Development and validation of the HLS-EU-Q12

Karin Waldherr, Tobias Alfers, Sandra Peer, and Jürgen M. Pelikan

Content

Vienna, 2019

1. Background

The European Health Literacy Questionnaire (HLS-EU-Q47) consists of 47 items, which according to the underlying conceptual model, address a matrix of 3 by 4 domains resulting in 12 elements of the health literacy (HL) conceptual matrix (cf. Sørensen et al., 2013; Sørensen et al., 2015; Pelikan & Ganahl, 2017). Accordingly, the 47 items assess self-reported difficulties in the four cognitive domains accessing, understanding, appraising and applying information relevant for taking decisions in the three health domains healthcare, disease prevention and health promotion (Sørensen et al., 2013; Sørensen et al., 2015). Participants are asked to rate each item on a 4-point Likert like scale (very easy, fairly easy, fairly difficult, very difficult). Furthermore, they have the option to choose "don't know".

The items were developed in English and then translated into Bulgarian, Dutch, German, Greek, Polish and Spanish. The psychometric properties of the questionnaire was investigated using Principal Component Analysis (PCA) and reliability analysis using data from a field test conducted in Ireland and the Netherlands (for more details on the development process see Sørensen et al., 2013; Sørensen et al., 2015). The HLS-EU-Q47 was applied in the first wave of the European Health Literacy Survey in eight countries (HLS-EU-8): Austria (AT), Germany (only North-Rhine-Westphalia, DE), Spain (ES), Ireland (IE), The Netherlands (NL), Bulgaria (BG), Poland (PL), and Greece (EL). Data was collected either by Computer Assisted Personal Interviewing (CAPI) or Paper Assisted Personal Interviewing (PAPI). Recruitment strategies varied between countries (cf. Pelikan & Ganahl, 2017).

Using data from HLS-EU-8, four main index scores were constructed for "general HL" (comprising all 47 items), "healthcare literacy", "disease prevention literacy" and "health promotion literacy", and reliability for these indexes was assessed using Cronbach's α. The Cronbach α's for all four indexes across all eight countries were at least 0.87 and the item correlations with the total scales exceeded 0.30 (HLS-EU Consortium, 2012). Furthermore, in order to justify the usage of an overall sum score, Item Response Theory (IRT) analysis was applied to examine unidimensionality of the HLS-EU-Q47. The Rating Scale Model (RSM; Andrich, 1978) was used with the four-point scale and the Rasch Model (RM) with dichotomized data (very easy / fairly easy vs. fairly difficult / very difficult). The RSM analysis showed poor model fit. To test the fit of the RM to the data Likelihood Ratio Tests (Andersen, 1973) using the split criteria median test score, gender and dichotomized educational level were conducted for each of the eight countries. As result from these analyses a 16-item version was proposed (HLS-EU-Q16; only unpublished manuscript available; for more details see Pelikan & Ganahl, 2017). Correlations between the indexes of this short version and the 47-item version varied between r = 0.73 and r = 0.88 in the different countries (cf. Pelikan & Ganahl, 2017). However, the HLS-EU-Q16 does not include an item of the element „apply information" in the "health promotion" domain of the HL conceptual matrix.

In the last years, both in Norway (HLS-Q12; Finbråten et al., 2017; Finbråten et al., 2018) and in Taiwan (HL-SF12; Duong et al., 2017) 12-item versions of the HLS-EU were developed in which each of the elements of the HL conceptual matrix is represented by one item. Whereas Finbråten et al. (2017) and Finbråten et al. (2018) applied Confirmatory Factor Analysis (CFA) and IRT, Duong et al. (2017) used only CFA to examine the psychometric properties of their 12-item version. However, only 50% of the items of these two 12-item versions are overlapping (see Table 1). Four of this six items are also contained in the HLS-EU-Q16.

Table 1: Items included in the HLS-EU-Q16, the HLS-Q12 (Finbråten et al., 2017; Finbråten et al., 2018) and the HL-SF12 (Duong et al., 2017)

	Item number in the HLS-EU-Q47		
HL conceptual matrix element	HLS-EU-Q16	HLS-Q12	HL-SF12
1 (acess information, healthcare)	2, 4	2	2
2 (understand information, healthcare)	5, 8	7	6
3 (appraise information, healthcare)	11	10	10
4 (apply information, healthcare)	13, 16	14	15
5 (acess information, disease prevention)	18	18	18
6 (understand information, disease prevention)	21, 23	23	23
7 (appraise information, disease prevention)	28	28	26
8 (apply information, disease prevention)	31	30	30
9 (acess information, health promotion)	33	32	33
10 (understand information, health promotion)	37, 39	38	39
11 (appraise information, health promotion)	43	43	43
12 (apply information, health promotion)	-	44	45

A short version representing all 12 elements of the HL conceptual matrix by one item which sufficiently meets the requirements of a unidimensional IRT model is highly desirable for several reasons. Therefore, in preparation of the second wave of the European Health Literacy Survey (HLS_{19}) new IRT-analyses using data from HLS-EU-8 were conducted with the goal to select a subsample of items - the HLS-EU-Q12 - which should fulfill the following criteria.

The HLS-EU-Q12 should

1. represent all 12 elements of the HL conceptual matrix by one item,
2. include as many items from the HLS-EU-Q16 as possible (cf. Table 1),
3. show the greatest possible overlap with the HLS-Q12 (Finbråten et al., 2017; Finbråten et al., 2018; cf. Table 1), and
4. represent a close to optimal 12-item solution, i.e. the solution with the lowest deviance from the assumptions of the Partial Credit Model (PCM; Masters, 1982) when analyzed separately for each HLS-EU-8 country.

In the following, the development of the HLS-EU-Q12 based on HLS-EU-8 data and its validation using data from HLS_{19} is described.

2. Development of the HLS-EU-Q12

2.1 Methods

Participants

Analyses are based on data from all eight countries of the HLS-EU-8 study collected in 2011. A detailed description of the HLS-EU-8 recruitment strategies in the different countries can be found elsewhere (e.g. Sørensen et al., 2015; Pelikan & Ganahl, 2017). Across all HLS-EU-8 countries data from n = 8102 persons were available whereby sample sizes varied between n = 1000 and n = 1057 in the individual countries (see Table 2).

Table 2: Sample sizes in the 8 HLS-EU countries

		Sample size
Country	AT	1015
	BG	1002
	EL	1000
	ES	1000
	IE	1005
	NL	1023
	PL	1000
	DE	1057
	Total	8102

Data analysis

The data set was divided randomly into a training data set (n = 4054) and a test data set (n = 4048). An iterative IRT analysis approach combined with expert judgement on content validity was chosen, including the following steps:

1) PCM analysis of HLS-EU-Q47 on the training data set across all HLS-EU-8 countries (n = 4054): The goal was to find additional items on top of HLS-EU-Q16 which could be used for item selection for the HLS-EU-Q12.
2) Selection of additional items based on the results of the PCM analysis (Step 1) and expert judgement on content validity (exclude low priority items).
3) PCM analysis of HLS-EU-Q16 plus additional items chosen in Step 2 using the test data set (n = 4048) for each HLS-EU-8 country separately: The aims of this step were to evaluate the item selection of Step 2 for each of the HLS-EU-8 country and to find the HLS-EU-Q12 solution with the best fit to the PCM.
4) PCM analysis of the selected 12 items (from Step 3) on the same test data set (n = 4048) for each HLS-EU-8 country. Since in Step 3 some items have been removed, the remaining 12 items were retested to evaluate if the scale has been affected (cf. Robinson et al., 2019).
5) Comparison of PCM model fit of the different questionnaire versions (HLS-EU-Q47, HLS-EU-Q16, HLS-EU-Q12, HLS-Q12, HL-SF12) using the test data set (n = 4048) for each HLS-EU-8 country, examination of the correlations of the HLS-EU-Q12 with the HLS-EU-Q47, the HLS-EU-

Q16 and the Newest Vital Sign test (NVS; Weiss et al., 2005[1]), and calculation of Cronbach's α as well as item-total correlations for the HLS-EU-Q12. In order to calculate the correlations of the HLS-EU-Q12 with the HLS-EU-Q47, HLS-EU-Q16 and the NVS, indices of HL were constructed as described in Sørensen et al. (2015).

PCM analysis:

All analyses were conducted in R 3.5.1 (https://cran.r-project.org/) using the packages TAM 3.1-45 (Robitzsch, Kiefer & Wu, 2019, 2020), sirt 3.3.-26 (Robitzsch, 2019), and mirt (Chalmers, 2012; version 1.30). Persons with more than 3 missing values were excluded. The PCM with ConQuest parametrization was used (Robitzsch, Kiefer & Wu, 2019).

In Steps 1, 3 and 4, item infit statistics and corresponding *t*-statistics were calculated for each item. The expected value is 1; values > 1 indicate that the item is less predictable than what would be expected according to the IRT model (underfit), values < 1 mean that the item is more predictable than what would be expected according to the expectations of the IRT model (= overfit; Linacre & Wright, 1994, p. 360). Underfitting items may severely degrade the measurement, whereas overfitting items may overestimate raw score differences (Smith et al., 2008). The Holm procedure was applied to adjust the *p*-values for multiple testing (cf. Robitzsch, Kiefer & Wu, 2019). Items were interpreted as over-/underfitting if the adjusted *p*-value was ≤ 0.05. The Nominal Categories Model was applied to check whether the expected ordering of response categories is supported by the data (Thissen, Cai & Bock, 2010; Chalmers et al., 2019, p. 100). Differential item functioning (DIF) analyses were conducted using gender and the dichotomized criteria age (median split) and education (< higher education entrance qualification vs. at least higher education entrance qualification). A facets analysis was conducted. The criteria were set up as facets (e.g. for gender, item+gender+item*gender), and the IRT analysis was rerun (Robitzsch, Kiefer & Wu, 2019). The interaction term item*gender yields the DIF magnitude.

For the comparison of PCM model fit of the different questionnaire versions (Step 5), *SRMSR* (standardized root mean square residual; Maydeu-Olivares, 2013) was calculated for each of the questionnaire versions (cf. Robitzsch, Kiefer & Wu, 2019). *SRMSR* is a global fit statistic based on the comparison of residual correlations of item pairs. Maydeu-Olivares suggests a cutoff of ≤ 0.05 for well-fitting IRT models. A less conservative value of 0.08 often is used as acceptable (cf. Hu & Bentler, 1999). Furthermore, the combined PCA / t-test protocol to examine unidimensionality (cf. Smith, 2002; Hagell, 2014) was applied to the different versions. Two subsets of items are formed based on a PCA of standardized item residuals pursuant to the loadings of the item residuals on the first principal component (cf. Hagell, 2014). Person parameter estimation is conducted in each of the two item subsets and the resulting person parameter estimates from the two subsets are compared by means of paired *t*-tests (cf. Hagell, 2014). Under the assumption of unidimensionality, the proportion of individuals with significantly different person parameters in the two item subsets is small, i.e. ≤ 5% of the *t*-tests are significant, or the lower bound of a 95% confidence interval (CI) of the observed proportion overlaps 5% (Hagell, 2014). In our analysis the Agresti-Coull CI was used. *WLE* reliability and *EAP* (expected a posteriory) reliability coefficients were calculated according to Adams (2005) (cf. Robitzsch, Kiefer & Wu, 2019). Additionally, deviance, Akaike's Information Criterion (*AIC*; Akaike, 1973), the *AIC* correction for small samples (*AICc*; Hurvich & Tsai, 1989), Bozdogan's (1987) consistent

[1] The NVS is a 6-item screening instrument for functional health literacy and is based on the ability to read, understand and apply information from a nutrition label.

AIC (*CAIC*) and the Bayesian Information Criterion (*BIC*; Schwarz, 1978) were calculated to compare the data-model fit for the different versions. Lower values indicate better data-model fit.

2.2 Results

Step 1:

No unordered response categories were observed. Seven items of the HLS-EU-Q16 had significant infit statistics (see Table A1 in the Appendix). Overfit was observed for six items (items 13, 21, 23, 33, 39, 43), and underfit was observed for item 28 with an infit statistic of 1.10 (t = 4.60, p < 0.001). Another 17 items of the remaining items of the HSL-EU-Q47 had significant infit statistics. DIF was observed for 12 items of the HLS-EU-Q16 (items 5, 18, 39 for age; items 8, 21, 23, 28, 31, 37 for education, and items 2, 11 and 33 for age and education). For 6 items of the HLS-EU-Q47 which are not included in the HLS-EU-Q16 neither over-/underfit nor DIF was observed (see Table A1). As in previous analyses, the most problematic subdomain was "health promotion"; only for two items of this subdomain neither DIF nor over-/underfit was observed.

Step 2:

Six items were candidates to be selected as additional items on top of the 16 items of the HLS-EU-Q16 according to the results of Step 1. Two of them were judged as low priority items and were not considered. Thus, four items were selected: 7, 10, 24, 44. Furthermore, it was decided to include two additional items from the health promotion domain, although they showed DIF in the training data set (items 36 and 42), such that each of the four cognitive domains (access, understand, appraise and apply) was represented by two items. This resulted in six additional items, whereby three of them are included in the HLS-Q12 (Finbråten et al., 2017).

Step 3:

Using the test data set, only for items 28 (infit: 1.23, t = 3.47, p = 0.032) and 36 (infit: 1.46, t = 5.75, p < 0.001) significant underfit was observed in Germany (see Table A2 in the Appendix). DIF for age was observed for item 2 in four countries (AT, EL, ES, NL), for item 33 in two countries (BG, EL), and for item 23, 39 and 42 in one country. DIF for gender was found for items 5 and 43 in one country, and DIF for education was found for item 8 in two countries as well as for items 11, 21 and 31 in one country.

The proposed solution fulfilling the abovementioned criteria (represent all 12 elements of the HL conceptual matrix by one item, include as many items of the HLS-EU-Q16 as possible, show the greatest possible overlap with the HLS-Q12 and showing the lowest deviance from the assumptions of the PCM across all HLS-EU-8 countries) consisted of the items 4, 7, 10, 16, 18, 23, 24, 31, 33, 37, 42, 44.

Step 4:

PCM analysis of the selected 12 items in each of the 8 countries revealed no significant infit statistics, however DIF for education for item 31 in Austria and for item 33 in Bulgaria, and DIF for age for item 33 in Bulgaria and Greece as well as for item 42 in The Netherlands (see Table A3 in the Appendix).

Step 5:

SRMSR-values for the HLS-EU-Q12 were < 0.08 in the individual countries and thus are acceptable (Table 3). For the HLS-Q12 values > 0.08 were observed in two countries and for the HL-SF12 in three countries.

Table 3: SRMSR values for the different versions in the eight countries

Country	HLS-EU-Q47	HLS-EU-Q16	HLS-EU-Q12	HLS-Q12	HL-SF12
AT	**0.0920**	**0.0828**	0.0712	0.0768	0.0775
BG	**0.0891**	0.0773	0.0696	0.0713	**0.0900**
EL	**0.1031**	**0.0912**	0.0769	**0.0936**	**0.1031**
ES	**0.0920**	0.0748	0.0683	0.0653	0.0760
IE	**0.0949**	**0.0885**	0.0789	0.0776	0.0716
NL	**0.0947**	**0.0885**	0.0745	0.0742	0.0767
PL	0.0795	0.0643	0.0595	0.0543	0.0644
DE	**0.1129**	**0.0921**	0.0798	**0.1105**	**0.0898**

For each of the three 12-item versions the proportions of significant *t*-tests were > 5% in all countries, and only in one country the lower bound of the 95% CI included 5% for each of the versions (Table 4). In three countries the proportion of significant *t*-tests was lowest for the HLS-EU-Q12 (ES, IE, NL), and in one country for the HL-SF12 (BG). In AT, EL and DE the proportion was comparable for the HLS-EU-Q12 and HL-SF12, and in PL it was comparable for the HLS-Q12 and HL-SF12.

Table 4: Results of PCA/t-test procedure (proportion of significant t-tests and lower bound of CI)

Country	HLS-EU-Q47	HLS-EU-Q16	HLS-EU-Q12	HLS-Q12	HL-SF12
AT	**0.259 (0.224)**	**0.184 (0.154)**	**0.086 (0.065)**	**0.121 (0.096)**	**0.082 (0.061)**
BG	**0.310 (0.271)**	**0.158 (0.129)**	**0.125 (0.099)**	**0.104 (0.080)**	**0.087 (0.066)**
EL	**0.271 (0.234)**	**0.159 (0.130)**	**0.101 (0.077)**	**0.114 (0.089)**	**0.098 (0.076)**
ES	**0.253 (0.217)**	**0.179 (0.147)**	**0.101 (0.077)**	**0.113 (0.088)**	**0.143 (0.115)**
IE	**0.289 (0.251)**	**0.139 (0.111)**	**0.084 (0.062)**	**0.102 (0.078)**	**0.104 (0.080)**
NL	**0.216 (0.182)**	**0.141 (0.112)**	0.069 (0.049)	**0.107 (0.082)**	**0.074 (0.054)**
PL	**0.234 (0.199)**	**0.103 (0.079)**	**0.078 (0.057)**	0.066 (0.047)	0.063 (0.044)
DE	**0.291 (0.253)**	**0.146 (0.118)**	**0.078 (0.058)**	**0.143 (0.115)**	**0.076 (0.056)**

When comparing the three 12-item versions by means of deviance and information criteria, the HLS-EU-Q12 showed best fit to the PCM (i.e. consistently had the lowest values in seven of the eight countries); in Austria the HL-SF12 had the lowest values (see Table A4 in the Appendix). All three 12-item versions had acceptable *WLE* and *EAP* reliability coefficients > 0.77 in all eight countries (see Table 5).

Table 5: WLE and EAP reliability coefficients for the different questionnaire versions

Country	HLS-EU-Q47	HLS-EU-Q16	HLS-EU-Q12	HLS-Q12	HL-SF12
WLE reliability coeff.					
AT	0.948	0.868	0.830	0.828	0.835
BG	0.965	0.912	0.884	0.884	0.882
EL	0.952	0.886	0.846	0.851	0.850
ES	0.951	0.873	0.830	0.837	0.827
IE	0.945	0.874	0.838	0.839	0.839
NL	0.938	0.839	0.781	0.784	0.771
PL	0.962	0.901	0.870	0.881	0.873
DE	0.946	0.878	0.829	0.827	0.820
EAP reliability coeff.					
AT	0.955	0.877	0.840	0.836	0.845
BG	0.972	0.927	0.902	0.897	0.897
EL	0.965	0.906	0.867	0.872	0.871
ES	0.954	0.876	0.835	0.841	0.829
IE	0.960	0.897	0.868	0.864	0.865
NL	0.944	0.861	0.813	0.807	0.800
PL	0.970	0.917	0.893	0.898	0.889
DE	0.958	0.905	0.860	0.850	0.850

The correlation of the HLS-EU-Q12 and the HLS-EU-Q47 indices was high in the total sample of all eight countries (r = 0.957). In the individual countries the correlations varied between 0.938 and 0.967 (see Table 6). The correlations with the HLS-EU-Q16 were comparable. The correlation of the HLS-EU-Q12 index with the NVS was r = 0.26 in the total sample and the correlations in the individual countries varied between r = 0.13 and r = 0.269. These values are comparable to the correlations of the HLS-EU-Q47 index with the NVS (r = 0.25 for the total EU-8, and correlations between r = 0.14 and r = 0.38 in the individual countries; cf. Pelikan & Ganahl, 2017).

Table 6: Correlations of the HLS-EU-Q12 with HLS-EU-Q16 and NVS

	AT	BG	EL	ES	IE	NL	PL	DE	Total (EU-8)
HLS-EU-Q47	0,946	0,967	0,960	0,938	0,963	0,938	0,962	0,961	0,957
HLS-EU-Q16	0,930	0,970	0,952	0,931	0,953	0,929	0,965	0,945	0,951
NVS	0,153	0,399	0,322	0,210	0,269	0,190	0,392	0,130	0,26

Replacing item 33 by item 32 and evaluating model fit

Following a consortium decision, it was examined if item 33 could be replaced by item 32. Therefore, Steps 4 and 5 were applied to a version consisting of items 4, 7, 10, 16, 18, 23, 24, 31, 32, 37, 42, 44 (called HLS-EU-$Q12_{32}$ in the following) in order to evaluate its model fit.

For the HLS-EU-$Q12_{32}$ no significant infit statistics were observed (see Table A5 in the Appendix), as was the case for the version containing item 33 instead of item 32. DIF for age was observed for item 32 in two countries (BG, EL) and DIF for education in two countries (BG, IE). In three countries *SRSMR*-values > 0.08 were observed for the HLS-EU-$Q12_{32}$, and the proportion of significant t-tests was < 5% only in one country (see Table 7). *WLE* und *EAP* reliability coefficients were comparable for both test versions with values > 0.77 for HLS-EU-$Q12_{32}$ and > 0.78 for the HLS-EU-Q12 in all countries. Comparing the two versions by deviance and information statistics, the version containing item 32 shows consistently lower values across all statistics and across all countries.

Table 7: Comparison of HLS-EU-Q12 und HLS-EU-$Q12_{32}$

	Country							
	AT	**BG**	**EL**	**ES**	**IE**	**NL**	**PL**	**DE**
SRMSR								
HLS-EU-Q12	0.0712	0.0696	0.0769	0.0683	0.0789	0.0745	0.0595	0.0798
HLS-EU-$Q12_{32}$	0.0728	0.0699	**0.0806**	0.0693	**0.0852**	0.0739	0.0631	0.0881
PCA/t-test (proportion significant t-tests, CI)								
HLS-EU-Q12	**0.086 (0.065)**	**0.125 (0.099)**	**0.101 (0.077)**	**0.101 (0.007)**	**0.084 (0.062)**	0.069 (0.049)	**0.078 (0.057)**	**0.078 (0.058)**
HLS-EU-$Q12_{32}$	**0.095 (0.073)**	**0.079 (0.058)**	**0.122 (0.096)**	**0.079 (0.058)**	**0.104 (0.08)**	0.039 (0.025)	**0.086 (0.064)**	**0.104 (0.08)**
***WLE* reliability coeff**								
HLS-EU-Q12	0.8304	0.8838	0.8464	0.8298	0.8383	0.7806	0.8703	0.8293
HLS-EU-$Q12_{32}$	0.8271	0.8842	0.8444	0.8315	0.8357	0.7761	0.8710	0.8308
***EAP* reliability coff.**								
HLS-EU-Q12	0.8304	0.8838	0.8464	0.8298	0.8383	0.8137	0.8933	0.8601
HLS-EU-$Q12_{32}$	0.8368	0.9010	0.8654	0.8365	0.8666	0.8112	0.8939	0.8622

Deviance								
HLS-EU-Q12	13130.16	11309.64	11713.98	9370.28	10756.19	10749.74	9835.49	12001.58
HLS-EU-Q12$_{32}$	13074.86	11294.01	11694.90	9269.33	10570.27	10516.50	9792.51	11895.35
AIC								
HLS-EU-Q12	13204.16	11383.28	11787.98	9444.56	10830.19	10823.74	9909.49	12075.58
HLS-EU-Q12$_{32}$	13148.86	11368.01	11768.90	9343.33	10644.27	10590.50	9866.51	11969.35
AICc								
HLS-EU-Q12	13209.79	11389.25	11793.86	9450.79	10836.39	10829.85	9915.67	12081.51
HLS-EU-Q12$_{32}$	13154.49	11373.98	11774.77	9349.55	10650.47	10596.61	9872.69	11975.28
CAIC								
HLS-EU-Q12	13399.74	11576.88	11982.16	9636.75	11022.54	11016.53	10101.91	12269.40
HLS-EU-Q12$_{32}$	13344.44	11561.61	11963.08	9535.52	10836.62	10783.29	10058.93	12163.17
BIC								
HLS-EU-Q12	13362.74	11539.88	11945.16	9599.75	10985.54	10979.53	10064.91	12232.40
HLS-EU-Q12$_{32}$	13307.44	11524.61	11926.08	9498.52	10799.62	10746.29	10021.93	12126.17

The correlation of the HLS-EU-Q12$_{32}$ and the HLS-EU-Q47 indices was r = 0.955 in the total sample of all eight countries and therefore comparable with the correlation of the version containing item 33. In the individual countries the correlations varied between r = 0.935 and r = 0.966. The correlations with the HLS-EU-Q16 were also comparable. The correlation of the HLS-EU-Q12 index with the NVS was r = 0.263 in the total sample and the correlations in the individual countries varied between r = 0.13 and r = 0.385. Therefore, all correlations are comparable with the version containing item 33 instead of item 32.

Table 8: Correlations of the HLS-EU-Q12$_{32}$ with HLS-EU-Q16 and NVS

	AT	BG	EL	ES	IE	NL	PL	DE	Total (EU-8)
HLS-EU-Q47	0.949	0.966	0.959	0.938	0.954	0.935	0.956	0.959	0.955
HLS-EU-Q16	0.929	0.969	0.945	0.925	0.942	0.919	0.959	0.940	0.946
NVS	0.168	0.406	0.314	0.211	0.266	0.195	0.385	0.130	0.263

3. References

Adams, R.J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*, 162-172.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267-281). Budapest, Hungary: Akademiai Kiado.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123- 140.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48(6),* 1-29. doi:10.18637/jss.v048.i06

Duong, T.V., Aringazina, A., Balsunova, G., Nurjanah Pham, T.V., Pham, K.M., Truong, T.Q., … 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.

Finbråten, H.S., Pettersen, K.S., Wilde-Larsson, B., Norström, G., Trollvik, A., Guttersrud, Ø. (2017). Validating the European Health Literacy Survey Questionnaire in people with type 2 diabetes: Latent trait analyses applying multidimensional Rasch modelling and confirmatory factor analyis. *J Adv Nurs, 73,* 2730-2744.

Finbråten, H.S., Wilde-Larsson, B., Nordström, G., Pettersen, K.S., Trollvik, A. & Guttersrud, Ø. (2018). Establishing the HLS-Q12 short version of the European Health Literacy Survey Questionnaire: latent trait analyses applying Rasch modelling and confirmatory factor analysis. *BMC Health Services Research, 18,* 506. https://doi.org/10.1186/s12913-018-3275-7

Fischer, G.H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika, 46,* 59–77.

Fischer, G.H. & Scheiblechner, H. (1970). Algorithmen und Programme für das probabilistische Testmodell von Rasch. [Algorithms and programs for Rasch's probabilistic test model.] *Psychologische Beiträge, 12,* 23-51.

García-Pérez, M.A. (2018). Order-Constrained Estimation of Nominal Response Model Parameters to Assess the Empirical Order of Categories. *Educational and Psychological Measurement, 78(5),* 826-856.

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.

HLS-EU Consortium (2012): *Comparative report of health literacy in eight EU member states. The European Health Literacy Survey HLS-EU.* Online Publication: http:/www.health-literacy.eu

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika, 72,* 297-307.

Linacre, J.M. & Wright, B.D. (1994). *Rasch Measurement Transactions, Part 2, Volume 8(2).* https://www.rasch.org/rmt.

Mair, P., Hatzinger, R., Maier, M.J., Rusch, T., & Debelak, R. (2020). eRm: Extended Rasch Modeling. 1.0-1. http://cran.r-project.org/package=eRm

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

Maydeu-Olivares, A. (2013). Goodness-of-Fit Assessment of Item Response Theory Models. *Measurement, 11,* 71-101.

Pelikan, J.M. & Ganahl, K. (2017). Measuring Health Literacy in General Populations: Primary Findings from the HLS-EU Consortium's Health Literacy Assessment Effort. In R.A. Logan & E.R. Siegel (Eds.), *Health Literacy: New Directions in Research, Theory and Practice* (pp. 34-59). Amsterdam, Berlin, Washington, DC: IOS Press.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research.* (Expanded edition. Chicago: The university of Chicago Press).

Robinson, M., Johnson, A.M., Walton, D.M., McDermid, J.C. (2019). A comparison of the polytomous Rasch analysis output of RUMM2030 and R (ltm/eRM/TAM/lordif). *BMC Medical Research Methodology, 19*:36.

Robitzsch, A. (2019). Package `sirt'. Version 3.3-26. March 18, 2019.

Robitzsch, A., Kiefer, T., & Wu, M. (2020). TAM: Test Analysis Modules. R package version 3.5-19. https://CRAN.R-project.org/package=TAM

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461-464.

Smith Jr., E.V. (2002) Detecting and Evaluating the Impact of Multidimensionality Using Item Fit Statistics and Principal Component Analysis of Residuals. *Journal of Applied Measurement, 3,* 205-231.

Smith, A.B., Rush, R., Fallowfield, L.J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology, 8:*33.

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). Health literacy in populations: Illuminating the design and development process of the European Health Literacy Survey Questionnaire (HLS-EU-Q). *BMC Public Health, 13(948).*

Sørensen, K., Pelikan, J.M., Röthlin, 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. on behalf of the 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.

Thissen, D., Cai, L., & Bock, R.D: (2010). The Nominal Categories Model. In M. L. Nering & R. Ostini (Ed.), *Handbook of Polytomous Item Response Theory Models* (pp. 43-75), Abingdon, Oxon: Routledge.

Weiss, B., Mays, M.Z., Martz, W., Castro, K.M., DeWalt, D.A., Pignone, M.P., Mockbee, J., Hale, F.A. (2005). Quick assessment of literacy in primary care: the newest vital sign. *Ann Fam Med, 3(6),* 514-522.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145. doi: 10.1177/014662168400800201