MRS HANNE SØBERG FINBRÅTEN (Orcid ID : 0000-0001-8911-2102)

Article type : Original Research: Empirical research - quantitative

Validating the European Health Literacy Survey Questionnaire in people with type 2 diabetes. Latent trait analyses applying multidimensional Rasch modelling and confirmatory factor analysis.

Running head: Validating the HLS-EU-Q47 in people with diabetes

Hanne Søberg FINBRÅTEN (corresponding author), PhD student, MSc, RN, Inland Norway University of Applied Sciences and Karlstad University

Address:
Inland Norway University of Applied Sciences
Faculty of Public Health
PO Box 400
N-2418 Elverum
Norway
e-mail: hanne.finbraten@inn.no

Kjell Sverre PETTERSEN, PhD, Professor, Oslo and Akershus University College of Applied Sciences

Bodil WILDE-LARSSON, PhD, RN, Professor, Karlstad University and Inland Norway University of Applied Sciences

Gun NORDSTRÖM, PhD, RN, Professor, Inland Norway University of Applied Sciences

Anne TROLLVIK, PhD, RN, Associate professor, Inland Norway University of Applied Sciences

Øystein GUTTERSRUD, PhD, Associate professor, University of Oslo

Acknowledgements
The authors thank Kristine Sørensen for permission to use the HLS-EU-Q47 in this study and Jari Appelgren for guidance in using the software SPSS AMOS.

Conflict of interests
No conflict of interest has been declared by the authors.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/jan.13342

Funding
The data collection was funded by Norwegian Nurses' Organization (reference no. 14/0023) and Inland Norway University of Applied Sciences (reference no. 2013/1565).

Abstract

Aim. To validate the European Health Literacy Survey Questionnaire (HLS-EU-Q47) in people with type 2 diabetes mellitus.

Background. The HLS-EU-Q47 latent variable is outlined in a framework with four cognitive domains integrated in three health domains, implying 12 theoretically defined subscales. Valid and reliable health literacy measurers are crucial to effectively adapt health communication and education to individuals and groups of patients.

Design. Cross-sectional study applying confirmatory latent trait analyses.

Methods. Using a paper-and-pencil self-administered approach, 388 adults responded in March 2015. The data were analysed using the Rasch methodology and confirmatory factor analysis.

Results. Response violation and trait violation (multidimensionality) of local independence were identified. Fitting the 'multidimensional random coefficients multinomial logit' model, 1-, 3- and 12-dimensional Rasch models were applied and compared. Poor model fit and differential item functioning were present in some items and several subscales suffered from poor targeting and low reliability. Despite multidimensionality in the data, we did not observe any unordered response categories.

Conclusion. Interpreting the domains as distinct but related latent dimensions, the data fit a 12-dimensional Rasch model and a 12-factor confirmatory factor model best. Therefore, the analyses did not support the estimation of one overall 'health literacy score'. To support the plausibility of claims based on the HLS-EU score(s), we suggest: removing the health care aspect to reduce the magnitude of multidimensionality; rejecting redundant items to confine response dependency; adding 'harder' items and applying a six-point rating scale to improve

subscale targeting and reliability; and revising items to improve model fit.

Keywords: confirmatory factor analysis, health literacy, HLS-EU-Q47, multidimensional Rasch modelling, nursing research, type 2 diabetes mellitus.

Summary statements

Why is this research or review needed?

- This study provides insights into a neglected issue: the application of multidimensional Rasch modelling to evaluate the psychometric properties of subscales identified with each aspect of a latent variable.
- To support self-management, well-adapted health communication and education should be derived from valid and reliable information on patients' health literacy. Validating new health literacy scales in people with type 2 diabetes is imperative for making knowledge-based decisions.
- Health literacy in people with type 2 diabetes should be measured using broader perspectives than the available scales consider. The European Health Literacy Survey Questionnaire has the potential to drive this research forward.

What are the key findings?

- Applied to people with type 2 diabetes, the European Health Literacy Survey Questionnaire violated the item response theory assumption of unidimensionality.
- Considering this violation of unidimensionality, we should question the plausibility of claims about people's health literacy based on the sum score of the European Health Literacy Survey Questionnaire.
- More specifically, the European Health Literacy Survey Questionnaire data did not

meet the Rasch models' expectations. Several of the theoretically defined subscales were poorly targeted and suffered from low reliability. Dependent items collecting redundant information reduce the efficiency of the questionnaire as a measure of health literacy.

How should the findings be used to influence policy/practice/research/education?

- One should be careful when developing and implementing new health policies based on conclusions drawn from scores of the current version of the European Health Literacy Survey Questionnaire.
- It should be common practice among health literacy researchers to evaluate the dimensionality of composite scales, such as the European Health Literacy Survey Questionnaire, by applying multidimensional Rasch models.
- Health literacy researchers should join forces and carefully implement an empirically driven revision of the European Health Literacy Survey Questionnaire to solve the problems identified. The revised version should be validated across different cultures, such as languages, ethnicities, health conditions and patient groups.

INTRODUCTION

To manage type 2 diabetes mellitus (T2DM), people need tailored information from diabetes specialist nurses, as well as sufficient information processing skills associated with health literacy (HL) (see e.g., Bohanny *et al.* 2013, Bagnasco *et al.* 2014). Considering people's HL as a mediator for effective communication with healthcare professionals (Ishikawa & Kiuchi 2010, Rudd *et al.* 2012), it becomes clear that low HL is a barrier to optimal management of T2DM (Schillinger *et al.* 2002, Bohanny *et al.* 2013).

While some studies have linked low HL in people with T2DM to shallow knowledge

about diabetes, poor glycaemic control and diabetic retinopathy (Schillinger *et al.* 2002, Powell *et al.* 2007, Tang *et al.* 2008, Sarkar *et al.* 2010, van der Heide *et al.* 2014), other studies do not support such relationships or have found contradictory evidence (Bains & Egede 2011, Al Sayah *et al.* 2013). This lack of consensus could be traced back to the complex endeavour of measuring HL.

HL scales are typically developed by analysing the aspects of the HL construct, using a subscale identified with each aspect. Capturing the complexity of the HL construct by defining subscales increases the 'face validity' of an HL scale but violates the assumption of a unidimensional interval scale and, hence, the requirement of additivity. Given this explicitly multidimensional design, Altin *et al.* (2014) recently warned that most HL scales have a multidimensional structure, which implicitly suggests that we should question the plausibility of claims about people's HL based on the *sum score* of composite HL scales.

This paper seeks to evaluate the plausibility of the interpretations and uses of the European Health Literacy Survey Questionnaire (HLS-EU-Q47) sum score found in peer-reviewed publications (Sørensen *et al.* 2013, 2015, Nakayama *et al.* 2015, Palumbo *et al.* 2016). More specifically, we aim to evaluate the plausibility of the claims, assumptions and ... erences that have been made based on the HLS-EU-Q47 because these might influence health policies. We will compare the results from multidimensional Rasch modelling to results from multifactorial confirmatory factor analysis (CFA) – a combined approach that is virtually absent in health-related research. According to Nguyen *et al.* (2015), surprisingly few HL scales have been validated by applying appropriate methodological approaches.

Background

The HLS-EU-Q47 is the most up-to-date measure of HL (Haun *et al*. 2014). It was developed after an analysis of aspects related to the complex concept of HL and intends to reflect the conceptual model proposed by Sørensen *et al.* (2012). According to the HLS-EU Consortium (2012) and Sørensen *et al.* (2012, 2013), the HLS-EU-Q47 aims at integrating four cognitive domains (accessing [A], understanding [B], appraising [C] and applying [D] health information) in three health domains (healthcare [HC], disease prevention [DP] and health promotion [HP]). Table 4 provides the distribution of the 47 items across the three health domains and the four cognitive domains. The HLS-EU-Q47 uses a four-point rating scale, with integer response categories from 1 (very easy) to 4 (very difficult) where higher scores indicate lower HL. For a thorough discussion on the substantive theory of the latent variable HL, including its definition and operationalization, we refer the reader to Sørensen *et al.* (2012) and HLS-EU Consortium (2012).

The psychometric properties of the HLS-EU-Q47 have been investigated using principal component analyses (PCA; Sørensen *et al.* 2013, van der Heide *et al.* 2013) and using CFA (Duong *et al.* 2015, Nakayama *et al.* 2015, Duong *et al.* 2017). To our knowledge, the HLS-EU-Q47 has not yet been validated in people with T2DM or by applying ... ltidimensional Rasch modelling. The family of Rasch models assumes locally independent data, similar to item response theory (IRT)-models. Unlike IRT-models, the Rasch models satisfy the requirements of fundamental measurement, such as additivity (Perline *et al.* 1979, Andrich 1989), invariance (Andrich 1988), specific objectivity (Stenner 1994) and sufficiency (Andersen 1977). Therefore, we apply Rasch models in this study.

THE STUDY

Aim

The aim of the present study is to validate the HLS-EU-Q47 in people with T2DM. We can translate this aim into one broad research question:

> To what extent do the HLS-EU-Q47 main dimension score and subscale scores support claims about health literacy that go beyond the observed performances?

Design

Cross-sectional study applying confirmatory latent trait analyses.

Participants

A random sample of 999 members of the Norwegian Diabetes Association (NDA) was drawn from the member list. The inclusion criteria were adults above 18 years of age diagnosed with T2DM. The NDA offered a geographically stratified sample that reflected the members' places of residence.

Data collection

Data were collected in March 2015 using a self-administered paper-and-pencil questionnaire distributed by regular mail to addresses provided by the NDA. Two questionnaires were returned owing to unknown addresses. Thirty-one individuals reported health conditions that were not compatible with responding and a further 18 reported having type 1 diabetes. Excluding these responses, we based our analyses on 388 responses (response rate of 41% or 388/948).

Ethical considerations

Norwegian Social Science Data Services (ref. no. 38917) approved the study. Participation was voluntary and we had access to depersonalized demographic data.

Adaptation and translation of the instrument

The response scale of the HLS-EU-Q47 was reversed in the present study so high raw scores would reflect performance at high proficiencies. The response category 'don't know' was added and recoded as missing data.

The translation of the HLS-EU-Q47 followed Brislin's protocol (1970). The items were translated from English into Norwegian by three bilingual researchers. After reaching consensus, a professional translator performed a 'blind' back-translation. The back-translated version was compared with the original to help obtain semantically, technically and contextually equivalent versions. Since the Scandinavian languages share several similarities, our Norwegian version was compared with the Danish translation (unpublished) and the Swedish version of the HLS-EU-Q16 (Wångdahl *et al.* 2014).

To avoid misleading item wording and ambiguities (Drennan 2003), cognitive interviews were conducted from December 2013 - January 2014 with five males and eight females (age 21–72 years). Minor adjustments were made based on the feedback from the interviewees to ensure accurate interpretation of item content.

Pilot testing the instrument

The HLS-EU-Q47, including demographic variables, was piloted from March to April 2014. Based on a unidimensional Rasch analysis of data from 191 individuals (not included in the main study) diagnosed with T2DM (120 males and 65 females [six individuals with missing data], age 23–91 years, mean age 63 years), four under-discriminating items (3, 29, 36 and 38) and two items with unordered response categories (17 and 24) were observed. In addition,

30 pairs of statistically dependent items were identified. These items were slightly rephrased with the intention of improving the fit to the Rasch model and avoiding dependency and unordered response categories.

Data analysis

Data were analysed against both the partial credit parameterization (Masters 1982) of the 1-dimensional (unidimensional) polytomous Rasch model (Rasch 1980) and the partial credit parameterization of the multidimensional 'between-item' polytomous Rasch model fitted by the 'multidimensional random coefficients multinomial logit' (MRCML) model (Adams *et al.* 1997). In the following, the 'unidimensional approach' refers to a unidimensional Rasch analysis where the three health domains are treated as three independent subscales. The '3-dimensional approach' refers to an oblique or unrestricted multidimensional Rasch analysis, where the three health domains are allowed to covary. The '12-dimensional approach' refers to a similar analysis, where the four cognitive domains and three health domains define 12 correlated subscales. The 1-, 3- and 12-dimensional models were estimated using ConQuest 4 software (Adams *et al.* 2015).

Applying the 'consecutive unidimensional approach', where the HLS-EU-Q47 domains or subscales are specified to be independent, the trait underlying each of the subscales can be analysed without any influences from the other subscales. For simplicity, RUMM2030 software (Andrich *et al.* 2003) was used for the consecutive approach and the initial analyses of unidimensionality.

The average item-location estimate was set to 0.0 in all analyses. While ConQuest applies marginal maximum likelihood estimation (MMLE; Bock & Aitkin 1981) for item-location estimates, RUMM uses pairwise maximum likelihood estimation (PMLE; Katsikatsou *et al.* 2012). Both statistical packages apply Warm's mean weighted likelihood

estimation (WLE) for person-location estimates (Warm 1989). CFA was performed using SPSS 23 AMOS software, which applies maximum likelihood estimation (Scholz 2004).

Rasch modelling

Dimensionality

The unidimensional Rasch models imply that one latent trait or dimension (in our case HL) is measured. This means that the correlations between the items are accounted for by one trait (Ryan 1983, Tennant & Conaghan 2007, Marais & Andrich 2008). Trait violation of local independence (multidimensionality) implies that the responses relate to more than one latent trait, thereby reflecting multidimensionality.

Scales composed of several subscales measuring different aspects of a construct violate unidimensionality (Briggs & Wilson 2003, Marais & Andrich 2008), such as the HLS-EU-Q47. An initial analysis of multidimensionality was performed using RUMM by applying the combined PCA of residuals and paired *t*-test procedures along with creating subtest structures (RUMM Laboratory Pty Ltd 2009, Hagell 2014).

The PCA identifies whether there are further patterns in the data other than the 'Rasch factor' or HL (Smith 2002, Stewart-Brown *et al.* 2009). The results from the paired *t* tests (Carlin & Doyle 2001) help identify individuals whose person-location estimates are significantly different on related subscales. The hypothesis of unidimensionality is weakened when the proportion of individuals with significantly different person-location estimates on a pair of compared subscales exceeds 5% (Stewart-Brown *et al.* 2009).

The fractal indexes (*r*, *c* and *A*) are specific to a subtest structure in RUMM. The index *A* describes the amount of variance common to the subscales, *c* identifies the magnitude of unique subscale variance and *r* is the correlation between the subscales (Andrich 2009). High values for *A* and *r and* a low value for *c*, might indicate an approximately unidimensional

scale. A 'significant' drop in the person separation index (PSI) when forming a subtest structure, indicates violations of local independence (Marais & Andrich 2008).

When the dimensionality tests in RUMM (PCA of residuals, *t*-test procedure and subtest structure with fractal indexes) consistently indicate violations of unidimensionality, we must either treat the subscales as orthogonal and apply a consecutive approach, or allow for subscale covariances and apply multidimensional approaches. A multidimensional approach provides more precise estimates than a corresponding consecutive approach and is, therefore, preferred. To compare fit across nested models, such as 1-, 3- and 12-dimensional models, ConQuest reports the deviance statistics where smaller values indicate better model fit (Adams & Wu 2010).

Reliability and response dependence

The reliability index Cronbach's α was estimated using SPSS, the PSI using RUMM and the person separation reliability (PSR) index using ConQuest. Estimates of reliability and internal consistency indexes should exceed 0.85 or 0.65 if conclusions are to be drawn at the individual or group level, respectively (Frisbie 1988).

Response dependency (Marais & Andrich 2008), indicating inefficient measures that collect redundant information, violates local independence and occurs when the responses to an item influence the responses to a subsequent item. Response dependence was determined by estimating residual correlations in RUMM (RUMM Laboratory Pty Ltd 2009) applying both the unidimensional and the consecutive analyses. Andrich *et al.* (2012) identified residual correlations above 0.3 as possible indicators of 'significant' response dependence between two items.

Targeting

In a well-targeted scale, the distribution of the item threshold estimates, centred at 0.0 logits, will match the distribution of the person estimates (Tennant & Conaghan 2007). Poor targeting might increase the possibility of extreme person scores (ceiling effect), increase the risk of unordered response categories (Hagquist *et al.* 2009) and result in deflated reliability indexes owing to poor person separation caused by deflated variance in person estimates.

Item fit

Infit MNSQ or 'weighted mean square' is a variance-weighted *z*-fit residual (Smith 1995) with an expected value of 1 (i.e., model accounts for 100% of variations in the data). An infit MNSQ above/below its expected value, with a corresponding positive/negative *t* value, indicates that the item is under-/over-discriminating according to the Rasch model expectations. Any misfit item is recognized by an infit estimate that is 'significantly' different from the expected value of 1, meaning the estimate lies outside the 95% confidence interval (CI) with a corresponding absolute *t* value above 1.96. Therefore, strongly under-discriminating items are identified by infit MNSQ that exceeds 95% CI and *t* values above 1.96 (Adams & Wu 2010).

Differential item functioning

The differential item functioning (DIF) notion involves checking for the presence of within-item bias, which refers to different levels of a person factor such as males and females, having different probabilities of endorsing an item despite similar locations on the underlying trait (Andrich & Hagquist 2014). Uniform DIF refers to consistent systematic differences in responses across person factor levels, whereas non-uniform DIF may be identified when there is, for example, a 'gender–trait' interaction (Andrich & Hagquist 2001).

Using RUMM, we examined DIF by applying the 1-dimensional model, a consecutive approach for the three health domains and a consecutive approach for the 12 subscales. Analyses of DIF were performed for the person factors gender, age and education. To obtain equally sized person factor levels, 'age' was dichotomized according to its median (74 years), whereas the person factor 'education' was split by the levels 'higher education' and 'primary and secondary school'.

Ordering of response categories

Unordered response categories imply that response categories are not working as intended (e.g. Andrich *et al.* 1997) and reflect data at the nominal level. The ordering of response categories was examined by applying the 1-dimensional model, a consecutive approach for the three health domains and a consecutive approach for the 12 subscales.

Confirmatory factor analysis

Using CFA, the latent structure is determined by the covariance between the items (Brown 2015). CFA was performed to investigate model fit and further confirm the factor structure. The following fit indexes were taken into account: normed or relative χ^2 (χ^2/d.f), root-mean-squared error of approximation (RMSEA) and comparative fit index (CFI). Normed χ^2 should be below 3.0 (Kline 1998). RMSEA values below 0.06 indicate good model fit. However, RMSEA values of 0.08 could be acceptable if the sample size is small and other fit indexes suggest a good model fit. CFI above 0.95 indicates good model fit and 0.90–0.95 indicates reasonable fit, whereas values below 0.90 indicate poor fit (Brown 2015).

Handling missing data

On average, there were six missing and 29 'don't know' responses per item. The large proportion of retirees in our sample (85%) might explain the high ratio of missing data (21 missing and 110 'don't know') on item 36 ('find out about efforts to promote your health at work'). Rasch modelling was performed on incomplete data (except subtest analyses), whereas mean imputation was used to achieve complete data for the CFA (Polit & Beck 2012). The imputed values were calculated based on the individual's score on the completed items in the respective health domain. On average, there were 8.9 imputed values per item; six items had no imputed values and 24 items had five or fewer imputed values. As recommended by the HLS-EU Consortium (2012), individuals with fewer than 38 responses were excluded from the CFA. Consequently, CFA was based on responses from 318 individuals.

RESULTS

The sample had a slight predominance of males, had an average age of 73 years (Table 1). Approximately one third had completed education up to university or university college level and about the same proportion had completed only compulsory comprehensive school.

Construct validity

Rasch modelling

Testing rating scale unidimensionality

Using PCA, item loadings on the first factor (PC1) were inspected for each pair of subscales (HC, DP and HP). Since the eight items 12, 15 (subscale HC), 18, 28, 30, 31 (subscale DP), 32 and 43 (subscale HP) did not correlate with PC1 as expected from the theoretical assumptions, the PCA only *partially* confirmed the latent subscale structure of the HLS-EU-Q47 framework. Item 12 was of specific concern because it repeatedly did not load as

expected applying the PCA in RUMM. In terms of the *t* tests, the proportion of individuals with significantly different person-location estimates exceeded 16% (Table 2) combining each pair of subscales (HC, DP and HP).

The fractal indexes specific to each subtest were estimated by creating subtests of pairs of subscales (HC, DP and HP; Table 2). For example, by creating a subtest of the subscales HC and DP, the PSI, which provides an estimate of reliability based on an assumption of unidimensionality, dropped from 0.95 (inflated value owing to violations of local independence) to 0.87. A smaller drop in PSI was observed when adjusting for violations of local independence in the scale that combined the highly correlated (r = 0.95) subscales, DP and HP. Furthermore, the composite scale of DP and HP items was recognized by a low unique subscale variance index (c = 0.23) and a high proportion of common true score variance (A = 0.97). Moving top-down in Table 2, we observe a decreasing drop in PSI while adjusting for violations of local independence—the index c decreases, while the A and r indexes mainly increase. These results indicate violations of local independence in the HLS-EU-Q47 dataset and that the health domain HC seems to contribute the most to the magnitude of multidimensionality.

The PCA of residuals and the *t*-test procedures revealed that no health domain subscale was sufficiently unidimensional. However, viewing the four cognitive domains as integrated in the three health domains, the PCA and *t*-test procedure indicated that most of the 12 sub-dimensions could be considered unidimensional, with the exception of the following: appraise within HC, apply within DP and access and apply within HP. Hence, a 12-dimensional model was worth investigating.

Comparing modelling approaches, significant drops in deviance were observed (Table 3) from the 1- to the 3-dimensional model (χ^2 [df = 5] = 574, P <0.05, critical value = 11) and from the 3- to the 12-dimensional model (χ^2 [df = 69] = 3875, P <0.05, critical value = 89).

Reliability, response dependency and targeting

Accounting for violations of local independence, acceptable reliability indexes were observed for the 1- and 3-dimensional models (Table 3). Owing to few items (on average 47/12 ~ 4 items), low reliability indexes were observed for the sub-dimensions when applying the 12-dimensional model. The sub-dimensions 'understand within DP' (B-DP), 'apply within DP' (D-DP) and 'appraise within HP' (C-HP) had low reliability indexes. When applying the 1-, 3- and 12-dimensional approaches, response dependence was observed in 24, 7 and 0 pairs of items, respectively (Table 3).

Centring the item estimates to 0.0 logits on each health domain, the distribution of the HC and the DP person-location estimates could have been better targeted (Table 3). Well-targeted scales were observed only for the sub-dimensions C-HC, D-DP, A-HP and B-HP when the 12-dimensional model was applied.

Item fit, item discrimination and DIF

Table 4 shows the item-specific analyses for different modelling approaches. Items 11, 12, 15, 30, 31 and 45 under-discriminated when applying the consecutive approach for the three health domains and when applying the 3-dimensional model. In addition, items 29 and 47 under-discriminated in the 3-dimensional analysis. Only items 11 and 38 under-discriminated when the 12-dimensional model was applied. Finally, items 27 and 39 over-discriminated in the consecutive and 3-dimensional analyses owing to dependency.

Since the 12-dimensional model (final model) fit the data best, the item-location estimates were reported only for this model (Table 4). With the lowest location estimate (–1.99), item 9 was the easiest to endorse whereas item 28 had the highest location estimate (Table 4).

Applying the 1-dimensional model and the consecutive approach for the three health domains, no items displayed DIF and no unordered response categories were observed. Applying a consecutive approach for the 12 subscales, item 45 displayed uniform DIF, while items 16 and 19 displayed non-uniform DIF associated with gender. No DIF was observed for the age and education factors.

Confirmatory factor analysis

The 12-factor model had the lowest RMSEA value with a tight confidence interval and returned a lower normed χ^2 than the 1- and 3-factor models (Table 5). This strengthened the hypothesis that the HLS-EU-Q47 has a 12-dimensional structure. A consecutive analysis of the three health domains showed that the three independent 1-factor CFA models, did not achieve acceptable fit indexes.

DISCUSSION

Validating interpretations and use of scale scores mean evaluating the claims based on the scores (Kane 2013). Scores are typically used to support claims that a group of respondents have some standing on a latent trait, i.e., some level of achievement in a domain. Our research was driven by a broad question concerning whether there is empirical support for claims about people's HL based on the HLS-EU-Q47 main dimension score and subscale scores. In conclusion, we do not find that our data, collected using the HLS-EU-Q47, support either the assumption of local independence or the requirements of fundamental measurements—requirements supported only when data fit Rasch models. The severity of these measurement problems is exacerbated by the fact that the HLS-EU Consortium, by publishing cut-off scores for 'inadequate' and 'sufficient' HL, interprets the HLS-EU-Q47 sum score as expressing a disparity between an individual's HL and some quantitative value representing

'excellent' HL. To date, there is no consensus among researchers to defend such a definition.

Dimensionality

By combining Rasch modelling with CFA, we found that neither the main dimension HL nor the three health domain subscales of the HLS-EU-Q47 were sufficiently unidimensional. In accordance with an explorative analysis by van der Heide *et al.* (2013) and confirmatory analyses by Duong *et al.* (2015, 2017) and Nakayama *et al.* (2015), we found that a 12-dimensional model, reflecting the hypothesis of the four cognitive domains being integrated in the three health domains (HLS-EU Consortium 2012), fit the data better than the 1- and 3-dimensional models. Despite the theoretically funded 12-dimensional structure described in the conceptual framework (HLS-EU Consortium 2012) and the absence of empirical evidence supporting a unidimensional scale, researchers continually add up the scores to one general index (e.g., Sørensen *et al*. 2013, 2015, Toçi *et al*. 2015, Palumbo *et al*. 2016). This widespread malpractice of adding up the scores, assuming that the HLS-EU-Q47 variables conform to one unidimensional interval scale, cannot be defended from a mathematical point of view. Furthermore, comparing *consecutive* factor analyses of the three health domains, such as those published by Nakayama *et al.* (2015) and Duong *et al.* (2015), to our multidimensional oblique approaches that allow for subdomain covariances, helps elicit the advantage of our multidimensional approaches, as the fit between the data and the model improves. Empirical evidence points to rejecting the HC subscale, as that domain increases the magnitude of multidimensionality.

Our conclusions on HLS-EU-Q47 dimensionality have practical implications for the interpretation of the scale's main dimension score and subscale scores. For example, the rather bold claim, which states, 'half the respondents from eight EU-member countries show limited HL' (HLS-EU Consortium 2012, p. 3) is based on the HLS-EU-Q47 main dimension

score, might be invalid. This statement also takes for granted the cut-off scores for HL defined by the HLS-EU Consortium, for which the Consortium does not provide concrete evidence. Another important implication of our study is that the rather large batch of health-related measurement scales, where the presence of subscales captures the complexity of the underlying trait, might similarly suffer from invalid inferences based on the main scale scores.

At the time of writing, no other known peer-reviewed publications have validated the HLS-EU-Q47 by applying multidimensional Rasch modelling (or any IRT-models). Consequently, we cannot confirm or clarify our conclusions.

Reliability and targeting

Our in-depth analyses suggest that only half of the 12 subscales were well targeted at our sample of individuals. A poorly targeted scale reflects measurement problems and should not be mistakenly interpreted as individuals having high or low standing on a trait. Refining health reforms or adapting health information to meet people's needs, based on poorly targeted scales, may have unintended consequences for public health.

Due to few subscale items, none of the 12 sub-dimensions achieved sufficiently high reliability indexes for measuring aspects of HL at the individual level (see Table 3). However, nine subscales achieved an acceptable reliability index for use at the group level, i.e., for measuring and comparing populations.

The main dimension and the three health domains reached reliability indexes at similar magnitudes to those reported in other studies (HLS-EU Consortium 2012, Duong *et al*. 2015, Nakayama *et al.* 2015, Toçi *et al*. 2015, Plaumbo *et al*. 2016). However, reliability indexes provide values of internal consistency on the assumption of locally independence (unidimensionality and no response dependency)—an assumption that does not hold for the HLS-EU-Q47. Consequently, reliability indexes are invalid measures for the HLS-EU-Q47

main dimension and the three health domains. To improve the targeting of the HLS-EU-Q47, we suggest developing items with higher difficulty (harder to endorse). To improve the reliability, we suggest piloting a six-point rating scale.

Item fit, item discrimination and DIF

Items 11 and 38 under-discriminated relative to the Rasch model expectations when the 12-dimensional model was applied. These items seem to tap into other constructs that do not correlate positively with the latent trait. Item 11 (need for second opinion) might tap into individuals' loyalties to their general practitioners—an aspect that was brought up by interviewees during the cognitive interviews. Furthermore, recalling the high average age in our sample (73 years), the responses to item 38 (food labelling) may have been influenced by the font size on such labels. However, we did not observe DIF associated with age for item 38.

Eight items (11, 12, 15, 29, 30, 31, 45 and 47) under-discriminated when the 3-dimensional model was applied. For example, item 15 (dialling for an ambulance) could, as expressed during cognitive interviews, be interpreted as either the ease of picking up a phone or the difficulty in assessing whether a situation requires an ambulance. Item 29 (decide the need for flu vaccination) might reflect knowledge about potential side effects of vaccines, which could indicate proficiency in critically assessing health information.

All items operationalising the cognitive domain 'apply' in the DP-subscale under-discriminated when the 3-dimensional model was applied. The HLS-EU-Q47 could benefit from new items that refer to, for example, the competency to apply health information to manage unhealthy behaviours; and the ability to communicate interactively with healthcare professionals.

As the mean age of the respondents was 73 years, we inferred that the respondents, on

average, were retired. Therefore, we can effortlessly explain the relatively large proportion of missing responses to item 36 (promoting health at work). Re-wording the item to avoid the 'at work' aspect is one example of adapting the HLS-EU-Q47 to older respondents.

No items displayed DIF when analysing the three health domains applying a consecutive approach. We observed non-uniform DIF associated with gender for items 16 and 19 and uniform DIF for item 45 when the 12-dimensional model was applied. Based on our limited sample of people with T2DM, we suggest further in-depth analyses before rejecting items 16 and 19 from the HLS-EU-Q47.

Gender, as such, does not play a causal role in explaining the uniform DIF observed for item 45 (join sports club or exercise class). We believe that the gender effect possibly stems from variables *associated* with gender. A cursory glance at the information returned when conducting a Google search for 'exercise class' would probably make most people conclude that females easier enrol in exercise classes. For example, re-wording item 45 to 'do exercise compatible with your health condition' could avoid within-item bias associated with gender, age and patient groups, such as people with T2DM, cardiovascular diseases, brain strokes, psychological distress or arthritis (Comins *et al*. 2008).

Limitations

The rather small sample size is a limitation of this study and might lower the generalizability of the findings. As the power to detect misfit and bias increases with the sample size, future studies might detect further weaknesses of the HLS-EU-Q47.

With Rasch modelling, the main concern is sufficient numbers of respondents per threshold. Linacre (1994) recommends at least 250 individuals and 10 extra individuals per response category for polytomous data. One of the most frequently reported lower 'subjects-to-variables (STV)' ratios is five (e.g., Velicer & Fava 1998). Even our main dimension (388

individuals/47 items) has an SVT ratio >8. Therefore, we believe our sample size is sufficient for initiating important discussions and drawing trustworthy conclusions concerning our research question.

Applying CFA, not even the 12-factor model achieved a sufficiently high CFI index and only the 3- and the 12-dimensional approaches obtained acceptable RMSEA values. This could be further investigated after revising the Q47.

CONCLUSION

Our research question, which was concerned with whether scale scores support claims that go beyond the observed performances, is of international significance across contexts and important to everyone making decisions based on rating-scale data. The most important outcome from our study is that the HLS-EU-Q47 sum score, being an operational definition of the latent trait *health literacy*, meets neither the assumption nor the requirements for fundamental measurement. Therefore, our main conclusion is that claims about people's HL should not be inferred from the HLS-EU-Q47 main dimension score or the HLS-EU-Q47 health domain scores.

A 12-dimensional model fit the data best. However, from a clinical point of view, it is impractical to rely on 12 different but related scale scores. Further, some might wrongly compare scores across the 12 subscales and come to invalid conclusions about patients' HL. Ideally, a unidimensional measurement scale should be developed on the basis of the HLS-EU-Q47.

We recommend the following to strengthen the plausibility of claims based on the HLS-EU-Q47 score: removing the health care aspect to reduce the magnitude of multidimensionality; rejecting redundant items to avoid dependency; adding 'harder' items

and applying a six-point rating scale to improve subscale reliability and possibly targeting; and revising items to improve model fit.

Abbreviations

A: accessing health information; B: understanding health information; C: appraising health information; CFA: confirmatory factor analysis; CFI: comparative fit index; D: applying health information; DIF: differential item functioning; DP: disease prevention; HC: healthcare; HL: health literacy; HLS-EU-Q47: European Health Literacy Survey Questionnaire; HP: health promotion; IRT: item response theory; MMLE: marginal maximum likelihood estimation; MRCML: multidimensional random coefficients multinomial logit model; NDA: Norwegian Diabetes Association; PCA: principal component analysis; PMLE: pairwise maximum likelihood estimation; PSI: person separation index; PSR: person separation reliability; RMSEA: root-mean-squared error of approximation; T2DM: type 2 diabetes mellitus; WLE: weighted likelihood estimation.

Author Contributions:

All authors have agreed on the final version and meet at least one of the following criteria (recommended by the ICMJE*):

1) substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data;

2) drafting the article or revising it critically for important intellectual content.

* http://www.icmje.org/recommendations/

REFERENCES

Adams R. & Wu M. (2010) Multidimensional models. In *ConQuest Tutorial*. Available at:

https://www.acer.edu.au/conquest/notes-tutorials.

Adams R.J., Wu M.L. & Wilson M.R. (2015) ACERConQuest: Generalised item response modelling software (Version 4). Camberwell, Victoria: Australian Council for Educational Research.

Adams R.J., Wilson M. & Wang W.-C. (1997) The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement* **21**(1), 1–23.

Al Sayah F., Majumdar S.R., Williams B., Robertson S. & Johnson J.A. (2013) Health literacy and health outcomes in diabetes: a systematic review. *Journal of General Internal Medicine* **28**(3), 444–452. doi: 10.1007/s11606-012-2241-z.

Altin S.V., Finke I., Kautz-Freimuth S. & Stock S. (2014) The evolution of health literacy assessment tools: a systematic review. *BMC Public Health* **14**(1), 1207. doi: 10.1186/1471-2458-14-1207.

Andersen E.B. (1977) Sufficient statistics and latent trait models. *Psychometrika* **42**(1), 69–81.

Andrich D. (1988) *Rasch Models for Measurement*. SAGE Publications, Newsbury Park, CA.

Andrich D. (1989) Distinctions between assumptions and requirements in measurement in the social sciences. In *Mathematical and Theoretical Systems* (Keats J.A., Taft R., Heath R.A. & Lovibond S.H. eds), Vol. 4. North-Holland, Elsevier Science Publishers, Amsterdam, pp. 7–16.

Andrich D. (2009) *Interpreting RUMM2030 Part IV Multidimensionality and subtests in RUMM*. RUMM Laboratory, Perth, Australia.

Andrich D., De Jong J. H. A. L., & Sheridan B. E. (1997) Diagnostic opportunities with the Rasch model for ordered response categories. In *Applications of latent trait and latent class models in the social sciences* (Rost J. & Langeheine R. eds.), pp. 59–70.

Andrich D. & Hagquist C. (2001) *Taking Account of Differential Item Functioning through*

Principals on Equating. Research Report no. 12. WA: Social Measurement Laboratory, Murdoch University, Perth.

Andrich D. & Hagquist C. (2014) Real and Artificial Differential Item Functioning in Polytomous Items. *Educational and Psychological Measurement* **75**(2) 185–207. doi: 10.1177/0013164414534258.

Andrich D., Humphry S.M. & Marais I. (2012) Quantifying local, response dependence between two polytomous items using the. *Applied Psychological Measurement* **36**(4), 309–324.

Andrich, D., Sheridan B. & Luo G. (2003) Rasch Unidimensional Measurment Model. RUMM2030. RUMM Laboratory Pty Ltd, Australia.

Bagnasco A., Di Giacomo P., Da Rin Della Mora R., Catania G., Turci C., Rocco G. & Sasso L. (2014) Factors influencing self- management in patients with type 2 diabetes: a quantitative systematic review protocol. *Journal of Advanced Nursing* **70**(1), 187–200. doi: 10.1111/jan.12178.

Bains S.S. & Egede L.E. (2011) Associations between health literacy, diabetes knowledge, self-care behaviors and glycemic control in a low income population with type 2 diabetes. *Diabetes Technology & Therapeutics* **13**(3), 335–341. doi: 10.1089/dia.2010.0160.

Bock R.D. & Aitkin M. (1981) Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **46**(4), 443–459.

Bohanny W., Wu S.-F.V., Liu C.-Y., Yeh S.-H., Tsay S.-L. & Wang T.-J. (2013) Health literacy, self-efficacy and self-care behaviors in patients with type 2 diabetes mellitus. *Journal of the American Association of Nurse Practitioners* **25**(9), 495–502. doi: 10.1111/1745-7599.12017.

Briggs D.C. & Wilson M. (2003) An introduction to multidimensional measurement using

Rasch models. *Journal of Applied Measurement* **4**(1), 87–100.

Brislin R.W. (1970) Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology* **1**(3), 185–216.

Brown, T.A. (2015) *Confirmatory factor analysis for applied research.* Guilford Publications, New York.

Carlin J.B. & Doyle L.W. (2001) Basic concepts of statistical reasoning: Hypothesis tests and the t- test. *Journal of Paediatrics and Child Health* **37**(1), 72–77.

Comins J., Brodersen J., Krogsgaard M. & Beyer N. (2008) Rasch analysis of the Knee injury and Osteoarthritis Outcome Score (KOOS): a statistical re- evaluation. *Scandinavian Journal of Medicine & Science in Sports* **18**(3), 336–345.

Drennan J. (2003) Cognitive interviewing: verbal data in the design and pretesting of questionnaires. *Journal of Advanced Nursing* **42**(1), 57–63.

Duong V.T., Lin I.-F., Sørensen K., Pelikan J.M., Van Den Broucke S., Lin Y.-C. & Chang P.W. (2015) Health literacy in Taiwan a population-based study. *Asia–Pacific Journal of Public Health* **27**(8), 871–880. doi:10.1177/1010539515607962.

Duong T.V., Aringazina A., Baisunova G., Nurjanah, Pham T.V., Pham K.M., Truong T.Q., Nguyen K.T., Oo W.M., Mohamad E., Su T.T., Huang H.-L., Sørensen K., Pelikan J.M., Van den Broucke S., & Chang P.W. (2017) Measuring health literacy in Asia: Validation of the HLS-EU-Q47 survey tool in six Asian countries. *Journal of Epidemiology*, **27**(2), 80–86. doi: 10.1016/j.je.2016.09.005.

Frisbie D.A. (1988) Reliability of Scores From Teacher- Made Tests. *Educational Measurement: Issues and Practice* **7**(1), 25–35.

Hagell P. (2014) Testing rating scale unidimensionality using the principal component analysis (PCA)/t-test protocol with the Rasch model: The primacy of theory over statistics. *Open Journal of Statistics* **4**, 456–465. doi: 10.4236/ojs.2014.46044.

Hagquist C., Bruce M. & Gustavsson J.P. (2009) Using the Rasch model in nursing research: An introduction and illustrative example. *International Journal of Nursing Studies* **46**(3), 380–393. doi: 10.1016/j.ijnurstu.2008.10.007.

Haun J.N., Valerio M.A., McCormack L.A., Sørensen K. & Paasche-Orlow M.K. (2014) Health literacy measurement: An inventory and descriptive summary of 51 instruments. *Journal of Health Communication* **19**(suppl 2), 302–333. doi: 10.1080/10810730.2014.936571.

HLS-EU Consortium (2012) Comparative report of health literacy in eight EU member states. The European health Literacy Survey HLS-EU. HLS-EU Consortium, Maastricht.

Ishikawa H. & Kiuchi T. (2010) Health literacy and health communication. *BioPsychoSocial Medicine* **4**(1), 18. doi: 10.1186/1751-0759-4-18.

Kane M. T. (2013) Validating the interpretations and uses of test scores. *Journal of Educational Measurement* **50**(1), 1–73.

Katsikatsou M., Moustaki I., Yang-Wallentin F. & Jöreskog K.G. (2012) Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics & Data Analysis* **56**(12), 4243–4258.

...ne R. B. (1998) *Principles and practice of structural equation modeling.* Guilford Press, New York.

Linacre J.M. (1994) Sample Size and Item Calibration Stability. *Rasch Measurement Transactions* **7**(4), 32.

Marais I. & Andrich D. (2008) Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement* **9**(3), 200–215.

Masters G.N. (1982) A Rasch modelRasch model for partial credit scoring. *Psychometrika* **47**(2), 149–174.

Nakayama K., Osaka W., Togari T., Ishikawa H., Yonekura Y., Sekido A. & Matsumoto M. (2015) Comprehensive health literacy in Japan is lower than in Europe: a validated Japanese-language assessment of health literacy. *BMC Public Health* **15**(1), 505. doi: 10.1186/S12889-015-1835-X.

Nguyen T.H., Paasche-Orlow M.K., Kim M.T., Han H.-R. & Chan K.S. (2015) Modern measurement approaches to health literacy scale development and refinement: Overview, current uses and next steps. *Journal of Health Communication* **20**(suppl 2), 112–115. doi: 10.1080/10810730.2015.1073408.

Palumbo R., Annarumma C., Adinolfi P., Musella M. & Piscopo G. (2016) The Italian Health Literacy Project: Insights from the assessment of health literacy skills in Italy. *Health Policy* **120**(9), 1087–1094. doi: 10.1016/j.healthpol.2016.08.007.

Perline R., Wright B.D. & Wainer H. (1979) The Rasch model as additive conjoint measurement. *Applied Psychological Measurement* **3**(2), 237–255.

Polit D.F. & Beck, C.T. (2012) *Nursing Research: Generating and Assessing Evidence for Nursing Practice.* Wolters Kluwer Health, Lippincott Williams & Wilkins, Philadelphia, PA.

Powell C.K., Hill E.G. & Clancy D.E. (2007) The relationship between health literacy and diabetes knowledge and readiness to take health actions. *The Diabetes Educator* **33**(1), 144–151. doi: 10.1177/0145721706297452.

Rasch G. (1980) *Probabilistic Models for Some Intelligence and Attainment Tests*. University of Chicago Press, Chicago.

Rudd R.E., Rosenfeld, L. & Simonds, V.W. (2012) Health literacy: a new area of research with links to communication. *Atlantic Journal of Communication* **20**(1), 16–30.

RUMM Laboratory Pty Ltd (2009) *Extending the RUMM2030 Analysis.* RUMM Laboratory Pty Ltd, Duncraig, Australia.

Ryan J.P. (1983) Introduction to latent trait analysis and item response theory. In: *Testing in the schools. New Directions for Testing & Measurement.*, Vol. 19 (Hataway W.E. ed.) Jossey-Bass, San Francisco, CA, pp. 49–65.

Sarkar U., Karter A.J., Liu J.Y., Moffet H.H., Adler N.E. & Schillinger D. (2010) Hypoglycemia is more common among type 2 diabetes patients with limited health literacy: the Diabetes Study of Northern California (DISTANCE). *Journal of General Internal Medicine* **25**(9), 962–968.

Schillinger D., Grumbach K., Piette J., Wang F., Osmond D., Daher C., Palacios J., Sullivan G.D. & Bindman A.B. (2002) Association of health literacy with diabetes outcomes. *JAMA: the Journal of the American Medical Association* **288**(4), 475–482.

Scholz F.W. (2004) Maximum likelihood estimation. In: *Encyclopedia of Statistical Sciences*. John Wiley & Sons, Inc., Chichester.

Smith E.V. Jr (2002) Understanding Rasch measurement: Detecting and evaluating the impact of multidimenstionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement* **3**(2), 205–231.

Smith R.M. (1995) Using item mean squares to evaluate fit to the Rasch model. Annual Meeting of the American Educational Research Association, San Francisco, CA.

Stenner A.J. (1994). Specific objectivity – local and general. *Rasch Measurement Transactions* **8**(3), 374.

Stewart-Brown S., Tennant A., Tennant R., Platt S., Parkinson J. & Weich S. (2009) Internal construct validity of the Warwick-Edinburgh mental well-being scale (WEMWBS): a Rasch analysis using data from the Scottish health education population survey. *Health and Quality of Life Outcomes* **7**(1), 15–22. doi: 10.1186/1477-7525-7-15.

Sørensen K., Van den Broucke S., Fullam J., Doyle G., Pelikan J., Slonska Z. & Brand H. (2012) Health literacy and public health: A systematic review and integration of

definitions and models. *BMC Public Health* **12**(1), 1–13. doi: 10.1186/1471-2458-12-80.

Sørensen K., Van den Broucke S., Pelikan J.M., Fullam J., Doyle G., Slonska, Z., Kondilis, B., Stoffels, V., Osborne, R.H. & Brand, H. (2013) Measuring health literacy in populations: illuminating the design and development process of the European Health Literacy Survey Questionnaire (HLS-EU-Q). *BMC Public Health* **13**(1), 948. doi: 10.1186/1471-2458-13-948.

Sørensen K., Pelikan J.M., Rothlin F., Ganahl K., Slonska Z., Doyle G., Fullam J., Kondilis B., Agrafiotis D., Uiters E., Falcon M., Mensing M., Tchamov K., van den Broucke S., Brand H. & HLS-EU Consortium (2015) Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). *European Journal of Public Health* **25**(6), 1053–1058. doi: 10.1093/eurpub/ckv043.

Tang Y.H., Pang S.M.C., Chan M.F., Yeung G.S.P. & Yeung V.T.F. (2008) Health literacy, complication awareness and diabetic control in patients with type 2 diabetes mellitus. *Journal of Advanced Nursing* **62**(1), 74–83. doi: 10.1111/j.1365-2648.2007.04526.x.

Tennant A. & Conaghan P.G. (2007) The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied and what should one look for in a Rasch paper? *Arthritis Care & Research* **57**(8), 1358–1362.

Toçi E., Burazeri G., Sørensen K., Kamberi H. & Brand H. (2015) Concurrent validation of two key health literacy instruments in a South Eastern European population. *European Journal of Public Health* **25**(3), 482–486. doi: 10.1093/eurpub/cku190.

van der Heide I., Rademakers J., Schipper M., Droomers M., Sørensen K., & Uiters E. (2013) Health literacy of Dutch adults: a cross sectional survey. *BMC Public Health* **13**(1), 1.

van der Heide I., Uiters E., Rademakers J., Struijs J.N., Schuit A.J. & Baan C.A. (2014) Associations Among Health Literacy, Diabetes Knowledge and Self-Management

Behavior in Adults with Diabetes: Results of a Dutch Cross-Sectional Study. *Journal of Health Communication* **19**(suppl 2), 115–131.

Velicer W.F. & Fava J.L. (1998) Affects of variable and subject sampling on factor pattern recovery. *Psychological Methods* **3**(2), 231.

Warm T.A. (1989) Weighted likelihood estimation of ability in item response theory. *Psychometrika* **54**(3), 427–450.

Wångdahl J., Lytsy P., Mårtensson L. & Westerling R. (2014) Health literacy among refugees in Sweden – a cross-sectional study. *BMC Public Health* **14**(1), 1. doi: 10.1186/1471-2458-14-1030.

Table 1 Sample characteristics (*n* =388)

Characteristic	***n* (%)**
Gender	
- male	207 (53)
- female	167 (43)
- missing	14 (4)
Age	
- mean (sd)	73.0 (8.6)
- median	74
- range	50−92
- missing	13
Education	
- compulsory comprehensive school	111 (29)
- upper secondary school	85 (22)
- university/university college level	118 (30)
- other	51(13)
- missing	23 (6)

Table 2 Testing unidimensionality of the HLS-EU-Q47

Subscales combined in subtest structure	Item set	Proportion (%) of significant *t*-tests[a]	*n*	PSI	PSI[b]	*c*	*A*	*r*[c] (RUMM/ConQuest/AMOS)
HC DP	1–31	17	192	0.95	0.87	0.42	0.92	0.78/0.81/0.84
HC HP	1–16, 32–47	20	166	0.95	0.90	0.34	0.95	0.90/0.78/0.78
DP HP	17–47	16	190	0.95	0.92	0.23	0.97	0.95/0.83/0.88

DP: disease prevention; HC: healthcare; HP: health promotion

n = number of individuals with complete HLS-EU-Q47 responses; PSI (person separation index), *c* (unique subscale variance) and *A* (common subscale variance) were estimated on complete dataset. Large estimates for *A* and *r*, and a small estimate for *c* imply unidimensional data.

[a]Estimated using RUMM2030 (incomplete dataset, *n* = 388).

[b]Adjusted for violations of local independence (response dependency and multidimensionality) by creating subtest structure. Large drop in PSI implies 'significant' violations of local independence.

[c]Subscale correlation (*r*) estimates based on subtest structure in RUMM2030 (complete dataset), 3-dimensional Rasch model using ConQuest 4 (incomplete dataset) and 3-factor CFA using AMOS (imputed dataset), respectively.

Table 3 Global scale characteristics of the HLS-EU-Q47 with fit statistics and information criteria for the different analysis approaches

Analysis approach	Subscale(s)	No. of items	Items	α (SPSS)	PSI based on PMLE (RUMM)	PSR based on MMLE/ WLE	Mean person location (SE) in logits (CQ)	Response dependence between pairs of items	Deviance (CQ)	No. of para-meters (CQ)	Change in deviance from 1-dim.
Consecutive	HC	16	1–16	0.91	0.89	0.89/0.89	1.21 (0.10)	1/2			N/A
	DP	15	17–31	0.90	0.87	0.88/0.88	1.12 (0.09)	22/23, 28/31, 30/31			N/A
	HP	16	32–47	0.90	0.88	0.88/0.88	0.48 (0.08)	41/42, 44/46, 45/47			N/A
1-dimensional[a]	HC+DP+HP	47	1–47	0.96[b]	0.95[b]	0.95/0.95[b]	0.80 (NR)	24 pairs	26 566	140	baseline
3-dimensional	HC	47	1–47	N/A	N/A	0.89/0.90	1.25 (0.10)	N/A	25 992	145	574
	DP					0.88/0.88	1.15 (0.01)				
	HP					0.88/0.88	0.47 (0.08)				
12-dimensional[c]	A-HC	47	1-47	N/A	N/A	0.73/0.73	1.62 (0.13)	N/A	22117	214	3875
	B-HC					0.69/0.69	0.76 (0.09)				
	C-HC					0.68/0.68	0.38 (0.09)				
	D-HC					0.65/0.66	1.42 (0.10)				
	A-DP					0.69/0.69	1.19 (0.09)				
	B-DP					0.51/0.52	2.14 (0.10)				
	C-DP					0.78/0.78	0.98 (0.11)				
	D-DP					0.55/0.52	0.14 (0.06)				
	A-HP					0.70/0.70	0.25 (0.08)				
	B-HP					0.75/0.74	0.31 (0.11)				
	C-HP					0.61/0.63	1.30 (0.11)				
	D-HP					0.70/0.69	0.73 (0.00)				

Table showing statistics applying the consecutive approach (treating the three health domains as orthogonal or uncorrelated subscales), the 1-dimensional approach (treating the three health domains as three parallel subscales), the 3-dimensional approach (treating the three health domains as correlated subscales) and the 12-dimensional approach (treating the four cognitive domains within the three health domains as 12 correlated subscales) (software applied is reported in parentheses).

[a]Initial analysis using RUMM2030 confirmed multidimensional data. The 1-dimensional approach is reported only for comparison with the 3- and 12-dimensional approaches.

[b]Measures of reliability are valid only when data are unidimensional. Values are reported only for comparison.

[c]12-dimensional approach consists of the dimensions access, understand, appraise and apply within HC, access, understand, appraise and apply within DP, and access, understand, appraise and apply within HP.

Well-targeted scales are centred to zero logits.
Smaller values of deviance indicate better fit.
α: Cronbach's α (internal consistence reliability); A: access; B: understand; C: appraise; CQ: ConQuest 4 software; D: apply; Deviance: deviance statistics. DP: disease prevention (15 items); HC: healthcare (16 items), HP: health promotion (16 items); Mean person-location: mean person estimate when mean item estimate is set to zero; MMLE: marginal maximum likelihood estimate; N/A: not applicable; NR: not reported; PMLE: pairwise maximum likelihood estimate; PSI: person separation index; PSR: person separation reliability; RUMM: RUMM2030 software; SPSS: SPSS 23 software; WLE: Warm's mean likelihood estimate.

Table 4 Single item characteristics for different analysis approaches using the ConQuest 4 software.

			Consecutive analysis (three independent health domains)				3-dimensional analysis (three correlated health domains)				12-dimensional analysis (12 correlated sub-dimensions defined by the four cognitive domains integrated within the three health domains)						Consecutive analysis (12 independent subscales)
	CD	Item	Infit	CI lb	CI ub	T	Infit	CI lb	CI ub	T	Infit	CI lb	CI ub	T	Item estimate	SE	DIF
		On a scale from very difficult to very easy, how easy would you say it is to:															
HC	A	1) find information about symptoms of illnesses that concern you	1.03	0.83	1.17	0.4	1.04	0.83	1.17	0.5	1.05	0.83	1.17	0.6	0.02	0.10	
	A	2) find information on treatments of illnesses that concern you	0.99	0.84	1.16	–0.1	0.99	0.85	1.15	–0.2	1.00	0.84	1.16	0	0.28	0.10	
	A	3) find out what to do in case of a medical emergency	1.04	0.84	1.16	0.5	1.07	0.85	1.15	0.9	1.09	0.84	1.16	1.1	0.73	0.10	
	A	4) find out where to get professional help when you are ill	0.94	0.81	1.19	–0.6	0.96	0.80	1.20	–0.4	1.14	0.79	1.21	1.3	-1.04	0.17	
	B	5) understand what your doctor says to you	0.89	0.84	1.16	–1.4	0.95	0.84	1.16	-0.6	0.97	0.82	1.18	-0.3	0.05	0.09	
	B	6) understand the leaflets that come with your medicine	1.08	0.84	1.16	1.0	1.07	0.85	1.15	0.9	1.07	0.83	1.17	0.8	-0.28	0.08	
	B	7) understand what to do in a medical emergency	0.88	0.85	1.15	–1.7	0.91	0.86	1.14	–1.2	0.88	0.85	1.15	-1.5	0.21	0.08	
	B	8) understand your doctor´s or pharmacist´s instruction on how to take a prescribed medicine	0.83	0.81	1.19	–1.8	0.86	0.81	1.19	–1.5	0.85	0.80	1.20	-1.5	0.01	0.15	
	C	9) judge how information from your doctor applies to you	0.93	0.82	1.18	–0.8	0.94	0.82	1.18	–0.7	0.97	0.82	1.18	-0.3	-1.99	0.09	
	C	10) judge the advantages and disadvantages of different treatment options	0.89	0.85	1.15	–1.5	0.91	0.85	1.15	–1.1	0.87	0.84	1.16	-1.7	0.57	0.08	
	C	11) judge when you may need to get a second opinion from another doctor	**1.25**[a]	0.85	1.15	**3**[a]	**1.26**[a]	0.84	1.16	**3.0**[a]	**1.21**[a]	0.84	1.16	**2.5**[a]	0.26	0.08	

C	12) judge if the information about illness in the media is reliable	**1.27**[a]	0.85	1.15	**3.2**[a]	**1.24**[a]	0.84	1.16	**2.9**[a]	1.06	0.83	1.17	0.7	1.16	0.14	
D	13) use information the doctor gives you to make decisions about your illness	0.93	0.83	1.17	–0.9	0.92	0.84	1.16	–1.0	0.93	0.82	1.18	-0.8	-0.02	0.10	
D	14) follow the instructions on medication	0.97	0.82	1.18	–0.3	1.01	0.82	1.18	0.1	1.17	0.80	1.20	1.6	-0.16	0.10	
D	15) call an ambulance in an emergency	**1.23**[a]	0.84	1.16	**2.6**[a]	**1.22**[a]	0.84	1.16	**2.5**[a]	1.17	0.83	1.17	1.8	-0.56	0.09	
D	16) follow instructions from your doctor or pharmacist	0.99	0.82	1.18	–0.1	1.00	0.82	1.18	0	1.01	0.81	1.19	0.1	0.74	0.17	Gender[+]
DP																
A	17) find information about how to manage unhealthy behaviour such as smoking, low physical activity and drinking too much	1.05	0.85	1.15	0.6	1.06	0.85	1.15	0.8	1.09	0.83	1.17	1	-0.82	0.08	
A	18) find information on how to manage mental health problems like stress or depression	1.07	0.84	1.16	0.8	1.08	0.84	1.16	0.9	0.97	0.83	1.17	-0.4	1.28	0.08	
A	19) find information about vaccinations and health screening that you should have	0.92	0.84	1.16	–0.9	0.95	0.84	1.16	-0.6	0.98	0.83	1.17	-0.2	-0.34	0.08	Gender[+]
A	20) find information on how to prevent or manage conditions like being overweight, high blood pressure or high cholesterol	0.99	0.85	1.15	–0.1	1.02	0.85	1.15	0.3	0.96	0.83	1.17	-0.5	-0.12	0.14	
B	21) understand health warnings about behaviour such as smoking, low physical activity and drinking too much	0.89	0.85	1.15	–1.5	0.93	0.85	1.15	–0.9	1.01	0.82	1.18	0.1	-0.42	0.09	
B	22) understand why you need vaccinations	1.03	0.85	1.15	0.5	1.10	0.85	1.15	1.3	1.08	0.82	1.18	0.9	-0.21	0.09	
B	23) understand why you need health screenings	0.9	0.86	1.14	–1.4	0.92	0.86	1.14	–1.1	0.90	0.83	1.17	-1.2	0.63	0.13	
C	24) judge how reliable health warnings are, such as smoking, low physical activity and drinking too much	0.9	0.86	1.14	–1.4	0.92	0.85	1.15	–1.1	0.97	0.83	1.17	-0.3	0.18	0.09	

C	25) judge when you need to go to a doctor for a check-up	1.07	0.84	1.16	0.8	1.10	0.84	1.16	1.2	1.16	0.83	1.17	1.8	-0.69	0.09
C	26) judge which vaccinations you may need	0.88	0.86	1.14	–1.8	0.92	0.86	1.14	–1.1	1.02	0.85	1.15	0.2	0.29	0.09
C	27) judge which health screenings you should have	**0.84**[b]	0.85	1.15	**–2.2**[b]	**0.84**[b]	0.85	1.15	**–2.1**[b]	0.92	0.83	1.17	-1	-1.24	0.09
C	28) judge if the information on health risks in the media is reliable	0.95	0.84	1.16	-0.7	0.98	0.84	1.16	–0.3	1.12	0.83	1.17	1.4	1.45	0.18
D	29) decide if you should have a flu vaccination	1.12	0.85	1.15	1.6	**1.19**[a]	0.86	1.14	**2.4**[a]	1.01	0.85	1.15	0.2	-0.82	0.07
D	30) decide how you can protect yourself from illness based on advice from family and friends	**1.16**[a]	0.85	1.15	**2.0**[a]	**1.19**[a]	0.85	1.15	**2.3**[a]	0.88	0.84	1.16	-1.5	0.36	007
D	31) decide how you can protect yourself from illness based on information in the media	**1.16**[a]	0.85	1.15	**2.0**[a]	**1.17**[a]	0.85	1.15	**2.1**[a]	0.89	0.84	1.16	-1.5	0.47	0.10
A	32) find information on healthy activities such as exercise, healthy food and nutrition	0.97	0.83	1.17	–0.3	0.94	0.83	1.17	–0.6	0.94	0.82	1.18	-0.6	-1.17	0.08
A	33) find out about activities that are good for your mental well-being	0.88	0.84	1.16	–1.4	0.87	0.84	1.16	–1.6	0.89	0.83	1.17	-1.3	-0.91	0.08
A	34) find information on how your neighbourhood could be more health friendly	0.9	0.84	1.16	–1.2	0.95	0.84	1.16	–0.6	0.91	0.84	1.16	-1.1	0.41	0.08
A	35) find out about political changes that may affect health	1.1	0.84	1.16	1.2	1.13	0.84	1.16	1.6	1.03	0.84	1.16	0.4	1.09	0.08
A	36) find out about efforts to promote your health at work	0.99	0.83	1.17	–0.1	1.02	0.83	1.17	0.3	1.01	0.83	1.17	0.1	0.57	0.16
B	37) understand advice on health from family members or friends	1	0.84	1.16	0.1	1.01	0.84	1.16	0.2	0.97	0.83	1.17	-0.3	-0.01	0.08
B	38) understand information on food packaging	1.14	0.86	1.14	1.9	1.13	0.86	1.14	1.7	**1.21**[a]	0.85	1.15	**2.5**[a]	0.33	0.08
B	39) understand information in the media on how to get healthier	**0.84**[b]	0.85	1.15	**–2.2**[b]	**0.85**[b]	0.85	1.15	**–2.1**[b]	0.87	0.84	1.16	-1.7	-0.05	0.08
B	40) understand information on how to keep your mind healthy	0.85	0.85	1.15	–2.0	0.87	0.85	1.15	–1.7	0.97	0.84	1.16	-0.3	-0.27	0.14
C	41) judge how where you live affects your health and wellbeing	0.97	0.84	1.16	–0.3	1.01	0.84	1.16	0.1	1.09	0.83	1.17	1	0.53	0.09

C	42) judge how your housing conditions help you to stay healthy	0.98	0.83	1.17	–0.2	1.02	0.83	1.17	0.3	1.08	0.83	1.17	1	0.18	0.09	
C	43) judge which everyday behaviour is related to health	0.94	0.84	1.16	–0.7	0.93	0.84	1.16	–0.8	1.02	0.83	1.17	0.2	-0.71	0.12	
D	44) make decisions to improve your health	1.02	0.86	1.14	0.3	0.98	0.86	1.14	–0.2	0.95	0.85	1.15	-0.7	-0.25	0.07	
D	45) join a sports club or exercise class if you want to	**1.25**[a]	0.84	1.16	**2.8**[a]	**1.29**[a]	0.84	1.16	**3.3**[a]	1.11	0.83	1.17	1.3	0.26	0.07	Gender[#]
D	46) influence your living conditions that affect your health and wellbeing	1.12	0.85	1.15	1.6	1.10	0.85	1.15	1.3	1.05	0.85	1.15	0.7	-0.34	0.07	
D	47) take part in activities that improve health and wellbeing in your community	1.16	0.85	1.15	1.9	**1.22**[a]	0.85	1.15	**2.7**[a]	1.08	0.84	1.16	0.9	0.33	0.12	

The table shows item fit indices applying the consecutive approach (treating the three health domains as orthogonal or uncorrelated subscales), the 3-dimensional approach (treating the three health domains as correlated subscales) and the 12-dimensional approach (treating the four cognitive domains within the three health domains as 12 correlated subscales). Item-location estimates with standard errors are reported only for the 12-dimensional model (the final model). Note that the original response scale was reversed to obtain a response scale where 1 means 'very difficult' and 4 means 'very easy'. Analyses of differential item functioning (DIF) was based on a consecutive approach (an orthogonal or restricted model) where the 12 subscales were analysed independently

Theoretically defined domains in the EU-HLS framework; HD: health domain; CD: cognitive domain
Theoretically defined health domain subscales; DP: disease prevention (15 items); HC: healthcare (16 items); HP: health promotion (16 items)
Theoretically defined cognitive domain subscales: A: access (13 items); B: understand (11 items); C: appraise (12 items); D: apply (11 items)

CI lb: confidence interval lower bound, CI ub: confidence interval upper bound; Infit: infit mean-square fit statistics (above/below 1 indicates an under-/over-discriminating item relative to Rasch-model expectations); DIF: differential item functioning; SE: standard error; T: an infit value above/below the confidence interval has *t*–value <–1.96 or >+1.96.

[a]A *t*–value >1.96 indicates a poorly fitting item owing to under-discrimination relative to the Rasch-model.
[b]A *t*–value <–1.96 indicates an "over-fitting" item owing to over-discrimination relative to the Rasch-model.
Item estimate: item-location estimate in log-odds unit or "logit"(a large negative/positive value means 'easy/hard to endorse').
[#] uniform DIF
[+] non-uniform DIF
No items showed unordered response categories (not reported in the table).

Table 5 Fit statistics for different factor structures using confirmatory factor analyses

Analysis approach	Subscale(s)	χ^2/ d.f.	RMSEA (CI)	CFI
Consecutive	HC	4.7	0.108 (0.098–0.118)	0.821
	DP	6.3	0.129 (0.119–0.139)	0.760
	HP	6.3	0.129 (0.120–0.139)	0.752
1-dimensional	HC+DP+HP	3.5	0.088 (0.085–0.091)	0.654
3-dimensional	HC DP HP	2.9	0.078 (0.074–0.081)	0.733
12-dimensional	A, B, C, D in HC A, B, C, D in DP A, B, C, D in HP	2.4	0.067 (0.063-0.070)	0.815

Table showing fit statistics applying a consecutive approach (treating the three health domains as orthogonal or uncorrelated subscales, the 1-dimensional approach (treating the three health domains as three parallel subscales), the 3-dimensional approach (treating the three health domains as correlated subscales) and the 12-dimensional approach (treating the four cognitive domains within the three health domains as 12 correlated subscales).

A: access, B: understand; C: appraise; CI: confidence interval; D: apply; DP: disease prevention; HC: health care; HP: health promotion.

χ^2/ d.f. (normed chi square)**:** values <3.0 are deemed as acceptable; RMSEA (root-mean-squared error of approximation): values <0.06 indicate good model fit, values <0.08 could be deemed acceptable; CFI (comparative fit index): values >0.9 are considered acceptable. Analyses were performed using SPSS 23 AMOS.