrella. Factor analysis, of which there is a range of similar methods, is the method mainly used to establish dimension in questionnaires and it can be used to examine whether or not there are underlying dimensions to questionnaires (exploratory factor analysis) or to test whether or not an hypothesised set of dimensions exists in a questionnaires (confirmatory factor analysis).

An alternative set of methods exists to study the dimensional nature of questionnaires and these fall under the umbrella of IRT. These methods are so-called because, rather than analysing the relationship between items, they primarily analyse the behaviour of individual items and, based on their properties, they then investigate how they relate to other items. However, individual items must meet certain minimum criteria—to be discussed—to be included in a questionnaire. IRT can be seen to offer some advantages over CTT in that they establish a more precise relationship between the score on an item and the score on the latent trait. In other words, while items will respond across the whole range of a latent trait, they will most accurately measure a region of the latent trait. For example, take two items purporting to measure depression: 1. ‘I do not feel it is worth getting out of bed in the morning’ and 2. ‘I feel like ending my life’. Clearly, both are related to depression but item 1 measures a much lower range of the latent trait of depression than item 2 which represents a more serious level of danger to the individual. IRT posits that the relationship between the sore on an item and the score on the latent trait is stochastic, in other words based on probability and, in the case of the items above, there is a much higher probability that someone will score high on item 1 before they score on item 2. This indicates another aspect of IRT which follows this assumption, and that is that items are ordered along the latent trait. IRT thereby becomes useful as we should, in theory, be able to tell how far along the latent trait and individual lies by only knowing the score on a single item. CTT is insensitive to the relationship between items and the latent trait.

IRT describes to basic methods: parametric and non-parametric and these are represented by Rasch analysis and Mokken scaling analysis (MSA), respectively. The difference between the methods is that parametric methods predict and, therefore, depend on a specific relationship between the score on an item and the score on the latent trait and non-parametric methods do not. The relationship between the score on an item and the score on a latent trait is represented by the item characteristic curve where the x-axis represents and score on the item and the y-axis represents the probability of obtaining that score. In both methods, the ICC must be monotonously homogenous—in other words as the score on the trait increases, so does the score on the latent trait. However, the ICC in parametric IRT has a sigmoidal shape and in non-parametric IRT—provided the criterion of monotone homogeneity is met—it can assume any shape. Clearly the two methods have different analytical features but the virtue of non-parametric IRT, represented by MSA, is that it is less conservative and tends to retain more items in an analysis. The resulting scales have high clinical utility but lack the precision of scales obtained using Rasch analysis, which is more suitable to the analysis, for example, of educational tests where greater precision is required.

Mokken scaling

As explained above, Mokken scaling analyses the properties of individual items as described by the item characteristic curve (ICCs), which relates the score on an item to the level of the latent trait being measured. It makes no assumptions about the precise nature of that relationship requiring only that ICCs are monotonely homogeneous (they continuously increase across the range of the latent trait) and that they do not intersect (ie the are doubly monotonous) (Mokken & Lewis 1982). Mokken scaling assumes that the response of items to the level of the latent trait is locally stochastically independent, in other words, that the score on an item is purely a result of the level if the latent trait present and not to a score on any of the other items. Therefore, the score on one item is not dependent on the score on any other items. As stated, this is usually an assumption and is not formally tested in Mokken scaling and, currently, methods for assessment local stochastic independence are still under development.

However, inspection of items in terms of their wording can usually confirm that items are not stochastically dependent. IRT does not assume that all items have an equal level of difficulty—an assumption that is not held by classical test theory methods such as factor analysis (Mokken & Lewis 1982). ‘Difficulty’ means the extent to which items are endorsed by respondents with more extreme items at the upper end of the range of the latent trait being the more difficult. For example, in a scale measuring psychological morbidity, an item labelled ‘I want to end my life’ would be more difficult than an item labelled ‘I don’t feel like getting out of bed’. Therefore, items are arranged along the latent trait in terms of their difficulty and the properties of items can be measured using a scalability coefficient *H* (Loevinger’s coefficient) which measures the extent to which all items are arranged as expected by their mean values along the latent trait. A Loevinger’s coefficient > 0.3 is the minimum acceptable value of H indicating a weak scale; H > 0.4 indicates and moderate scale and H > 0.5 indicates a strong scale. Items can also be analysed for violations of monotone homogeneity and the reliability of sets of items purporting to form Mokken scales can be calculated and expressed in a reliability coefficient *Rho*. The coefficient Rho is preferred in Mokken scaling due to some wellk-known problems with Cronbach’s alpha. Admittedly, Cronbach’s alpha is commonly used to assess reliability in scales but it not independent of the number of items in the scale (Agbo 2010)) and may not be accurate for relatively small numbers of respondents (Sijtsma 2009). Rho—also known as the Molenaar Sijtsma statistic—was especially developed for use in Mokken scaling (van der

Ark et al 2018). Finally, a desirable although not essential feature of a Mokken scale is invariant item ordering (IIO) whereby the order of items along the latent trait is the same for all respondents at all levels of the latent trait. This is investigated primarily by plotting ICCs and inspecting for non-intersection—which clearly violates IIO—and then by investigating IIO mathematically to look for significant violations and then calculating the accuracy of IIO as expressed in a coefficient Htrans or H^T. Values of H and H^T exceeding 0.3 indicate acceptably strong scales and acceptable accuracy of IIO, respectively. For both coefficients, values exceeding 0.4 indicate moderate levels and values exceeding 0.5 high levels of strength and accuracy (Mokken & Lewis 1982, Watson et al 2012).

BACKGROUND

The MOLES index

The MOLES index is an instrument designed to test the motivation of individuals to self-examine their skin for lesions which may indicate that the have skin cancer (Cowdell & Dyson 2014). The MOLES index is comprised of 20 items and was developed from the perspective of the Theoretical Domains Framework which is designed to make behaviour change accessible to health practitioners other than psychologists.

The MOLES index was developed, as described by (Dyson & Cowdell 2014) through a combination of literature review, qualitative work and psychometric testing. As such, the MOLES index resulted from a three stage process with a sample of members of the public and involving: i) identifying items from the barriers to SSE identified in the literature and through a survey of members of the general population (N=261); ii) categorisation of barriers to theoretical framework by experts in the fields of dermatology and psychology (N=11); and iii) validity and reliability testing (face validity, internal consistency, factor analysis and test retest reliability) (N=314).

Examples of items in the MOLES index include: ‘I believe examining my skin leads to better health’; ‘If I examine my skin I may prevent cancer’; and ‘I am able to make checking my skin a regular routine’. Four items are negatively worded, for example: ‘Remembering to check my skin is difficult’; and ‘I cannot be bothered with skin self-examination’. The items are scored on a 7-point Likert type scale running from ‘Strongly agree’ to ‘Strongly disagree’. Therefore, higher scores indicate lower endorsement of skin self-examination and the negatively worded items are reverse scored before using the total score on the scale and before the analysis conducted in this study.

The result of the initial psychometric analysis of the MOLES (Dyson & Cowdell 2014) was a five-factor structure: (i) Outcome expectancies; (ii) Intention; (iii) Self-efficacy; (iv) Social influences; (v) Memory), 20-item instrument which tested well for reliability and construct validity.

The value of this theoretically based instrument is the ease with which behaviour change techniques can be mapped (Michie et al 2013) to the five factors (behavioural determinants) allowing theory based pragmatic and tailored interventions to be developed to support SSE (Cowdell & Dyson 2014). In this paper we build on the existing MOLES index in two ways: 1. We translated the MOLES index into another language (Italian) and we analyse the MOLES index (English and Italian versions) exploratory using Mokken scaling. We suspected that the items in the MOLES index may be suitable to MSA because they were likely to form a hierarchy. For example, some of the questions require only a belief (eg 'I believe examining my skin leads to better health'), whereas some require knowledge (eg 'I could explain the correct method for skin self-examination') and some require commitment (eg 'I am able to make checking my skin a regular routine'). Therefore, it is possible that people endorse beliefs (which are relatively easy and require no action) before they endorse knowledge and actions and, indeed, that belief and knowledge are prerequisites to action.

Research question

How well can Mokken scaling be used to compare to version of the same scale (the MOELS) in two languages (English and Italian) and how do these versions compare when analysed using Mokken scaling?

METHODS

The translation process of the MOLES index

The developers of the MOLES were part of the present team. One member of the team is bilingual and local Italian experts were on hand to assist. Two expert native Italian translators separately conducted the English-Italian forward translation. The two Italian versions were compared and the differences between the two versions were resolved following a discussion by the research team. The resulting Italian version was then back-translated into English by a third expert bilingual English-Italian translator. The differences between the original English version and the English translated version of the Moles Index were discussed and resolved directly with the original authors.

Face Validity of the Italian version of the Moles Index

In September 2015, the final draft of the Italian version of the Moles Index was piloted with 30 2nd and 3rd year nursing students to check face validity and language clarity. All the students easily understood the questionnaire and no further amendment was required.

Data collection

Italian data were collected in October 2016, after presenting the study and illustrating the MOLES index to all the 1st year nursing students, during a general assembly on their first day at a university in the north of Italy. The students were given the URL to an online version of the MOLES index and invited to complete the questionnaire by the end of October and to encourage their family members' friends to do the same thing. The questionnaire was anonymous, and its completion was voluntary. By accepting to complete the questionnaire, respondents automatically expressed their consent to take part in the study. Privacy was ensured, and data were handled exclusively for use in this study. The UK data were collected in 2014 and 2016 and ethical permission obtained as previously described (Dyson & Cowdell 2014).

Analysis

Package ‘mokken’ (https://cran.r-project.org/web/packages/mokken/mokken.pdf last accessed 20 May 2017) from the online public domain statistical software *R* (https://www.r-project.org/ last accessed 12 April, 2016) was used to analyse the data. Data were entered into *R* by converting from SPSS files into .Rdata files using package ‘foreign’ in *R* and then analysed in the following sequence: the automated item selection procedure ‘*aisp*’ was used, with default settings, to investigate how many putative scales were present in the data; the resulting scales were then analysed to see if the items were likely to form a Mokken scale using ‘*coefH*’ to establish the scalability of items, item pairs and the total scales; items were then checked to exclude any items violating montonicity using ‘*check.montonicity*’; items pairs were then plotted using ‘*plot(check.iio(FileR))*’ and the item pairs examined for intersection, floor and ceiling items and any items lying far from the main cluster to decide if they were suitable for analysis of IIO using ‘*iio.results <- check.iio(FileR)*’ followed by ‘*summary(check.iio(FileR, item.selection = FALSE))*’; and reliability of resulting scales was checked using ‘*check.reliability*’. SPSS version 22.0 was used to perform an independent samples t-test.

Ethical approval

The original ethical application in the UK referred to above was subject to a minor modification in 2015 and then this study was approved by the Academic Board of the Italian university.

RESULTS

Demographics

The total number of participants in the present study was 1086: 620 from Italy (340 females; 278 males (2 non-responses); age range 18-70) and 466 from the UK (381 females; 85 males; age range 18-85). Any items with non-responses were removed before running the analysis.

Mokken scaling analysis

The outcome of the *aisp* indicated for the Italian and UK samples showed that in both Italy and the UK eight items clustered on a single scale; the remainder either did not scale or formed other clusters with too few items to form a meaningful scale. The focus of the subsequent analysis was, therefore, on the items clustering on scale 1 in both the Italian and the English samples. Inspection of the relative item ordering by mean values suggested that the Italian and UK samples were insufficiently similar to merit combining the samples; the two scales only have one item in common. From both samples, one further item was removed from the scale due to violating monotonicity, leaving seven items in each scale.

All 20 questions from Section B of the MOLES index are shown in the order in which they appear in the questionnaire along with their mean values for the Italian and the UK samples (Table 1). The difference in total mean scores – tested using a t-test – was not significantly different between the Italian and UK samples. For clarity, the values of *Hi* and the respective standard deviations are only shown for the items which scale. The values of *Hs* along with their respective standard deviations, the values of H^T and the values of *Rho* are given at the foot of each column. Inspection of item pair plots for the combined sample showed that items were quite closely clustered with minimal intersection and no items showing either a ‘floor’ or a ‘ceiling’ effect or lying far from the cluster. None of the seven items remaining in either the Italian or UK scales violated IIO. Using the standard errors, the 95% confidence intervals around *Hs* and *Hi* were inspected and they did not include the lowerbound value of 0.30. The seven items from the Italian data formed a moderate Mokken scale which was reliable, but H^T was not strong enough to show IIO. The seven items from the UK data formed a weak Mokken scale which was reliable, and H^T was strong enough to show weak IIO.

Items are ordered according to their mean value in Table 2. Higher mean scores indicate lower endorsement of the item and, therefore, greater difficulty. In this light, the least difficult item in the Italian data was ‘I believe examining my skin leads to better health’ and the most difficult item was ‘I would be able to explain the benefits of skin self-examination to somebody else’ and, in the UK data the least difficult item was ‘I feel confident that (with the help of someone else if needed) I could examine my skin thoroughly’ and the most difficult item was ‘My doctor/nurses encourages me to examine my skin regularly’. Only one item: ‘I believe examining my skin leads to better health’ was common to both scales.

DISCUSSION

The results show that Mokken scales exist in both the Italian and the English versions of the MOLES index. The same number of items formed a Mokken scale in the Italian and the English versions. There was only one item in common between the English and the Italian versions meaning that the two scales were insufficiently similar to combine the samples and analyse for a single Mokken scale.

The two scales indicate that different constructs within the MOLES Index are important in Italy and the UK. Items ordered by Mokken scaling in Italy relate mainly to belief about the value of SEE in terms of the 'Outcome expectations' and 'Intentions' factors previously identified (Dyson & Cowdell 2014). Items in the UK scale mainly relate to the 'Social influences' and 'Memory' factors previously identified (Dyson & Cowdell 2014). Both scales share items from the 'Self-efficacy' scale. There is no overall significant difference in the total scale scores and looking for significant differences between individual items is prone to type I error, therefore, an explanation must be sought for the very different Mokken scales formed in the two samples and what the implications are for the use of the MOLES index.

First, the differences in the items included in the scales could indicate differences in the perception of risk of melanoma between the Italian and the UK samples. Items that are ordered in Mokken scales are likely to be those that respondents largely respond to consistently relative to one another. Therefore, it appears that respondents in the Italian sample more consistently responded to a set of items related to belief about SSE and the UK sample responded more consistently to a set of items about actions related to SSE. Due to the there being no statistically significant difference between the two samples and no consistent in difference is the pattern of responses to the MOLES items, the apparent difference in the two scales probably does not indicate the importance ascribed to any particular aspects of SSE. Thus, the differences may not have utility in designing interventions or targeting specific aspects of SSE. However, the potential utility of the scales is that these items may also respond consistently to health education and health promotion about melanoma and SSE. Thus, they may have utility – separately – in measuring the outcome of SSE interventions in Italy and the UK, respectively.

It is possible that larger sample sizes may lead to inclusion of more items and greater congruence between the two scales. Thus, a future line of research is suggested by repeating the study with larger samples, possibly in the region of n = 1000 per country (Straat et al 2014). It would also be valuable to replicate the confirmatory factor analysis in an Italian sample. A useful indication of the utility of the MOLES – which is about motivation – would be to relate actual practices related to SSE with the MOLES index in individuals. In that light, the present study suggests a clear line of research related to SEE in different populations.

Limitations

Fewer than 50% of the items in the MOLES were included in either of the Mokken scales. This raises the question of the purpose of the remaining items and the possibility of construct underrepresentation. The implication could be that some items in the MOLES are redundant, but it should also be noted that the sample sizes in the present study are relatively small according to our most recent understanding of sample size requirements for Mokken scaling (Straat et al 2014).

Conclusion

The significance of this study lies in its originality in applying Mokken scaling to the MOLES index according to rigorous analytical criteria. The study provides additional psychometric insight into the MOLES index and augments the original work which used factor analysis. An immediate line of inquiry is suggested that could further test the construct validity of the MOLES index by comparing the latent structure that is apparent in the Mokken scales with a measurement of actual practices—frequency and efficacy—of skin self-examination.

Conflict of interest

No authors have any conflict of interest to declare.

Funding

The study was not funded.

Ethical permission and consent

These are specified in the manuscript.

REFERENCES

Agbo, A.A (2010) Cronbach's alpha: review of limitations and associated recommendations. *Journal of Psychology in Africa* **20**, 233-239

Cowdell, F. and Dyson, J. (2014) A novel intervention for skin cancer prevention. *Dermatol Nurs*, ***13***(3), 45-49.

Dyson, J. and Cowdell, F. (2014) Development and psychometric testing of the 'Motivation and Self-Efficacy in Early Detection of Skin Lesions' index. *J Adv Nurs* **70**, 2952–2963.

Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., Eccles, M.P., Cane, J. and Wood, C.E. (2013) The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *AnnBehav Med*, ***46***, pp.81-95.

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *App Psychol Measur,* **6**, 417-430.

Sijtsma, K. (2009) On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika **74**, 107-120.

Ind Diff, **50**, 31-37.

Straat, J. H., van der Ark, L. A., and Sijtsma, K. (2014) Minimum sample size requirements for Mokken scale analysis. *Ed Psychol Meas, 74,* 809-822.

Van der Ark, A., Straat, J.H., Koopan, L. (2018) Package 'mokken'(https://cran.r-project.org/web/packages/mokken/mokken.pdf; accessed 10 February 2019)

Watson, R., van der Ark, L. A., Lin, L-C., Fieo, R., Deary, I. J., and Meijer, R. R. (2012) Item response theory: How Mokken scaling can be used in clinical practice. *J Clin Nurs,* **21**, 2736-2746.

Table 1 Mokken scaling of Italian and UK MOLES data

Item	Descriptor	Mean item scores [*Hi* (SE)			
		Italy (n=619)		UK (n=460	
1.	I believe examining my skin leads to better health	**2.06**†	**[0.46 (0.023)]**	**4.40**†	**[0.37 (0.031)]**
2.	I could describe the moles and marks on my skin	3.02		**3.85**†	**[0.40 (0.028)]**
3.	My doctor/nurse encourages me to self-examine my skin regularly	3.61		**5.37**†	**[0.31 (0.033)**
4.	If I examine my skin I may prevent cancer	**2.36**†	**[0.47 (0.026)]**	2.38	
5.	Remembering to check my skin is difficult	4.23		**5.25**†	**[0.42 (0.030)]**
6.	I can make the effort to examine my skin each month	2.48†		4.41	
7.	My friends encourage me to examine my skin regularly	4.34		3.87	
8.	It does not occur to me to examine my skin	4.24		**5.06**†	**[0.47 (0.027)]**
9.	The risk of skin cancer is exaggerated by the medical profession	4.29		2.58	
10.	I could make a habit of skin self-examination	**2.58**†	**[0.52 (0.024)]**	2.90†	
11.	If I had a skin lesion, self-examination and early reporting may prevent it getting worse	**3.23**†	**[0.53 (0.022)]**	2.68	
12.	I am able to make checking my skin a regular routine	**2.74**†	**[0.52 (0.023)]**	2.91	
13.	I could explain the correct method for skin self-examination	3.68†		4.52	
14.	I know someone who had skin cancer	3.86		2.54	
15.	I cannot be bothered with skin	4.51		**4.15**†	**[0.34 (0.030)]**

self-examination				
16. Examining my skin will make me feel more control over my health	**2.53**†	**[0.50 (0.031)]**	2.79	
17. I would be able to explain the benefits of skin self-examination to somebody else	**3.11**†	**[0.33 (0.033)]**	3.13	
18. I am confident about my ability to examine my skin	3.66		2.45	
19. I feel confident that (with the help of someone else if needed) I could examine my skin thoroughly	2.45		**2.45**†	**[0.31 (0.033)]**
20. My family encourages me to examine my skin regularly	3.57		4.67	
Mean total	3.32		3.60	
Hs†	0.47 (0.021)		0.38 (0.023)	
$H^{T\dagger}$	0.27		0.35	
Rho†	0.85		0.79	

† = for items included in scale 1

Table 2 Items in scale 1 ordered by increasing mean value

Item	Italy	Item	UK
1	I believe examining my skin leads to better health[1]	19	I feel confident that (with the help of someone else if needed) I could examine my skin thoroughly[3]
11	If I had a skin lesion, self-examination and early reporting may prevent it getting worse[1]	2	I could describe the moles and marks on my skin[3]
4	If I examine my skin I may prevent cancer[1]	15	I cannot be bothered with skin self-examination*[5]
16	Examining my skin will make me feel more control over my health[1]	1	I believe examining my skin leads to better health[1]
10	I could make a habit of skin self-examination[2]	8	It does not occur to me to examine my skin[5]
12	I am able to make checking my skin a regular routine[2]	5	Remembering to check my skin is difficult*[5]
17	I would be able to explain the benefits of skin self-examination to somebody else[3]	3	My doctor/nurses encourages me to examine my skin regularly[4]

NB: higher means lower endorsement; * - reverse scored

1 – Outcome expectations factor

2 – Intentions factor

3 – Self-efficacy factor

4 – Social influences factor

5 – Memory factor

Thanks for the feedback, some of this has helped to revise the manuscript but there is a limit to what we can address without changing the nature of the study. There is also some difference of opinion about how best to run and present Mokken scaling but we have followed the standard process that others use and that we have used in many studies.

Reviewer(s)' Comments to Author:

Reviewer: 1

Comments to the Author

GENERAL COMMENTS

Thank you for your successful efforts in performing this very interesting piece of scholarly work. This is a very important topic for all health care professionals and also the journal´s diverse readers. This reviewer enjoyed the opportunity to appraise this paper,

This reviewer has a few questions and/or requests for clarifications that could increase the quality of the manuscript.

With regards to specific comments please include the following or reflect up on in the manuscript:

SPECIFIC COMMENTS

Introduction/Background

- The introduction could be strengthened by further contextualizing the study.

From the comments below – which imply that this reviewer wanted to see a study that focused more on the method than on skin self-examination – we have focused on Mokken scaling in the Introduction.

- In the background the conceptual framework could be enhanced, thus addressing theories that can clarify the underlying mechanisms pertaining to the scientific problem; ii) expanding the critical synthesis of knowledge from the empirical literature identifying what is already known and what is not known; and iii) the researcher's individual thoughts and ideas.

Likewise we have focused more on Mokken scaling here explaining what the method is and what it can do but the study was a straightforward comparison of two scales to see how they behaved under the methods; it would be very contrived for us to try to turn this around too much towards being a study of fundamental properties of the method as the study was not designed to do this.

- This section could benefit from a stronger rational and justification how using Mokken scaling in this study can contribute the scientific community with new knowledge. Thus, more convincingly and better define the gap in current knowledge in this area and how the study will address it. So, the author/s could better rationalize their choice of analysis.

We have tried to justify this method better but it was not designed to address anything fundamental about the method.

- The manuscript should be focused more on the analytic technique, not the MOLES index.

As above – we have tried to do this by, essentially, omitting most of the material on the MOLES.

- Expand the description of the MOLES index e.g. quantity of items, reversed scoring, type of items (statements vs questions), response categories.

We have done this – some of this was already there but we have added the reverse scoring.

- Why would you expect that latent traits/constructs captured with the MOLES index would fit the requirement of a Mokken scale?

This is an excellent point to make and we have now address it as follows:

'We suspected that the items in the MOLES index may be suitable to MSA because they were likely to form a hierarchy. For example, some of the questions require only a belief (eg 'I believe examining my skin leads to better health'), whereas some require knowledge (eg 'I could explain the correct method for skin self-examination') and some require commitment (eg 'I am able to make checking my skin a regular routine'). Therefore, it is possible that people endorse beliefs (which are relatively easy and require no action) before they endorse knowledge and actions and, indeed, that belief and knowledge are prerequisites to action.'

- Please address that empirical studies have relied comprehensively on classical test theory, and therefore the theoretical foundations of IRT need to be elaborated.

We have compared and contrasted the method with classical test theory by adding some material.

- The wording when describing the development of MOLES is for this reviewer lightly peculiar. Was a confirmatory factor analysis used in the exploratory phase of instrumental development?

This seemed to be clear to us already but hopefully we have clarified better.

- Maybe the detailed specific result from the referenced validity study is redundant. Otherwise, assist the presumptive reader (maybe not familiar with structural equation modelling) by explaining the interpretation, e.g. RMSEA values ≤ 0.08 (if possible ≤ 0.05) are considered to designate acceptable fit and CFI are scaled to range between 0-1,with values above 0.90 suggesting a good model fit while values ≥ 0.85 can be considered to indicate acceptable model fit.

This material has been removed.

- The description of Mokken scaling is rather shallow. Please address that Mokken belongs to the class of non-parametric item response theory and aimed at assessing unidimensional scales of dichotomous or polytomous items.

We have already boosted the description of the method.

- Please address both non-parametric models, thus MHM och DMM

Addressed as follows:

‘It makes no assumptions about the precise nature of that relationship requiring only that ICCs are monotonely homogeneous (they continuously increase across the range of the latent trait) and that they do not intersect (ie the are doubly monotonous) (Mokken & Lewis 1982)’

- In the present paper it is not clear if the Mokken sacling is used in a confirmatory or in an exploratory way.

Address as follows:

‘We translated the MOLES index into another language (Italian) and we analyse the MOLES index (English and Italian versions) exploratory using Mokken scaling’

- Elucidate and elaborate on ALL four assumptions underlying the models: unidimensionality, monotonicity, local independence and invariant item ordering.

The monotone model already addressed as is invariant ordering – unidimensionality is assumed in all scaling work (classical test and item response theories) and does not require re-stating.

Local independence addressed as follows:

'Mokken scaling assumes that the response of items to the level of the latent trait is locally stochastically independent, in other words, that the score on an item is purely a result of the level if the latent trait present and not to a score on any of the other items. Therefore, the score on one item is not dependent on the score on any other items. As stated, this is usually an assumption and is not formally tested in Mokken scaling and, currently, methods for assessment local stochastic independence are still under development.'

- Explain the ICC for readers who are not familiar with the concept, thus how discrete items in a scale perform in relation to the latent trait.

We have expanded on the ICC.

- Explain how Loevinger's coefficient H is used at different levels of the analysis.

We think this means to provide the levels related to weak, moderate and strong scales which we have now added.

- Even though this reviewer also (sometimes) use H (trans), it is not without problems. Sijtsma and Meijer (1992) recommends in the article of Ligtvoet (which is quite natural) to first exclude items with a flat IRF, ie most of those who have low Hij. Violations and low Hij are related so I think it's enough to identify items with poor scalability. Furthermore, items that are close to each other (regarding item difficulty) will yield a low H (trans). An assessment of H (trans) also requires local independence, which was not evaluated.

We don't agree with this point, leading Mokken scalers recommend the use of Htrans and we understand that low Hij items may violate – essentially, the reviewer is giving a description of what Htrans does.

- Argue for why Mokken´s Rho rather than Cronbach's alpha was used as an estimator of reliability.

We have added this and supporting references.

The study

- Should not the heading be "Methods"?

Yes – changed.

- Please elaborate on the process of forward and backward translation.

Mostly this was already done as follows, we added the underlined text:

'The developers of the MOLES were part of the present team. One member of the team is bilingual and local Italian experts were on hand to assist. Two expert native Italian translators separately conducted the English-Italian forward translation. The two Italian versions were compared and the differences between the two versions were resolved following a discussion by the research team. The resulting Italian version was then back-translated into English by a third expert bilingual English-Italian translator. The differences between the original English version and the English translated version of the Moles Index were discussed and resolved directly with the original authors. '

- It is often recommended in guidelines for developing, translating, and validating a questionnaire that an expert committee is suggested to produce the pre-final version of the translation. Members of the committee are often including an expert who is familiar with the construct of interest, a methodologist, both the forward and backward translators, and if possible the developers of the original questionnaires. Was any such of strategy performed?

We have added a sentence to cover this.

- Cognitive interviewing is meant to identify and analyze sources of response error in instruments by focusing on the cognitive processes respondents use to answer items. The purpose of cognitive interviews is to focus on the survey items, not the person answering the items. Which technique (concurrent probing, retrospective probing, think aloud methodology) was used and please elaborate on item discrepancies.

This is unnecessary – we had a bilingual expert on the team.

- The period of time of data collection for the Italian and UK data is not entirely clear.

We think it is.

- Please state the analytical strategy in a step-by-step manner and be more detailed so the replication is possible.

This is already done under the method.

- R commands are redundant and can be added as appendices.

It is quite conventional in reporting MSA papers to include them in-line in the methods section.

- Was data handled according to the Declaration of Helsinki and was informed consent obtained from respondents/participants?

Ethical permission was obtained in tow countries that adhere to this – not necessary to add this.

- How were items with reversed scoring treated?

Now explained.

- Were any subgroup analyses performed? If demographic data was collected it would be interesting to investigate discrepancies (or if not) using Mokken scaling.

No.

Result

- This section is not very comprehensive and this reviewer think that more data and variables could be reported, such as instrument response rate, item response rate, result tables including the entire scalability analysis and also non-scalable items.

We are sorry but we do not agree – the results are explained in a very systematic and clear way in terms of the steps in the analysis and what we found. We already report the response rates at the start of the results as follows:

'The total number of participants in the present study was 1086: 620 from Italy (340 females; 278 males (2 non-responses); age range 18-70) and 466 from the UK (381 females; 85 males; age range 18-85). Any items with non-responses were removed before running the analysis.'

There is no such thing as the ‘entire scalability’ to which the reviewer refers – only the scales we found in each set of data – and these are reported comprehensively.

- Please elaborate why referring to “missing” rather the “non-response”.

We have described ‘missing’ now as ‘non-response’.

- Why were no imputations made?

The missing data were very few and imputation remains a controversial area with several methods available and there is no track-record in the literature of imputation of data in MSA studies.

- Justify why parametric statistical inference (t-test) was used on ordinal data.

When such ordinal data are summed then the resulting scales usually show interval level data.

- Why is SD rather than SEM reported?

We are not trying to describe the population here, merely to indicate the spread around the mean.

- Consider to create separate tables or report distinctly Mokken Rho and HT.

This would not be parsimonious and would also not be conventional.

- Please elaborate the sentence regarding “item difficulty”. See Watson et al., 2012 for a more inclusive discussion.

We already elaborate on the concept of difficulty as follows:

‘‘Difficulty’ means the extent to which items are endorsed by respondents with more extreme items at the upper end of the range of the latent trait being the more difficult. For example, in a scale measuring psychological morbidity, an item labelled ‘I want to end my life’ would be more difficult than an item labelled ‘I don’t feel like getting out of bed’. Therefore, items are arranged along the latent trait in terms of their difficulty and the properties of items can be measured using a scalability coefficient *H* (Loevinger’s coefficient) which measures the extent to which all items are arranged as expected by their mean values along the latent trait.’

- Table 2 need to be further explained.

With apologies but we simply don't see how much clearer this table could be. It is not clear what needs clarifying.

Discussion

- It might help if you discussed your results using the five strands more clearly i) a synopsis and a brief review of the most important findings, ii) relating your findings to the empirical and theoretical literature, iii) discussing methodological limitations, iv) raising your findings to a more meta-level, discussing practical, clinical, educational and academic implications and v) finish with where to go next (future directions). With such a structure readers could consider merits of your position, quality of supportive evidence and whether your contribution advances understanding inside and outside the scientific community.

This is an entirely empirical study and we achieved what we set out to achieve and explained it in the results and discussion in a systematic and logical way.

- Discuss the pro and cons using Mokken scaling in this study but also from a more general point of view.

The purpose behind using the MSA was clearly elucidated in the Introduction and Background and we found scales; different methods may have produced different results but we have absolutely no other MOLES data with which to compare these.

- Argue for why a non-parametric IRT method was used rather than parametric IRT models such as the Rasch model.

Addressed as follows:

'Clearly the two methods have different analytical features but the virtue of non-parametric IRT, represented by MSA, is that it is less conservative and tends to retain more items in an analysis'

Conclusion

- Please create a better alignment between the research question including the aim of the study and the Mokken scaling method used and the conclusions drawn upon the findings.

WE did adjust the title and aims of the study – however, the study was of the MOLES and the existent of Mokken scales in these two versions; we found that – and explained it and our conclusions are based on that; we cannot draw conclusions from a study we did not carry out.

Reviewer: 2

Comments to the Author

1. the title and the aims of the study didn't match. what are the real aims of the paper?

Both altered to accommodate.

2. the significance of comparing the Italian and the UK MOLEX was not clear at all.

This is less the focus of the study now in the light of the comments of Reviewer 1.

3.the demographics of Italian and the UK sample were missing, so the base of comparing didn't exist.

No, they were not:

'The total number of participants in the present study was 1086: 620 from Italy (340 females; 278 males (2 non-responses); age range 18-70) and 466 from the UK (381 females; 85 males; age range 18-85). Any items with non-responses were removed before running the analysis.'

4.the reference systems were inconsistent and important reference could not be traced.

Checked.

5. how the UK data and what characteristics of the sample were not presented

This is presented.

6.page 13, the floor and ceiling effect was not supported by any data

It was – the inspection of the item plots which we did not present – this is a conventional step in MSA.

7. why the author used t-test was not clear. and the t-test was not useful in comparing the scale

We disagree – we tested the difference in the total scores on the scales to see if there was any fundamental difference in the level of the latent trait between the samples – there was nonw.

8.in data analysis process, if the ICC was adjusted to explore a better solution was not clear.

You cannot adjust ICCs.

9. the results in the two samples were totally different, what does this mean and imply?

This is partly the basis of the study and is explained as follows:

'The two scales indicate that different constructs within the MOLES Index are important in Italy and the UK. Items ordered by Mokken scaling in Italy relate mainly to belief about the value of SEE in terms of the 'Outcome expectations' and 'Intentions' factors previously identified (Dyson & Cowdell 2014). Items in the UK scale mainly relate to the 'Social influences' and 'Memory' factors previously identified (Dyson & Cowdell 2014). Both scales share items from the 'Self-efficacy' scale. There is no overall significant difference in the total scale scores and looking for significant differences between individual items is prone to type I error, therefore, an explanation must be sought for the very different Mokken scales formed in the two samples and what the implications are for the use of the MOLES index.

First, the differences in the items included in the scales could indicate differences in the perception of risk of melanoma between the Italian and the UK samples. Items that are ordered in Mokken scales are likely to be those that respondents largely respond to consistently relative to one another. Therefore, it appears that respondents in the Italian sample more consistently responded to a set of items related to belief about SSE and the UK sample responded more consistently to a set of items about actions related to SSE. Due to the there being no statistically significant difference between the two samples and no consistent in difference is the pattern of responses to the MOLES items, the apparent difference in the two scales probably does not indicate the importance ascribed to any particular aspects of SSE. Thus, the differences may not have utility in designing interventions or targeting specific aspects of SSE. However, the potential utility of the scales is that these items may also respond consistently to health education and health promotion about melanoma and SSE. Thus, they may have utility – separately – in measuring the outcome of SSE interventions in Italy and the UK, respectively.'

10. the writing needs extensive proofreading.

Done.

11.it is difficult to understand table 1, it may be better to put all Hi in one column.

We agree – it was presented in portrait in error and it should be in landscape which aligns all values.

12 the HT of Italian version was 0.27, not significant, so there is no need to compare the seven items. and the HT should be put in table 2, instead of table 1.

The reviewer makes a reasonable point about the size of the Ht (it is NOT the significance as you cannot test that) but this is only one measure of a Mokken scale and it was close to 3 and it merely indicates the extent of IIO – discarding items can lead to construct underrepresentation and this sample size here was quite small; we indicate that larger samples in subsequent studies may be required. This research was described as promising – not definitive.