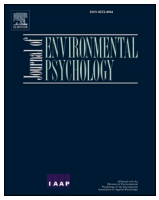
Journal of Environmental Psychology 101 (2025) 102520

Contents lists available at ScienceDirect

Journal of Environmental Psychology

journal homepage: www.elsevier.com/locate/jep



An item response theory approach to measurement in environmental psychology: A practical example with environmental risk perception

Fanny Lalot [a,*], Juulia Räikkönen [b], Sanna Ahvenharju [c]

[a] Faculty of Psychology, University of Basel, Basel, Switzerland
[b] Biodiversity Unit, University of Turku, Turku, Finland
[c] Finland Futures Research Centre, University of Turku, Turku, Finland

ARTICLE INFO

Keywords:
Environmental risk perception
Item response theory
Modern test theory
Scale development

ABSTRACT

Environmental psychology heavily relies on psychometric scales to approach relevant psychological constructs. Traditionally, these scales have most often been developed using classical test theory, despite the availability of more advanced methods like Item Response Theory (IRT)—a specific form of "modern test theory". The increasing capabilities of statistical software and the growing availability of open-source tools such as *R* packages have made IRT analyses more accessible and easier to implement. Adopting such approach would significantly benefit the field by enhancing the rigour and precision of our measurement instruments. In this short note, we present a practical example of applying IRT to developing a short scale of environmental risk perception (assessing perceived likelihood, seriousness, and concern about threats related to biodiversity loss and climate change). We use data from a large-scale survey of the views of the population of Finland about biodiversity and other environment-related issues ($N = 2005$). In a dual-step process of confirmatory factor analysis followed by IRT, we demonstrate evidence of validity and reliability of the 6-item environmental risk perception scale in the context of a national study. We illustrate how IRT offers a more informative and comprehensive evaluation of specific items (assessing their location, discrimination, and information) and, therefore, of the overall scale (information and conditional reliability) compared to classical test theory. We advocate for the broader adoption of IRT within environmental psychology to improve the quality of the instruments we rely upon as a field.

1. Introduction

Environmental psychology relies heavily on psychometric scales to approach relevant psychological constructs (such as perceived risk, attitudes, etc.). Many researchers put real effort into developing instruments that are valid and reliable. However, they often resort to *classical test theory* (or "conventional" psychometric testing) instead of more advanced methods such as item response theory (IRT; a specific case of *modern test theory*; see Crocker & Algina, 1986).[1] Yet, these methods have become considerably more accessible to non-statistician researchers (e.g., with several open-access packages accessible on *R*).

With this short note, we aim to provide a concrete example of applying item response theory to developing a short scale of environmental risk perception. We illustrate how IRT can provide more information about specific items and the overall scale than classical test theory. We hope to convince other researchers to consider using item response theory to improve the quality of the instruments we rely upon as a field.

1.1. Item response theory: a brief introduction

More than a theory, Item Response Theory (IRT) is a collection of models which aim to explain the process by which individuals respond to items (Edwards, 2009; Hambleton et al., 1991; Veldkamp, 2005; Wirth & Edwards, 2007) with foundations dating as far back as the 1940–1950s (Guttman, 1944; Lord, 1952; Rasch, 1960). Any answer

* Corresponding author. Fakultät für Psychologie, Missionsstrasse 64a, 4055, Basel, Switzerland
E-mail address: fanny.lalot@unibas.ch (F. Lalot).

[1] For example, a search for "item response theory" in all articles from the *Journal of Environmental Psychology* (22.08.2024) only yielded 17 hits published between 2003 and 2024, of which 13 were indeed examples of using IRT analysis (and four were co-authored by a single researcher) – see Discussion for the full list. A noteworthy exception in the field is the Campbell paradigm, which relies on a classical Rasch model (i.e., a one-parameter logistic item response theory model; see e.g., Kaiser et al., 2010, 2011; 2018).

https://doi.org/10.1016/j.jenvp.2025.102520
Received 24 October 2024; Received in revised form 4 January 2025; Accepted 4 January 2025
Available online 6 January 2025

Table 1
Descriptive statistics of the six items forming the environmental risk perception scale.

Item	Label	*M*	*SD*
bd_risk1	I consider it likely that sometime during my life, I will experience serious threats to my health or overall wellbeing, as a result of biodiversity loss.	4.04	1.58
bd_risk2	I am highly concerned about biodiversity loss.	4.52	1.65
bd_risk3	The impacts of biodiversity loss cause a serious threat in Finland.	4.43	1.62
cc_risk1	I consider it likely that sometime during my life, I will experience serious threats to my health or overall wellbeing, as a result of climate change.	4.19	1.66
cc_risk2	I am highly concerned about climate change.	4.51	1.75
cc_risk3	The impacts of climate change cause a serious threat in Finland.	4.42	1.71

Note. All items use a 7-point Likert scale.

provided on a scale can be conceptualised as a function of person properties and item properties. Within an IRT framework, the unit of analysis is the individual item response.

Different models have been developed that can handle either dichotomous or polytomous response formats. These models allow us to describe items in terms of their *difficulty* (or *location*; as the relationship between the probability of ticking a specific answer and the 'real' level of the latent ability) and *discrimination* (how quickly the probability of ticking a specific answer changes as the latent ability increases). More complex models can also include a guessing parameter (Baker & Kim, 2017).

A unique characteristic of IRT is thus that item properties are expressed as a function of the latent trait. It allows, for example, to assess how people respond to items at different levels of the latent trait and how well the items (and the test) may capture differences between people at different levels of this continuum (e.g., whether they can reliably differentiate individuals high versus very-high or low versus very-low on the trait). This implies that measurement precision may vary at different levels of the latent trait, which IRT assesses with a test information function (thus providing more accurate information than a single value coefficient such as α). IRT therefore allows for a more informative and thorough evaluation of the items (and, from there, the person's latent trait). Within this framework, scale scoring can rely on expected a posteriori scores (EAP; Thissen et al., 1995), which reflect the estimated ability levels (θ) of respondents calculated as the mean of the posterior distribution of θ given their item responses (Brown & Croudace, 2015). In other words, unlike sum scores, EAP provides an estimation of the latent trait that is weighted on item parameters (discrimination and location) and thus directly represents the underlying psychometric structure of the scale.

As such, IRT may address the limitations of classical test theory, including its test-level approach and the fact that scores depend on the test items and sample. Still, there are some downsides to consider: IRT requires greater mathematical computations (although this is less and less problematic given the current capacity of different widely available software) and larger sample sizes (Paek & Cole, 2019). IRT also adopts a confirmatory approach in nature and requires different assumptions to be respected and thus verified beforehand (e.g., number of dimensions; see below). It is beyond the scope of this short note to provide a thorough background for IRT methods and calculations. Still, we point to relevant references and tutorials in the general discussion.

1.2. Measuring environmental risk perception

In the present study, we aim to test the psychometric properties of a short environmental risk perception scale, relying on an IRT approach. Environmental risk perception refers to the subjective evaluation and interpretation of potential environmental hazards and their impacts. It has been recognised as a key determinant of pro-environmental intentions (Xie et al., 2019) and actions (Bradley et al., 2020), as well as adaptive and preventive behaviour (De Dominicis et al., 2015; Tan & Xu, 2019). Risk perception also increases support for government-implemented environmental policies (O'Connor et al., 1999). It remains a central construct in contemporary environmental psychology research (e.g., Gilbert & Lachlan, 2023; Hornsey & Pearson, 2024; van der Linden, 2015; Xue et al., 2014).

Many existing scales of risk perception focus on one environmental issue, such as climate change (van der Linden, 2014), flood (De Dominicis et al., 2015), etc. Some distinguish risk subtypes, for example, personal versus societal (van der Linden, 2015), while others consider risk globally (Gilbert & Lachlan, 2023). In the present project, we sought to develop a risk perception measure that (i) was short enough to suit a large-scale survey having to compromise between the number of constructs assessed and the number of items, and (ii) would potentially capture perceived risk from two environmental issues that are at the core of the research project: climate change and biodiversity loss.

1.3. The present study

We rely here on data collected in a larger-scale project aiming to investigate the views of the population of Finland about biodiversity and other environment-related issues (https://biodiful.fi). In line with the general objectives of the research project, we developed items that asked about perceived risk from climate change and biodiversity loss. Assuming high face validity, one might expect that the two subscales represent distinct types of risk and form separate indicators. However, we suspected respondents might respond to these items by expressing a general degree of perceived environmental risk. The items might then measure not different facets but different degrees of risk. In this case, a single scale might better capture different levels of perceived risk. The current analyses aimed to test this possibility.

As a validity criterion, we also investigated the relationship between perceived risk and support for government-led environmental policies. We expected perceived risk to be positively related to policy support. This analysis provides us with the opportunity to compare (and comment on) expected a posteriori (EAP) scores, directly derived from the IRT model, with the more traditional sum scores (see below). The research was conducted in compliance with the Declaration of Helsinki and ethical considerations for research with human participants. We preregistered the study's aim, hypothesis, analysis plan, sample size, and rules for exclusions: https://aspredicted.org/pqgs-t98q.pdf.[2] Data, materials, and code for analysis are publicly available on the OSF: https://osf.io/h6ru2/.

2. Method

2.1. Participants and procedure

We recruited a representative sample of Finland's population with quotas set to match demographics by gender, age group, and region. The target sample size was determined by available funding ($N = 2000$). An independent polling company collected the data and was directly responsible for data exclusion. Low-quality responses (more than 10% missing values, unrealistically quick completion time, and failed attention checks) were removed, with slots reopening to new participants on a rolling basis. The final sample included 2005 participants (931 men, 1049 women, 17 non-binary or other, and 6 undisclosed) with a mean age of 47.32 years ($SD = 16.12$). The study was an online questionnaire

[2] One may note that the preregistration refers to "environmental concern". However, we later recognised that the measure and items correspond to perceived *risk* rather than *concern*. Therefore, we consistently refer here to "perceived risk".

Table 2
Goodness-of-fit of the three confirmatory factor analysis (CFA) models.

Model	χ^2	df	χ^2/df	AIC	BIC	CFI	RMSEA, 90% CI	SRMR
Unifactorial	1086	9	120.7	35801	35903	.907	.244 [.232, .257]	.042
Unifactorial (modified)	582	8	72.7	35299	35405	.951	.189 [.176, .202]	.020
Two-factor	985	8	123.1	35702	35808	.916	.247 [.234, .260]	.043
Two-factor (modified)	345	7	49.3	35064	35176	.971	.155 [.141, .169]	.014
Bi-factor	368	5	73.6	35079	35169	.969	.190 [.174, .207]	.040
Bi-factor (modified)	22	4	5.5	34735	34830	.998	.047 [.029, .068]	.039

Note. Models indicated as "modified" include one additional covariance between two items that share a common wording. CFA used a maximum likelihood estimator.

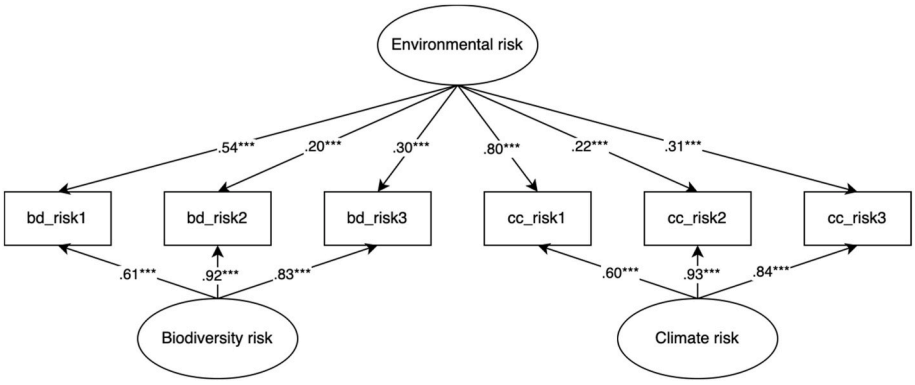


Fig. 1. A Bi-Factor Model of Environmental Risk Perception (CFA Results)
Note. Loadings are standardised estimates. This figure represents the modified bi-factor model with one additional covariance (not drawn). Item wording is reported in Table 1. ***$p < .001$.

that included measures of perceived environmental risk, attitudes towards environmental policies, and demographics.[3]

2.2. Materials

Perceived environmental risk. We developed two sets of three items inspired by van der Linden (2014) to capture risks related to both climate change and biodiversity loss.[4] We aimed to test whether a scale shorter than the original (one 8-item set) could still be used reliably. Items assessed how *likely* participants consider they could personally experience threats to their overall wellbeing, how *serious* they think the threat is to their country, and how *concerned* they are about it (Table 1). All items used a 7-point Likert scale (1 = Strongly disagree, 7 = Strongly agree). Participants were first presented with the biodiversity items and then with the climate change items, interspersed with a few other measures.

Acceptability of environmental policies. Participants rated how acceptable four general types of environmental policies were to them (e.g., "Activities that significantly harm nature are prohibited by law") on a 7-point Likert scale (1 = Very low acceptability, 7 = Very high acceptability). Exploration of the 4-item scale's factor structure is reported in Supplementary Materials (SM1).

Demographics and political orientation. In addition to age (in years) and gender (man/woman/non-binary or other/prefer not to say), we also assessed participants' *education* (highest level of education achieved; 10 levels), *income* (monthly household income; 10 bands), and *political orientation* (2 items on a 11-point scale: "Where would you place yourself on a scale from … " (i) 0 = Left, 10 = Right; (ii) 0 = Liberal, 10 = Conservative). As both items were strongly correlated ($r(1991) = .52$, $p < .001$), we aggregated them in a single indicator of political orientation.

2.3. Analysis strategy

There were no missing values on any of the key variables (environmental risk perception or policy acceptability) and only a few missing values on demographics (ranging $n = 2$ [gender and education] to $n = 16$ [income]). We did not use any imputation method and focused on complete cases only (i.e., full sample for the main analysis and $n = 1953$ for analyses using demographics).

We first conducted a confirmatory factor analysis (CFA; i.e., classical test theory approach) to test the goodness-of-fit of different models: single-factor, two-factor, and bi-factor structure, with maximum likelihood estimator. In the single-factor model, all six items load on a single factor. In the two-factor model, items load on their respective dimension (climate risk vs. biodiversity risk) and the two dimensions are allowed to covary. Finally, the bi-factor model defines one primary dimension of interest on which all items load and secondary dimensions (or subdomains) on which subsets of items load (Gibbons, 2014). This secondary dimension is often a methodological factor (e.g., positively and negatively worded questions) or a content domain (here: climate and biodiversity-related items). In this model, items load on their respective dimension (the "secondary" dimension) *and* all items also load on a primary dimension (here: general environmental risk).

The bi-factor model is useful when testing a model where one assumes the existence of a single underlying construct but where the classical method may identify two factors because of higher intercorrelations between subsets of items. This step is important because the number of dimensions must be clearly prespecified in IRT models. To assess the model fit, we considered the root mean square error of approximation (RMSEA, Steiger & Lind, 1980) and standardised root mean residual (SRMR, Bentler, 1995; see also Diamantopoulos & Siguaw, 2000; Hu & Bentler, 1999).

[3] The questionnaire also included other measures that are not discussed here (see preregistration for the full list).

[4] The three items were selected based on secondary data analysis of a previous study of ours on a similar population ($N = 1000$) which had included the full van der Linden's (2014) scale as well as other measures. A CFA identified these items (see Table 1 for wording) as having the highest loadings while being clearly distinguishable from a related construct of importance of nature.

Table 3
Graded response model parameter estimates for the environmental risk perception scale.

Item	Discrimination	Category boundary parameters						Location
	a (*SE*)	b_1 (*SE*)	b_2 (*SE*)	b_3 (*SE*)	b_4 (*SE*)	b_5 (*SE*)	b_6 (*SE*)	
bd_risk1	2.111 (.075)	−1.953 (.070)	−1.119 (.047)	−0.490 (.037)	0.333 (.036)	1.143 (.048)	1.950 (.071)	−0.023
bd_risk2	3.673 (.129)	−1.926 (.061)	−1.211 (.042)	−0.625 (.033)	−0.076 (.029)	0.594 (.032)	1.183 (.041)	−0.344
bd_risk3	4.198 (.151)	−1.764 (.054)	−1.185 (.040)	−0.601 (.032)	0.011 (.028)	0.659 (.032)	1.237 (.041)	−0.274
cc_risk1	2.952 (.100)	−1.710 (.056)	−1.018 (.040)	−0.523 (.033)	0.151 (.031)	0.843 (.037)	1.507 (.051)	−0.125
cc_risk2	4.594 (.173)	−1.612 (.049)	−1.065 (.038)	−0.605 (.031)	−0.032 (.028)	0.506 (.030)	1.028 (.037)	−0.297
cc_risk3	5.054 (.198)	−1.566 (.047)	−1.031 (.036)	−0.553 (.030)	−0.011 (.028)	0.575 (.030)	1.178 (.039)	−0.235

Note. The slope or discrimination parameter a represents the 'peakedness' of the response functions: the greater the value, the quicker the probability of choosing one response category changes when moving on the latent construct. Peakedness can be seen graphically in Fig. 2: higher a parameters result in steeper response characteristic curves. Category boundary parameters b_s (or threshold parameters estimates) show the threshold (50% likelihood) for choosing different response categories. b_1 is the level of (latent) perceived environmental risk when the likelihood of choosing category 2, 3, 4, 5, 6, or 7 is 50% (threshold); b_2 is the level of perceived environmental risk when the likelihood of choosing category 3, 4, 5, 6, or 7 is 50%, etc.

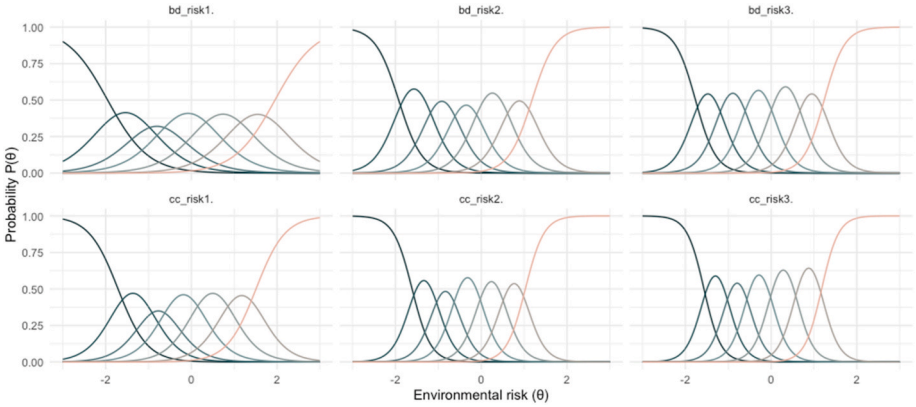


Fig. 2. Response Characteristic Curves
Note. The lines represent the probability curve for each possible answer on the Likert scale, from 1 (darker line on the left) to 7 (lighter line on the right).

We also report the comparative fit index (CFI, Bentler, 1990) and χ^2. Typically, CFI >.95, RMSEA <.07, and SRMR <.08 indicate good model fit (MacCallum et al., 1996; Steiger, 2007). These analyses were conducted on *R* with the *lavaan* package (Rosseel, 2012).

As this first step concluded to a better fit of the bi-factor model (as detailed below), we then pursued IRT analyses to explore the characteristics of the different items. As items are measured on a several-point scale, we relied on a polytomous model, specifically the *graded response model* (GRM, Samejima, 1968, 2010). The GRM assumes unidimensionality (which we had just verified), local independence, monotonicity, and a normally distributed latent trait (Edwards, 2009; Wirth & Edwards, 2007). It estimates two parameters for each item: slope or discrimination parameter a and threshold parameters b_s.

We computed and analysed: (i) *Response characteristic curve* and related coefficients, (ii) *Item information function* (and its graphical representation, the item information curve), and (iii) *Test characteristic function* (or information function) as well as standard errors and reliability. These analyses were conducted on *R* with package *mirt* (Chalmers, 2012) and additional graphs were produced with *ggmirt* (Masur, 2022).

We extracted the expected a posteriori (EAP) scores from the IRT model. In the context of IRT analyses, EAP is more appropriate than a sum score since it directly builds on the factorial structure we sought to test and validate (for similar considerations about scale validation in general, see, e.g., McNeish & Wolf, 2020). We used EAP to examine the correlation with demographics and support for environmental policies as a validity criterion. We finally tested for Differential Item Functioning (DIF) and Differential Test Functioning (DFT) to ensure against a systematic assessment bias across groups (men vs. women and left vs. right-wing respondents).

3. Results

3.1. Confirmatory factor analysis

We compared the goodness-of-fit of the three CFA models and inspected modifications indices. In each case, these indicated that adding one covariance between two items that shared a common wording (i.e., bd_risk1 and cc_risk1) would greatly improve model fit ($MI_s > 481$). Fit indices are reported in Table 2 for models with and without this additional covariance.

All models yielded reasonable fit except for RMSEA, which exceeded acceptable values for the unifactorial and two-factor models (with or without the additional covariance). The bi-factor model, in contrast, yielded a much lower RMSEA value and, overall, excellent fit statistics (CFI = .998, RMSEA = .047, 90% CI [.029, .068], SRMR = .039). We therefore concluded that the 6-item environmental risk scale was best represented by a bi-factor model: a single underlying construct organised in two subdimensions (Fig. 1).

3.2. Item response theory analysis

Having asserted that the six items represent a single dimension of general environmental risk, we turned to IRT to investigate their specific properties.[5]

[5] We initially ensured that the GRM, which assesses two parameters a and b, provided a better fit than the simpler constrained model that assumes equal discrimination parameters across items (i.e., $a_i = a$). Model comparison revealed a better fit of the unconstrained model, suggesting that discrimination parameters differ across items, $\chi^2(11) = 10974$, $p < .001$.

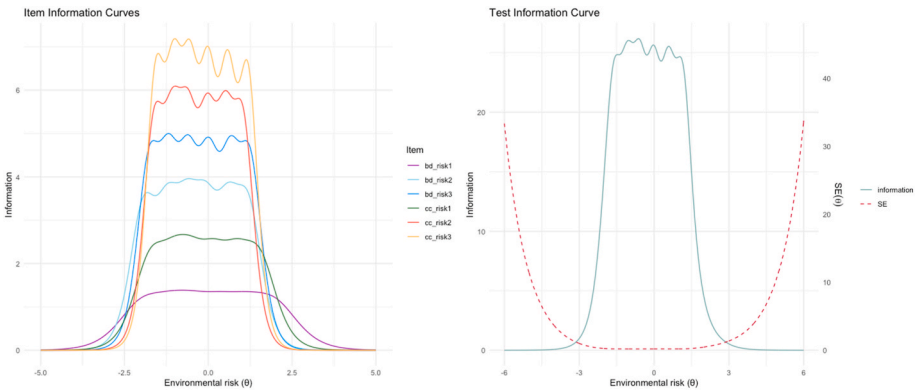


Fig. 3. Item Information Curves and Test Information Curve for the Environmental Risk Perception Scale.

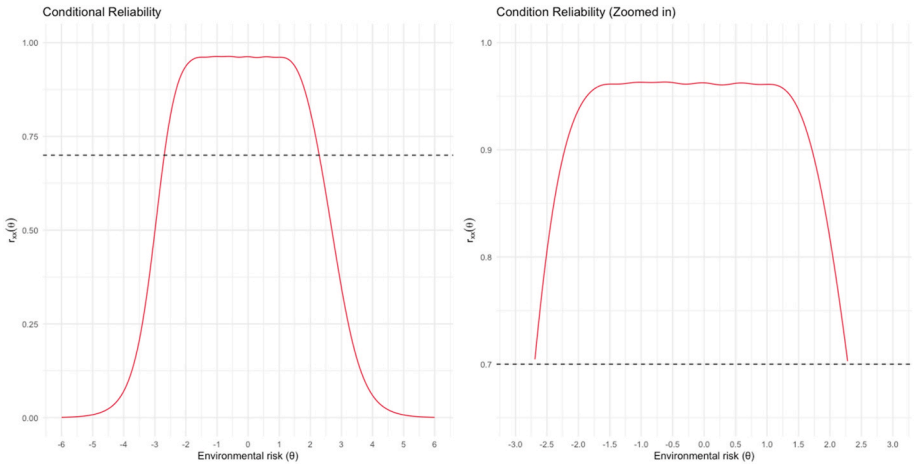


Fig. 4. Conditional Reliability Curve
Note. The dashed line marks the cut-off value of .70 for satisfactory measurement.

Response characteristic curve. Response characteristics for each item are reported in Table 3, including their discrimination parameter (or slope a), category boundary parameters (b_1-b_6), and item location (or difficulty). The discrimination parameter reflects how well an item differentiates between individuals with different levels of the latent trait: higher discrimination values indicate that the items are highly sensitive to differences in the latent construct of perceived risk around their difficulty threshold. In the present case, all discrimination parameters were >1.70, indicating very high discrimination power (Baker, 2001).[6] While high discrimination is generally desirable, extremely high values ($a > 4$) could indicate that some items are overly sensitive, which may limit their generalisability across the latent trait continuum.

Fig. 2 shows the response characteristics curves: this plots, for every level of "real" perceived risk (i.e., the underlying latent construct, plotted along the x-axis), the probability that a participant ticks each potential answer, here the number from 1 to 7 on the Likert scale. At any given point, the probability must equal 100%. This graphically represents each item's location: the point on the latent construct at which a participant is most likely to endorse a particular response option. Specifically, it indicates the level of the underlying latent trait required for a given response category to become the most probable choice. For example, a person with a latent risk score of +2SD has only a 50% chance of endorsing the highest possible answer on item bd_risk1, whereas the same latent risk score would lead to an 80% probability of endorsing the highest possible answer (7) on item cc_risk1. This indicates that bd_risk1 is *more difficult* than cc_risk1 because it requires a higher level of the latent construct (perceived risk) for participants to select the most extreme response option. Overall location (or difficulty) is also formally calculated and reported in the last column of Table 3.

Item information function. Item information quantifies how much precision an item provides in measuring the latent trait (θ) at different points along the latent trait continuum. Often, items do not discriminate well across the entire range of the construct but only at certain levels. As illustrated in Fig. 3 (left), most of our items provide more information between −2 and +2SD of perceived risk. Some items (e.g., cc_risk3) also provide more information than others (e.g., bd_risk1). Interestingly, while bd_risk1 seems to provide less information than other items, it is the one item that can better discriminate at higher levels of perceived risk (above +2SD). Therefore, this item might be especially useful for research that focuses on samples likely to express high levels of risk. However, when it is expected that most of the sample falls between −2 and +2SD of risk, this item may not provide much additional information and could potentially be discarded for brevity purposes.

Test information function. Similarly, the test information function quantifies how much precision the entire test or scale provides in measuring the latent trait along the latent trait continuum. The test curve is a sum of the individual item curves, so information values are necessarily bigger. As pictured in Fig. 3 (right), the test information curve is the inverse function of the standard error (*SE*), which expresses the standard error attached to each score along the x-axis of latent environmental risk. Here, we see that the environmental risk scale

[6] Baker (2001) provides guidelines for interpreting discrimination parameters as follows: $a \leq 0.34$: very low; $0.35 \leq a \leq 0.64$: low; $0.65 \leq a \leq 1.34$: moderate; $1.35 \leq a \leq 1.69$: high; $a \geq 1.70$: very high.

Table 4
Criterion-related validity: Effect of environmental risk perception on environmental policy acceptability.

Variables	Model 1				Model 2			
	b (*SE*)	*t*-test	*p*-value	β	*b* (*SE*)	*t*-test	*p*-value	β
(Intercept)	−0.00 (.017)	−0.01	.99	–	−0.37 (.082)	−4.48	<.001	–
Environmental risk	0.57 (.017)	33.11	<.001	.59	0.55 (.019)	29.43	<.001	.57
Gender					0.04 (.017)	2.10	.036	.04
Age					0.00 (.001)	3.36	<.001	.06
Education					0.04 (.008)	4.99	<.001	.10
Income					−0.00 (.004)	−0.46	.64	−.01
Political orientation					−0.00 (.007)	−0.52	.60	−.01
Model fit	$F(1, 2003) = 1096$, $p < .001$; $R^2_{adj} = .35$				$F(6, 1946) = 187.2$, $p < .001$; $R^2_{adj} = .36$			

Note. Gender is coded −1 = men, +1 = women, non-binary or undisclosed = missing values. Political orientation is coded so that higher values indicate a more right-wing orientation. Lower degrees of freedom in Model 2 are due to missing values for some demographics. Environmental risk and policy acceptability are both measured as expected a posteriori (EAP) scores.

provides more information or allows for capturing scores with relatively low standard errors, from roughly −2 to +2SD. Beyond values of |3SD|, however, the test cannot discriminate between individuals anymore.

One issue with the test information curve is that absolute information values are difficult to interpret. To provide a more precise interpretation, some authors have proposed to derive a conditional reliability coefficient (theoretically ranging from 0 to 1) and use a cut-off of .70 to determine the threshold for satisfactory measurement (Nicewander, 2018, 2019). Fig. 4 plots the reliability function of the environmental risk scale. From this graph (zoomed in on the right panel), we can more precisely assert that the scale reliably captures environmental risk perception (reliability >.70) between scores of −2.75 and +2.25SD.

For comparison purposes, we also generated the conditional reliability curve of a five-item scale that would exclude bd_risk1, the item that seemed to discriminate high values of environmental risk better. Without this item, the scale achieved satisfactory reliability between −2.63 and +2.13SD, showing a small decrease in the capacity to capture very high and very low risk values.

3.3. DIF/DTF and criterion-related validity

DIF and DTF analysis showed that the items and the test itself were invariant across gender groups and political groups. For brevity purposes, these results are reported in SM2.

We finally turned to criterion-related validity, regressing policy acceptability on risk perception (using expected a posteriori (EAP) scores for both constructs)[7] and adding the set of demographics as control variables in a second step. This analysis confirmed that environmental risk perception was strongly related to policy acceptability, regardless of whether demographics were introduced (Table 4).

A separate analysis showed that respondents with greater environmental risk perception were more educated, left-wing oriented, and more likely women (Table 5). In light of the DIF/DTF results, we can be confident these differences in mean levels across groups do not hide different ways of responding to the items.

4. Discussion

This short note aimed to provide a concrete example of how IRT could be applied to develop reliable instruments in environmental psychology research. We relied on data from a large-scale representative survey, which included six items to assess risk perception related to climate change and biodiversity loss.

[7] Details of the calculation of the EAP score of policy acceptability is reported in SM1. In the present case, the average score and EAP scores were very highly correlated (risk perception: $r = .98$, policy acceptability: $r = .99$) and results were virtually unchanged when considering one or the other. The full output of the comparison between calculations is reported in SM3.

Table 5
Relationships between environmental risk perception, demographics, and political orientation.

Variables	*b* (*SE*)	*t*-test	*p*-value	β
(Intercept)	−0.01 (.101)	−0.07	.95	–
Gender	0.12 (.021)	5.47	<.001	.12
Age	0.00 (.001)	0.96	.34	.02
Education	0.07 (.010)	7.17	<.001	.16
Income	0.00 (.005)	0.42	.67	.01
Political orientation	−0.10 (.009)	−12.00	<.001	−.26
Model fit	$F(5, 1947) = 53.83$, $p < .001$; $R^2_{adj} = .12$			

Note. Gender is coded −1 = men, +1 = women, non-binary or undisclosed = missing values. Political orientation is coded so that higher values indicate a more right-wing orientation. Lower degrees of freedom are due to missing values for some demographics. Environmental risk is measured as expected a posteriori (EAP) scores.

The analysis concluded that, despite a face validity argument, the six items measure a single construct of general environmental risk perception. This was assessed in the first step of data analysis relying on CFA (bi-factor model). In the second step, IRT analysis provided a thorough evaluation of the items. We could precisely determine the range of values on the latent trait (risk perception) that the scale would reliably capture. We could also qualify the items regarding their difficulty, discrimination, and amount of information they individually provide. In the present case, we can conclude that the 6-item scale reliably captures environmental risk perception over a wide range of values (−2.75 to +2.25SD). Including items with different levels of difficulty (or location) also ensures that the scale can discriminate adequately across this range of values.

We identified one item that provides less information overall and could be removed in future studies that aim to minimise item numbers with no major loss of reliability (bd_risk1). However, we also note that this item provides the most information regarding high levels of risk perception. Therefore, future research that aims to study samples where risk perception is expected to be high overall (e.g., among climate activists) would be well advised to retain this item for better measurement at the high end of the continuum.

We investigated criterion-related validity by testing the relationship between risk perception and pro-environmental policy support (see O'Connor et al., 1999), relying on the expected a posteriori scores (EAP) directly derived from the IRT model. This approach is generally preferable to the more traditional sum scores since it aligns with how the scale was validated (although in the present case EAP and sum scores were strongly correlated and yielded virtually identical results; see SM3). We finally tested for measurement invariance with Differential Item/Test Functioning (DIF/DTF), which ensured against systematic measurement bias between men/women and left/right-wing oriented respondents. This final step was crucial to ensure the instrument's validity as it helps confirm that observed group differences reflect true

variations in the trait of interest rather than artifacts of biased items.

We hope this applied example proves useful and may convince other researchers to try IRT. We conclude by pointing to the following relevant resources for readers wanting to familiarise themselves with the technique: for textbooks and chapters, see Nering and Ostini (2010); Sadler and Stokes (2022); for tutorials focusing on *R*: Baker and Kim (2017); Chalmers (2012); Paek and Cole (2019); Rizopoulos (2006); for other example papers: Edwards (2009); Mangold (2024); and for previous papers published in the *Journal of Environmental Psychology* specifically: Alisat and Riemer (2015); Bauske et al. (2022); Cologna et al. (2024); Holt et al. (2023); Irvine et al. (2023); Kaiser et al. (2003, 2007); Krettenauer et al. (2024); Miceli et al. (2008); Oinonen and Paloniemi (2023); Rodríguez-Casallas et al. (2020); Smolders et al. (2012); Zhu and Lu (2017).

CRediT authorship contribution statement

Fanny Lalot: Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Juulia Räikkönen:** Writing – review & editing, Project administration, Methodology, Funding acquisition, Conceptualization. **Sanna Ahvenharju:** Writing – review & editing, Methodology, Investigation, Conceptualization.

Data availability and open science

The study was preregistered: https://aspredicted.org/pqgs-t98q.pdf. Data, materials, and code for analysis are publicly available on the OSF: https://osf.io/h6ru2/

Funding

This research was supported by the Strategic Research Council of the Academy of Finland ("Biodiversity-respectful leadership"; grant number 345885). FL is also supported by the Swiss National Science Foundation (Grant Number: PZ00P1_216373/1).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jenvp.2025.102520.

References

Alisat, S., & Riemer, M. (2015). The environmental action scale: Development and psychometric evaluation. *Journal of Environmental Psychology, 43*, 13–23. https://doi.org/10.1016/j.jenvp.2015.05.006

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.

Baker, F. B., & Kim, S.-H. (2017). *The basics of item response theory using R*. Cham: Springer. https://doi.org/10.1007/978-3-319-54205-8

Bauske, E., Kibbe, A., & Kaiser, F. G. (2022). Opinion polls as measures of commitment to goals: Environmental attitude in Germany from 1996 to 2018. *Journal of Environmental Psychology, 81*, Article 101805. https://doi.org/10.1016/j.jenvp.2022.101805

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M. (1995). *EQS structural equations program manual*. Multivariate Software.

Bradley, G. L., Babutsidze, Z., Chai, A., & Reser, J. P. (2020). The role of climate change risk perception, response efficacy, and psychological adaptation in pro-environmental behavior: A two nation study. *Journal of Environmental Psychology, 68*, Article 101410. https://doi.org/10.1016/j.jenvp.2020.101410

Brown, A., & Croudace, T. J. (2015). Scoring and estimating score precision using multidimensional IRT models. In S. P. Reise, & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 307–333). Routledge/Taylor & Francis Group.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Cologna, V., Berthold, A., Kreissel, A. L., & Siegrist, M. (2024). Attitudes towards technology and their relationship with pro-environmental behaviour: Development and validation of the GATT scale. *Journal of Environmental Psychology, 95*, Article 102258. https://doi.org/10.1016/j.jenvp.2024.102258

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory* (1st ed.). Wadsworth Publishing.

De Dominicis, S., Fornara, F., Ganucci Cancellieri, U., Twigger-Ross, C., & Bonaiuto, M. (2015). We are at risk, and so what? Place attachment, environmental risk perceptions and preventive coping behaviours. *Journal of Environmental Psychology, 43*, 66–78. https://doi.org/10.1016/j.jenvp.2015.05.010

Edwards, M. C. (2009). An introduction to Item Response Theory using the need for cognition scale. *Social and Personality Psychology Compass, 3*(4), 507–529. https://doi.org/10.1111/j.1751-9004.2009.00194.x

Gibbons, R. (2014). Bi-factor analysis. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 386–394). Netherlands: Springer. https://doi.org/10.1007/978-94-007-0753-5_207.

Gilbert, C., & Lachlan, K. (2023). The climate change risk perception model in the United States: A replication study. *Journal of Environmental Psychology, 86*, Article 101969. https://doi.org/10.1016/j.jenvp.2023.101969

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139–150. https://doi.org/10.2307/2086306

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications, Inc.

Holt, J. R., Bui, D.-P., Chau, H., Wang, K., Trevisi, L. M., Jerdy, A. C. R., Lobban, L., Crossley, S., & Feltz, A. (2023). Development of an objective measure of knowledge of plastic recycling: The outcomes of plastic recycling knowledge scale (OPRKS). *Journal of Environmental Psychology, 91*, Article 102143. https://doi.org/10.1016/j.jenvp.2023.102143

Hornsey, M. J., & Pearson, S. (2024). Perceptions of climate change threat across 121 nations: The role of individual and national wealth. *Journal of Environmental Psychology, 96*, Article 102338. https://doi.org/10.1016/j.jenvp.2024.102338

Hu, L., & Bentler, P. M. (1999). Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Irvine, K. N., Fisher, J. C., Bentley, P. R., Nawrath, M., Dallimer, M., Austen, G. E., Fish, R., & Davies, Z. G. (2023). BIO-WELL: The development and validation of a human wellbeing scale that measures responses to biodiversity. *Journal of Environmental Psychology, 85*, Article 101921. https://doi.org/10.1016/j.jenvp.2022.101921

Kaiser, F. G., Byrka, K., & Hartig, T. (2010). Reviving Campbell's paradigm for attitude research. *Personality and Social Psychology Review, 14*(4), 351–367. https://doi.org/10.1177/1088868310366452

Kaiser, F. G., Doka, G., Hofstetter, P., & Ranney, M. A. (2003). Ecological behavior and its environmental consequences: A life cycle assessment of a self-report measure. *Journal of Environmental Psychology, 23*(1), 11–20. https://doi.org/10.1016/S0272-4944(02)00075-0

Kaiser, F. G., Hartig, T., Brügger, A., & Duvier, C. (2011). Environmental protection and nature as distinct attitudinal objects: An application of the Campbell paradigm. *Environment and Behavior, 45*(3), 369–398. https://doi.org/10.1177/0013916511422444

Kaiser, F. G., Merten, M., & Wetzel, E. (2018). How do we know we are measuring environmental attitude? Specific objectivity as the formal validation criterion for measures of latent attributes. *Journal of Environmental Psychology, 55*, 139–146. https://doi.org/10.1016/j.jenvp.2018.01.003

Kaiser, F. G., Oerke, B., & Bogner, F. X. (2007). Behavior-based environmental attitude: Development of an instrument for adolescents. *Journal of Environmental Psychology, 27*(3), 242–251. https://doi.org/10.1016/j.jenvp.2007.06.004

Krettenauer, T., Lefebvre, J. P., & Goddeeris, H. (2024). Pro-environmental behaviour, connectedness with nature, and the endorsement of pro-environmental norms in youth: Longitudinal relations. *Journal of Environmental Psychology, 94*, Article 102256. https://doi.org/10.1016/j.jenvp.2024.102256

Lord, F. (1952). *A theory of test scores (Vol. Psychometric Monograph No. 7)*. Psychometric Corporation.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130–149. https://doi.org/10.1037/1082-989X.1.2.130

Mangold, F. (2024). Improving media trust research through better measurement: An item response theory perspective. *Journal of Trust Research, 14*(1), 8–38. https://doi.org/10.1080/21515581.2023.2229791

Masur, P. K. (2022). ggmirt: Plotting functions to extend "mirt" for IRT analyses. https://github.com/masurp/ggmirt.

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods, 52*(6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Miceli, R., Sotgiu, I., & Settanni, M. (2008). Disaster preparedness and perception of flood risk: A study in an alpine valley in Italy. *Journal of Environmental Psychology, 28*(2), 164–173. https://doi.org/10.1016/j.jenvp.2007.10.006

Nering, M. L., & Ostini, R. (2010). *Handbook of polytomous item response theory models*. Routledge.

Nicewander, W. A. (2018). Conditional reliability coefficients for test scores. *Psychological Methods, 23*(2), 351–362. https://doi.org/10.1037/met0000132

Nicewander, W. A. (2019). Conditional precision of measurement for test scores: Are conditional standard errors sufficient? *Educational and Psychological Measurement, 79*(1), 5–18. https://doi.org/10.1177/0013164418758373

O'Connor, R. E., Bord, R. J., & Fisher, A. (1999). Risk perceptions, general environmental beliefs, and willingness to address climate change. *Risk Analysis, 19*(3), 461–471. https://doi.org/10.1023/A:1007004813446

Oinonen, I., & Paloniemi, R. (2023). Understanding and measuring young people's sustainability actions. *Journal of Environmental Psychology, 91*, Article 102124. https://doi.org/10.1016/j.jenvp.2023.102124

Paek, I., & Cole, K. (2019). *Using R for item response theory model applications*. Routledge. https://doi.org/10.4324/9781351008167

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1–25. https://doi.org/10.18637/jss.v017.i05

Rodríguez-Casallas, J. D., Luo, W., & Geng, L. (2020). Measuring environmental concern through international surveys: A study of cross-cultural equivalence with item response theory and confirmatory factor analysis. *Journal of Environmental Psychology, 71*, Article 101494. https://doi.org/10.1016/j.jenvp.2020.101494

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. http://www.jstatsoft.org/v48/i02/.

Sadler, B. P., & Stokes, S. L. (2022). Item Response Theory and Fisher information for small tests. In H. K. T. Ng, & D. F. Heitjan (Eds.), *Recent advances on sampling methods and educational statistics: In honor of S. Lynne Stokes* (pp. 233–250). Springer International Publishing. https://doi.org/10.1007/978-3-031-14525-4_12.

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. *ETS Research Bulletin Series, 1*, 1–169. https://doi.org/10.1002/j.2333-8504.1968.tb00153.x

Samejima, F. (2010). The general graded response model. In M. L. Nering, & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (1st ed., pp. 77–108). Routledge.

Smolders, K. C. H. J., de Kort, Y. A. W., Tenner, A. D., & Kaiser, F. G. (2012). Need for recovery in offices: Behavior-based assessment. *Journal of Environmental Psychology, 32*(2), 126–134. https://doi.org/10.1016/j.jenvp.2011.12.003

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences, 42*(5), 893–898. https://doi.org/10.1016/j.paid.2006.09.017

Steiger, J. H., & Lind, J. C. (1980). *Statistically-based tests for the number of common factors annual spring meeting of the psychometric society*. Iowa City.

Tan, H., & Xu, J. (2019). Differentiated effects of risk perception and causal attribution on public behavioral responses to air pollution: A segmentation analysis. *Journal of Environmental Psychology, 65*, Article 101335. https://doi.org/10.1016/j.jenvp.2019.101335

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item Response Theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*(1), 39–49. https://doi.org/10.1177/014662169501900105

van der Linden, S. (2014). On the relationship between personal experience, affect and risk perception:The case of climate change. *European Journal of Social Psychology, 44*(5), 430–440. https://doi.org/10.1002/ejsp.2008

van der Linden, S. (2015). The social-psychological determinants of climate change risk perceptions: Towards a comprehensive model. *Journal of Environmental Psychology, 41*, 112–124. https://doi.org/10.1016/j.jenvp.2014.11.012

Veldkamp, B. P. (2005). Optimal test construction. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 933–941). Elsevier. https://doi.org/10.1016/B0-12-369398-5/00447-3.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*(1), 58–79. https://doi.org/10.1037/1082-989x.12.1.58

Xie, B., Brewer, M. B., Hayes, B. K., McDonald, R. I., & Newell, B. R. (2019). Predicting climate change risk perception and willingness to act. *Journal of Environmental Psychology, 65*, Article 101331. https://doi.org/10.1016/j.jenvp.2019.101331

Xue, W., Hine, D. W., Loi, N. M., Thorsteinsson, E. B., & Phillips, W. J. (2014). Cultural worldviews and environmental risk perceptions: A meta-analysis. *Journal of Environmental Psychology, 40*, 249–258. https://doi.org/10.1016/j.jenvp.2014.07.002

Zhu, X., & Lu, C. (2017). Re-evaluation of the New Ecological Paradigm scale using item response theory. *Journal of Environmental Psychology, 54*, 79–90. https://doi.org/10.1016/j.jenvp.2017.10.005

Diamantopoulos, A., & Siguaw, J.A. (2000). Introducing LISREL. doi:10.4135/9781849209359.