METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

Translation Symmetry and Familiarity of End-Point Anchor Points in Likert Scales

Ajda Šulc [1], Vanja Erčulj [1], Anže Mihelič [1]

[1] *Faculty of Criminal Justice and Security, University of Maribor, Ljubljana, Slovenia.*

Methodology, 2025, Vol. 21(3), 161–179, https://doi.org/10.5964/meth.17093

Received: 2025-02-22 • **Accepted:** 2025-07-17 • **Published (VoR):** 2025-09-30

Handling Editor: Belen Fernández, UNED | Universidad Nacional de Educación a Distancia, Madrid, Spain

Corresponding Author: Ajda Šulc, Faculty of Criminal Justice and Security, University of Maribor, Kotnikova ulica 8, 1000 Ljubljana, Slovenia. E-mail: ajda.sulc@um.si

Abstract

Likert scales are widely used to measure agreement levels, typically on a structurally and linguistically symmetrical scale. In non-English settings, literal translations of scale anchors often produce awkward or asymmetrical phrases that deviate from everyday language, potentially affecting data validity. This study examines the impact of translations through an online experiment with 532 Slovenian smartphone users, randomly assigned to two groups. One group used structurally symmetrical translations (GroupSA), while the other used more natural but asymmetrical translations (GroupCA) to measure constructs related to information security within Protection Motivation Theory. GroupCA showed significant correlations with the dependent variable for all predictors, consistently higher Composite Reliability and Average Variance Extracted, slightly higher means, and lower skewness and kurtosis. Additionally, "Completely agree" was chosen more often than the less familiar "Strongly agree." These results highlight the influence of translation choices on survey outcomes, emphasizing the need for careful linguistic adaptation in cross-cultural research.

Keywords

Likert scale, end-point anchor, familiarity hypothesis, survey translation, Likert extreme anchors

The widespread adoption of the Likert scale in research has led to extensive debates regarding its suitability and the reliability of the so-collected data. The main dilemma of treating an otherwise ordinal scale as an interval scale has been widely discussed (Kampen, 2019; Lalla, 2017). Empirical research on its applicability with parametric and non-parametric methods suggests that, given a sufficiently large sample size, i.e., more

than 15 though Sangthong (2020) recommends a minimum of 100, and comparable distributions across subgroups, treating the Likert scale data with parametric tests yields results similar to those obtained with non-parametric approaches (Mircioiu & Atkinson, 2017). However, despite considerations, some evidence indicates that treating ordinal data as numeric can introduce systematic errors, including spurious effect detection (Type I errors) and failures to identify effects (Type II errors) (Liddell & Kruschke, 2018), since unequal intervals between scale anchors may introduce systematic error into statistical analyses (Spector, 1976). This can be affected by both the number of response categories in the selected scale, as Wakita et al. (2012) found that psychological distance deviates more in the 7-point scale than in the 4- or 5-point scale suggesting that increasing the number of responses options does not necessarily result in evenly perceived intervals, as well as by the naming of the anchor points. Studies show that the manner in which anchor points are labelled can significantly affect how respondents perceive the distances between response options. This has been especially questioned in scales that measure relative frequency, such as those using labels like: *Never*, *Seldom*, *Occasionally*, *Frequently*, and *Always*. Psychologically, respondents might perceive the distance between *Never* and *Seldom* as either shorter or longer than the distance between *Seldom* and *Occasionally* (Casper et al., 2020). Evidence suggests that for the agreement scales, the following set of anchors can be used as a relatively equal-interval scale: *Disagree*, *Somewhat disagree*, *Neither agree nor disagree*, *Moderately agree*, and *Very much agree*. However, from a semantic point of view, this scale does not seem to be symmetric, as opposed to the more intuitive one: *Disagree*, *Moderately disagree*, *Neutral*, *Moderately agree*, and *Agree*. Surprisingly, results show that respondents perceived the latter set as less evenly spaced, despite its apparent structural and linguistic balance (Casper et al., 2020).

Additionally, the assumption of equidistance of Likert-type scales has been further questioned by an important finding that the perceived distances between anchors on a five-point Likert scale vary depending on label placement. Specifically, when labels are applied only to the endpoints, respondents tend to perceive the intervals at the scale's extremes as larger, whereas middle categories appear closer together. Conversely, when all five anchors are labelled, respondents perceive a greater distance between middle categories than between those at the scale's endpoints (Lantz, 2013). Nevertheless, some studies conclude that labelling all the response options enhances test-retest reliability compared to labelling only the endpoints which have been associated with less consistent responses (Menold et al., 2014; Weng, 2004). On the contrary, Weijters et al. (2010) found that the scale with labels only at the endpoints performs better in terms of criterion validity than the scale with all the anchors labelled.

The primary requirement for naming Likert scale anchors is ensuring that all respondents understand the labels and interpret them clearly and consistently (Tourangeau et al., 2000). However, lexical miscomprehension — individuals ascribing different mean-

ings to the same word — can lead to significant inconsistencies in responses (Hardy & Ford, 2014). Such discrepancies may arise due to respondents' individual characteristics and can be encouraged by poorly named or set anchors introducing additional confusion (Casper et al., 2020). Empirical research on the effect of the scale wording confirms the importance of considering the labels' formation. Empirical results, obtained by scales of different formats, were proven to be non-comparable (Weijters et al., 2010), which is particularly a challenge for multilingual surveys conducted in diverse cultural contexts and translated into different languages, where the labels will necessarily be different. This will, therefore, lead to systematic differences in scale use among the respondents from different cultural or language groups, also called *scale usage heterogeneity* (Weijters et al., 2016). One of the possible reasons for non-equal usage or understanding of seemingly the same scale across languages is supported by an important finding by Weijters et al. (2013), confirming a familiarity hypothesis that assumes respondents are more likely to select response options featuring phrases commonly used in their daily language. It builds on the concept that certain word combinations (collocations or formulaic sequences) are more frequently encountered in specific languages, allowing for quicker cognitive processing and greater perceived reliability due to their familiarity (Conklin & Schmitt, 2008). Consequentially, respondents will demonstrate a preference and trust towards collocations or familiar phrases, and when those are presented in the form of possible answers on a scale, they will also choose them more confidently and more often. Results of the study, confirmed for several languages, show that the endpoints of Likert scales will attract more responses if the label used for naming them is more familiar to respondents. For example, in the English language, *Completely agree* has been found to be more familiar than *Strongly agree*, resulting in a higher selection frequency by respondents (Weijters et al., 2013). These findings highlight the influence of linguistic and cultural factors on anchor perception and the challenge of ensuring comparability across different respondent groups.

Another semantics-related issue originates from the intensity hypothesis, which suggests that more intense labels marking extreme positions (such as *Never* and *Always* or *Terrific* or *Superior*) will be less likely selected by the respondents due to their perceived extremity (Weijters et al., 2013). This might reduce the response variance, potentially lower reliability, and weaken correlations between items (Nunnally & Bernstein, 1994). Research indicates that when using less absolute endpoint labels (i.e., *Agree* and *Disagree*), responses tend to be distributed more evenly across the five anchors compared to when more extreme labels are presented (Wyatt & Meyers, 1987). This is directly related to the intensity of amplifiers used, as those multiply the extremity of the endpoints. For example, the adjective *Extremely* has a stronger effect than *Very*, and *Very* is more extreme than *Decidedly* (Cliff, 1959). Similarly, other commonly used adjectives for endpoints also differ in their perceived intensity: *Completely, Definitely, Strongly,* and *Very much* (Smith et al., 2008). For example, *Completely agree* is perceived as a more intense

expression of agreement than *Strongly agree*. What is more, the choice of amplifier not only affects how extreme the endpoint labels appear but also influences the perceived spacing between all response options on the scale (Weijters et al., 2013).

On the other hand, respondents might be inclined towards selecting the extreme endpoints of the scale (*Extreme response style*), or towards consistently agreeing with the positive response regardless of the content of the question (*Acquiescence response style*); both of which are frequently observed in cross-cultural research (Fischer, 2004; Ilgun Dıbek, 2020). To mitigate potential semantic biases in Likert-scale measurements, Funke and Reips (2012) propose integrating traditional ordinal rating scales with visual analogue scales. This hybrid approach allows respondents to adjust their ratings visually, thereby improving response precision and enhancing data quality. However, this is not always applicable since it imposes additional cognitive and time-consuming work for respondents. Therefore, addressing the issue of anchor labelling remains critical, as such concerns are primarily associated with textual scales (as expected from the Likert scale), in contrast to the purely numerical scales, as intervals between the numbers are mathematically equally distanced from each other and do not depend on semantic interpretations. To reduce semantic biases, some researchers advocate for a combination of verbal and numerical anchors (Casper et al., 2020), but even those are not a guarantee for perceiving them as equally distanced. Another suggestion for mitigating bias is the *calibrated sigma method*. This procedure avoids assuming equal interval perception across all research groups. Instead, respondents first complete a set of control items and only after that, the numerical values for anchors in the actual questionnaire are assigned based on the distribution of responses to these preliminary items. This adaptive approach aims to enhance measurement accuracy by calibrating the scale according to respondents' perceived intervals (Weijters et al., 2016).

Motivation and Aim

In addition to the challenges already discussed, an additional difficulty arises when these scales are adapted to languages other than English. Literal translations of extreme Likert scale anchors — such as "Strongly disagree" and "Strongly agree" — often produce awkward or unfamiliar phrases that deviate from natural, everyday language. This incongruence can compromise the clarity, interpretability, and validity of responses, particularly when respondents struggle to relate to or visualize the extremities of these anchors. For instance, in the Slovenian language, the literal translation of "Strongly disagree" and "Strongly agree" as "*Močno se ne strinjam*" and "*Močno se strinjam*, " respectively, sounds somewhat unnatural and is rarely encountered in everyday conversations. However, phrases like "*Sploh se ne strinjam*" ("Do not agree at all") *and* "*Se popolnoma strinjam*" ("I completely agree") are more commonly used and may be perceived as more intuitive and relatable. However, this translation is both linguistically asymmetrical and thus potentially lacks intra-scale symmetry, as the lower extremity ("*Sploh se ne strinjam*")

and the upper extremity ("*Se popolnoma strinjam*") differ in their linguistic structure, as well as stray away from the direct translation from the original (English language). Similar issues have been observed in other languages, such as German, where natural expressions like "*Stimme überhaupt nicht zu*" and "*Stimme ganz und voll zu*" deviate from the structurally symmetrical form of Likert anchor points. Furthermore, a lack of symmetry between Likert scale extremities can introduce bias by affecting how respondents perceive and interpret the scale. If the lower and upper extremities differ in intensity, this can also result in a difference in the perceived distance between anchor points. Unequal intervals between response options may lead to inconsistent responses and challenges in accurately measuring constructs. These issues are especially problematic in cross-linguistic research, where natural but asymmetrical translations might exacerbate these effects and result in greater perceived distances between anchor points.

This paper aims to answer the following research question:

> **RQ:** How do structurally and linguistically symmetrical and more familiar but asymmetrical translations of Likert anchor points affect survey responses?

Addressing this research question is important for ensuring the validity and reliability of cross-linguistic research, since the choice of anchor points may inadvertently introduce systematic biases or influence the interpretation of results. To explore this, we designed an online experiment to explore how varying anchor points influence responses and, consequently, the results of the study.

We addressed this problem in the context of information security, drawing upon Protection Motivation Theory (PMT). PMT was originally introduced by Rogers (1975) and is a widely recognized theoretical framework that explains how individuals are motivated to adopt protective behaviours in response to perceived threats – in relation to the perceived severity of a threat, their vulnerability to it, and the efficacy of certain protective behaviour. Several studies confirmed the correlation of the said constructs with security motivation intention (Li et al., 2022; Mou et al., 2022; Zuwita & Rahmatullah, 2021).

Method

To answer the research question, we designed an online experiment to evaluate how different translations of Likert scale anchor points influence survey responses. First, we have evaluated which phrases used in the Slovenian Likert end-points are more familiar to the Slovenian population. We analysed Google search engine results (using Google.si for the Slovene language, with personalization of the results disabled) for each phrase by calculating the familiarity score as suggested by Weijters et al. (2013), i.e., by dividing the count for each endpoint label (e.g., "strongly agree") by the count of the label name

Table 1

Anchor Levels and Corresponding Descriptions According to the Groups in the Original Form (Slovenian) and English Translations for Reference

Anchor level	GroupCA [Slovenian]	GroupCA [English]	GroupSA [Slovenian]	GroupSA [English]
1	Sploh se ne strinjam	Do not agree at all	Močno se ne strinjam	Strongly disagree
2	Se ne strinjam	Disagree	Se ne strinjam	Disagree
3	Sem nevtralen	Neutral	Sem nevtralen	Neutral
4	Se strinjam	Agree	Se strinjam	Agree
5	Se popolnoma strinjam	Completely agree	Močno se strinjam	Strongly agree

without the modifier (e.g., "agree") and computing the natural logarithm of the ratio: $(\ln \frac{hits(\text{"strongly agree"})}{hits(\text{"agree"})})$. The endpoints used in GroupCA exhibited higher familiarity score (−2.7 for "Do not agree at all" and −5 for "Completely agree") compared to their (positive or negative) counterparts used in GroupSA (− 4.1 for "Strongly disagree" and − 6.6 for "Strongly agree").

The items used in the survey were adapted from Thompson et al. (2024) and adapted for secure smartphone use. The items were translated from English to Slovene language using a standard back-translation process (Brislin, 1980). Perceived vulnerability, perceived severity, and self-efficacy were measured by six items, and response efficacy, and security intentions by four items. Items were measured on a 5-point Likert scale.

To investigate the effects of structurally symmetrical versus linguistically natural but asymmetrical translations of anchor points on participant answers, we conducted an experiment in which a survey randomly assigned participants to one of the two groups: GroupSA for the linguistically less natural but structurally symmetrical translations of anchor points, and GroupCA for the linguistically natural but asymmetrical translations (see Table 1).

Sampling and Sample Description

Non-probability, purposeful sampling was used for data collection. In November 2020 link to the online survey was posted in various Facebook groups to which the request to become a member was approved. The sample included 532 respondents who provided answers to all questions in the questionnaire. They were divided into two groups according to the type of scale used in the questionnaire (as explained in Table 1). Sample characteristics for the two study groups are summarized in Table 2. Random allocation to the two groups was used and no statistically significant differences between the study groups in sample characteristics were found.

Table 2

Sample Description

Variable	GroupSA (n = 256)	GroupCA (n = 276)	p
Sex			.612
Male	116 (45.7)	120 (43.5)	
Female	138 (54.3)	156 (56.5)	
Age			.055
< = 21	48 (19.1)	57 (20.7)	
22–23	42 (16.7)	51 (18.5)	
24–26	40 (15.9)	45 (16.3)	
27–30	34 (13.5)	45 (16.3)	
31–39	56 (22.3)	33 (12)	
40+	31 (12.4)	45 (16.3)	
Education			.224
Elementary or less	5 (2.1)	6 (2.3)	
Vocational or high school	118 (48.8)	105 (39.6)	
University or master's degree	110 (45.5)	141 (53.2)	
More than a master's degree	9 (3.7)	13 (4.9)	
Employment status			.728
Student	93 (38.3)	110 (41.5)	
Employed or self-employed	129 (53.1)	135 (50.9)	
Unemployed	11 (4.5)	13 (4.9)	
Retired	0 (0)	0 (0)	
Other	10 (4.1)	7 (2.6)	

Data Analysis

Categorical variables were described with frequencies and percentages, numerical with arithmetical means, standard deviations, coefficients of skewness, and kurtosis. The value of the coefficient of kurtosis and skewness between -2 and 2 indicated approximately normally distributed variables. Comparison in sample characteristics between the study groups was performed by using the Chi-Square Test. The familiarity hypothesis as proposed by Weijters et al. (2013) was tested using Independent Sample *T*-test. The total number of high-anchor responses ("Strongly agree" and "Completely agree") chosen across all 26 items was computed and compared between groups. Confirmatory Factor Analysis with a maximum likelihood estimation method was used to test the questionnaire validity (separately performed for each study group).

Convergent validity was assessed by factor loadings higher than 0.50 and statistically significant loading on the factor they are supposed to be measuring, as proposed by Anderson and Gerbing (1988). It was further supported by the value of the Average Variance Extracted (AVE) above 0.50 (Fornell & Larcker, 1981) and the good overall fit of the model.

The examined goodness-of-fit measures were the statistical significance of the Chi-Square value and its ratio to degrees of freedom. Acceptable values for the latter were between 1 and 3 (Vieira, 2011). In addition, the Comparative Fit Index (CFI), Incremental Fit Index (IFI), Non-Normed Fit Index (NNFI), the Root Mean Square Error of Approximation (RMSEA), and the Standardised Root Mean Square Residual (SRMR) were examined to evaluate the model fit. Values ≥ 0.95 or at least .90 for CFI, NNFI, and IFI, and values ≤ .08 for RMSEA and SRMR, indicate a good model fit (Hu & Bentler, 1998).

The Composite Reliability Measure was calculated to examine the measurement reliability. Values above 0.7 indicate sufficient measurement reliability, as proposed by Nunnally (1978). Multigroup analysis was performed to test the measurement invariance across the two groups. When metric invariance was not met, further examination by the chi-square difference test was performed to identify the loadings that were not equal in the two groups. At least partial metric and scalar invariance was established for all factors — with at least two indicators showing invariant loadings and intercepts per factor — we proceeded to compare latent means across groups. This level of partial invariance provides sufficient identification for meaningful comparison, as supported by Byrne et al. (1989) and Brown (2015). Finally, the structural model — referring to the relationship between independent latent variables and dependent latent variable — was compared across groups using a chi-square difference test. All statistical testing was done at the significance level $\alpha = .05$. Statistical analysis was performed by computer software SPSS, v. 27, and LISREL 8.80.

Results

Descriptive statistics per item in each study group are summarized in Table 3. There are no major differences in means or standard deviations per item between the study groups. Slightly higher means are present in the GroupCA for the perceived vulnerability. Slightly higher variation is present in some of the items in the GroupCA. The distribution of answers to all items is approximately normal.

Table 3

Descriptive Statistics of Items (Adjusted for Smart Phones' Use) by Study Groups

	GroupSA (n = 256)				GroupCA (n = 276)			
Construct/Indicator	***M***	***SD***	**Skew**	**Kurt**	***M***	***SD***	**Skew**	**Kurt**
Perceived vulnerability								
I could be subject to a serious information security threat (PV-1).	3.4	1.0	-0.5	-0.4	3.6	1.1	-0.5	-0.5
I am facing more and more information security threats (PV-2).	3.2	1.1	-0.3	-0.8	3.4	1.1	-0.3	-0.9
I feel that my smart phone could be vulnerable to a security threat (PV-3).	3.4	1.0	-0.5	-0.5	3.6	1.1	-0.6	-0.3
It is likely that my smart phone will be compromised in the future (PV-4).	2.7	1.0	0.2	-0.6	2.8	1.1	0.3	-0.6
Information and data on my smart phone is vulnerable to security breaches (PV-5).	3.1	1.0	-0.2	-1	3.2	1.2	-0.2	-1
I could fall victim to a malicious attack if I fail to follow good security practices (PV-6).	3.4	1.0	-0.6	-0.1	3.6	1.2	-0.5	-0.6
Perceived severity								
A security breach on my smart phone would be a serious problem for me (PS-1).	3.6	1.1	-0.6	-0.4	3.6	1.2	-0.6	-0.5
Loss of information resulting from hacking would be a serious problem for me (PS-2).	3.6	1.2	-0.6	-0.5	3.7	1.3	-0.7	-0.7
Having my confidential information on my smart phone accessed by someone without my consent or knowledge would be a serious problem for me (PS-3).	3.9	1.1	-0.9	0.3	3.9	1.1	-0.9	0.1
Having someone successfully attack and damage my smart phone would be very problematic for me (PS-4).	3.7	1.1	-0.8	0.1	3.8	1.1	-0.9	0.1
I view information security attacks on me as harmful (PS-5).	3.8	1	-0.9	0.6	3.8	1.2	-0.9	-0.1
I believe that protecting the information on my smart phone is important (PS-6).	4.2	0.9	-1.3	2	4.3	0.9	-1.4	1.9
Response efficacy								
Technical security measures help preventing security breaches on smart phones (RE-1)	3.9	0.8	-0.6	0.5	3.9	0.9	-0.9	1.2
Implementing security measures on my smart phone is an effective way to prevent security breaches (RE-2).	3.6	0.8	-0.2	-0.1	3.5	1	-0.4	-0.1
Enabling technical security measures would prevent hackers to steal personal information from smart phones (RE-3).	3.5	1	-0.6	0	3.6	1	-0.6	0.2
Available preventive measures are effective to stop people from getting confidential information from smart phones (RE-4).	3.5	0.8	-0.3	-0.1	3.5	1	-0.4	-0.1

	GroupSA (n = 256)				GroupCA (n = 276)			
Construct/Indicator	***M***	***SD***	**Skew**	**Kurt**	***M***	***SD***	**Skew**	**Kurt**
Self-efficacy								
I feel comfortable taking measures to secure my smart phone (SE-1).	3.7	0.9	-0.8	0.8	3.7	1	-0.4	-0.4
Taking the necessary security measures is entirely under my control (SE-2).	3.4	1	-0.3	-0.4	3.3	1.1	0	-0.9
I have the resources and the knowledge to take the necessary security measures (SE-3).	3.3	1.1	-0.4	-0.6	3.3	1.1	-0.1	-0.8
Taking the necessary security measures is easy (SE-4).	3.3	1	-0.2	-0.2	3.2	1	-0.1	-0.5
I can protect my smart phone by myself (SE-5).	3.5	1.1	-0.7	0	3.3	1.2	-0.3	-0.8
I can enable security measures on my smart phone (SE-6).	3.5	1.1	-0.6	-0.1	3.3	1.1	-0.3	-0.6
Security intentions								
I am likely to take security measures on my smart phone (SI-1).	3.9	0.9	-0.9	1.2	3.8	1	-0.8	0.4
It is possible that I will take security measures to protect my smart phone (SI-2).	4	0.8	-0.7	0.8	3.9	0.9	-0.8	0.6
I am certain that I will take security measures to protect my smart phone (SI-3).	3.8	1	-0.7	0	3.8	1	-0.5	-0.2
It is my intention to take measures to protect my smart phone (SI-4).	3.9	0.9	-0.8	0.7	3.9	1	-0.7	0.1

To test the familiarity hypothesis (Weijters et al., 2013), we computed the total number of high-anchor responses ('Strongly agree' and 'Completely agree') across all 26 items and compared them using an Independent Samples *T*-Test. The mean number of "Strongly agree" responses per item was $M = 51.7$, $SD = 23.9$, and the mean number of "Completely agree" responses was $M = 66.2$, $SD = 28.0$. The difference in means (*T*-Test for Independent Variables) is statistically significant, $t = 2.01$, $p < .05$. Hence, we conclude that respondents more frequently chose the anchor point that uses phrases more commonly used in their daily language.

Next, to assess how different anchor points influence the validity and reliability of the questionnaire, we conducted a Confirmatory Factor Analysis (CFA). The results of CFA per study group are summarized in Table 4. The results are very similar. Both CFA models exhibit good overall fit, factor loadings above 0.5 and AVE above 0.50, supporting convergent validity for both models. The reliability, as measured by composite reliability measure, indicates sufficient measurement reliability for all PMT scales in both study groups. However, slightly higher values for CR and AVE were observed in GroupCA.

Table 4

Results of CFA by Study Group

Construct/Indicator	Std. Weights (GroupSA)	Std. Weights (GroupCA)
Perceived Vulnerability	CR = .86; AVE = .51	CR = .88; AVE = .56
PV-1	.79	.84
PV-2	.71	.71
PV-3	.80	.89
PV-4	.59	.65
PV-5	.67	.75
PV-6	.69	.60
Perceived Severity	CR = .90; AVE = .61	CR = .92; AVE = .65
PS-1	.76	.83
PS-2	.77	.81
PS-3	.78	.82
PS-4	.84	.87
PS-5	.84	.87
PS-6	.68	.61
Response Efficacy	CR = .82; AVE = .54	CR = .84; AVE = .56
RE-1	.75	.76
RE-2	.81	.81
RE-3	.68	.71
RE-4	.68	.71
Self-Efficacy	CR = .90; AVE = .60	CR = .91; AVE = .63
SE-1	.63	.66
SE-2	.61	.70
SE-3	.79	.78
SE-4	.68	.75
SE-5	.93	.91
SE-6	.93	.93
Security Intentions	CR = .93; AVE = .78	CR = .92; AVE = .75
SI-1	.91	.91
SI-2	.78	.89
SI-3	.93	.84
SI-4	.91	.82
	$\chi^2(289) = 644.8$, $p < .001$, $\chi^2/df = 2.2$; RMSEA = .07; NFI = .92; NNFI = .95; IFI = .95; CFI = .95; SRMR = .07	$\chi^2(289) = 695$, $p < .001$, $\chi^2/df = 2.4$; RMSEA = .07; NFI = .93; NNFI = .96; IFI = .96; CFI = .96; SRMR = .07

Multigroup analysis was performed to test the measurement invariance of PMT scales across the two groups (Table 5). The results indicate that the global latent factor structure is the same across groups (configural invariance). Metric invariance was not supported by the data, suggesting that factor loadings differ between the two groups. Further analysis showed that measurement non-invariance is due to SI-2 and SI-3 indicators

measuring security intentions. Partial measurement invariance was supported in the model, allowing the two indicators to have different factor loadings across groups. Scalar invariance was supported by the overall model fit and an insignificant chi-square difference test between the partial metric and scalar invariance model. Finally, the model's fit presupposing different latent means across groups was not significantly better than the model with the same latent means, suggesting the five latent factors do not differ in the mean values between the two groups.

Table 5

Test of Measurement Invariance for PMT

Level/Type	*df*	χ^2	χ^2/df	RMSEA	NFI	NNFI	CFI	IFI	SRMR	$\Delta\chi^2$	Δdf	*p*
Config.	578	1339.8*	2.3	0.07	0.93	0.95	0.96	0.96	0.07	—	—	—
Metric	599	1388.9*	2.3	0.07	0.93	0.95	0.96	0.96	0.08	49.1	21	.001
Partial metric	597	1363.3*	2.3	0.07	0.93	0.95	0.96	0.96	0.08	23.5	19	.215
Scalar	616	1379.3*	2.2	0.07	0.96	0.96	0.96	0.96	0.08	15.9	19	.662
Mean	621	1387.6*	2.2	0.07	0.93	0.96	0.96	0.96	0.08	8.4	5	.138

*$p < .001$.

Finally, we tested the structural equation model in the two groups (Table 6). The model presupposing the same structural loading between each independent variable (perceived vulnerability, perceived severity, response efficacy, and self-efficacy) and security intentions did not differ significantly from the model suggesting equal structural weights across the two groups ($\Delta\chi^2 = 2.59$; $p = .629$). However, the model fit was slightly better for the model with different structural weights across groups (RMSEA = 0.07; CFI = 0.96; IFI = 0.96, SRMR = 0.09, NFI = 0.93, NNFI = 0.96). Although non-significantly different, the weight between perceived vulnerability and security intentions is higher and statistically significant in the CA group, while in SA group, the relationship is not statistically significant. It could be due to chance, but the relationship should be further explored when using SA or CA scale.

Table 6

Relationship Between PTM Constructs – Multigroup Comparison (Results of SEM)

	SA (n = 256)			CA (n = 276)		
Path	**Std reg coef.**	***t***	***p***	**Std reg coef.**	***t***	***p***
PV —> SI	0.01	0.14	.889	0.13	2.07	.039
PS —> SI	0.22	3.2	.002	0.22	3.22	.001
RE —> SI	0.33	4.8	< .001	0.25	3.62	< .001
SE —> SI	0.35	5.53	< .001	0.35	5.56	< .001

Discussion

According to the finding from the *T*-Test that "Completely agree" was, on average, selected more often than "Strongly agree", we confirm the familiarity hypothesis assuming that respondents will feel more comfortable choosing the more familiar endpoint, despite its semantically higher extremity. This is in accordance with the research findings (Weijters et al., 2013) that respondents tend to choose more familiar labels more frequently, even if these labels convey greater intensity (and should, theoretically, be chosen less often for that reason).

When comparing the mean values for the constructs, perceived vulnerability values were statistically significantly higher in the GroupCA compared to the GroupSA with greater variability in the GroupCA. This finding also aligns with the familiarity hypothesis (Weijters et al., 2013), suggesting that idiomatic and relatable translations encourage respondents to engage more with the extreme points of the scale. However, mean differences for other constructs were not statistically significant, though the GroupCA showed consistently higher variances across several scales. Furthermore, differences in skewness and kurtosis also emphasize the impact of anchor wording on response distributions. The GroupCA showed flatter distributions with reduced kurtosis and skewness values closer to zero for some scales, such as self-efficacy. These patterns suggest that idiomatic anchors in the GroupCA promoted more balanced and diverse response distributions.

CFA (Table 3) showed good model fit and convergent validity across both groups. Both groups achieved factor loadings above 0.5, AVE values above 0.50, and CR values above 0.7. This indicates sufficient measurement reliability and validity for both groups. However, nearly all CR and AVE values are slightly higher in the GroupCA compared to the GroupSA. While these differences are not large, their consistency suggests a slight advantage in measurement reliability and validity for the GroupCA. These results may reflect the influence of more idiomatic and familiar anchor translations in the GroupCA, which could have improved the respondents' ability to consistently interpret and respond to the scale. This aligns with prior research that emphasized the role of anchor familiarity in enhancing response fluency since more familiar word combinations allow respondents to process them more quickly and perceive them as more reliable (Conklin & Schmitt, 2008; Weijters et al., 2013). Although the differences in CR and AVE values are very minor, their near-uniformity across constructs highlights the potential benefits of selecting anchors that resonate more naturally with the target population.

The test of measurement invariance using multigroup analysis provided sufficient support for configural invariance, but not for full metric and scalar invariance across GroupSA and GroupCA. Metric and scalar invariance were established for all factors but Security intentions, where two items showed differing loadings and intercepts across the groups. Partial metric and scalar invariance was achieved for this factor, as the remaining two items were invariant across groups. This suggests that, in some cases, differences in response scale wording may influence how certain items are understood and answered.

However, having at least two invariant items per factor, enables meaningful comparisons of latent means across groups, as conducted in our study. We recommend caution when using different response scales across languages, especially when constructs are measured with fewer than three items. Such practices may threaten measurement invariance and lead to potentially erroneous conclusions about cross-cultural differences, which could steam from differences in item interpretation rather than true construct differences.

The results of multigroup SEM models indicate that for both groups, perceived severity, response efficacy, and self-efficacy were significantly associated with security intentions, consistent with the predictions of Protection motivation theory. The standardized path coefficients are stable across the groups, suggesting that these constructs consistently influence behavioural intentions regardless of the scale format used. However, the relationship between perceived vulnerability and security intentions was significant only in the GroupCA. This suggests that the anchor points in GroupCA may have enabled respondents to visualize their responses regarding their perceived vulnerability better. The difference may have caused it to have a significant impact on respondents' behavioural intentions, compared to GroupSA, where this association was not statistically significant. The differences in regression results between the groups may also reflect the influence of anchor familiarity on response behaviour. As indicated in Weijters et al. (2013), more familiar anchors can enhance cognitive fluency, which leads to more confident and potentially extreme responses. Theoretically, all PMT constructs are recognized as the key determinants of protective behaviours or security intentions. Consequently, significant correlations among these constructs are expected, as they have been consistently empirically confirmed in previous studies (Li et al., 2022; Mou et al., 2022; Zuwita & Rahmatullah, 2021). This suggests that results from GroupCA yielded more valid conclusions.

The consistent differences observed between the groups show the nuanced influence of anchor phrasing on the measurement properties of survey instruments. The higher variance and more normally distributed response patterns in the GroupCA suggest that idiomatic phrasing may encourage respondents to engage more intensively with the full spectrum of response options. This might also potentially contribute to mitigating biases like central tendency or acquiescence bias. These findings support the idea that familiar anchors create cognitive ease and enable respondents to better differentiate between response categories and express their attitudes with higher precision. This may also explain the stronger association between perceived vulnerability and security intentions in the GroupCA in the regression model. When respondents are presented with familiar and intuitive phrasing, they may feel more confident in articulating their perceptions of vulnerability, resulting in a more pronounced connection between their attitudes and behavioural intentions.

We should note that even though the observed differences in means, variances, and regression outcomes were not large, their consistency across multiple constructs and items may indicate that the linguistic relatability of anchors does not merely affect isolated responses but can influence the overall structure of the data. For example, skewness and kurtosis values of the constructs in the GroupCA reflect a tendency toward more balanced engagement with the scale. Participants have spread their responses more evenly across the available options. These findings suggest that anchor phrasing should not be considered a neutral design choice but an important part of survey design that shapes how respondents interpret and interact with survey items. This can impact results on the item level, construct level, and the associations between constructs.

Limitations and Future Work

First, this study was conducted only in the Slovenian language. This limits the generalizability of the findings to other languages. Future research should explore these effects in additional languages to examine whether the observed patterns hold universally or are specific to languages with similar syntactic and semantic norms. Second, the study did not evaluate the differences between widely used English-language end-point anchors, such as "*Completely disagree*" to "*Completely agree*" versus "*Strongly disagree*" to "*Strongly agree*". Future work should directly compare these anchor sets in English using the same items and constructs as those employed in our research to distinguish whether the observed effects are due to differences in anchor phrasing or specific to the translation into Slovenian. This would help explain if the results stem from properties of the anchor points themselves or from linguistic factors introduced during the translation. Third, the assessments of intensity and familiarity of analysed endpoints were based on semantic assessment and Google hits familiarity score. Additional research is needed to empirically measure the ratings as perceived by respondents. Fourth, the study focused on a single topic — security-related constructs measured using protection motivation theory. Therefore, the findings may not be generalizable to other domains or psychological constructs. Future research should, therefore, include domains other than information security. Fifth, even though we found perceived vulnerability to be a significant predictor in one group and not significant in the other, the overall model comparison was not statistically significant, so this difference should be interpreted with caution and commends future work with larger sample sizes. Additionally, the lack of full scalar invariance implies that observed differences in mean scores between groups should be interpreted with caution, as they may reflect measurement bias rather than true differences in the underlying constructs. Finally, while the study indicated slight differences in reliability, validity, and response distributions, it did not investigate cognitive mechanisms that could drive these effects. Future work should investigate whether these differences are indeed attributable to cognitive fluency and linguistic familiarity or a combination of these and other factors.

Funding: The authors have no funding to report.

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: The authors have declared that no competing interests exist.

Ethics Statement: The study received appropriate ethical review and clearance, thus meeting all ethical standards for research (The Ethics Commission of the Faculty of Criminal Justice and Security, 20. 02. 2025, No.: 2002-2025).

Data Availability: The dataset collected and analysed for this publication is available upon reasonable request from the corresponding author.

Supplementary Materials

Type of supplementary material	Availability/Access
Data	
The dataset can be requested from the corresponding author.	—
Preregistration	
Study was not preregistered.	—
Code	
No code was provided from the study.	—
Material	
No material was provided from the study.	—

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*(3), 411–423. https://doi.org/10.1037/0033-2909.103.3.411

Brislin, R. W. (1980). Translation and content analysis of oral and written materials. In H. C. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology* (pp. 389–444). Allyn & Bacon

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Casper, W. C., Edwards, B. D., Wallace, J. C., Landis, R. S., & Fife, D. A. (2020). Selecting response anchors with equal intervals for summated rating scales. *Journal of Applied Psychology, 105*(4), 390–409. https://doi.org/10.1037/apl0000444

Cliff, N. (1959). Adverbs as multipliers. *Psychological Review, 66*(1), 27–44. https://doi.org/10.1037/h0045660

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics, 29*(1), 72–89. https://doi.org/10.1093/applin/amm022

Fischer, R. (2004). Standardization to account for cross-cultural response bias. *Journal of Cross-Cultural Psychology, 35*(3), 263–282. https://doi.org/10.1177/0022022104264122

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39–50. https://doi.org/10.1177/002224378101800104

Funke, F., & Reips, U.-D. (2012). Why semantic differentials in web-based research should be made from visual analogue scales and not from 5-point scales. *Field Methods, 24*(3), 310–327. https://doi.org/10.1177/1525822X12444061

Hardy, B., & Ford, L. R. (2014). It's not me, it's you. *Organizational Research Methods, 17*(2), 138–162. https://doi.org/10.1177/1094428113520185

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424–453. https://doi.org/10.1037/1082-989X.3.4.424

Ilgun Dıbek, M. I. (2020). Effect of extreme and acquiescence response style in Timss 2015. *Eurasian Journal of Educational Research, 20*(87), 199–220. https://doi.org/10.14689/ejer.2020.87.10

Kampen, J. K. (2019). Reflections on and test of the metrological properties of summated rating, Likert, and other scales based on sums of ordinal variables. *Measurement, 137*, 428–434. https://doi.org/10.1016/j.measurement.2019.01.083

Lalla, M. (2017). Fundamental characteristics and statistical analysis of ordinal variables: A review. *Quality & Quantity, 51*(1), 435–458. https://doi.org/10.1007/s11135-016-0314-5

Lantz, B. (2013). Equidistance of Likert-type scales and validation of inferential methods using experiments and simulations. *Electronic Journal of Business Research Methods, 11*(1), 16–28.

Li, W., Liu, R., Sun, L., Guo, Z., & Gao, J. (2022). An investigation of employees' intention to comply with information security system — A mixed approach based on regression analysis and fsQCA. *International Journal of Environmental Research and Public Health, 19*(23), Article 16038. https://doi.org/10.3390/ijerph192316038

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology, 79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Menold, N., Kaczmirek, L., Lenzner, T., & Neusar, A. (2014). How do respondents attend to verbal labels in rating scales? *Field Methods, 26*(1), 21–39. https://doi.org/10.1177/1525822X13508270

Mircioiu, C., & Atkinson, J. (2017). A comparison of parametric and non-parametric methods applied to a Likert scale. *Pharmacy, 5*(2), Article 26. https://doi.org/10.3390/pharmacy5020026

Mou, J., Cohen, J., Bhattacherjee, A., & Kim, J. (2022). A test of protection motivation theory in the information security literature: A meta-analytic structural equation modeling approach in search advertising. *Journal of the Association for Information Systems, 23*(1), 196–236. https://doi.org/10.17705/1jais.00723

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Nunnally, J., & Bernstein, I. (1994). *Psychological theory*. McGraw-Hill.

Rogers, R. W. (1975). A protection motivation theory of fear appeals and attitude change. *Journal of Psychology, 91*(1), 93–114. https://doi.org/10.1080/00223980.1975.9915803

Sangthong, M. (2020). The effect of the Likert point scale and sample size on the efficiency of parametric and nonparametric tests. *Thailand Statistician, 18*(1), 55–64.

Smith, T. W., Mohler, P. P., Harkness, J., & Onodera, N. (2008). Methods for assessing and calibrating response scales across countries and languages. In M. Sasaki (Ed.), *New frontiers in comparative sociology* (pp. 45–94). BRILL. https://doi.org/10.1163/ej.9789004170346.i-466.29

Spector, P. E. (1976). Choosing response categories for summated rating scales. *Journal of Applied Psychology, 61*(3), 374–375. https://doi.org/10.1037/0021-9010.61.3.374

Thompson, N., McGill, T., & Narula, N. (2024). "No point worrying" — The role of threat devaluation in information security behavior. *Computers & Security, 143*, Article 103897. https://doi.org/10.1016/j.cose.2024.103897

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

Vieira, A. L. (2011). *Interactive LISREL in practice: Getting started with a SIMPLIS approach*. Springer. https://doi.org/10.1007/978-3-642-18044-6

Wakita, T., Ueshima, N., & Noguchi, H. (2012). Psychological distance between categories in the Likert scale. *Educational and Psychological Measurement, 72*(4), 533–546. https://doi.org/10.1177/0013164411431162

Weijters, B., Baumgartner, H., & Geuens, M. (2016). The calibrated sigma method: An efficient remedy for between-group differences in response category use on Likert scales. *International Journal of Research in Marketing, 33*(4), 944–960. https://doi.org/10.1016/j.ijresmar.2016.05.003

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing, 27*(3), 236–247. https://doi.org/10.1016/j.ijresmar.2010.02.004

Weijters, B., Geuens, M., & Baumgartner, H. (2013). The effect of familiarity with the response category labels on item response to Likert scales. *Journal of Consumer Research, 40*(2), 368–381. https://doi.org/10.1086/670394

Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64*(6), 956–972. https://doi.org/10.1177/0013164404268674

Wyatt, R. C., & Meyers, L. S. (1987). Psychometric properties of four 5-point Likert type response scales. *Educational and Psychological Measurement, 47*(1), 27–35. https://doi.org/10.1177/0013164487471003

Zuwita, R. M., & Rahmatullah, B. (2021). Relationship between PMT appraisals and Security Practice: Analysis of prevention of insider threat in organization success factor. *Elementary Education Online, 20*(4), 1118–1126.

EUROPEAN ASSOCIATION OF METHODOLOGY

Methodology (METH) is the official journal of the European Association of Methodology (EAM).

zpid Leibniz-Institut für Psychologie

PsychOpen GOLD is a publishing service provided by the Leibniz Institute for Psychology (ZPID), Germany.