Naigaga Desire Alice is a PhD candidate studying Health Sciences, with a background in Public Health Nutrition (MSc) and Human Nutrition and Dietetics (BSc). She has a keen interest in health literacy and nutrition literacy of adolescents in both developed and developing economies.

Kjell Sverre Pettersen is a professor in health communication, with experience in health science teaching, supervising students, and researching on topics related to health literacy and nutrition literacy.

Sigrun Henjum is an associate professor with extensive research experience within the field of nutritional epidemiology, particularly dietary assessment, maternal and infant nutrition and research on environmental determinants of dietary behaviour among adolescents.

Øystein Guttersrud is an associate professor in science education. He has experience in physics/physics education, quantitative measurement, and supervision of students. He is currently working on different evaluation projects in education.

DESIRE ALICE NAIGAGA (corresponding author)

PhD student, Oslo Metropolitan University, Norway
dnaiga@oslomet.no

KJELL SVERRE PETTERSEN

Professor, Oslo Metropolitan University, Norway
spetters@oslomet.no

SIGRUN HENJUM

Associate professor, Oslo Metropolitan University, Norway
shenjum@oslomet.no

ØYSTEIN GUTTERSRUD

Associate professor, University of Oslo, Norway
oystein.guttersrud@naturfagsenteret.no

Assessing adolescent ‘self-efficacy in body and health’- Exploring the psychometric properties of the SEBH scale

Abstract

Self-efficacy beliefs are significant predictors of achievement in education. However, majority of existing self-efficacy measures are rather ‘general’ and assess aggregated perceptions of students’ proficiencies within broad academic disciplines. Applying Rasch analysis, the present study explored the psychometric properties of the five-item ‘self-efficacy in body and health’ (SEBH) scale as administered to more than 1600 tenth-graders aged 15-16years in Norway. Based on our sample, the SEBH

scale stood out as well targeted and reliable with acceptable overall fit to the partial credit parameterization of the polytomous unidimensional Rasch model. Except for a slightly reversed threshold in item 1, which could be explained by few persons located at low levels of self-efficacy, the locally independent items had ordered response categories and functioned in the same way for the different levels of relevant person factors. Adapting this scale to different fields of education would contribute to development of more specific measures of perceived capability.

INTRODUCTION

Does it matter whether you have the will and belief that you can? Of course it does! Self-efficacy signifies a person's belief that he or she is able to execute successfully the behaviours required to produce a specific outcome. Thus, self-efficacy is the person's belief in his or her capability to control and execute actions in spite of potential obstacles. A person's perceived self-efficacy has a direct influence on the choice of activities and settings, and the stronger the perceived self-efficacy, the more active the efforts to cope with the task at hand (Bandura, 1977). Therefore, self-efficacy affects individuals' decisions concerning the effort and endurance they will put into a task. In general, higher self-efficacy is linked with greater effort, perseverance and resilience (Van Dinther, Dochy, & Segers, 2011; Zeegers, 2004).

In school achievement, self-efficacy refers to an individual's belief in his or her ability to successfully accomplish academic tasks or to achieve academic goals (Schunk, 1991). Scales measuring academic self-efficacy evaluate the extent to which students perceive they can accomplish established academic goals (Marsh, Hau, Artelt, Baumert & Peschar, 2006; Pastorelli et al, 2001). However, according to Bong and Skaalvik (2003), majority of the existing academic self-efficacy measures are 'wide-ranging', aiming at school proficiency in general, thus making them more reflective of 'academic self-concept'. Self-efficacy is a specific view of one's capacities in a given domain and it follows that efficient self-efficacy measures be tailored to the particular domain of interest (Bandura, 2006).

Studies have found that adolescents do display autonomy associated with health and self-efficacy (Schunk & Meece, 2006; Taylor, Adelman & Kaser-Boyd (1984). Today, adolescents are exposed to diverse sources of health and nutrition information that shape their lifestyle attitudes (Gray, Klein, Noyce, Sesselberg & Cantrill, 2005; Neuhauser, Rothschild & Rodríguez, 2007). Notable among these sources are schools, through the health topics offered (Brown, Tufel & Birch, 2007).

In health literacy research, studies have shown that self-efficacy is a significant predictor of health behaviour and achievement (Gutiérrez-Doña, Lippke, Renner, Kwon, & Schwarzer, 2009; Rayane & Achterberg, 1997; Schwarzer, 2008). As a result, various health-specific self-efficacy assessment tools have been developed (see, for example, Schwarzer & Renner, 2009; Young, Mills, Woolmore, Hawkins, & Tennant, 2012).

In spite of the advantages that item response theory (IRT) models, and Rasch models in particular, have over classical test theory (CTT), few health-related and health literacy scales have been evaluated using IRT and Rasch models (see, for example, Davidson, Keating & Eyres, 2004; Escobar et al., 2015; Huang et al. 2018; Nguyen, Paasche-Orlow, Kim, Han and Chan, 2015). One such advantage is concerned with the assumption of item-sample independence, which is strongly emphasized in IRT and Rasch models. While violations of local independence in IRT and Rasch models, and 'error correlations' in confirmatory factor models (CFM), might refer to similar 'problems' in the data, there is no direct link between the probabilistic IRT and Rasch models and the correlation-based CFM. Unlike descriptive IRT-models, the family of prescriptive Rasch models satisfy the requirements of fundamental measurement (Andrich, 1988).

To fill in the gaps identified, there is a need for a measurement scale for the evaluation of how adolescents perceive their proficiency in accomplishing specific academic tasks within health, which meets the assumptions and satisfies the requirements of fundamental measurement. To exemplify this in the field of science education, the current study focuses on the subject area of 'body and health' in the Norwegian compulsory school science curriculum. This subject area; which focuses on the structure of our bodies, how the body is affected by nutrition and lifestyle and how the body changes over time; will play a vital part in the new and forthcoming interdisciplinary school topic 'public health and wellbeing' (KD, 2016). A self-efficacy in 'body and health' scale might be efficient for evaluating the proficiency with which adolescents perceive they can apply that knowledge to solve complex problems in new and unfamiliar contexts and adopting critical thinking skills associated with 'deeper learning' (Paakkari, L. & Paakkari, O., 2012; Pellegrino & Hilton 2012; KD, 2016).

The main objective of the current study is therefore to, applying Rasch-analysis, validate a five-item measurement scale tailored towards assessing adolescent self-efficacy in 'body and health' at the end of compulsory school (tenth grade). We will test the following hypotheses:

H1) The 'self-efficacy in body and health' (SEBH) scale has acceptable overall fit to the rating scale parameterization of the polytomous unidimensional Rasch model, consists of locally independent items, and represents a well-targeted and reliable measurement scale.

H2) Each SEBH-scale item has ordered response categories, is functioning in the same way for the different levels of relevant person factors, and shows acceptable fit to the rating scale parameterization of the polytomous unidimensional Rasch model.

Our first hypothesis is concerned with the overall SEBH-scale psychometric properties, while our second hypothesis refers to the psychometric properties at the individual item level. With the goal of estimating as few parameters as possible (parsimony rule), we hypothesized a unidimensional scale with items sharing the same set of thresholds.

METHOD - SAMPLE

A sample of 200 Norwegian lower secondary schools was randomly selected, and the school principals were contacted by email and telephone seeking consent to volunteer. Fifty-eight schools (30%) accepted the invitation. From April to May 2015, 1622 students in the tenth grade (47% girls) responded by using an electronic assessment tool.

The substantive theory of the SEBH latent variable

The SEBH-scale is a revised and further developed version of a self-efficacy scale reported by Guttersrud & Pettersen (2015), which was based on self-efficacy measures in science and the control expectation scale applied in PISA (Organization for Economic Co-operation and Development [OECD], 2001). The items were reworded to reflect competencies within 'body and health', with one additional item (Table 1): *'I am confident that I can apply the knowledge that I have in Body and Health in new and unfamiliar situations'*. This item reflects aspects of adaptability—the transferability of self-efficacy beliefs to novel and changing situations (Martin, Nejad, Colmar, & Liem, 2013; Pellegrino & Hilton, 2012) and deeper learning – the mastering of core academic content at high levels (Pellegrino & Hilton, 2012).

Table 1. The wording of the items in the self-efficacy in body and health (SEBH) scale (originally stated in Norwegian). A six-point rating scale with the extreme response categories anchored with a phrase 1 = 'strongly disagree' and 6 = 'strongly agree' was used.

Item	Item wording
1	I am confident that if I have to learn something very thoroughly in Body and Health, I will be able to manage it.
2	I am confident that I can do an excellent job with difficult tasks in Body and Health
3	I am confident that I can do very well in tests in Body and Health.
4	I am confident that I can understand difficult learning material in Body and Health
5	I am confident that I can apply the knowledge that I have in Body and Health in new and unfamiliar situations.

Person factor levels and data processing

Students reported the following five person factors (with levels indicated in parentheses); gender (male/female); age at the time of the survey (15 or 16 years old); language predominately spoken at home (Norwegian, Danish/ Swedish (i.e., Scandinavian languages) or 'other'); student's, mother's and father's place of birth (Norway, Denmark/ Sweden or 'other'); and the number of books at home (five categories). A picture showing how different numbers of books might appear like on shelves was included to improve validity or 'response accuracy'.

The variables for birthplace were re-coded into a new variable named 'cultural background' with the levels 'majority' (if at least the student or one of the parents were born in any of the Scandinavian countries i.e., Norway, Denmark or Sweden) and 'minority'. This classification is valid as countries within Scandinavia share strong cultural and linguistic similarities. The five levels of 'number of books at home' were merged into the categories 'less than 100 books' and '100 or more books'. These two levels reflected the largest difference in SEBH-scale score (cf. DIF analysis). The number of books was used as an indicator of socioeconomic status (SES), as research on SES and family resources shows that children's initial reading competency is correlated with the home literacy environment and number of books owned; with children from poor households often having less access to learning materials, including books, computers and skill-building lessons to create a positive literacy environment (Aikens & Barbarin, 2008; Bergen, Zuijen, Bishop, & Jong, 2016; Bradley, Corwyn, McAdoo, & García Coll, 2001; Orr, 2003). As a consequence, research indicates that children from low-SES households develop academic skills slower than children from higher SES groups (Morgan, Farkas, Hillemeier, & Maczuga, 2009).

SEBH-scale response characteristics

A six-point rating scale with the extreme response categories anchored with a phrase 1 = 'strongly disagree' and 6 = 'strongly agree' was applied for all SEBH-scale items. Out of the 1622 student responses there were 1568 valid responses: There were 166 extreme scorers of which 12 students attained the lowest possible raw score on the SEBH-items responded to and 154 students attained the highest possible score (ceiling effect) on the items responded to. There were a total of 36 missing responses to the five items, with item 5 having the highest number of these (15) and item 1 having the least (2). We have no evidence weakening the hypothesis stating that 'data are missing completely at random' (MCAR; Allison, 2001).

The unidimensional Rasch model—a rationale for the methodological decisions

The prescriptive Rasch models estimate the probability of endorsing an item based on the difference between the person location (proficiency or attitude) and item location (difficulty or affective level)

(Rasch, 1960; Shaw, 1991). Person and item location estimates refer to the point estimate of a person's or an item's location on the latent trait scale, respectively (Harris, 1989). In the current study, person location refers to an individual's self-reported perceived proficiency in body and health. The different threshold locations reflect the locations at which the probability of a response in two adjacent categories is equal. For example, a dichotomously scored item has one threshold, and the threshold location refers to the location at which the probability of a response in the two adjacent categories is 0.5. In this paper, we applied RUMM 2030 for all analyses (Andrich, Lyne, Sheridan, & Lou, 2010). RUMM uses pairwise maximum likelihood estimation (PLME) and Warm's mean weighted likelihood estimation (WLE) for estimating item location estimates and person locations respectively (Katsikatsou, Moustaki, Yang-Wallentin, & Joreskog, 2012; Warm, 1989).

The concept 'item discrimination' refers to the degree with which an item separates individuals with higher person location estimates from those with lower location estimates. An under-discriminating item differentiates weaker between such respondent groups than the RM expects, given the item location.

Using the Rasch Model (RM), raw scores at the ordinal level (presumes 'ordered response categories' otherwise nominal) are transformed into interval implying additivity (Andrich, 1989; Perline, Wright & Wainer, 1979; Salzberger, 2010). Fit to Rasch models implies that the property of *invariance* holds meaning that the item-trait relationships are stable for the different person locations along the latent trait scale (Andrich, 1988). Rasch models satisfy *specific objectivity* which refers to the requirement of item-person independence; any person location estimate must be independent of the specific measurement device or items applied (Stenner, 1994). As the raw scores contain all the information needed to estimate Rasch models parameters i.e., item and person locations, the raw score is a *sufficient statistic* for Rasch models (Andersen, 1977).

While both Rasch models and other IRT models assume locally independent data—unidimensional and statistically independent data, only the family of Rasch models ensure additivity, invariance, specific objectivity and sufficiency as described above. Therefore, we applied prescriptive Rasch models and not descriptive IRT models in this study.

Overall model fit

The parameters of the rating scale parameterization (RSM; Andrich, 1978) of the RM are a subset of the parameters of the partial credit parameterization (PCM; Masters, 1982) of the RM, so the RSM is nested in the PCM. We compare data-model fit for nested models using likelihood ratio test (LRT). The LRT test statistic – the change in deviance (D) – is asymptotically χ^2 distributed (i.e., for large samples) with degrees of freedom (df) equal to the difference in model estimated parameters (Wilks, 1938, p. 62). A 'significant' χ^2 value implies rejecting the 'null hypothesis' stating that the less complex and nested model, describing the data using fewer threshold estimates, is preferred (cf. hypothesis 1). Compared to RSM, the df of PCM is larger and the PCM therefore usually accounts better for the observed data.

Individual item and person fit

To account for our somewhat large sample size ($N = 1622$), we drew five random samples of 250, 500 and 750 persons from the SPSS file storing the data – a total of fifteen samples. These sample sizes correspond to 10, 20 and 30 persons per thresholds (Andrich, 2010). We estimated individual item χ^2 and overall χ^2 for each sample, and we reported the mean values. To account for the significance testing of k individual items, we Bonferroni-adjusted the individual item χ^2 p -values by the number of χ^2 tests performed: $0.05/k = 0.01$ (see Bland & Altman, 1995).

Person z-fit shows how well a person's response pattern conforms to the 'Guttman structure' (Andrich, 1978). The difference in difficulty of the items caused by dependence is reported as a z-fit residual statistic at a conservative 1% level of significance ($z = 2.56$), a positive z-fit > 2.56 indicates an unexpected response pattern (Andrich & Kreiner, 2010).

Local independence–response independency and unidimensionality

Once we have extracted the Rasch factor—the unidimensional underlying latent trait “self-efficacy”, we assume there are no further patterns in the residuals (Wright, 1996). This assumption is tested by checking for response dependency and multidimensionality. *Response dependency* implies that items are linked in such a way that the responses to one item influence the responses to other items, and we identify this phenomenon by inspecting the item residual correlation matrix. The commonly used conservative item residual correlation of < 0.30 , has recently come under criticism for being too conservative. Therefore, Yen (1984) proposed exploring local dependence based on comparing the item residual correlation values up against the average item residual correlation with values 0.2 above the average item residual as displaying dependency.

Unidimensionality means that only one latent trait - self-efficacy - explains all the covariances between the items (cf. partial correlations). A combined principal component analysis (PCA) of residuals and paired *t*-tests procedure is applied to check for unidimensionality (Hagell, 2014). If approximately 5% or less of the dependent *t*-tests comparing respondents' location estimates on two distinct subscales are significant, then unidimensionality is assumed (Smith Jr, 2002; Tennant & Pallant, 2006).

Furthermore, by creating a 'subtest structure' for a pair of item subsets identified, we can estimate fractal indices (r , c and A) specific to the 'subtest structure'. The index A describes the amount of common variance among the two subsets or subscales identified, c identifies the magnitude of unique subscale variance, and r is the correlation between the two subsets (RUMM, 2009). High values for both A and r , and a low value for c , might therefore indicate an approximately unidimensional scale (Andrich, 2016; Andrich, 2015).

Targeting, reliability, ordering of response categories and differential item functioning

In a well-targeted scale, the distribution of the person estimates matches the distribution of the item threshold estimates centred at 0.0 logits. Poor targeting might increase the risk of unordered response categories and disordered thresholds, large standard errors, extreme person scores, and therefore deflated reliability indices and poor information at certain locations along the latent trait scale.

The internal consistency reliability of the latent trait measurement scale is reported as Person Separation Index (PSI), which is analogous to Cronbach's alpha, and indicates the capacity to separate persons with higher location estimates from those with lower location estimates on the latent trait (Andrich, 1982). Different criteria are suggested for PSI, with values > 0.70 , > 0.80 and > 0.90 indicating 'acceptable', 'good' and 'excellent' reliability respectively (Duncan, Bode, Lai & Perera, 2003). Often 0.7 is used as the minimum value for group and 0.85 as the minimum value for assessments at the individual item level (Cronbach, 1951).

Differential item functioning (DIF) or 'within-item bias' might occur when different 'levels' or 'groups' of a person factor, such as males and females, at equivalent levels of the underlying construct have different probabilities of endorsing an item (Holland & Wainer, 1993; Walker, Beretvas, & Ackerman, 2001). When persons belonging to a particular 'level' show a consistent systematic difference in their responses to an item, uniform DIF is implied. In cases where the differences vary across levels of the attribute between the person factor groups, non-uniform DIF is indicated. Items that display non-uniform DIF are discarded from the instrument.

A procedure in RUMM2030, allows for the resolution of uniform DIF by resolving the item into multiple items, one for each group levels and comparing the estimates of the item parameters from the different 'levels'.

RESULTS

We found that the SEBH items did not share the same set of threshold difficulties. A significant likelihood ratio test statistic LRT χ^2 ($p = 0.000019$; $df = 11$) indicated that the PCM (partial credit parameterization) of the polytomous unidimensional Rasch model described the data 'significantly' better than the RSM (rating scale parameterisation).

In Table 2, we report the overall adjusted mean χ^2 value for each of the amended sample sizes estimated from five random samples reflecting 10, 20 and 30 individuals per scale threshold, as χ^2 is a sample size dependent fit statistic. The PCM of the polytomous unidimensional Rasch model was applied.

Table 2. Overall mean χ^2 fit statistics for the SEBH scale using amended sample sizes.

df (estimated parameters)	Amend sample size (N)	χ^2	$p(\chi^2)$	Scale thresholds
45	750	71.09	0.015	25
35	500	53.45	0.159	
15	250	20.81	0.197	

To sum up, Table 2 indicates that hypothesis 1 (the SEBH data is sufficiently described by RSM) is not fully supported.

Individual person residuals showed that 20 and 89 students had z-fit above/below the cut-off criterion of ± 2.56 , respectively (Andrich & Kreiner, 2010). Concerns were raised about values above the $+2.56$ threshold, as these indicate response patterns that are unlikely i.e., deviate significantly from the Guttman pattern given the self-efficacy score sum. However, removing these few responses did not significantly change any fit parameter estimates.

The assumption of a locally independent scale holds for the SEBH-scale as no response dependence between any pair of items was observed, and only 6.2% of paired t -tests were significant. The t -test structure was based on two subsets of items empirically indicated by the PCA of residuals procedure (the easily endorsable items 1–3 (subscale 1) versus items 4 and 5 (subscale 2), see Table 3). A subset analysis indicated that these two subscales measured strongly related latent traits (high subscale common variance $A = 0.90$, subscale correlation $r = 0.97$ and low subscale unique variance $c = 0.17$).

Table 3. Individual mean item χ^2 fit statistics for the SEBH scale using the amended sample sizes.

Item	Loc	SE	χ^2 (N=750)	$p(\chi^2)$	χ^2 (N=500)	$p(\chi^2)$	χ^2 (N=250)	$p(\chi^2)$
1	-0.61	0.04	15.82	0.13	11.24	0.19	5.11	0.19
2	0.16	0.04	17.09	0.07	9.63	0.27	5.36	0.17
3	-0.03	0.04	10.48	0.33	8.12	0.40	3.04	0.44
4	0.26	0.04	6.27	0.70	8.03	0.44	2.54	0.51
5	0.22	0.04	21.43	0.02	16.43	0.08	4.76	0.26

Note. The location estimates with the standard errors are based on the full sample. Each χ^2 value is the mean value estimated from five random samples of sample sizes corresponding to 10, 20 and 30 persons per thresholds respectively ($N = 250, 500, 750$).

For all random sample sizes of 250 and 500, all the chi-square values were insignificant ($p(\chi^2)$). For the random samples of 750, the chi-square value for item 5 was significant in two of the five random samples.

When centering the average item location at 0.0 logits, the resulting average person proficiency was at 1.4 logits, pointing to a scale that could have been better targeted. The positively skewed distribution of person self-efficacy estimates deviates somewhat from the locations at which the items measure most efficiently.

The above results suggest that the SEBH-scale is a rather *valid* measure of self-efficacy in tenth graders. Sufficiently high reliability indices indicated a *reliable* measure (PSI = 0.88 for original and complete data sets and Cronbach's alpha = 0.88 (excluding extremes) and 0.92 (including extremes) for the complete data set where the 36 respondents with missing data for one or more SEBH-items were discarded). Hence, the SEBH-scale is an accurate and precise measure of self-efficacy.

Moving from the overall analyses to the single item level, the slightly disordered response categories observed for item 1 (Figure 1) is explained by the somewhat poorly targeted SEBH-scale (Figure 2). The curves in Figure 1 show the probability of endorsing each of the six response categories (1 = 'strongly disagree' and 6 = 'strongly agree') versus person location. The second category does not function as intended. The dotted line is the upper limit asymptote, where probability equals 100%.

Figure 1. Category probability curves for item 1.

Figure 2. Histogram showing the distributions of person and item threshold locations including Fisher's information function (curve) for the SEBH scale

A skewed distribution toward higher locations of self-efficacy indicates that the items could have been better targeted at the sample. This skewed distribution leaves few persons located at the lower end of the continuum – the trait locations at which the easily endorsable item 1 (item location at -0.61 according to Table 3) has its lower thresholds. We therefore interpret the SEBH-scale raw score as a sufficient statistic at the ordinal level.

Finally, we investigated DIF using the amended sample sizes based on the rule of thumb of 10, 20 and 30 persons per threshold (Andrich, 2011), with a total of 25 thresholds (5 items with 5 thresholds). No DIF was observed for any person factor (gender, age, cultural background, language at home and books at home) using the amended sample sizes of $N = 250$, $N = 500$ and $N = 750$ based on a total of 25 thresholds (5 items with 5 thresholds).

DISCUSSION

Empirical data partially support our two composite hypotheses. The first hypothesis was strengthened except for a deviation from our ideal of parsimony: The partial credit parameterization (PCM), estimating one set of threshold parameters for each item, described the data better than the less complex rating scale parameterization (RSM) estimating one set of threshold difficulties common for all items.

Furthermore, the targeting of the SEBH-scale was not optimal with few items at higher locations. The lack of items providing information at higher levels of the latent trait is a well-known problem in health-literacy measurement (Nguyen et al., 2015). One of few exceptions is the 'Claim Evaluation Tools' developed by the Informed Health Choices group. The second hypothesis was strengthened except for slightly disordered thresholds observed for item 1. The disordering of response categories for item 1 has a simple explanation: The distribution of person estimates is skewed toward higher locations thereby locating few persons at the lower end of the continuum – the locations at which we find the lower threshold parameters for item 1.

Since the SEBH-scale built on a self-efficacy scale published by Guttersrud & Pettersen (2015), the scale seems to easily translate to different fields of education improving the generalizability and ex-

ternal validity of our findings. We interpret this as a serious strength of our study. A limitation to our study is the low school participating rate (58 out of 200 or 30%). This might result in responses from students enrolled in classes taught by above average motivated and enthusiastic teachers—teachers more likely to see the benefits of external assessment resources like the one we developed. This possible difference between the target sample and the accessed sample might explain the high mean self-efficacy estimate in our sample, which again could cause the skewed distribution of self-efficacy person location estimates and the disordering observed for item 1.

CONCLUSIONS

The present paper provides insights into an issue that seems to have passed health literacy research by: the application of Rasch analysis to evaluate the psychometric properties of measurement scales. By fitting the Rasch model, our findings indicate that the SEBH-scale meets the assumptions and satisfies the requirements for fundamental measurement.

The SEBH-scale presented in the study exemplifies that Rasch analysis is a powerful tool for evaluating construct validity of measurement instruments. This is indicated by the absence of construct-irrelevant variance, as all five items fit the Rasch model, implying that the items don't capture unrelated constructs that affect responses in a manner irrelevant to the construct. On the other hand, by meeting the assumption of unidimensionality albeit with the presence of strongly correlated sub-dimensions, the SEBH-scale points to the absence of construct underrepresentation—another threat to construct validity, in which the assessment is too narrow and fails to capture different facets and sub dimensions of the construct.

Furthermore, the total score on the SEBH-scale can be viewed as one of several possible sets of indicators of the construct—perceived self-efficacy in a science subject. An important recommendation is to include more items in the SEBH-scale in order to improve the preciseness with which the abilities of persons that fall between successive items along the hypothesized unidimensional continuum are measured.

The positive effect of perceived self-efficacy on management of diseases is well documented, developing and validating equivalent measures for 'non-sick' individuals particularly adolescents in different domains, as exemplified in the present study, will go a long way in providing measurement tools to inform, design successful health literacy policies and interventions within public health and education.

Acknowledgements

The authors would like to thank the students who responded and the extraordinary support from school heads and teachers who selflessly facilitated the data collection process and the reviewers who provided constructive comments and suggestions.

Disclosure statements

Conflict of interests: None.

Funding: This research received no specific grant for the work done.

Ethics of human subject participation: Ethics approval was not required.

REFERENCES

- Aikens, N. L. & Barbarin, O. (2008). Socioeconomic differences in reading trajectories: the contribution of family, neighborhood and school contexts. *Journal of Educational Psychology*, 100(2), 235-251. doi:10.1037/0022-0663.100.2.235
- Allison, P. D. (2001). Missing Data: Sage University Papers Series on Quantitative Applications in the Social Sciences (7-136). Thousand Oaks, CA.
- Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, 42(1), 69-81. doi:http://dx.doi.org/10.1007/BF02293746
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561-573. doi:http://dx.doi.org/10.1007/BF02293814
- Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR. 20 index, and the Guttman scale response pattern. *Education Research and Perspectives*, 9(1), 95-104. Retrieved from http://www.rasch.org/erp7.htm
- Andrich, D. (1988). Rasch Models for Measurement. Beverly Hills: Sage Publications, Inc.
- Andrich, D. (1989). Distinctions between assumptions and requirements in measurement in the social sciences. In J.A. Keats, R. Taft, R.A. Heath, & S.A. Lovibond (Eds.) *Mathematical and Theoretical Systems* (Vol.4, pp. 7-16). North Holland: Elsevier Science Publishers BV.
- Andrich, D. & Kreiner, S. (2010). Quantifying response dependence between two dichotomous items using the Rasch model. *Applied Psychological Measurement*, 34(3), 181 - 192. doi: https://doi.org/10.1177/0146621609360202
- Andrich, D., Lyne, A., Sheridan, B., & Luo, G. (2010). RUMM2030, Computer software and manual. Perth, Australia: RUMM Laboratory Pty Ltd.
- Andrich, D. (2011). *EDUC5606: Advanced course in Rasch Measurement Theory [lecture handout]*. Faculty of Education. The University of Western Australia, Australia.
- Andrich, D. (2015). Components of Variance of Scales with a Subscale Structure Using Two Calculations of Coefficient alpha. (*Pensamiento Educativo*). *Journal of Latin American Educational Research*, 52(2), 6-33.
- Andrich, D. (2016). Components of variance of scales with a Bifactor subscale structure from two calculations of α . *Educational Measurement: Issues and Practice*, 35(4), 25-30. doi:10.1111/emip.12132
- Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 191-215. doi: http://dx.doi.org/10.1037/0033-295X.84.2.191
- Bandura, A. (1997). Self-efficacy: The exercise of self-control. New York: W.H. Freeman and Company.
- Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.) *Self-efficacy beliefs of adolescents* (Vol 5, pp. 307-337). Greenwich, CT: Information Age Publishing.
- Bergen, E., Zuijlen, T., Bishop, D., & Jong, P. F. (2017). Why are home literacy environment and children's reading skills associated? What parental skills reveal. *Reading Research Quarterly*, 52(2), 147-160. doi:10.1002/rrq.160
- Bland J.M., & Altman, D.G. (1995). Multiple significance tests: the Bonferroni method. *British Medical Journal*, 310, 170. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2548561/pdf/bmj00576-0038.pdf
- Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational psychology review*, 15(1), 1- 40. doi:10.1023/A:1021302408382
- Bradley, R. H., Corwyn, R. F., McAdoo, H. P., & García Coll, C. (2001). The home environments of children in the United States part I: Variations by age, ethnicity, and poverty status. *Child Development*, 72(6), 1844-1867. doi:10.1111/1467-8624.t01-1-00382
- Brown, S. L., Teufel, J. A., & Birch, D. A. (2007). Early adolescents' perceptions of health and health literacy. *Journal of School Health*, 77(1), 7-15. doi:10.1111/j.1746-1561.2007.00156.x
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334. doi:10.1007/BF02310555

- Davidson, M., Keating, J. L., & Eyres, S. (2004). A low back-specific version of the SF-36 Physical Functioning scale. *Spine*, 29(5), 586-594. doi:10.1097/01.BRS.0000103346.38557.73
- Duncan, P. W., Bode, R. K., Lai, S. M., Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: the stroke impact scale. *Archives of Physical Medicine and Rehabilitation*, 84(7), 950-963. doi:https://doi.org/10.1016/S0003-9993(03)00035-2
- Dunn, O. J. (1959). Estimation of the medians for dependent variables. *Annals of Mathematical Statistics*, 30(1), 192-197. doi:10.1214/aoms/1177706374
- Escobar, A., del Mar Trujillo-Martín, M., Rueda, A., Pérez-Ruiz, E., Avis, N. E., & Bilbao, A. (2015). Cross-cultural adaptation, reliability and validity of the Spanish version of the Quality of Life in Adult Cancer Survivors (QLACS) questionnaire: application in a sample of short-term survivors. *Health and Quality of Life Outcomes*, 13(1), 182-195. doi: http://doi.org/10.1186/s12955-015-0378-2
- Gray, N. J., Klein, J. D., Noyce, P. R., Sesselberg, T. S., & Cantrill, J. A. (2005). The Internet: a window on adolescent health literacy. *Journal of adolescent health*, 37(3), 243.e1-243.e7. doi:10.1016/j.jadohealth.2004.08.023
- Gutiérrez-Doña, B., Lippke, S., Renner, B., Kwon, S., & Schwarzer, R. (2009). Self-Efficacy and Planning Predict Dietary Behaviors in Costa Rican and South Korean Women: Two Moderated Mediation Analyses. *Applied Psychology: Health and Well-Being*, 1(1), 91-104. doi:10.1111/j.1758-0854.2009.01004.x
- Guttersrud, Ø., & Pettersen, K. S. (2015). Young adolescents' engagement in dietary behaviour – the impact of gender, socio-economic status, self-efficacy and scientific literacy. Methodological aspects of constructing measures in nutrition literacy research using the Rasch model. *Public Health Nutrition*, 18(14), 2565-2574. doi: https://doi.org/10.1017/S1368980014003152
- Harris, D. (1989). Comparison of 1-, 2-, and 3-Parameter IRT Models. *Educational Measurement: Issues and Practice*, 8(1), 35-41. doi: 10.1111/j.1745-3992.1989.tb00313.x
- Hofer, B. K., & Pintrich, P. R. (1997). The Development of Epistemological Theories: Beliefs about knowledge and knowing and their relation to learning. *Review of Educational Research*, 67(1), 88-140. doi:10.3102/00346543067001088
- Holland, P., & Wainer, H. (1993). *Differential Item Functioning*. Hillsdale (NJ): Lawrence Erlbaum Associates, Publishers.
- Huang, Y.-J., Chen, C.-T., Gong-Hong, L., Wu, T.-Y., Chen, S.-S., Lin, L.-F., ... Hsieh, C.-L. (2018). Evaluating the European Health Literacy Survey Questionnaire in patients with stroke: A latent trait analysis using Rasch modeling. *The Patient - Patient-Centered Outcomes Research*. 11(1), 83-96. doi:10.1007/s40271-017-0267-3
- 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. Retrieved from https://www.diva-portal.org/smash/get/diva2:458843/FULLTEXT01.pdf
- KD (2016). *Fag – Fordypning – Forståelse. En fornyelse av Kunnskapsløftet* [No English title available]. The Office of the Prime Minister. Retrieved from https://www.regjeringen.no/contentassets/e8e1f41732ca4a64b003fca213ae663b/no/pdfs/stm20152016002800odddpdfs.pdf
- Linacre, J. M. (1998). Detecting multidimensionality: which residual data-type works best? *Journal of outcome measurement*, 2(3), 266-283. Retrieved from https://pdfs.semanticscholar.org/d970/94395b593b123607ba2fc40376edd5130f51.pdf#page=99
- Marsh, H. W., Hau, K. T., Artelt, C., Baumert, J., & Peschar, J. L. (2006). OECD's Brief self-report measure of educational psychology's most useful affective constructs: cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing*, 6(4), 311-360. Retrieved from http://dx.doi.org/10.1207/s15327574ijt0604_1
- Martin, A. J., Nejad, H. G., Colmar, S., & Liem, G. A. D. (2013). Adaptability: How students' responses to uncertainty and novelty predict their academic and non-academic outcomes. *Journal of Educational Psychology*, 105(3), 728. doi:10.1037/a0032794

- Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174. doi: 10.1007/BF02296272
- Morgan, P. L., Farkas, G., Hillemeier, M. M., & Maczuga, S. (2009). Risk factors for learning-related behavior problems at 24 months of age: Population-based estimates. *Journal of Abnormal Child Psychology*, 37(3), 401-413. doi:10.1007/s10802-008-9279-8
- Neuhauser, L., Rothschild, R., & Rodríguez, F. M. (2007). MyPyramid. gov: assessment of literacy, cultural and linguistic factors in the USDA food pyramid web site. *Journal of nutrition education and behavior*, 39(4), 219-225. doi:10.1016/j.jneb.2007.03.005
- 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(sup2), 112-115. doi: 10.1080/10810730.2015.1073408
- Organization for Economic Co-operation and Development, Programme for International Student Assessment (2001). Knowledge and skills for life. Paris: OECD.
- Paakkari, L., & Paakkari, O. (2012). Health literacy as a learning outcome in schools. *Health Education*, 112(2), 133-152. doi: <https://doi.org/10.1108/09654281211203411>
- Pastorelli, C., Caprara, G. V., Barbaranelli, C., Rola, J., Rozsa, S., & Bandura, A. (2001). The structure of children's perceived self-efficacy: A cross-national study. *European Journal of Psychological Assessment*. 17(2). doi: 10.1027//1015-5759.17.2.87
- Pellegrino, J. W., & Hilton, M. (2012). Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century. Washington, DC: The National Academies Press.
- Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, 3(2), 237-255. doi:<http://dx.doi.org/10.1177/014662167900300213>
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (Expanded edition 1983 ed.). Copenhagen: Danish Institute for Educational Research.
- Rayane, A., & Achterberg, C. (1997). Review of self-efficacy and locus of control for nutrition and health-related behavior. *Journal of the academy of nutrition and dietetics*. 97(10), 1122-1132. doi: 10.1016/S0002-8223(97)00273-3
- RUMM (2009). *Interpreting RUMM2030 Part IV Multidimensionality and subtests in RUMM*. 1. ed. Perth, WA: RUMM Laboratory Pty Ltd.
- Salzberger, T. (2010). Does the Rasch model convert an ordinal scale into an interval scale? *Rasch Measurement Transactions* 24(2) 1273-1275. Retrieved from <https://www.rasch.org/rmt/rmt242a.htm>
- Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist*, 26(3-4), 207-231. doi:10.1207/s15326985ep2603&4_2
- Schunk, D. H., & Meece, J. L. (2006). Self-efficacy development in adolescence. In F. Pajares & T. C. Urdan (Eds.) *Self-efficacy beliefs of adolescents* (pp. 72-96). Greenwich, Connecticut: Information Age Publishing.
- Schwarzer, R. (2008). Modeling health behavior change: How to predict and modify the adoption and maintenance of health behaviors. *Applied Psychology*, 57(1), 1-29. doi:10.1111/j.1464-0597.2007.00325.x
- Schwarzer, R., & Renner, B. (2009). Health-specific self-efficacy scales. Retrieved from http://www.academia.edu/2391923/Health-Specific_Self-Efficacy_Scales
- Shaw, F. (1991). Descriptive IRT vs. prescriptive Rasch. *Rasch Measurement*, 5(1), 131. Retrieved from <http://www.rasch.org/rmt/rmt51f.htm>
- 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. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/12011501?dopt=Abstract>
- Stenner, A. J. (1994). Specific objectivity-local and general. *Rasch Measurement Transactions*, 8(3), 374. Retrieved from <http://www.rasch.org/rmt/rmt83e.htm>

- Taylor, L., Adelman, H. S., & Kaser-Boyd, N. (1984). Attitudes toward involving minors in decisions. *Professional Psychology: Research and Practice*, 15(3), 436-49. Retrieved from <http://psycnet.apa.org/fulltext/1984-25481-001.pdf>
- Tennant, A. and Pallant, J. (2006) Unidimensionality Matters! (A Tale of two Smiths?). *Rasch Measurement Transactions*, 20(4), 1048-1051. Retrieved from <https://rasch.org/rmt/rmt201.pdf>
- Tsai, C.-C., Ho, H. N. J., Liang, J.-C., & Lin, H.-M. (2011). Scientific epistemic beliefs, conceptions of learning science and self-efficacy of learning science among high school students. *Learning and Instruction*, 21(6), 757-769. doi:10.1016/j.learninstruc.2011.05.002
- Walker, C.M., Beretvas, S.N. & Ackerman, T. (2001). An Examination of Conditioning Variables Used in Computer Adaptive Testing for DIF Analyses. *Applied Measurement in Education*, 14(1), 3-16. Retrieved from <http://www.learntechlib.org/p/93401>
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(3), 427-450. doi:<https://doi.org/10.1007/BF02294627>
- Wright, B. D. (1996). Local dependency, correlations and principal components. *Rasch Measurement Transactions*, 10(3), 509-511. Retrieved from <https://www.rasch.org/rmt/rmt103b.htm>
- Young, C. A., Mills, R. J., Woolmore, J., Hawkins, C. P., & Tennant, A. (2012). The unidimensional self-efficacy scale for MS (USE-MS): developing a patient based and patient reported outcome. *Multiple Sclerosis Journal*, 18(9), 1326-1333. doi:[10.1177/1352458512436592](https://doi.org/10.1177/1352458512436592)
- Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, 25(1), 82-91. doi:10.1006/ceps.1999.1016