Evaluating short versions of the European Health Literacy Survey Questionnaire (HLS-EU-Q47) for health checkups

Azumi MAIE¹ Sachiyo KANEKUNI¹ Yuki YONEKURA² Kazuhiro NAKAYAMA² Rieko SAKAI^{1,3}

¹Center for Preventive Medicine, St. Luke's International Hospital Affiliated Clinic

²Graduate School of Nursing Science, St. Luke's International University

³Department of Preventive Health Care for Women, Aikku Hospital, Maternal and Child Health Center

ABSTRACT

Objective: The benefit of health literacy (HL) on adopting a healthy lifestyle is well established and enhancing HL as part of regular health checkups contributes to the prevention of lifestyle-related diseases. Tuned health advices requires preliminary assessment of HL. To balance evaluation precision and cumbersomeness, we aimed to find the most parsimonious questionnaire for our checkups.

Methods: Participants (n = 714 with 55% female, 51.2 ± 10.9 years old) were recruited among the examinees who visited our checkups, and a cross-sectional survey was conducted by using the Japanese version of the European Health Literacy Survey Questionnaire (HLS-EU-Q47). Three short versions of HLS-EU-Q47 (HLS-Q12, HL-SF12, and HLS-EU-Q16), were explored using confirmatory factor analysis (CFA) to assess dimensionality and factorial validity together with Rasch analysis (RA) to provide detailed information at the item level.

Results: Using likelihood-ratio tests, we found that all three short versions were better fit under a three-dimensional Rasch model (health care, disease prevention and health promotion) than one-dimensional one. All of them were above threshold for person separation reliability (PSR > 0.65), with only moderate loss of separation compared to HLS-EU-Q47. CFA supported the factor structure of the three domains, and HLS-Q12 has the best goodness-of-fit indices.

Conclusions: Among the parsimonious questionnaire that we compared, HLS-Q12 is the best candidate, meeting the assumptions and the requirements of objective measurement while offering clinically feasible screening during regular checkups. In the next phase, we will evaluate the benefit of using individual's HL to tune life-style improvement advice.

(HEP. 2021; 48: 351-358.)

Key words health literacy questionnaire, parsimonious survey, checkup, confirmatory factor analysis, Rasch analysis

INTRODUCTION

Life-style related diseases, such as cancer and cardiovascular disease, cause about 60% of deaths and 30% of public medical expenses in Japan¹⁾. As part of a prevention strategy, it is important to provide screenings and health checkups with a focus on life-style change, such as the “specific health checkups” together with “specific health guidance” programs launched in 2008 in Japan²⁾. With the increase in health awareness, people tend to more actively manage their own health and are more involved in various health and medical decision making. This crucially depends on access to high-quality health information.

Until recently, medical and health information were mainly

sourced from specialists such as primary care doctor, however, patient/examinee are increasingly bombarded with health information and, more worryingly, health misinformation. Thus, health literacy (HL) is now recognized as a critical aspect enabling the patient/examinee to identify and utilize reliable health information.

A comprehensive definition of HL has been developed by Sørensen *et al.*³⁾ as follows: “Health literacy is linked to literacy and entails people's knowledge, motivation and competences to access, understand, appraise, and apply health information in order to make judgments and take decisions in everyday life concerning healthcare, disease prevention and health promotion to maintain or improve quality of life during the life course”. It has been reported⁴⁾ that inadequate HL is associated with lower use of health care services and poorer health outcomes. In a survey of individuals with cardiovascular disease, ability to understand health information and ability to actively engage with healthcare providers were associated with improved health behaviors and health status⁵⁾. For maximum efficacy, it thus seems important that health guidance is tuned to the measured level of HL of each patient/examinee. Among the multiple scales available for measuring HL, the European Health Literacy Survey Questionnaire (HLS-EU-Q47) appears to be the most suitable for use in preventive medicine^{6, 7)}. This is a scale consisting of 47 questions, evaluating four abilities (“accessing”, “understand-

Received: August 7, 2020, Accepted: December 15, 2020

J-STAGE Advance published: February 20, 2021

Corresponding author: Rieko Sakai

¹Address: St. Luke's International Hospital Affiliated Clinic, Center for Preventive Medicine

St. Luke Tower, Akashi-cho 8-1, Chuo-ku, Tokyo 104-6591, Japan

TEL: +81-3-5550-2400, FAX: +81-3-5550-2404

E-mail: launey-ty@umin.ac.jp

³Present Address: Department of Preventive Health Care for Women, Aikku Hospital, Maternal and Child Health Center

1-16-10 Shibaura, Minato-ku, Tokyo 105-8321, Japan

TEL: +81-3-6453-7300, FAX: +81-3-6453-7301

E-mail: rsakai@aikku.net

ing”, “appraising”, and “applying”) in three domains identified as “health care (HC)”, “disease prevention (DP)”, and “health promotion (HP)”. Translated versions are used in many European⁶⁾ and Asian^{7,8)} countries, with high level of validity and reliability. However, the large number of items is rather cumbersome for patient/examinee and thus may not be well-suited for systematic inclusion in a health-check course. Several short versions of HLS-EU-Q47 have been suggested, such as the HLS-Q12⁹⁾, HL-SF12¹⁰⁾ and HLS-EU-Q16¹¹⁾, that might be better suited to this context. Additionally, all of these scales have been developed overseas and it thus seems necessary to examine whether they are suitable for Japanese health checkup recipients. The aim of the present study was to identify the most suitable questionnaire for our check-up system, by comparing the reliability and validity of the HLS-Q12, HL-SF12 and HLS-EU-Q16.

SUBJECTS AND METHODS

Participant

We recruited 751 Japanese participants willing to participate in the survey, among the examinees who visited our checkups between July 30th and August 10th, 2018. Participants gave an informed consent and the study was approved by our institutional review board. Participants answered anonymously a self-administered paper-and-pencil questionnaire of HLS-EU-Q47 validated Japanese version⁷⁾ and indicated their gender and age.

Measures

Japanese version of the HLS-EU-Q47

We used the Japanese version of the HLS-EU-Q47 which was translated into Japanese and confirmed its reliability and validity, previously⁷⁾. The answers of questionnaire were ranked on a four-point Likert-type scale (1 = very difficult, 2 = fairly difficult, 3 = fairly easy, 4 = very easy and 5 = don’t know/not applicable). As this survey was a self-administered questionnaire, we included the response “don’t know/not applicable” to avoid blocking questions. These responses were recoded as missing values. We derived a general HL index comprising all items as well as three subsets indices representative of HC, DP and HP. Index scores, as with the original scale, were standardized⁶⁾ on a metric between 0 and 50, using the formula: $(\text{MEAN}-1) \times (50/3)$. MEAN is the mean of all item responses for each participant. All values are reported as mean $\pm$ standard deviations.

Data analysis

Rasch analysis

The Rasch model is known as item response model, in which the relationship between the examinee’s ability and the difficulty of the test item is defined by a logistic model. Likelihood-ratio tests (LRT) were used to compare data-model fit for nested models, such as the one-, three- and twelve-dimensional Rasch models, and obtained statistic values of deviance (for nested models) and Akaike’s information criterion (AIC, for non-nested models) were used to compare the data-model fit, with lower values indicating better data-model fit. Using a Multi-dimensional Partial Credit model, the difficulty parameters (item-location estimate, Loc) were evaluated; negative values indicate easier questions. Items were also interpreted as under-/over-discriminating by infit

statistics, infit Mean Square (infit MSQ) and infit t value. Infit MSQ value between 0.70 and 1.30 indicates a good fit, in accordance with Linacre and Wright¹²⁾, where a value of 1 is a perfect fit. Values of Infit MSQ less than 0.70 and greater than 1.30 are termed mis-fitting and over-fitting, respectively. When infit t value is non-significant (between -1.96 and 1.96), it also indicates good model fit. Differential Item Functioning (DIF) was explored using the Many Facet Rasch model for person factors of gender and age: age was dichotomized as above or below the sample mean (51 years old). Items were considered as DIF when the effect size was > 0.5 logits¹³⁾ and absolute z value was > 1.96 . Negative z indicates that older examinees find item more difficult; similarly for gender, negative z indicates that female examinees find the item more difficult. The ordering of the response categories was also examined¹⁴⁾. Significantly different thresholds in correct order were used as evidence for ordered response categories. The analysis was performed on the data without any missing values using R 3.5.2 and package TAM 3.0-21.

Confirmatory factor analysis

For assessing construct validity, a confirmatory factor analysis (CFA) was conducted consisting of the three dimensions of health (HC, PD and HP). Selected items which are from twelve hypothetical components³⁾ were loaded onto the three factors. The subsets of items are different in the three questionnaires.

The adequacy of the data to the model was examined by goodness-of-fit (GOF) indices¹⁵⁾; comparative fit index (CFI), Tucker-Lewis index (TLI), root mean square error of approximation (RMSEA) and chi-square/degrees of freedom ratio (χ^2/df ratio). A CFI and TLI value of .90 or larger is generally considered to indicate acceptable model fit. RMSEA value of less than 0.05 represents good fit, and a value < 0.08 is acceptable. More satisfied indices would indicate better construct validity of the questionnaire. The AIC was also taken into account to compare the various short versions, which lower values indicate better data-model fit. The missing value were replaced by the mean of each item, and those analyses were performed using SPSS Statistics 24 and Amos 24.0.0 (IBM Corp; Armonk, NY, USA).

Reliability

Cronbach’s alphas were calculated to examine internal consistency, a reliability coefficient of 0.70 or higher is considered “acceptable”. Person separation reliability (PSR) were also used as measures of reliability, values exceeding 0.85 (0.65) are sufficient if conclusions are to be drawn at the individual (group) level, respectively.

Ethical considerations

The study received prior approval from the Research Ethics Committee of St. Luke’s International University, Japan (18-R048, 19-R049).

Results

Out of 751 participants, 743 returned the questionnaire. 714 respondents validly answered at least 80% of the items (response rate of 95%). Among the 655 respondents who answered all questions, 497 answered without any “don’t know/not applicable”.

Participant characteristics

The sample ($n = 714$) distributions for gender and age were in

accordance with the distribution of these metrics in the population of examinees in our facility. The mean age of participants was 51.2 ± 10.9 years old, with 54.8% women (**Table 1**). About 90% of them have had two or more previous annual check-up in our facility. More than 65% were referred to our check-up as part of their employer welfare program and are mostly white-collar, with higher than average social status and education level. HL score of our examinees using HLS-EU-Q47 was 27.3 ± 7.3 .

Table 1 Characteristics of study participants (N = 714)

Variables	n	%	General Health Literacy score
age	51.2 ± 10.9		27.3 ± 7.3
Age group			
20–39 years	113	16.7	26.0 ± 7.4
40–49 years	183	27.1	26.9 ± 7.6
50–59 years	232	34.4	27.5 ± 6.8
60+ years	147	21.8	28.5 ± 7.4
missing	39		
Gender			
male	305	45.2	26.6 ± 7.1
female	370	54.8	27.9 ± 7.4
missing	39		

Table 1: Variables are given as means $\pm$ standard deviations. General health literacy scores are obtained using the European Health Literacy Survey Questionnaire (HLS-EU-Q47).

Table 2 Rasch analysis of HLS-EU-Q47, fit incidence and PSR

			One-dimensional	Three-dimensional	Three-dimensional (consecutive)	Twelve-dimensional
Likelihood Ratio Test						
Deviance			45,896	44,784	45,771	43,436
AIC			46,180	45,078	46,059	43,874
ep			142	147	144	219
Reliability	Health Domains	Cognitive Domains				
		access				0.775/0.77
		understand				0.825/0.823
	Health Care	appraise		0.906/0.908	0.9/0.902	0.743/0.73
		apply				0.628/0.607
		access				0.757/0.755
	Disease Prevention	understand	0.954/0.956	0.895/0.897	0.892/0.894	0.515/0.529
		appraise				0.798/0.792
		apply				0.637/0.609
	Health Promotion	access				0.763/0.758
		understand		0.904/0.907	0.901/0.904	0.783/0.78
		appraise				0.798/0.793
		apply				0.764/0.758

Table 2: Overall fit statistics for the one-, three- and 12-dimensional approaches (treating the theoretical subscales as correlated) of Rasch model fit to the HLS-EU-Q47.

For the three-dimensional approach, the consecutive approach (treating the three health domains as uncorrelated) was also examined. The deviance (for nested models) and Akaike's information criterion (AIC, for non-nested models) were used to compare the data-model fit, in which lower values indicate better data-model fit. PSR (person separation reliability): Maximum Likelihood Estimate (MLE)/Warm's mean weighted Likelihood Estimate (WLE). Values exceeding 0.85 (0.65) are sufficient at the individual (group) level, based on a marginal maximum likelihood estimate. ep: number of estimated parameters.

Analysis of the HLS-EU-Q47

First, we examined the factor structure and psychometric properties of HLS-EU-Q47 to identify the sources of the discrepancies between previous studies and ours ($n = 497$ with 53% of female, 51.0 ± 10.7 years old). **Table 2** shows fitting parameters using LRT. The observed data were more likely under a twelve-dimensional Rasch model than a one- or three-dimensional model. The twelve-dimensional approach refers to the four cognitive domains and the three health domains defining twelve correlated subscales (see Methods). The data-model fit improved with the addition of dimensions from the one-dimensional through the twelve-dimensional model. The three-dimensional model showed better fit than when the three health domains, HC, DP and HP were applied consecutively, with a lower AIC ($\Delta AIC = 987$). Regarding reliability however, we observed low PSR when applying the twelve-dimensional approach, most likely owing to the low number of items per scale. Consequently, we considered that the three-dimensional approach offered an acceptable trade-off between model fit and reliability indices.

Table 3 shows item level analysis of HLS-EU-Q47, including item fit (Infit) and DIF. As misfit, one item (item 45: "Join a sports club or exercise class if you want to?") was over-discriminating by both criteria of Infit MSQ > 1.3 and Infit $t > 1.96$. DIF, which indicates a potential source of bias, were not found in gender and/or age categories by our criterion. Unordered responses, which can cause item misfit as a result of inconsistent responses, were not found either.

Infit				DIF				Infit				DIF				Infit				DIF			
MSQ (t value)				effect size (z value)				MSQ (t value)				effect size (z value)				MSQ (t value)				effect size (z value)			
				gender age								gender age								gender age			
1	1.20 (3.08)	0.16 (2.46)	0.11 (-1.71)	13	0.96 (-0.62)	0.17 (-2.71)	0.01 (0.13)	25	0.92 (-1.21)	0.07 (-1.04)	0.28 (4.39)	37	0.96 (-0.53)	0.06 (0.91)	0.16 (-2.55)								
2	1.17 (2.66)	0.06 (0.88)	0.07 (-1.10)	14	1.03 (0.42)	0.05 (-0.73)	0.17 (-2.58)	26	0.93 (-1.07)	0.06 (-0.88)	0.08 (1.23)	38	1.06 (0.96)	0.14 (2.13)	0.07 (1.10)								
3	1.08 (1.27)	0.17 (2.75)	0.04 (-0.59)	15	1.28 (4.33)	0.28 (-4.39)	0.07 (1.00)	27	0.90 (-1.70)	0.03 (-0.42)	0.21 (3.30)	39	0.91 (-1.41)	0.15 (2.39)	0.08 (-1.23)								
4	1.13 (2.07)	0.18 (-2.84)	0.01 (0.04)	16	1.03 (0.38)	0.09 (-1.34)	0.16 (-2.34)	28	0.94 (-0.83)	0.01 (-0.04)	0.09 (1.39)	40	0.83 (-2.89)	0.11 (1.68)	0.17 (-2.75)								
5	0.88 (-1.82)	0.18 (-2.88)	0.05 (0.68)	17	0.94 (-0.81)	0.06 (0.91)	0.06 (-0.82)	29	1.26 (3.98)	0.19 (-2.91)	0.02 (0.19)	41	0.97 (-0.44)	0.03 (-0.39)	0.02 (-0.33)								
6	0.90 (-1.51)	0.03 (-0.42)	0.03 (-0.42)	18	0.91 (-1.51)	0.09 (1.39)	0.04 (-0.52)	30	0.89 (-1.73)	0.04 (0.65)	0.07 (1.13)	42	0.93 (-1.10)	0.06 (0.94)	0.03 (0.39)								
7	0.90 (-1.60)	0.02 (0.30)	0.06 (0.91)	19	0.85 (-2.49)	0.15 (2.36)	0.08 (1.30)	31	0.87 (-2.10)	0.03 (0.42)	0.02 (0.33)	43	0.93 (-1.14)	0.07 (1.13)	0.14 (2.13)								
8	0.91 (-1.20)	0.02 (-0.29)	0.08 (-1.19)	20	0.86 (-2.19)	0.06 (0.91)	0.02 (-0.23)	32	0.87 (-2.03)	0.15 (2.30)	0.07 (-1.10)	44	0.94 (-0.89)	0.04 (-0.55)	0.06 (0.88)								
9	0.88 (-1.93)	0.05 (-0.75)	0.03 (0.42)	21	1.14 (1.96)	0.06 (0.78)	0.18 (-2.49)	33	0.92 (-1.13)	0.24 (3.72)	0.13 (-1.94)	45	1.40 (6.09)	0.04 (-0.65)	0.07 (1.13)								
10	0.92 (-1.35)	0.02 (0.26)	0.01 (0.13)	22	1.09 (1.41)	0.14 (-2.10)	0.09 (-1.37)	34	0.98 (-0.32)	0.08 (-1.20)	0.08 (-1.17)	46	1.22 (3.33)	0.06 (0.84)	0.06 (0.94)								
11	1.10 (1.57)	0.04 (-0.54)	0.18 (2.72)	23	0.94 (-0.85)	0.08 (1.18)	0.05 (-0.71)	35	1.01 (0.11)	0.05 (-0.71)	0.13 (2.04)	47	1.26 (4.10)	0.10 (-0.24)	0.12 (-0.28)								
12	1.02 (0.33)	0.02 (-0.29)	0.14 (2.16)	24	1.10 (1.47)	0.02 (0.25)	0.03 (-0.40)	36	1.09 (1.31)	0.11 (-1.78)	0.01 (0.10)												

Table 4 Data-model fit and reliability applying Rasch modelling to the various short versions

		HLS-Q12		HLS-SF12		HLS-EU-Q16	
		One-dimensional	Three-dimensional	One-dimensional	Three-dimensional	One-dimensional	Three-dimensional
Log-likelihoods							
	Deviance	12,192.44	12,124.29	12,775.30	12,696.50	15,307.35	15,127.93
	AIC	12,266	12,208	12,849	12,780	15,405	15,236
Reliability							
	PSR	0.862	0.709/0.691/0.706	0.844	0.68/0.698/0.655	0.883	0.806/0.721/0.942

Since we are interested in implementing shorter versions of the HLS-EU-Q47 questionnaire, we evaluated HLS-Q12, HL-SF12 and HLS-EU-Q16, with detailed analysis of dimensionality, data-model fit, response dependence and reliability. As with HLS-EU-Q47, we observed slightly better fit for the subscales of the three-dimensional approaches in HLS-Q12 and HL-SF12 when comparing deviances and AIC (**Table 4**).

models are more fine-grained in identifying HL problems and therefore better suited to give advises targeted to particular domains. We thus evaluated the construct validity of those short three-dimensional versions using CFA (714 respondents, **Table 5**). The RMSEA and other GOF indices demonstrated good model data fit of HLS-Q12. HL-SF12 also had best GOF indices, while HLS-EU-Q16 displayed the weakest overall fit indices but had higher Cronbach's alpha for internal consistency reliability.

Table 5 Overall fit and reliability using confirmatory analyses of the short various versions, three-dimensional

	GOF index					Reliability
	CFI	TLI	RAMSEA (90%CI)	chi-square/df	AIC	Cronbach's α
HLS-Q12	0.903	0.875	0.079 (0.070–0.089)	5.48	240.49	0.85
HLS-SF12	0.932	0.912	0.061 (0.052–0.071)	3.66	333.61	0.83
HLS-EU-Q16	0.793	0.754	0.105 (0.099–0.112)	8.85	963.35	0.88

Goodness of fit (GOF) indices, log-likelihood and reliability using confirmatory factor analysis (CFA) when treating the various short versions as three-dimensional.

CFA was performed using SPSS Amos software. CFI (comparative fit index) and TLI (Tucker-Lewis index): value of .90 or larger is generally considered to indicate acceptable model fit. RMSEA (root-mean-squared error of approximation): value of less than .05 represents good fit, and a value < .08 is acceptable. AIC (Akaike's information criterion): lower values indicate a better overall fit. Cronbach's alpha is a measure of internal consistency, a reliability coefficient of .70 or higher is considered "acceptable".

Table 6 Item-fit indices applying the three-dimensional approach to the various short versions

Item No.	HLS-Q12				HLS-SF12				HLS-Q16			
	Infit		DIF		Infit		DIF		Infit		DIF	
	Loc	MSQ (t)	effect size (z) gender	age	Loc	MSQ (t)	effect size (z) gender	age	Loc	MSQ (t)	effect size (z) gender	age
2	0.40	1.17 (2.70)	0.01 (–0.14)	0.08 (1.20)	0.37	1.11 (1.83)	0.03 (–0.40)	0.07 (1.10)	0.40	1.14 (2.18)	0.05 (–0.74)	0.04 (0.64)
4									0.70	1.10 (1.65)	0.18 (2.94)	0.03 (–0.50)
5									–0.67	0.88 (–1.98)	0.18 (3.00)	0.07 (–1.14)
6					–0.46	0.94 (–0.81)	0.06 (0.87)	0.03 (0.50)				
7	0.70	0.93 (–1.16)	0.03 (0.47)	0.05 (–0.80)								
8									–1.49	0.89 (–1.54)	0.03 (0.46)	0.05 (0.78)
10	0.77	0.96 (–0.54)	0.03 (0.49)	0.01 (–0.04)	0.71	0.92 (–1.25)	0.01 (0.14)	0.01 (–0.10)				
11									1.19	1.08 (1.19)	0.04 (0.65)	0.20 (–3.13)
13									–0.17	0.96 (–0.65)	0.17 (2.84)	0.04 (–0.60)
14	–2.39	1.09 (1.35)	0.10 (1.47)	0.18 (2.72)								
15					–0.43	1.18 (2.91)	0.29 (4.80)	0.06 (–0.94)				
16									–2.07	1.06 (0.84)	0.10 (1.49)	0.13 (1.85)
18	–0.19	0.98 (–0.29)	0.04 (–0.67)	0.04 (0.64)	–0.17	0.92 (–1.27)	0.06 (–0.94)	0.04 (0.57)	–0.19	0.99 (–0.10)	0.08 (–1.24)	0.01 (0.04)
21									–2.36	1.15 (2.07)	0.04 (–0.55)	0.15 (2.03)
23	–2.36	1.04 (0.61)	0.03 (–0.46)	0.06 (0.77)	–2.24	0.97 (–0.44)	0.05 (–0.64)	0.05 (0.76)	–2.38	0.98 (–0.27)	0.07 (–0.98)	0.02 (0.24)
26					0.01	0.94 (–0.85)	0.08 (1.27)	0.08 (–1.20)				
28	0.89	0.97 (–0.38)	0.05 (0.78)	0.08 (–1.23)					0.88	0.99 (–0.15)	0.01 (0.17)	0.11 (–1.78)
30	0.06	0.90 (–1.55)	0.01 (0.10)	0.07 (–1.04)	0.05	0.92 (–1.29)	0.02 (–0.20)	0.07 (–1.10)				
31									0.45	0.90 (–1.66)	0.02 (–0.30)	0.05 (–0.77)
32	–0.89	0.93 (–1.04)	0.10 (–1.55)	0.08 (1.20)								
33					–1.25	0.91 (–1.29)	0.20 (–3.23)	0.13 (2.04)	–1.33	1.01 (0.08)	0.23 (–3.59)	0.10 (1.52)
37									–1.19	0.99 (–0.06)	0.05 (–0.75)	0.13 (2.07)
38	0.04	1.05 (0.83)	0.09 (–1.40)	0.06 (–0.97)								
39					–0.63	0.92 (–1.24)	0.12 (–1.94)	0.08 (1.30)	–0.67	0.92 (–1.19)	0.14 (–2.27)	0.05 (0.80)
43	–0.54	0.97 (–0.47)	0.03 (–0.40)	0.13 (–2.04)	–0.50	0.97 (–0.50)	0.04 (–0.67)	0.13 (–2.07)	–0.53	1.06 (0.87)	0.06 (–0.24)	0.16 (–0.64)
44	–0.40	1.04 (0.60)	0.08 (0.39)	0.05 (–0.22)								
45					–0.15	1.31 (4.91)	0.07 (0.33)	0.07 (–0.32)				

Table 6: Item-fit indices when applying three-dimensional approach to the various short versions. Figures in bold indicate values outside criterion. LOC (item-location estimate): negative values indicate easier questions. Infit MSQ (mean square): values between 0.70 and 1.30 are within acceptable limits, Infit t: t-test values between 1.96 and –1.96 are within acceptable limits.

DIF (differential item functioning): value of logit > 0.5, and t: z-test values between 1.96 and –1.96 are used as a criterion for a significant DIF.

Item level analyses of the short versions, HLS-Q12, HL-SF12 and HLS-EU-Q16

For analysis at the item level, Loc, Infit and DIF (Table 6)

were evaluated.

High Loc, which suggest the hard to endorse items, was detected for item 10 ("Judge the advantages and disadvantages of

different treatment options?") in HLS-Q12 and HL-SF12, and item 28 ("Judge if the information on health risks in the media is reliable?") in HLS-Q12 and HLS-EU-Q16. For all versions, item 23 ("Understand why you need health screenings?") showed the lowest Loc suggesting it was the easiest to endorse. Regarding detection of under-/over-discrimination by Infit statistics, all the items are good fits, except one item in HL-SF12 (item 45). We also tested for DIF and unordered responses and none of them exceeded threshold in any items of the three short versions. We noted however that effect size for DIF has a narrower spread for HLS-Q12 than for the other two versions, with less than 0.1 logits for gender and less than 0.2 logits for age. Both HL-SF12 and HLS-EU-Q16 showed a few items with gender and/or age DIF that, albeit being less than the 0.5 logit threshold, showed higher spread.

DISCUSSION

Importance of measuring health literacy at health checkups

According to a national survey in 2016¹⁶⁾, among adult Japanese in the age of 50 to 59 years, 79.9% of the male population and 71.0% of the female population received some kind of checkups during the past one year, with 87.0% of the male population and 85.5% of the female population assessing themselves as "in good health". While an increased awareness of health related life-style was observed over this decade, this self-assessment and evaluation by medical practitioner may differ. For instance, 60% of people listed "nutritionally balanced diet" as the main target of their diet¹⁾, while the proportions of obesity (BMI: ≥ 25) has not significantly decreased over this period¹⁷⁾, suggesting that information and/or knowledge is insufficient. It is also reported that health-related behaviors such as diet, exercise and drinking, were significantly associated with HL¹⁸⁾ measured by HLS-EU-Q47 and that people with high HL were more willing to receive health guidance¹⁹⁾.

During health checkups, it is customary to include interviews and patient-filled forms to document medical history and life-style. It is arguable that systematically adding a health literacy questionnaire would allow to evaluate the HL of examinee and make it possible for medical practitioners to provide better targeted advices and information on life-style related diseases. Thus, our aim here is to identify suitable tools to evaluate health literacy in the context of a health checkup setting.

Validation of HLS-EU-Q47

It has been pointed out that compared to similar countries, Japanese population has a low health literacy score (24.4 for male and 26.2 for female) when it is measured by HLS-EU-Q47⁷⁾. Here, we also observed a low general health literacy score of 27.3, lower than reported in Europe (score of 33.8)⁶⁾ and in some Asian countries⁸⁾. This is likely due to differences in the medical system and cultural background. Note first that although we make reference to the original English language questions in this report, the test given was the previously validated Japanese translation of HLS-EU-Q47⁷⁾, making it more likely that ambiguities arise from cultural rather than translation differences. For instance, item 11 ("Judge when you may need to get a second opinion from

another doctor?") was most often answered as "very difficult" (31.9%: 209/655) and even reached 75% together with "fairly difficult". This may correspond to low utilization of second medical opinion in Japan²⁰⁾, creating a cultural bias in the answers. Similarly, some question formulation may be culturally ambiguous, as in item 15 "Calling an ambulance in an emergency?" which may be interpreted as "whether to call" or "how to call". For all our recipients, the "how" is probably trivial. In contrast, if the question is understood as asking to judge the necessity of an ambulance, it might be perceived as difficult to answer.

On the other hand, in our health checkup context, some questions appear non-discriminating. For example, item 21 ("Understanding health warnings about behavior such as smoking, low physical activity and excessive drinking?") gets almost 60% (373/655) of "very easy" responses (93% "very" + fairly easy) and item 23 had the highest endorsement. This is expected since enrolment is mandatory for only a small fraction of our participants; those for whom it is part of their company's labor health regulations. Examinees who voluntarily enroll in medical checkups are generally more aware of preventive behavior. Similarly, it has been reported²¹⁾ that HL from patients with chronic disease tend to show better health awareness for Health Care (HC) domain. Finally, our Rasch analysis confirmed previous reports⁹⁾ that HLS-EU-Q47 fits better to a multi-dimensional model (Table 2) with subclass for different HL domains. Thus, the overall HL score provided by HLS-EU-Q47 may not be the most informative metric for our Japanese checkup examinees. One additional aspect to consider in the present context is the cost/benefit ratio for the examinee, striking a balance between filling cumbersome questionnaire and receiving improved health advices. This discussion is thus focused on comparing three shorter versions of the HLS-EU-Q47.

Validation of three short versions of HLS-EU-Q47

We explored whether one of the short versions, HLS-Q12, HL-SF12 and HLS-EU-Q16, might be more suitable for our checkups. For this analysis, we adopted Infit MSQ statistics rather than t-statistics since it is reported that the latter are highly sensitive to sample size for polytomous data in contrast with mean square statistics that remained relatively stable²²⁾ over wide range of sample size. While our sample size ($n = 497$) is adequate for both approaches, with at least 10 endorsements for each potential response category for each item, we nonetheless chose Infit MSQ for its robustness. Note also that fitting may depend on the approaches taken by the statistical programs²³⁾. Here, all data were analyzed with a R package utilizing likelihood ratio chi-square test to compare the models (see methods).

Rasch model analysis of the short versions showed that HLS-Q12 with three-dimensional approach has the smallest deviance and AIC, suggesting best model fit overall while the three health domain subscales each returned sufficiently large reliability indices of PSR (Table 4). By CFA, we confirmed that the HLS-Q12 and HL-SF12 have higher validity than HLS-EU-Q16, which has the smallest CFI and TLI together with a larger RMSEA (> 0.08) (Table 5). Thus, we confirm with the present study that HLS-Q12 is the best-targeted scale⁹⁾. In our particular setting, however, it is

not free from under-discriminating items and DIF. At the item level analysis (Table 6), HLS-Q12 showed one item (item 2) as under-discriminating with $\text{Infitt} > 1.96$, however Infitt MSQ was within acceptable range. Regarding DIF, we observed that two items had slightly bigger effect size of DIF for age (item 14: “Following the instructions on medication?” and 43: “Judge which everyday behaviour is related to your health?”) in HLS-Q12, which were not entirely unexpected. For instance, it seems that answer to item 14 could be biased for older examinees who were more likely already used to take medication regularly.

Finally, item 23, may need collapsing of the categories to improve the overall fit to the model together with item 14 ($\text{Loc} = -2.39$) since 92.8% of the health checkup examinees predictably answered “easy” to this question ($\text{Loc} = -2.36$).

Limitations

The original English version HLS-EU-Q47⁶⁾ was conducted face-to-face, without a formal “don’t know /not applicable” answer. Here, since the test was self-administered and anonymous, we used this previously validated modification introduced by Nakayama⁷⁾. In face-to-face settings, the presence of an examiner ensures that the questionnaire is both completed and returned, which increases response rate. On the downside, it also biases the responses since it is not anonymous and since the presence of an interviewer may influence the respondent’s answers²⁴⁾. Thus ideally, the questionnaire should be filled alone but validated immediately when returned and be completed if necessary. Among the 655 participants who returned a fully answered questionnaire, 158 (24.1%) answered “don’t know/not applicable” at least once. Note however that on a per item basis, the occurrence of “don’t know/not applicable” was only 1.7% on average, lower than in previous studies^{6,7)}. Note also that one specific item (Q36, “Find out about efforts to promote your health at work?”) had disproportionally high percentage of “don’t know/not applicable” answer (7.3% overall, 10.8% among women) owing to the demographic of our respondents (house-wives) for whom the item is clearly “not applicable”. Overall, with less than 2% of “don’t know/not applicable” answer per item, this had low influence on the study.

Conclusion

Our results indicated that among the three scales, HLS-Q12 is the best candidate for our checkup and would be a valid tool to comprehensively assess health literacy in our clinical settings. While scaling properties would probably be improved by rephrasing/replacing the problematic items, it seems equally important to preserve the common international framework offered by the original HLS-Q12, to benefit from the insights of future work done by others. Our future contribution will be to evaluate the impact of health literacy assessment on successful disease prevention and improved health service.

Acknowledgments

The authors would like to thank the staff at the Center for Preventive Medicine, St. Luke International Hospital, especially Mrs. Aya KAMEDA, Asako SHIMOZAWA, Mizuki OKAZAKI

for their help in collecting the subject data and planning-stage discussion, Dr. Katsunori MASUDA for his support and Dr. Thomas LAUNEY, RIKEN Institute of Physical and Chemical Research, for critical discussion and English editing.

The authors state that they have no Conflict of Interest (COI).

REFERENCES

- 1) Ministry of Health, Labor and Welfare (Japan). Annual Health, Labour and Welfare Report for the Realization of a Society of Health and Longevity -First Year of Health and Prevention- 2014. (Accessed May 11, 2020, at <https://www.mhlw.go.jp/english/wp/wp-hw8/dl/summary.pdf>)
- 2) Ministry of Health, Labor and Welfare (Japan). Annual Health, Labour and Welfare Report 2008-2009: Specific Health Checkups and Specific Health Guidance. (Accessed May 11, 2020, at <https://www.mhlw.go.jp/english/wp/wp-hw3/dl/2-007.pdf>)
- 3) Sørensen K, Van den Broucke S, Brand H, Fullam J, Doyle G, Pelikan J, et al. Health literacy and public health: a systematic review and integration of definitions and models. *BMC Public Health*. 2012; 12: 80-92.
- 4) Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Crotty K. Low health literacy and health outcomes: an updated systematic review. *Ann Intern Med* 2011; 155(2):97-107.
- 5) Aaby A, Friis K, Christensen B, Rowlands G, Maindal HT. Health literacy is associated with health behaviour and self-reported health: A large population-based study in individuals with cardiovascular disease. *Eur J Prev Cardiol* 2017; 24: 1880-8.
- 6) HLS-EU CONSORTIUM (2012): Comparative Report of Health Literacy in Eight EU Member States. The European Health Literacy Survey HLS-EU (second revised and extended version, date July 22th, 2014), Online Publication. (Accessed October 12, 2020, at <http://www.health-literacy.eu>)
- 7) Nakayama K, Osaka W, Togari T, Ishikawa H, Yonekura Y, Sekido A, et. al. Comprehensive health literacy in Japan is lower than in Europe: a validated Japanese-language assessment of health literacy. *BMC Public Health* 2015; 15: 505-16.
- 8) Duong TV, Aringazina A, Baisunova G, Nurjanah, Pham TV, Pham KM, et. al. Measuring health literacy in Asia: Validation of the HLS-EU-Q47survey tool in six Asian countries. *Journal of Epidemiology* 2017; 27: 80-86.
- 9) Finbråten HS, Wilde-Larsson B, Nordström G, Pettersen KS, Trollvik A, Guttersrud Ø. 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* 2018; 18: 506-22.
- 10) Duong TV, Chang PWS, Yang SH, Chen MC, Chao WT, Chen T, et. al. A new comprehensive short-form health literacy survey tool for patients in general. *Asian Nursing Research* 2017; 11: 30-5.
- 11) Tiller D, Herzog B, Kluttig A, Haerting J. Health literacy in an urban elderly East-German population-result from the population-based CARLA study. *BMC Public Health* 2015; 15: 883-91.
- 12) Wright BD, Linacre JM. Reasonable mean-square fit values. *Rasch Measurement Transactions*. 1994; 8: 370. (Accessed June 22, 2020, at <https://www.rasch.org/rmt/rmt83b.htm>)
- 13) Myers ND, Wolfe EW, Feltz DL, Penfield RD. Identifying Differential item functioning of rating scale items with the Rasch model: An introduction and an application. *Measurement in Physical Education and Exercise Science* 2006; 10: 215-40.
- 14) Linacre JM. Demarcating Category Intervals. *Rasch Measurement Transactions* 2006; 19: 1041-3. (Accessed June 26, 2020, at <https://www.rasch.org/rmt/rmt194f.htm>)

- 15) Schumacker RE, Lomax RG. A beginner's Guide to Structural Equation Modeling. 4th ed. New York: Routledge 2015.
- 16) Ministry of Health, Labor and Welfare (Japan). Summary Report of Comprehensive Survey of Living Conditions 2016. (Accessed May 11, 2020, at https://www.mhlw.go.jp/english/database/db-hss/dl/report_gaikyo_2016.pdf)
- 17) Ministry of Health, Labor and Welfare (Japan). National Health and Nutrition Survey 2018 (in Japanese). (Accessed May 20, 2020, at <https://www.mhlw.go.jp/content/10900000/000615383.pdf>)
- 18) Goto E, Ishikawa H, Nakayama K, Kiuchi T. Comprehensive health literacy and health-related behaviors within a general Japanese population: Differences by health domains. *Asia Pacific Journal of Public Health* 2018; 30: 717-26.
- 19) Kimura N, Obara K, Akibayashi N, Miyamoto T. Association between health literacy and behavior regarding health checkups and health counseling in Japanese employees: A comprehensive health literacy survey of a Japanese railway company. *Sangyo Eiseigaku Zasshi* 2019; 61: 123-32.
- 20) Ministry of Health, Labor and Welfare (Japan). Patient's Behavior Survey 2011 (in Japanese). (Accessed May 11, 2020, at <https://www.mhlw.go.jp/toukei/saikin/hw/jyuryo/11/dl/kakutei-kekka-gaiyo.pdf>)
- 21) Finbråten HS, Pettersen KS, Wilde-Larsson B, Nordström G, Trollvik A, Guttersrud Ø. Validating the European health literacy survey questionnaire in people with type 2 diabetes. Latent trait analyses applying multidimensional Rasch modelling and confirmatory factor analysis. *J Adv Nurs*. 2017; 73: 2730-44.
- 22) Smith AB, Rush R, Fallowfield LJ, Velikova G, Sharpe M. Rasch fit statistics and sample size considerations for polytomous data. *BMC Med Res Methodol*. 2008; 8: 33-43.
- 23) Robinson M, Johnson AM, Walton DM, MacDermid JC. A comparison of the polytomous Rasch analysis output of RUMM2030 and R (ltm/eRm/TAM/lordif). *BMC Med Res Methodol*. 2019; 19: 36-47.
- 24) de Leeuw ED, Hox J, Dillman D. *International Handbook of Survey Methodology*. New York: Routledge, 2008.