MISS Measurement Instruments for the Social Sciences

PsychOpen GOLD

Validation of Measurement Instruments

Check for updates

Of Paper, Panels, and Pedagogy: A Multi-Sample Analysis to Establish Measurement Invariance of the Academic Risk-Taking Scale Across Questionnaire Modes and Academic Disciplines

Vanessa Hübner [1], Maximilian Pfost [1]

[1] *Department of Educational Research, University of Bamberg, Bamberg, Germany.*

Measurement Instruments for the Social Sciences, 2026, Vol. 8, Article e20721, https://doi.org/10.5964/miss.20721

Received: 2025-11-05 • **Accepted:** 2026-08-10 • **Published (VoR):** 2026-09-04

Handling Editor: Lena Roemer, Michigan State University, East Lansing, MI, USA

Corresponding Author: Vanessa Hübner, Department of Educational Research, University of Bamberg, Markusplatz 3, DE-96045 Bamberg, Germany. Telephone: +49 (0)951-863 2653. E-mail: vanessa.huebner@uni-bamberg.de

Supplementary Materials: Code, Data [see Index of Supplementary Materials]

Abstract

Academic risk-taking refers to students' engagement under uncertainty about the correctness of their contributions. It reflects a learning strategy of accepting potential errors to deepen understanding and is associated with higher learning success. The Academic Risk-Taking Scale measures such behaviors, but its validity depends on demonstrating measurement invariance across relevant groups. This study examined measurement invariance across administration modes, recruitment strategies, and academic disciplines using MG-CFA on data from N = 1,259 students. Strict measurement invariance was established for recruitment mode and academic disciplines, and partial measurement invariance, from the scalar level onward, for administration mode. This level of measurement invariance permitted meaningful latent mean comparisons, revealing group differences across administration modes, recruitment strategies, and academic disciplines. These results indicate that the scale can be reliably applied across these groups, supporting its use in diverse research contexts and providing a foundation for further investigations into patterns of academic risk-taking.

Keywords

academic risk-taking, measurement invariance, higher education

This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International License, CC BY 4.0, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Academic Risk-Taking and its Measurement

Academic risk-taking refers to a specific type of behavioral engagement in which students act under uncertainty about the correctness of their contributions (Clifford, 1991), the risk of making errors, and appearing less competent in front of others (Beghetto, 2009). Such uncertainty arises when the difficulty of a given task challenges or even slightly exceeds students' current abilities. For instance, students may contribute ideas to a class discussion on a challenging topic even when they are unsure whether their reasoning is correct. In such situations, assumptions may be confirmed and integrated into existing knowledge structures. However, the assumptions may also be incorrect. When errors occur and are subsequently addressed, they can foster learning and generate negative knowledge, i.e., knowledge about how something does not work (Oser & Spychiger, 1999). Academic risk-taking can thus be viewed as a learning strategy that entails accepting possible errors to deepen understanding (Krochmal & Roth, 2017). At the same time, its expression is inherently contingent on situational contexts and the stakes involved. Even students with a high disposition towards academic risk-taking may not display such behavior if the learning environment does not provide opportunities for it. Similarly, in high-stakes situations such as formal examinations, the potential costs of errors may outweigh potential learning benefits, making academic risk-taking less likely or less adaptive.

Various approaches have been developed to measure academic risk-taking. Clifford (1988) established the *Academic Risk-Taking Task*, which consists of task sets in mathematics, spelling, and vocabulary with predetermined and visible difficulties. The respondents select a certain number of tasks, and academic risk-taking is operationalized based on the difficulty level of the chosen tasks. Later, Tan et al. (2017) adapted this approach for university students, again operationalizing academic risk-taking through task choice. However, both approaches can be criticized for ignoring students' true ability levels, meaning that identical choices may reflect different subjective risks (Bran & Vaidis, 2020). Furthermore, two self-report scales have been proposed. The *School Failure Tolerance Scale* (Clifford, 1988) consists of three subscales which assess emotional reactions to errors, attraction to difficult tasks, and behavioral responses to difficulty. It therefore shows a strong focus on errors that have already occurred, rather than the anticipation and acceptance of potential errors that is central to the conceptualization of academic risk-taking. The *Intellectual Risk-Taking Scale* (Beghetto, 2009) better aligns with current conceptualizations of academic risk-taking, as its six items capture behavior types under uncertainty, which may result in error. However, the instrument is confined to science classroom contexts and does not capture potential multidimensionality that can arise from different social learning contexts. In summary, existing measures have advanced the understanding of academic risk-taking but remain limited in scope, context sensitivity, and theoretical alignment.

PsychOpen GOLD

Against this background, the *Academic Risk-Taking Scale* (Hübner & Pfost, 2023) was developed for use in higher education settings. Ten items assess behaviors reflecting academic risk-taking as conceptualized in prior theoretical work (Beghetto, 2009), such as asking critical questions, proposing hypotheses, and contributing to discussions on challenging topics despite uncertainty. Recognizing that students' behavior varies by social context (Lund Dean & Jolly, 2012), the instrument adopts a multidimensional structure, distinguishing between (1) common seminar situations involving both instructors and peers, and (2) those involving peers only. In addition, the scale can be applied to both general study contexts and specific seminar situations, thereby capturing both a relatively stable trait component and a context-dependent state component. The instrument has shown consistent and strong psychometric properties in several samples such as a stable two-factor structure with good model fit, and high internal consistency (see Hübner, 2025). Moreover, measurement invariance has been established across gender, socioeconomic status (Hübner & Pfost, 2025), and degree level (Hübner & Pfost, 2024b), supporting the instrument's comparability across key demographic groups.

Evidence for *construct validity* has also been reported. Initial analyses (Hübner & Pfost, 2023) showed moderate correlations between the seminar group and peer dimensions (r = .24 and r = .27), indicating that they represent related but distinct aspects of academic risk-taking. Furthermore, the correlations between the state and trait component (r = .76 and r = .77) suggest that while academic risk-taking reflects a relatively stable tendency, it may also fluctuate due to situational conditions. Consistent with this interpretation, considerable variation in students' academic risk-taking across different seminars has been observed (Hübner & Pfost, 2024a), indicating that contextual characteristics of learning environments may shape the expression of this behavior. Regarding *discriminant validity* Hübner and Pfost (2023) found moderate correlations between students' general engagement and the seminar group dimension of academic risk-taking (r = .31 and r = .43), suggesting that while related, academic risk-taking and academic engagement are not identical. Moreover, both dimensions of academic risk-taking show consistent negative associations with performance-avoidant goal orientation (r = -.17 to r = -.53) or the desire to hide one's own shortcomings from others (Hübner & Pfost, 2022, 2023, 2024a). Similar patterns have been reported in U.S. samples using alternative measures of academic risk-taking (Abercrombie et al., 2022), supporting the theoretical expectation that performance-avoidant goal orientation is detrimental for academic risk-taking, but not the same construct. Finally, *predictive validity* is indicated by positive associations between academic risk-taking and learning success (Hübner & Pfost, 2025).

The present study extends this work by examining measurement invariance across different recruitment and administration modes as well as academic disciplines.

Measurement Invariance and Mean Differences

Testing measurement invariance is a core step in psychometric validation, especially when comparing latent constructs across groups. If an instrument exhibits the same latent structure across groups, observed mean differences can be interpreted as reflecting true variation in the construct rather than systematic measurement bias. For methodological grouping variables, such as administration or recruitment mode, systematic response biases may be an issue, leading to artifactual variance. Regarding these variables, researchers typically do not expect true mean differences in the latent construct but rather seek to ensure that any observed differences are not driven by measurement artifacts, particularly when combining data from multiple sources or comparing results across studies. Concerning substantive grouping variables such as academic discipline, differences in the latent construct are often theoretically expected and of primary interest. Still, such interpretations require that the measurement model operates equivalently across groups. Establishing measurement invariance is therefore a prerequisite for distinguishing true group mean differences from potential measurement bias, even in cases where such differences are not explicitly anticipated.

Administration Mode

Examining measurement invariance across different administration modes is crucial, as it has become common practice in psychological research to jointly analyze data from paper-and-pencil and online questionnaires to combine their specific advantages (Martinez-Gomez et al., 2017). However, the mode of administration may elicit systematic response differences that may result in measurement non-invariance. While digital literacy effects are unlikely among university students (Lynch, 2022), the presence of researchers during paper-based administration may increase social desirability bias (Vereecken & Maes, 2006). Social desirability may lead to systematic differences in item responses across modes, potentially biasing results when data from different administration formats are combined or when findings are compared across studies using different modes. Overall, empirical evidence points to robust psychometric stability of many scales across administration modes, such as the Beck Depression Inventory and State-Trait Anxiety Inventory (Alfonsson et al., 2014), measures of well-being (Zager Kocjan et al., 2022), transformational leadership (Cole et al., 2006), the persuasiveness of health messages (Lewis et al., 2009), and the SCHNAPP spelling test (Schöfl et al., 2025). In contrast, Carini et al. (2003) found that college experience scales varied by administration mode. Given these mixed findings, measurement equivalence between paper-based and online formats should be empirically scrutinized.

Recruitment Mode

Recruitment via access panels has become a widespread alternative to self-recruitment. When individual researchers and research groups follow a self-recruitment strategy of sampling (often labeled convenience sampling), they typically invite their study participants directly. For example, they may distribute questionnaires among students at their own university via mailing lists, courses, and other institutional channels, or they may draw on the support of closer personal social networks, such as colleagues. Participation is usually voluntary and may or may not involve incentives such as course credit. When recruiting via an access panel, researchers provide an online questionnaire and sampling criteria to an access panel provider, which recruits participants from its existing pool. Participants recruited through such access panels typically differ from self-recruited participants in two main respects. First, they receive a monetary incentive for participation, which demonstrably increases response rates (Edwards, 2005; Robb et al., 2017) but does not seem to impact data quality (Dykema et al., 2024; Stanley et al., 2020). Second, access panel members are often highly experienced in completing questionnaires, which may lead to fatigue effects and satisficing strategies such as straightlining, which increase with panel tenure (Schonlau & Toepoel, 2015). In turn, such response patterns may introduce artifactual differences in the measurement model, thereby biasing results when data from different recruitment modes are combined or compared. Because inattentive responding is a well-documented issue in access panel research, various quality control procedures are implemented in these contexts. Approaches include the use of attention-check items, minimum response time thresholds, and methods for straightlining detection (Ward & Meade, 2023). In contrast, self-recruited samples are generally assumed to exhibit higher intrinsic motivation and attentiveness. Therefore, such procedures are applied less frequently. Due to an increasing practice of using access panel providers in psychology and education research, it is particularly important to empirically test whether scales perform equivalently in self-recruited and access panel samples. To our knowledge, however, no studies have yet systematically examined measurement invariance between these two recruitment modes.

Academic Disciplines

In contrast to administration and recruitment mode, true mean differences in academic risk-taking between students from different disciplines are conceptually significant, as disciplines vary considerably in their didactic approaches, teaching traditions, and disciplinary cultures (Abercrombie et al., 2022). For instance, and as opposed to other disciplines, psychology and law students often have distinct experiences with project-based learning formats (Schaeper, 2008). Furthermore, Scharlau and Huber (2019) report that a highly structured, instructional learning environment predominates in chemistry, whereas in the humanities, open discussion and reflection formats are more common. Such contextual variation is likely to result in true differences in the latent construct. Empirical work supports this view as environments emphasizing memorization and reproduction tend to reduce academic risk-taking (Hübner & Pfost, 2024a), whereas those promoting autonomy and exploration enhance it (Dachner et al., 2017). To investigate and interpret mean differences between disciplines in a theoretically sound manner, however, it is essential that the underlying measurement instrument operates invariantly across academic disciplines. Establishing measurement invariance is therefore essential to evaluate whether observed mean differences reflect variation in academic risk-taking rather than potential measurement artifacts, even in the absence of strong a priori expectations of bias.

Research Aims

The aim of the present study is to examine measurement invariance of the Academic Risk-Taking Scale across administration modes (paper-based vs. computer-based), recruitment modes (self-recruited vs. access panel), and four disciplinary groups (humanities, social sciences, cultural sciences, and STEM). If measurement invariance is supported, group differences in academic risk-taking can be interpreted as reflecting true differences in the underlying construct rather than artifacts. Consequently, latent mean differences between the groups are further explored.

Method

Sample

The sample consisted of five subsamples collected over a four-year period between 2021 and 2025. Sample 1 (n = 159, female = 72.3%, Md_{age} = 24 years) was collected at the authors' university between May and July 2021 with 116 students responding to an electronic questionnaire, and 43 students responding to a paper-pencil version. During this period, German universities were largely closed due to the COVID-19 pandemic and teaching took place almost exclusively online. Sample 2 (n = 381, female = 70.6%, Md_{age} = 22 years) was obtained at the same university between November and December 2021 and again between May and July 2022. Data was collected exclusively via paper-and-pencil questionnaires. Most universities had returned to in-person teaching at that time, and the survey was administered during face-to-face seminars towards the end of a seminar session. Sample 3 (n = 245, female = 64.9%, Md_{age} = 24 years) was collected between April and June 2024 using a mixed recruitment strategy. 76 students from the authors' university were recruited via academic mailing lists. An additional 37 students were recruited through the platform SurveyCircle. These participants were excluded from analyses because they could not be clearly assigned to either the self-recruited or access panel group used in the measurement invariance analyses. In addition, 132 students were recruited through the access panel provider Bilendi. All students responded to an electronic questionnaire. Sample 4 (n = 170, female = 71.2%, Md_{age} = 20 years) was obtained at the authors' university during several face-to-face seminars in the summer semester 2024 and the winter semester 2024/2025 using an electronic questionnaire administered during seminar sessions. Finally, sample 5 (n = 341, female = 62.1%, Md_{age} = 24 years) was collected between May and July 2025. 64 students were recruited at the authors' university using an electronic questionnaire, while 277 students were recruited through the access panel provider Bilendi using a computer-based survey. The comparatively high proportion of women in Samples 1, 2 and 4 likely reflects the university's strong focus on programs in the humanities and social sciences.

Overall, data was used from N = 1,259 German university students, 409 of which were recruited through the access panel provider Bilendi. The remaining 850 participants were recruited by the authors, of which 424 participants completed a paper-based questionnaire, while 426 participated in a computer-based online survey.

Sample characteristics are summarized in Table 1. Most participants identified as female. The self-recruited sample included a higher proportion of younger students, whereas the access panel sample was older on average. There was substantial variation in academic discipline across the sample groups. In the self-recruited samples, programs in educational sciences, psychology, and teacher education (summarized within the category humanities) predominated. The access panel sample displayed a broader distribution, with higher proportions of students from economics, natural sciences, and engineering (social sciences, STEM).

Table 1

Sample Characteristics

	Self-recruited		**Access panel**
	PBQ ***n* = 424**	**CBQ** ***n* = 426**	**CBQ** ***n* = 409**
Gender			
Female	72.2%	72.1%	58.2%
Male	25.2%	22.8%	40.6%
Diverse	1.2%	1.6%	1.0%
Age			
≤ 20	29.8%	36.9%	10.0%
21-24	47.9%	36.3%	35.2%
≥ 25	21.4%	22.6%	54.8%
Academic disciplines			
Humanities	54.7%	84.5%	22.3%
Social sciences	20.7%	6.1%	21.5%
Cultural sciences	13.0%	4.3%	13.2%
STEM	5.9%	0.5%	30.1%
Other fields of study	2.1%	0.4%	9.3%

Note. Percentages are column-wise for the three criteria gender, age, and academic discipline. PBQ = paper-based questionnaire, CBQ = computer-based questionnaire.

Instruments

Academic Risk-Taking

We used the Academic Risk-Taking Scale by Hübner and Pfost (2023), which consists of ten items and captures two dimensions. The seminar group dimension focuses on behaviors displayed in the presence of both peers and instructors (e.g., "To participate in seminar discussions even on difficult topics"). The peer dimension targets behaviors that occur exclusively in the presence of fellow students (e.g., "To ask fellow students to proofread my written work, even though I am unsure about its quality"). In the current study, only the general version of the scale, which does not focus on one specific seminar but relies on the students' studies in general, was used (instruction: "Please estimate how likely the following behaviors are for you in your studies in general"). Items are rated on a five-point Likert scale (1 = *very unlikely*, 5 = *very likely*). For the total sample, the internal consistency was $\alpha = .85$ for the seminar group dimension, and $\alpha = .77$ for the peer dimension.

Academic Discipline

Individual study programs were grouped into broader disciplinary categories, to ensure sufficient cell sizes for the analyses and to reflect similarities in disciplinary orientation and learning contexts. The grouping was based on thematic proximity and distinguishes between fields characterized by similar thematic areas, theoretical and epistemological traditions as well as similar teaching practices. Programs in educational science, psychology, and teacher education were classified as humanities (n = 697). Social sciences (n = 214) included business administration, economics, business education, sociology, political science, and communication studies. Cultural sciences (n = 130) comprised linguistics, languages and literary studies, theology, religious studies, philosophy, history, archaeology, heritage studies, art, and related disciplines. Finally, the STEM category (n = 154) encompassed the corresponding degree programs in the natural, technical, and computer sciences. Degree programs that could not be clearly assigned to one of these groups (e.g., law) were classified as other (n = 51); due to the small number of cases and a lack of conceptual coherence, this category was excluded from analyses when observing measurement invariance between disciplines.

Analysis Strategy

Analyses were conducted using *R Version 4.5.1* (R Core Team, 2025)[1]. First, a latent measurement model with two correlated factors of academic risk-taking was specified using *lavaan Version* 0.6-19 (Rosseel, 2012). For model identification, the factor loadings of the

1) A minimal dataset and codebook as well as the analysis script are available on PsychArchives (see Hübner & Pfost, 2026a, 2026b)

first item in each group were fixed to one. The methodology for testing measurement invariance involves estimating the latent measurement model separately for each group and introducing model constraints in a stepwise process, each step allowing for different conclusions (Rohrer & Paulewicz, 2025; van de Schoot et al., 2012; Vandenberg & Lance, 2000): In the *configural invariance model*, the same factorial structure is specified for all groups, but no equality constraints are imposed. In the next step, the *metric invariance model* constrains factor loadings to be equal across groups. Establishing metric invariance is often considered a crucial threshold as it allows meaningful comparisons of relationships between the latent variable and other variables, such as correlations or regression coefficients. The *scalar invariance model* additionally constrains item intercepts to be equal across groups. Item intercepts represent the expected agreement to an item when the latent construct is held constant and thus, scalar invariance implies that individuals with the same level of the latent trait are expected to provide equivalent mean agreement across groups. By contrast, a violation of scalar invariance indicates that group membership is associated with systematic differences in mean level agreement to an item beyond differences in the latent trait itself. If scalar invariance holds, observed group differences in item responses can therefore be interpreted as reflecting differences in the underlying latent construct, making comparisons of latent means across groups possible. Finally, the *strict invariance model* imposes equality constraints on factor loadings, intercepts, and residual variances. When this level of invariance is achieved, comparisons of latent variances and covariances across groups are considered meaningful. In practice, full measurement invariance at a given level is not always achieved. In such cases, *partial measurement invariance* may be established by freeing the parameters that differ across groups while retaining invariance for the remaining parameters. For example, partial scalar invariance indicates that most items have equivalent intercepts across groups, whereas a limited number of items show group-specific differences in the intercept despite respondents exhibiting the same level on the latent trait. Provided that the construct is measured by multiple indicators and that most parameters remain invariant, meaningful comparisons of structural relations or latent means may still be possible (Byrne et al., 1989; Putnick & Bornstein, 2016).

Measurement invariance can be assumed if the χ^2 difference between a less restrictive and a more restrictive model is not statistically significant (Pendergast et al., 2017). However, because this criterion is highly sensitive in large samples and may lead to over-rejection (Marsh et al., 2004), we also relied on established cutoff values and considered measurement invariance to be supported if changes in RMSEA were < 0.015, in CFI < 0.01, and in SRMR < 0.03 (Putnick & Bornstein, 2016). To test measurement invariance across administration modes, the paper-based self-recruited sample was compared with the computer-based self-recruited sample. For recruitment mode, the computer-based access panel sample was compared with the computer-based self-recruited sample. For academic discipline, comparisons were conducted across the four disciplinary groups of humanities, social sciences, cultural sciences, and STEM.

Latent means were explored when (partial) scalar measurement invariance was established. For each mean comparison, a reference group was defined with its latent mean fixed to 0. The latent means of the comparison groups were then estimated relative to this reference. Negative estimates indicate that the comparison group has lower values than the reference group, whereas positive estimates indicate higher values. In addition, fully standardized latent mean differences (*d*) are reported, which can be interpreted analogously to effect sizes.

Given the very low number of missing item responses, missing data was handled using listwise deletion. The number of available observations is indicated for each individual model.

Results

Measurement Invariance

The model fit indices for the measurement invariance tests across administration modes are presented in Table 2. The structural model of academic risk-taking showed an excellent fit to the data (N = 815, df = 68, χ^2 = 122.256, p < .001, RMSEA = 0.037, TLI = 0.983, CFI = 0.987, SRMR = 0.040). Comparing the configural model with the metric model revealed no significant changes in model fit indices. The transition from the metric to the scalar model showed a significant change in the χ^2 value ($\Delta\chi^2$ = 47.976, Δdf = 8, p < .001) and a slight exceedance of the predefined CFI cutoff (ΔCFI = 0.01). Closer inspection indicated that two item intercepts ("To express an opinion that differs from that of the majority during seminar discussions", and "To engage in seminar discussions actively, even if I feel that the content is above my level of competence") varied significantly across groups. By releasing these parameters, an adjusted model was obtained that did not differ significantly from the metric model. Since partial scalar invariance holds, latent means for students completing the paper-based and computer-based questionnaires can be meaningfully compared.

Next, we tested measurement invariance across recruitment modes. The structural model again demonstrated very good fit (N = 821, df = 68, χ^2 = 163.493, p < .001, RMSEA = 0.046, TLI = 0.982, CFI = 0.986, SRMR = 0.044). The comparison between the configural and metric models showed no significant changes in model fit. Between the metric and scalar models, the χ^2 value changed significantly ($\Delta\chi^2$ = 22.873, Δdf = 8, p = .001); however, changes in the further fit indices were well below the predefined cutoff values. Thus, latent mean comparisons can be considered meaningful. No significant changes in model fit were observed between the scalar and strict invariance models.

PsychOpen GOLD

Table 2

Model Fit Indices for the Measurement Invariance Tests

	χ^2	*df*	*p*	CFI	TLI	RMSEA	SRMR
Administration Mode							
Configural model	122.256	68	< .001	0.987	0.983	0.037	0.040
Metric model	119.868	76	.001	0.988	0.986	0.033	0.041
Scalar model	167.844	84	< .001	0.978	0.976	0.044	0.049
Adjusted scalar model	129.042	82	.001	0.988	0.986	0.033	0.042
Adjusted strict model	138.276	92	.001	0.987	0.988	0.032	0.045
Recruitment Mode							
Configural model	163.493	68	< .001	0.986	0.982	0.046	0.044
Metric model	147.499	76	< .001	0.988	0.986	0.041	0.044
Scalar model	170.372	84	< .001	0.986	0.985	0.042	0.047
Strict model	183.581	94	< .001	0.984	0.985	0.042	0.050
Academic disciplines							
Configural model	214.775	136	< .001	0.991	0.998	0.035	0.037
Metric model	209.314	160	.005	0.993	0.992	0.030	0.041
Scalar model	249.784	184	.001	0.990	0.991	0.032	0.044
Strict model	271.344	214	.005	0.991	0.992	0.028	0.046

Note. Scaled χ^2 values and *p*-values, and robust values for CFI, TLI und RMSEA are displayed. For administration mode, two item intercepts were allowed to differ across groups in the adjusted scalar/strict model (partial measurement invariance). $N_{\text{administration mode}} = 815$, $N_{\text{recruitment mode}} = 821$, $N_{\text{academic disciplines}} = 1{,}178$

Lastly, measurement invariance analyses across academic disciplines were conducted. The structural model showed excellent fit ($N = 1{,}178$, $df = 136$, $\chi^2 = 214.775$, $p < .001$, RMSEA = 0.035, TLI = 0.988, CFI = 0.991, SRMR = 0.037). No significant changes in model fit were observed between the configural and metric models. Between the metric and scalar models, the χ^2 value changed significantly ($\Delta\chi^2 = 40.470$, $\Delta df = 8$, $p = .002$), but changes in the other fit indices remained below the cutoff values. Therefore, comparing latent means across academic disciplines is possible. No significant changes in model fit were observed between the scalar and strict models.

Latent Mean Comparisons

For the comparison between the self-recruited paper-based and self-recruited computer-based surveys, the partially invariant scalar model was used. Compared with students who completed the paper-based questionnaire, students who completed the computer-based questionnaire took fewer academic risks on the seminar group dimension (ΔM =

https://doi.org/10.5964/miss.20721

-0.252, $p < .001$, $d = -0.320$). Regarding the peer dimension, the two groups did not differ significantly ($\Delta M = 0.091$, $p = .093$, $d = 0.131$).

With respect to recruitment mode, participants recruited via the access panel exhibited less academic risk-taking compared with self-recruited students, both on the seminar group dimension ($\Delta M = -0.172$, $p = .006$, $d = -0.193$) and on the peer dimension ($\Delta M = -0.172$, $p = .005$, $d = -0.226$)[2].

Concerning latent mean differences of academic risk-taking on the seminar group dimension, students in the humanities displayed higher values than students in STEM fields ($\Delta M = 0.239$, $p = .003$, $d = 0.318$). Furthermore, students of cultural sciences showed more academic risk-taking on the seminar group dimension than students in STEM ($\Delta M = 0.230$, $p = .045$, $d = 0.267$). Regarding the peer dimension, students in the humanities engaged in more academic risk-taking in the presence of peers compared with students of social sciences ($\Delta M = 0.220$, $p = .001$, $d = 0.314$) as well as compared to students of cultural sciences ($\Delta M = 0.232$, $p = .006$, $d = 0.332$). Further significant mean differences between disciplines were not observed.

Discussion

Interpretation of Results

This study examined measurement invariance of the Academic Risk-Taking Scale across two administration modes, two recruitment modes, and four academic disciplines, and compared their latent mean differences.

For administration mode, metric measurement invariance was established, indicating that associations between academic risk-taking and other variables can be meaningfully compared across paper-based and computer-based formats. Partial measurement invariance was observed for the scalar and the strict measurement invariance models. The change in CFI slightly exceeded the predefined cutoff. Specifically, the intercepts of two items describing particularly demanding forms of academic risk-taking in seminar settings, namely expressing a dissenting opinion (higher in the computer-based sample) and

2) Additional post-hoc analyses were conducted to explore whether these differences are carried by academic disciplines, as the access panel sample does include a substantial amount of STEM students, while the self-recruited sample predominantly consists of students from the humanities and cultural sciences. Restricting the analysis to students from the humanities ($n = 360$ self-recruited students vs. $n = 91$ access panel participants) revealed that latent means were descriptively slightly lower in the access panel group compared to the self-recruited group. This pattern was observed for both the seminar-group dimension ($\Delta M = -0.068$, $p = .559$, $d = -0.071$), and the peer dimension ($\Delta M = -0.124$, $p = .184$, $d = -0.179$), although neither difference reached statistical significance. This result suggests that disciplinary composition may account in parts for the originally observed recruitment-mode differences. However, it must be noted that the groups in the additional analysis are highly imbalanced and especially the smaller size of the access panel subgroup results in reduced statistical power and larger standard errors, limiting the ability to detect small effects.

actively participating despite feeling insufficiently competent (higher in the paper-based sample), varied across groups. This indicates that, at the same level of latent academic risk-taking, students differed systematically in their endorsement of these items depending on administration mode. One possible explanation is that students responded to the questionnaire in different social contexts. The anonymity or perceived distance associated with online questionnaires may reduce concerns about social evaluation, making behaviors involving disagreement with others, such as expressing a dissenting opinion, appear more acceptable. Conversely, paper-and-pencil administration conducted in the presence of researchers may increase situational awareness and reflection on one's own competence, potentially affecting responses to items involving feelings of inadequacy.

Given the overall excellent model fit in a large sample of over 800 observations, it is unclear whether this deviation is of practical relevance. Such minor fluctuations are typical in large samples and may not indicate substantive differences. Relatedly, recent methodological work emphasizes that, beyond binary decisions based on cutoff criteria, it is informative to examine the magnitude of non-invariance. Approaches proposed by Gunn et al. (2020) and Nye et al. (2019) quantify the extent to which non-invariant parameters affect latent mean differences or other substantive conclusions. Such evaluations help to distinguish between trivial deviations and practically meaningful bias, thereby providing a more nuanced interpretation of invariance results. While these analyses were beyond the scope of the present paper, it may be an interesting avenue for future research. Against this background, observed latent mean differences between administration modes should be interpreted as indicative rather than definitive. Specifically, we observed participants completing the paper-based questionnaire to report higher academic risk-taking in peer-related situations than those in the online survey, an effect that was weak to moderate in size. This difference may reflect motivational and social selection effects since the paper-based questionnaires were distributed in seminar and campus settings, and course attendance is related to variables such as conscientiousness and engagement (Credé et al., 2010), which may cause the observed differences.

For recruitment mode, strict measurement invariance was observed between self-recruited and access panel samples. This aligns with prior findings, which have shown that quality of response behavior was unaffected by monetary incentives (Dykema et al., 2024; Stanley et al., 2020). Further exploration revealed that access panel participants showed lower academic risk-taking on both the seminar-group and peer dimension, again with weak to moderate effect sizes. On the one hand, this may be due to disciplinary differences as the access panel sample contained more students from STEM fields, where risk-avoidant learning behavior tends to be more prevalent. Additional exploratory analyses restricted to only students of the humanities (see footnote 1) further strengthens this interpretation. On the other hand, previous research has shown that access panel samples and self-recruited samples can differ substantially across various participant characteristics beyond recruitment mode itself (Douglas et al., 2023; Necka et

https://doi.org/10.5964/miss.20721

al., 2016). Therefore, additional sources of sample heterogeneity, such as differences in participants' background characteristics or participants' interest in the specific topic of the survey cannot be entirely ruled out.

Finally, strict measurement invariance was observed across academic disciplines. Furthermore, students in STEM reported lower academic risk-taking in seminar-group contexts than students in humanities and cultural sciences with weak to moderate effect sizes. This pattern suggests that disciplinary learning environments may shape how students engage with uncertainty. In line with previous assumptions (Schaeper, 2008; Scharlau & Huber, 2019), more structured and instructor-centered formats common in STEM education may provide fewer opportunities for students to take intellectual risks in group settings. For the peer dimension, students in humanities reported higher academic risk-taking than students in social sciences, as well as students in cultural sciences with weak to moderate effect sizes. One possible explanation is that the humanities sample included psychology students, whose programs are often characterized by comparatively high workload and performance demands, which are factors that have previously been associated with more academic risk-taking in peer contexts (Hübner & Pfost, 2024a). Previous research has highlighted the importance of peer networks in higher education, but has largely focused on samples that were rather homogeneous with respect to academic discipline (e.g., Brouwer et al., 2022). Although the differences observed in our sample were subtle, they are not trivial and our results therefore extend this previous research, suggesting that academic discipline may be considered as an additional contextual factor when investigating how students interact with their peers.

Limitations and Implications for Future Research

Aggregating individual study programs into broader disciplinary groups was necessary to ensure sufficiently large group sizes. However, this approach also results in a loss of variance and may obscure further fine-grained discipline-specific differences. Furthermore, it would be desirable to acquire a larger sample and test whether the scale measures academic risk-taking invariantly at the level of individual study programs. Given the well-established link between academic risk-taking and learning outcomes (Hübner & Pfost, 2025; Özbay & Köksal, 2021; Varışoğlu & Ekinci Çelikpazu, 2019), this desideratum goes along with questions on how learning cultures shape the relationship between academic risk-taking and learning outcomes. The conditions under which academic risk-taking enhances learning and when other forms of engagement may be more adaptive (Perels et al., 2020; Schmitz & Wiese, 2006) still remain to be examined, particularly in rule-bound and precision-oriented fields such as engineering or the natural sciences.

Further, it should be noted that the present study was conducted in the German-speaking context and used the German version of the Academic Risk-Taking Scale. As higher education systems differ in their instructional traditions, classroom interaction patterns, and expectations regarding student participation, these contextual factors may

PsychOpen GOLD

influence opportunities for and expressions of academic risk-taking (Hecht & Kahrens, 2021). Consequently, the findings cannot be assumed to generalize directly to other national or institutional contexts.

Finally, regarding analyses of mean differences across groups, we cannot rule out that the alpha-level may be inflated due to multiple testing. Since our analyses were exploratory in nature, and adjustment procedures such as Bonferroni correction increase the likelihood of type II errors while making statistical conclusions highly dependent on the number of tests included in the correction, we decided against such adjustment procedures.

Conclusion

To obtain reliable results in latent group comparisons, researchers must employ instruments that demonstrate measurement invariance for the specific comparison at hand. The present study provides substantial evidence for the psychometric robustness of the Academic Risk-Taking Scale, particularly through the demonstration of strict measurement invariance across recruitment modes and academic disciplines, as well as partial measurement invariance across administration modes. These findings enhance confidence in the scale and suggest that it can be reliably used in diverse contexts of higher education research.

Funding: This research was supported by grants from the Ständige Kommission für Forschung und wissenschaftlichen Nachwuchs (FNK) of the University of Bamberg, Germany.

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

Ethics Statement: Ethical review and approval were not required for the study on human participants in accordance with the local legislation and institutional requirements. All individuals provided informed consent prior to participating in the study.

Use of AI Tools: During the preparation of this work the authors used ChatGPT-4o and DeepL to improve readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Data Availability: For this article, data is freely available (see Hübner & Pfost, 2026a).

Supplementary Materials

For this article, the following Supplementary Materials are available:

- Dataset and codebook (see Hübner & Pfost, 2026a)
- Analysis script (see Hübner & Pfost, 2026b)

Index of Supplementary Materials

Hübner, V., & Pfost, M. (2026a). *Dataset for: Of paper, panels, and pedagogy: A multi-sample analysis to establish measure-ment invariance of the Academic Risk-Taking Scale across questionnaire modes and academic disciplines* [Data, codebook]. PsychArchives. https://doi.org/10.23668/psycharchives.22379

Hübner, V., & Pfost, M. (2026b). *Code for: Of paper, panels, and pedagogy: A multi-sample analysis to establish measure-ment invariance of the Academic Risk-Taking Scale across questionnaire modes and academic disciplines* [Analysis script]. PsychArchives. https://doi.org/10.23668/psycharchives.22380

References

Abercrombie, S., Bang, H., & Vaughan, A. (2022). Motivational and disciplinary differences in academic risk taking in higher education. *Educational Psychology, 42*(7), 895–912. https://doi.org/10.1080/01443410.2022.2076810

Alfonsson, S., Maathz, P., & Hursti, T. (2014). Interformat reliability of digital psychiatric self-report questionnaires: A systematic review. *Journal of Medical Internet Research, 16*(12), Article e268. https://doi.org/10.2196/jmir.3395

Beghetto, R. A. (2009). Correlates of intellectual risk taking in elementary school science. *Journal of Research in Science Teaching, 46*(2), 210–223. https://doi.org/10.1002/tea.20270

Bran, A., & Vaidis, D. C. (2020). Assessing risk-taking: What to measure and how to measure it. *Journal of Risk Research, 23*(4), 490–503. https://doi.org/10.1080/13669877.2019.1591489

Brouwer, J., De Matos Fernandes, C. A., Steglich, C. E. G., Jansen, E. P. W. A., Hofman, W. H. A., & Flache, A. (2022). The development of peer networks and academic performance in learning communities in higher education. *Learning and Instruction, 80*, Article 101603. https://doi.org/10.1016/j.learninstruc.2022.101603

Byrne, B., Shavelson, R., & 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

Carini, R., Hayek, J., Kuh, G., Kennedy, J., & Ouimet, J. (2003). College student responses to web and paper surveys: Does mode matter? *Research in Higher Education, 44*(1), 1–19. https://doi.org/10.1023/A:1021363527731

Clifford, M. M. (1988). Failure tolerance and academic risk-taking in ten- to twelve-year-old students. *The British Journal of Educational Psychology, 58*(1), 15–27. https://doi.org/10.1111/j.2044-8279.1988.tb00875.x

Clifford, M. M. (1991). Risk taking: Theoretical, empirical, and educational considerations. *Educational Psychologist, 26*(3–4), 263–297. https://doi.org/10.1080/00461520.1991.9653135

Cole, M. S., Bedeian, A. G., & Feild, H. S. (2006). The measurement equivalence of web-based and paper-and-pencil measures of transformational leadership: A multinational test. *Organizational Research Methods, 9*(3), 339–368. https://doi.org/10.1177/1094428106287434

Credé, M., Roch, S. G., & Kieszczynka, U. M. (2010). Class attendance in college: A meta-analytic review of the relationship of class attendance with grades and student characteristics. *Review of Educational Research, 80*(2), 272–295. https://doi.org/10.3102/0034654310362998

Dachner, A. M., Miguel, R. F., & Patena, R. A. (2017). Risky business: Understanding student intellectual risk taking in management education. *Journal of Management Education, 41*(3), 415–443. https://doi.org/10.1177/1052562917695775

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS One, 18*(3), Article e0279720. https://doi.org/10.1371/journal.pone.0279720

Dykema, J., Stevenson, J., Jones, C. P., & Day, B. (2024). Guaranteed incentives and prize drawings: Effects on participation, data quality, and costs in a web survey of college students on sensitive topics. *Social Science Computer Review, 42*(3), 777–793. https://doi.org/10.1177/08944393231189853

Edwards, P. (2005). Meta-analysis of randomised trials of monetary incentives and response to mailed questionnaires. *Journal of Epidemiology and Community Health, 59*(11), 987–999. https://doi.org/10.1136/jech.2005.034397

Gunn, H. J., Grimm, K. J., & Edwards, M. C. (2020). Evaluation of Six Effect Size Measures of Measurement Non-Invariance for Continuous Outcomes. *Structural Equation Modeling, 27*(4), 503–514. https://doi.org/10.1080/10705511.2019.1689507

Hecht, J., & Kahrens, M. (2021). How does national culture affect the teaching style in higher education and what are the implications for the student experience? *Tertiary Education and Management, 27*(2), 107–127. https://doi.org/10.1007/s11233-021-09068-2

Hübner, V. (2025). *Zwischen Fehlern und Lernpotenzial: Eine Untersuchung individueller und kontextueller Prädiktoren akademischer Risikobereitschaft Studierender* [Between errors and learning potential: An investigation of individual and contextual predictors of university students‘ academic risk-taking]. Otto-Friedrich-Universität Bamberg. https://doi.org/10.20378/irb-108354

Hübner, V., & Pfost, M. (2022). University students' beliefs about errors predict their willingness to take academic risks. *Frontiers in Education, 7*, Article 992067. https://doi.org/10.3389/feduc.2022.992067

Hübner, V., & Pfost, M. (2023). Operationalization of academic risk-taking in university students. *Journal for Educational Research Online, 15*(1), 74–94. https://doi.org/10.31244/jero.2023.01.04

Hübner, V., & Pfost, M. (2024a). Academic risk taking and teaching quality in higher education. *Learning and Instruction, 90*, Article 101877. https://doi.org/10.1016/j.learninstruc.2024.101877

Hübner, V., & Pfost, M. (2024b). *Supplementary material (measurement invariance) for: Leap, learn, earn: Exploring academic risk taking and learning success across gender and socioeconimic groups.* PsychArchives. https://doi.org/10.23668/PSYCHARCHIVES.15388

Hübner, V., & Pfost, M. (2025). Leap, learn, earn: Exploring academic risk taking and learning success across gender and socioeconomic groups. *Higher Education, 90*(1), 49–67. https://doi.org/10.1007/s10734-024-01307-w

Krochmal, A. R., & Roth, T. C. (2017). From comfort to confidence: Modeling science as a process of risk-taking in the classroon. In R. Kelty & B. A. Bunten (Eds.), *Risk-taking in higher education: The importance of negotiating intellectual challenge in the college classroom* (pp. 103–118). Rowman & Littlefield.

Lewis, I., Watson, B., & White, K. M. (2009). Internet versus paper-and-pencil survey methods in psychological experiments: Equivalence testing of participant responses to health-related messages. *Australian Journal of Psychology, 61*(2), 107–116. https://doi.org/10.1080/00049530802105865

Lund Dean, K., & Jolly, J. P. (2012). Student identity, disengagement, and learning. *Academy of Management Learning & Education, 11*(2), 228–243. https://doi.org/10.5465/amle.2009.0081

Lynch, S. (2022). Adapting paper-based tests for computer administration: Lessons learned from 30 years of mode effects studies in education. *Practical Assessment, Research & Evaluation, 27*(1), Article 22. https://doi.org/10.7275/pare.1317

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

Martinez-Gomez, M., Marin-Garcia, J. A., & Giraldo O'Meara, M. (2017). Testing invariance between web and paper students satisfaction surveys: A case study. *Intangible Capital, 13*(5), 879–901. https://doi.org/10.3926/ic.1049

Necka, E. A., Cacioppo, S., Norman, G. J., & Cacioppo, J. T. (2016). Measuring the prevalence of problematic respondent behaviors among MTurk, campus, and community participants. *PLoS One, 11*(6), Article e0157732. https://doi.org/10.1371/journal.pone.0157732

Nye, C. D., Bradburn, J., Olenick, J., Bialko, C., & Drasgow, F. (2019). How big are my effects? Examining the magnitude of effect sizes in studies of measurement equivalence. *Organizational Research Methods, 22*(3), 678–709. https://doi.org/10.1177/1094428118761122

Oser, F., & Spychiger, M. (1999). Lernen aus Fehlern: Zur Psychologie „negativen" Wissens [Learning from errors: On the psychology of "negative" knowledge]. In W. Althof & F. Oser (Eds.), *Fehlerwelten: Vom Fehlermachen und Lernen aus Fehlern; Beiträge und Nachträge zu einem interdisziplinären Symposium aus Anlaß des 60. Geburtstages von Fritz Oser* (pp. 11–41). Leske + Budrich.

Özbay, H. E. & Köksal, M. S. (2021). Middle school students' scientific epistemological beliefs, achievements in science and intellectual risk-taking. *Science & Education, 30*(5), 1233–1252. https://doi.org/10.1007/s11191-021-00217-y

Pendergast, L. L., Von Der Embse, N., Kilgus, S. P., & Eklund, K. R. (2017). Measurement equivalence: A non-technical primer on categorical multi-group confirmatory factor analysis in school psychology. *Journal of School Psychology, 60*, 65–82. https://doi.org/10.1016/j.jsp.2016.11.002

Perels, F., Dörrenbächer-Ulrich, L., Landmann, M., Otto, B., Schnick-Vollmer, K., & Schmitz, B. (2020). Selbstregulation und selbstreguliertes Lernen [Self-regulation and self-regulated learning]. In E. Wild & J. Möller (Eds.), *Pädagogische Psychologie* (3rd ed., pp. 45–66). Springer.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Robb, K. A., Gatting, L., & Wardle, J. (2017). What impact do questionnaire length and monetary incentives have on mailed health psychology survey response? *British Journal of Health Psychology, 22*(4), 671–685. https://doi.org/10.1111/bjhp.12239

R Core Team. (2025). *R: A language and environment for statistical computing* (Version 4.5.1) [Software]. R Foundation for Statistical Computing. https://www.R-project.org.

Rohrer, J. M., & Paulewicz, B. (2025). Rethinking measurement invariance causally. *Current Research in Ecological and Social Psychology, 9*, Article 100241. https://doi.org/10.1016/j.cresp.2025.100241

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Schaeper, H. (2008). Lehr-/Lernkulturen und Kompetenzentwicklung: Was Studierende lernen, wie Lehrende lehren und wie beides miteinander zusammenhängt [Teaching and learning cultures and competence development: What students learn, how teachers teach, and the interplay of both]. In K. Zimmermann, M. Kamphans, & S. Metz-Göckel (Eds.), *Perspektiven der Hochschulforschung* (pp. 197–214). VS Verl. für Sozialwissenschaften.

Scharlau, I., & Huber, L. (2019). Welche Rolle spielen Fachkulturen heute? Bericht einer Erkundungsstudie [What role do disciplinary cultures play today? Report of an exploratory study]. *Die Hochschullehre, 5*, 315–354.

Schmitz, B., & Wiese, B. S. (2006). New perspectives for the evaluation of training sessions in self-regulated learning: Time-series analyses of diary data. *Contemporary Educational Psychology, 31*(1), 64–96. https://doi.org/10.1016/j.cedpsych.2005.02.002

Schöfl, M., Steinmair, G., Zepnik, S., & Gnambs, T. (2025). Equivalence of computer- and paper-based administrations of the SCHNAPP spelling test in German for six-year-old children. *European Journal of Psychological Assessment, 41*(3), 183–190. https://doi.org/10.1027/1015-5759/a000793

Schonlau, M., & Toepoel, V. (2015). Straightlining in web survey panels over time. *Survey Research Methods, 9*(2), 125–137. https://doi.org/10.18148/SRM/2015.V9I2.6128

Stanley, M., Roycroft, J., Amaya, A., Dever, J. A., & Srivastav, A. (2020). The effectiveness of incentives on completion rates, data quality, and nonresponse bias in a probability-based internet panel survey. *Field Methods, 32*(2), 159–179. https://doi.org/10.1177/1525822X20901802

Tan, E. W. S., Lim, S. W. H., & Manalo, E. (2017). Global-local processing impacts academic risk taking. *The Quarterly Journal of Experimental Psychology, 70*(12), 2434–2444. https://doi.org/10.1080/17470218.2016.1240815

PsychOpen GOLD

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–70. https://doi.org/10.1177/109442810031002

Varışoğlu, B., & Ekinci Çelikpazu, E. (2019). Secondary school students' academic risk-taking levels in turkish lesson. *International Journal of Progressive Education, 15*(4), 241–258. https://doi.org/10.29329/ijpe.2019.203.18

Vereecken, C. A., & Maes, L. (2006). Comparison of a computer-administered and paper-and-pencil-administered questionnaire on health and lifestyle behaviors. *The Journal of Adolescent Health : Official Publication of the Society for Adolescent Medicine, 38*(4), 426–432. https://doi.org/10.1016/j.jadohealth.2004.10.010

Ward, M. K., & Meade, A. W. (2023). Dealing with careless responding in survey data: Prevention, Identification, and recommended best practices. *Annual Review of Psychology, 74*(1), 577–596. https://doi.org/10.1146/annurev-psych-040422-045007

Zager Kocjan, G., Lavtar, D., & Sočan, G. (2022). The effects of survey mode on self-reported psychological functioning: Measurement invariance and latent mean comparison across face-to-face and web modes. *Behavior Research Methods, 55*(3), 1226–1243. https://doi.org/10.3758/s13428-022-01867-8

PsychOpen GOLD is a publishing service by
Leibniz Institute for Psychology (ZPID), Germany.
www.leibniz-psychology.org