


Validation of two test anxiety scales for physics undergraduate courses through confirmatory factor analysis and Rasch analysis

Agostino Cioffi¹,¹ Silvia Galano¹,¹ Raffaella Passeggia²,² and Italo Testa^{1,*}

¹Department of Physics “E. Pancini”, University Federico II, Naples, Italy

²Department of Humanity Sciences, University Federico II, Naples, Italy

 (Received 28 October 2023; accepted 7 March 2024; published 11 April 2024)

The assessment of test anxiety has received increasing attention in educational research due to the potential negative effects of anxiety on student performance. Traditionally, test anxiety scales have been developed for mathematics, but few studies have focused on physics. In this study, we validated two test anxiety scales for undergraduate physics courses: the Test Anxiety Inventory for Physics (TAIP) and the Abbreviated Test Anxiety Inventory for Physics scale (ATAIP), which were adapted from existing instruments. A convenience sample of 361 engineering students enrolled in a first-semester introductory physics course participated in the study. Confirmatory factor analysis and Rasch analysis were used to establish the construct validity of both scales. Convergent validity for the TAIP scale was established by examining its correlation with a scale adapted from the math anxiety scale. Criterion-related validity for both TAIP and ATAIP was established by analyzing the relationship between students’ Rasch scores on the two scales and their performance on two conceptual tests. Finally, measurement invariance of TAIP and ATAIP scales was established using both multigroup and differential item functioning analyses to reliably investigate gender differences in the corresponding Rasch measures. The study confirms a robust four-factor structure of the TAIP. The four subscales, Worry, Emotionality, Interference, and Lack of Confidence, demonstrate good reliability (McDonald’s $\omega = 0.78, 0.86, 0.87, 87$, respectively). Rasch analysis also confirms that, for each subscale, the rating scale functioning was consistent with the item difficulty and person measures. The TAIP also demonstrates adequate convergent and criterion-related validity, as well as measurement invariance with respect to gender. The ATAIP also demonstrates good reliability (McDonald’s $\omega = 0.84$), a well-functioning rating scale, and sufficient criterion-related validity. Additionally, it exhibits measurement invariance with respect to gender. Overall, the study supports that both the TAIP and ATAIP scales are reliable instruments for measuring students’ test anxiety in an undergraduate physics course. Implications for physics instruction at the university introductory level are briefly discussed.

DOI: [10.1103/PhysRevPhysEducRes.20.010126](https://doi.org/10.1103/PhysRevPhysEducRes.20.010126)

I. INTRODUCTION

The role of emotions in the learning process has been increasingly documented in the science education literature since the 1980s and 1990s [1–3]. Researchers have specifically investigated the influence of emotions during science lessons [4] and examinations [5]. Many studies have shown that positive emotions can act both as direct precursors of performance [6,7] and as mediators of motivation [8]. Recent studies have also shown that positive emotions affect students’ engagement at university [9,10].

Conversely, negative emotions can impair academic performance and self-regulatory learning processes [11–13].

Among negative emotions, the role of anxiety has been extensively researched [14,15]. From a psychological perspective, the construct of anxiety is not intended as a pathological syndrome but refers to a set of manifestations of a stressful nature that may occur as a result of negative expectations and thoughts regarding a possible failure related to the specific tasks an individual is asked to perform [16]. Literature on anxiety has shown that people who suffer from it perceive themselves as unable to cope with challenges and experience continuous changes in emotional state, feelings of pain, communication, and relationship difficulties [17]. Moreover, anxiety may be a precursor of psychopathologies such as depression and distress [18], which may interfere with developmental and formative processes [19,20].

Research has shown that studying a particular discipline, such as math, physics, biology or chemistry can also be

*Corresponding author: italo.testa@unina.it

Published by the American Physical Society under the terms of the *Creative Commons Attribution 4.0 International license*. Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI.

challenging as students may experience forms of psychological distress and anxiety that, if not properly monitored, can lead to more severe forms of psychological syndromes [21–26]. Discipline-related anxiety refers to the anxiety and unpleasant feelings experienced during class or when dealing with a particular task in that discipline [27,28]. Thus, students' experience of discipline-related anxiety may be reflected in lower expectations in that discipline, lower motivation to pursue a career in that field, and a higher likelihood of dropping out of a university course related to that discipline [29,30]. For such reasons, research on anxiety has mainly been conducted with undergraduate students [31–34]. However, most studies have focused on math and science anxiety [35–38] and only a few recent studies have begun to investigate the role of anxiety on physics performance [30,39,40]. Remarkably, these studies have tended to focus on gender differences and their consequences for student performance, but with low or no attention to the theoretical conceptualization of anxiety, the consistency of the dimensions that should be assessed, and the psychometric properties of the measures used.

To address this issue, in this paper, we focus on a specific form of discipline-related anxiety called *test anxiety*. More specifically, we report on the validation of two test anxiety scales for physics using confirmatory factor analysis and Rasch analysis.

In the following section, we describe the theoretical framework adopted and review previous instruments for measuring test anxiety. Being the aim of this review also to familiarize the interested physics education researchers with existing conceptualizations of the test anxiety construct, we will limit it only to the most relevant work in the research field.

II. BACKGROUND

A. Conceptualizations of test anxiety

As noted above, negative emotions associated with anxiety include the fear of an inadequate performance [41] or a poor evaluation in an examination or specific task such as a test [42,43]. The term *test anxiety* refers to a set of personal emotional, physical, and behavioral responses to possible negative consequences or failure in examinations or other testing situations [44–46]. In particular, test anxiety may manifest as (i) a conscious feeling of anticipated fear that makes the subject unable to identify actual threats; (ii) a pattern of physiological arousal and tension experienced during a test situation; (iii) a lack of cognitive control and mental organization that makes difficult for the subject to cope with the testing situation [47,48]. As such, test anxiety is a situation-specific trait that can threaten the validity of a test, as students' performance may be not in line with their actual ability [49,50], especially in high-stakes tests [51]. Among the other factors that may affect test anxiety, literature in educational psychology has

identified gender, self-beliefs, personality traits, and task-related value [52]. For this study, we are more interested in the gender differences in test anxiety. Literature suggests that girls often put more pressure on themselves to achieve academic success and this may lead to a greater fear of the consequences of failing a test and hence to a greater test anxiety [53].

A traditional line of research conceptualizes test anxiety as a construct comprising the dimensions of “worry” and “emotionality” during or just before a test situation [54]. The worry dimension refers to the cognitive act of having negative thoughts about the possible consequences of failing the task or to the fear of not being able to understand and clearly think about the test. The emotionality dimension refers to the affective response to the task, such as the sense of nervousness, tension, rapid heartbeats, or cold sweat. While many studies suggest that test anxiety is significantly and negatively associated with test performance [55], self-regulated learning, and motivation [29,30,56], some authors have shown a differential influence of the two dimensions of test anxiety. In particular, the worry dimension appears to be more strongly associated with lower test scores and lower expectations than the emotionality dimension [57–59].

A four-dimensional conceptualization of test anxiety has been proposed in Germany by Hodapp [60]. This model adds to “worry” and “emotionality” two further dimensions, “interference,” which refers to the perceived distraction from the test by personal thoughts not relevant to the required task, and “lack of confidence,” which refers to the feeling of not being able to cope with the required task because of ones' perceived shortcomings. The latter two dimensions were found to be negatively correlated with both performance and self-efficacy in mathematics and language [56].

Finally, test anxiety has also been conceptualized as a construct that includes the following three dimensions: behavioral, cognitive, and physical [61–63]. In this tripartite model, the behavioral component refers to the subject's intention to postpone the exam or avoid studying for the task. The cognitive and physical components roughly correspond to the worry and emotionality dimensions of the previous model. However, in this model, the cognitive dimension also includes intrusive thoughts and worries about possible failure and social concerns, namely the fear of suffering a reduction in the subject's self-image in the eyes of relatives and peers as a consequence of poor performance in a test [47]. The physical dimension refers to both negative and positive features of anxiety, such as a faster heartbeat (negative) or a slight nervousness that improves performance (positive).

B. Existing instruments to measure test anxiety

The different conceptualizations of test anxiety have traditionally led to the use of different instruments to

measure the dimensions of the construct. Since anxiety had been historically considered psychopathology to be cured, earlier instruments were developed to help mental health professionals and behavioral researchers identify whether a subject was affected by anxious symptomatology when taking a test. For example, the items of the Suinn Test Anxiety Behavior Scale (STABS) [64] featured experiences that may cause fear or apprehension (see Table I). As such, the scale was unidimensional and intended as a means to assess if a subject should be assigned to a specific clinic treatment, as the desensitization [65]. Later, researchers in the field became more interested in measures of the psychological mechanisms underlying test anxiety. One of the first instruments to use the two-dimensional model of test anxiety described in the previous section is the *Worry and Emotionality Questionnaire* (WEQ) [54]. The WEQ has 10 items, equally divided between the Worry and Emotionality subscales. Although the WEQ is important because it was one of the first scales to be developed to measure the psychological mechanisms underlying test anxiety, it lacked sound psychometric validation of the two subscales. It also included other aspects of test anxiety in the worry dimension, such as lack of confidence.

Because of these limitations, the WEQ has been replaced over the years by the 20-item *Test Anxiety Inventory* (TAI, [66]), which also conforms to the two-dimensional conceptualization of test anxiety. The TAI has two subscales, one for the worry dimension and the other for the emotionality dimension. The reported reliability of the 20 items is good (see Table I), as well as its convergent validity and criterion-related validity [42]. A brief five-item version of the TAI was later developed [67].

Although psychometrically sound, the TAI has also been criticized because of the high correlation between the two subscales and the inclusion of items relating to intrusive and disturbing thoughts in the worry dimension [45]. To address these issues and improve the measure of different facets of the test anxiety construct, starting from the *Reactions to Test Scale* [68], Hoddapp *et al.* proposed a 20-item instrument based on the four-dimension model described above, the *German Test Anxiety Inventory* (in German: *Prüfungs Angst Fragebogen*, PAF) [60,69]. Exploratory and confirmatory factor analysis confirmed the factor structure of the PAF, with five items for each scale [70]. Subsequent studies provided also evidence for convergent and criterion-related validity [71]. The PAF has

TABLE I. Existing instruments to measure test anxiety. See text for the more details about each scale.

Name of the scale	Acronym	Items	Subscales (number of items)	Example items	Reported reliability	Reference
Suinn Test Anxiety Behavior Scale	STABS	50	None	<i>Having a test returned or waiting for a test to be handed out</i> <i>Seeing a test question and not being sure of the answer.</i>	0.74	[64]
Worry and Emotionality Questionnaire	WEQ	10	Worry (5)	<i>I do not feel very confident about my performance on this test</i>	0.79–0.88	[54]
Test Anxiety Inventory	TAI	20	Emotionality (5) Worry (8)	<i>I feel my heart beating fast</i> <i>During tests I find myself thinking about the consequences of failing</i>	0.80	[42,45,66–68]
			Emotionality (8)	<i>While taking the examinations I have an uneasy upset feeling</i>		
			No subscales (4)	<i>Thoughts of doing poorly interfere with my concentration on tests</i> <i>I feel confident and relaxed while taking tests</i>		
German Test Anxiety Inventory (in German: Prüfungs Angst Fragebogen)	PAF	20	Worry (5 items) Emotionality (5 items) Interference (5 items) Lack of Confidence (5 items)	<i>I think about what will happen if I don't do well.</i> <i>I feel anxious</i> <i>I am preoccupied by other thoughts that distract me.</i> <i>I am convinced that I will do well.</i>	0.72–0.83	[69–72]
FRIEDBEN Test Anxiety Scale	FTAS	23	Social Derogation (8) Cognitive Obstruction (9) Physiological Tenseness (6)	<i>If I fail a test I am afraid people will consider me worthless</i> <i>I feel I just can't make it in tests</i> <i>I am terribly scared of tests</i>	0.81–0.86	[47]

(Table continued)

TABLE I. (*Continued*)

Name of the scale	Acronym	Items	Subscales (number of items)	Example items	Reported reliability	Reference
Test Anxiety Measure	TAM	42	Worry (8)	<i>I continue to worry about a test though it is over</i>	0.80–0.94	[62,63]
			Physiological Hyperarousal (7)	<i>My heart pounds in my chest while I take a test</i>		
			Cognitive Interference (7)	<i>I have difficulty thinking clearly while I take a test</i>		
			Task Irrelevant Behaviours (6)	<i>I stay home on days when I am scheduled to take a test</i>		
			Social Concerns (7)	<i>I worry that my instructor will be upset with me if I do poorly on a test</i>		
			Facilitating Anxiety (7)	<i>I focus better on a test when I feel slightly anxious</i>		
French version of the Revised Test Anxiety	F-RTA	18	Worry (3)	<i>During tests I find myself thinking about the consequences of failing</i>	0.66–0.84	[73]
			Bodily Symptoms (3)	<i>Sometimes I find myself trembling before or during tests</i>		
			Test-irrelevant thinking (3)	<i>During tests I find I am distracted by thoughts of upcoming events</i>		
			Perceived control (5)	<i>During tests I believe in my ability to receive a good grade</i>		
			Tension (4)	<i>I worry a great deal before taking an important exam</i>		
Physics Anxiety Rating Scale	PARS	32	Physics course/exam anxiety	<i>I am usually very nervous when I study for a physics exam</i>	0.82	[74]
			Anxiety about lack of physics knowledge	<i>If my teacher asked me to explain a physical event from everyday life, I would be worried</i>		
			Mathematics anxiety	<i>When I open a physics book, I am afraid to see a page full of formulas without any explanation</i>		
			Physics laboratory anxiety	<i>I am very comfortable using laboratory materials</i>		

been translated into English and different languages, including Italian [72].

Almost in parallel to the TAI, the FRIEDBEN Test Anxiety Scale (FTAS) [47] was developed from the responses to open questions asking how a student who suffers from test anxiety behaves before, during, and after a test. The authors identify three dimensions, which roughly correspond to the tripartite model described above (see Table I). The first dimension refers to social fear or concern about the consequences of a poor test performance. The second dimension refers to worries related to being not able to succeed in the test. The third dimension refers to the physical discomfort perceived when taking the test. The three subscales have good reliability (Table I).

Two further instruments to measure test anxiety were more recently developed. The *Test Anxiety Measure* (TAM) for College Students [62] roughly refers to the three-dimensional model of test anxiety and features six subscales with good reliability (see Table I). The TAM subscales have also good convergent validity. Evidence of measurement invariance across gender and country has also been reported [63].

Another recently developed instrument is the French version of the *Revised Test Anxiety* (F-RTA) [73]. The 18-item instrument features five subscales, which roughly correspond to the model by Hoddap (see Table I). The authors report a robust factorial structure as well as the convergent and discriminant validity of the F-RTA.

C. Relationships between test and discipline anxiety

The scales developed to measure test anxiety tend to refer to “tests” in a generic way, i.e., the items are not formulated specifically for a particular discipline. As a result, discipline-specific test anxiety has traditionally been conceptualized as a subdimension of the more general construct of “subject anxiety.” Despite the focus on emotional and motivational aspects of science learning, most research has paid attention to the construct of math anxiety. Math anxiety refers to the feelings of tension, fear, and apprehension toward mathematics [35]. When students experience math anxiety, they are more likely to fail mathematics tasks and avoid mathematics courses as part of their higher education or career path [22,75–78]. In addition, math anxiety is associated with lower performance on mathematics tasks [35]. While highly correlated with each other (on average, $r \sim 0.70$) [77,79], math anxiety and test anxiety are considered distinct constructs, as the respective emotionality and worry dimensions differently correlate with math performance [80].

Several instruments have been developed to measure math anxiety, such as the *Mathematics Anxiety Scale* (MAS) [81] or the *Abbreviated Math Anxiety Scale* (AMAS) [82]. The latter scale is the most interesting for the present study because it has two factors, *anxiety about learning mathematics* and *anxiety about mathematics testing*. The first factor refers to anxiety about the process of learning mathematics (e.g., when listening to a lecture), and the second refers to the usual test anxiety specific to a mathematical task. The AMAS has good reliability for both factors (0.78 and 0.79, respectively) and has been translated into different languages, including Italian [35,36].

Despite the relevance of “discipline anxiety” in the learning process, very few studies have used a scale specifically related to physics. The *Physics Anxiety Rating Scale* (PARS) is a four-factor scale consisting of 32 items [74]. Although the authors report the results of an exploratory and confirmatory factor analysis and a good reliability of the instrument (see Table I), the scale lacks a solid theoretical justification for the four scales. Furthermore, the first factor does not distinguish between the worry and emotionality dimensions of test anxiety. Finally, the factors are highly correlated, in particular, the subscale *anxiety about lack of physics knowledge* has correlations with *mathematics anxiety* and *physics laboratory anxiety* that are greater than 0.80, suggesting that the scale lacks a strong factorial structure. Because of these shortcomings, the scale has not been used in subsequent studies.

A more recent study [39] investigated gender differences in test anxiety and self-efficacy in an introductory physics course. The authors used a four-item scale adapted from one of the subscales of the *Motivated Strategies for Learning questionnaire* [83]. However, this subscale does not distinguish between the dimensions of worry, emotionality, and interference. Furthermore, the exploratory factor analysis

reported in Ref. [84] shows that test anxiety, control of learning beliefs, task value, and intrinsic goal orientation load on the same factor. Thus, this subscale does not have sufficient psychometric strength to be used as a reliable measure of test anxiety.

III. AIMS OF THE STUDY

The literature reviewed so far on general test anxiety shows that several measures of test anxiety have been developed and validated. However, despite their usefulness, very few scales are available for specific disciplines such as physics.

The reasons for having a validated scale to measure physics anxiety in general, and physics test anxiety in particular, are manifold. First, test anxiety can be a precursor to the avoidance of undergraduate career paths or advanced secondary school courses where physics is a relevant subject [85], a pattern similar to the avoidance of mathematics [86]. Second, it may be relevant to examine test anxiety in introductory physics courses in science and engineering undergraduate programs, where physics exams are often offered as the first exams. Since anxiety is negatively related to exam performance, measuring test anxiety in these courses may be crucial for the early prevention of dropout [87]. Third, as test anxiety is negatively related to self-efficacy, which is a relevant variable in explaining gender differences in physics learning [88,89], it would be important to measure physics test anxiety adequately. Fourth, the scales that have been developed and used in previous studies in physics education research lack a sound theoretical background, as they do not reflect current conceptualizations of the multidimensional nature of test anxiety. Fifth, the used scales lack robust psychometric validation. In particular, these scales have never been validated with concurrent instruments or with statistical approaches that address typical measurement issues associated with the use of Likert scales. Furthermore, no study has yet analyzed the test anxiety scales using Rasch analysis. Finally, further research is needed to replicate the findings of previous test anxiety studies in a physics context or with diverse student populations, such as undergraduate students enrolled in an introductory physics course.

In an effort to address these issues, we selected the 20-item German Test Anxiety Inventory (PAF in German) and adapted them for use in an undergraduate calculus-based introductory physics course. The rationale for choosing the PAF instrument in the present study is threefold: (i) it has a strong theoretical background, targeting four different but important facets of the anxiety construct; (ii) it is a psychometrically robust instrument; (iii) it has been validated in numerous subsequent studies and, in particular, an Italian-language form is already available. However, although the PAF instrument has only 20 items, its administration may be demanding for an introductory university course, especially if it is recommended to be administered

before or immediately after an exam situation. Especially, in the latter situation, students' retention in answering the items may be higher with a longer scale. Moreover, the scoring of a multidimensional scale can be difficult to interpret. Finally, all previous studies in physics education have used short scales to measure test anxiety but they were not psychometrically validated. Therefore, we decided to also validate a short form of test anxiety for physics by adapting the five-item short form of the TAI. The reason for this choice was that it is a highly reliable instrument with items from the two main scales of the original TAI, worry and emotionality. Moreover, also the original TAI is available in the Italian language.

Consequently, we hypothesize that the adapted PAF for physics, which we will call the Test Anxiety Inventory for Physics (TAIP) has

- (H1) Adequate construct-related validity, namely it shows a robust four-factor structure;
- (H2) An adequate convergent validity, namely it correlates with an instrument that measures the same (or a very closely related) construct;
- (H3) Adequate criterion-related validity, namely it correlates with a variable that literature has shown to be correlated with the construct that is being measured;
- (H4) An adequate measurement invariance with respect to gender.

Similarly, for the Abbreviated form of TAI for Physics (ATAIP), we hypothesize that it shows:

- (H5) Adequate construct validity;
- (H6) Adequate criterion-related validity;
- (H7) Adequate measurement invariance with respect to gender.

The reason for testing the measurement invariance of the two instruments with respect to gender is that the literature has shown that test anxiety may differ significantly between women and men, so it is important that the measurement model of the two instruments does not introduce bias in the item formulation.

IV. METHODS

A. Sample and procedure

This study was carried out on a sample of undergraduate students enrolled in the biomedical engineering and computer science degree courses at a large State University in the South of Italy, who took an introductory calculus-based physics course in the first year, in the first semester. The first reason for such a choice is that physics is not only a fundamental subject for these courses, but it is also often considered one of the most difficult exams to pass in the first year, leading to feelings of anxiety and discomfort. The second reason is that during the semester of the physics course, the lecturers set an entrance test, which has only an informative role, and a written midterm exam, which consists of solving an open-ended problem (part 1, 5 pt) and a five-item multiple-choice questionnaire (part 2, 5 pt).

The two parts of the midterm exam are weighted equally in the evaluation, so a good performance in the multiple-choice questionnaire is necessary to pass the exam. Students who score at least 6 points have the possibility to access a shorter and easier form of the final written exam. For this reason, students consider their performance in this midterm exam to be very relevant to pass the exam. The structure of the midterm exam is known to the students in advance, as the instructors have to explain the exam procedures at the beginning of the course. For this study, the TAIP was submitted just after the entrance test in the same lecture. Note that a similar procedure was followed in the Italian validation of the PAF. Similarly, the ATAIP was submitted immediately after the end of the midterm exam, which took place 2 weeks after the submission of the TAIP.

The procedure to involve the sample was as follows: the last author of the study sent an email to the coordinators of the biomedical engineering and computer science courses, explaining the aims, objectives, data collection protocol, and return of results. The physics lecturers of the two courses were then contacted to ask for the participation of their students in the study. Once the number of lecturers who agreed to allow their students to participate in the study had been determined, the researchers involved in the study went to the classroom during one of the lectures in the first week of the semester and explained the purpose of the research and asked the students to cooperate by participating in the study. The researcher also explained the type of data that would be collected. The researcher carefully emphasized that the data collected would be used for research purposes in an anonymous form. Furthermore, he or she made sure that the instruments would be administered to all students, including those who did not participate in the study. In response to this concern, it was explained that the study would only use the responses of those who had given their consent for the data to be used for research purposes. In this way, it was not possible for the lecturers to identify those students who participated in the study and those who did not. The researcher then explained that at the end of the data collection, the answers would be analyzed only after being associated with an alphanumeric code and that from that point on, the data collected would be stored in an exclusively anonymous form and it would never be possible for the researchers to trace the author of a given set of answers. At the end of the meeting, the consent form for the use of the answers for research purposes and the privacy statement were distributed.

Students were asked about their gender at the end of the survey with the following item: "Please indicate your gender." The item had three options: *female*, *male*, *prefer not to say*. After inspection of the data, for the purpose of this study, we left out the students who ticked the *prefer not to say* option since they were less than 1% of the collected data. Overall, after removing also the respondents who left at least one blank response, the dataset featured a total number of 361 students (female students = 108). More specifically, 244

students (female students = 72) participated in the validation of the TAIP, while 117 students (female students = 37) participated in the validation of the ATAIP.

B. Instruments

The following instruments were used to collect evidence to support the study hypotheses.

- (a) The German Test Anxiety Inventory (PAF) consists of 20 items on a four-point Likert scale (1 = hardly ever; 2 = rarely; 3 = often; 4 = almost always). The scale consists of four subscales: Worry, composed of five items (e.g., *I think about how important the test is to me*), Emotionality, composed of five items (e.g., *I feel anxious*), Interference, which includes five items (e.g., *Suddenly thoughts cross my mind which inhibit me*) and lack of confidence, which consists of five reversed items (e.g., *I trust in my performance*). The Italian adaptation of the PAF is described in Ref. [72]. The authors report that the translation was carried out by two nonprofessional translators and finalized by the authors themselves. Their validation study was carried out with 326 high school Italian students (11–16 years old). Authors report adequate to good reliability values for the four scales: McDonald’s $\omega = 0.75$ (Worry); 0.85 (Emotionality); 0.80 (Interference); and 0.75 (Lack of Confidence). Gender invariance for the measurement model was also supported. The authors report that convergent validation evidence was collected by inspecting the correlation between the scores of the four subscales and the score in the Italian version of the Abbreviated Math Anxiety Scale subscales [36]. The authors report significant positive correlations, thus supporting the convergent validity of the Italian version of the PAF. In order to adapt the Italian version to physics without changing the items, we added at the beginning of the survey a brief statement: *Read the following statements and think about your experience with a PHYSICS test.*
- (b) The short form of the TAI [67] is a five-item scale, two for each dimension of the TAI with the addition of one of the items not included in the two subscales. The authors report that the brief scale shows good reliability (0.87) and good convergent validity. The translation of the items in Italian was provided by the authors in Ref. [72]. At the beginning of the survey, we added the statement: *Read the following statements and think about the PHYSICS exam that you have just taken.*
- (c) Abbreviated Anxiety Scale (AAS). This scale was adapted from the Italian version of the AMAS [35,36] specifically for this study. While the original AMAS features two subscales (see above), we used only the four-item evaluation anxiety scale, whose items measure the subject’s perception of discomfort about examination or homework (e.g., *“Having to solve many difficult problems for homework”*). For the purpose of

this study, at the beginning of the survey, we added the following statement: *Now imagine yourself in the situations described below for the PHYSICS course. Rate each situation in terms of how much anxiety you feel during the specified activities using the following scale: 1 = non-negative feeling; 2 = somewhat negative feeling; 3 = moderately anxious or nervous; 4 = negative feeling; 5 = very negative feeling.*

- (d) In order to assess the physics performance of the participating students, we used two instruments specifically developed for this study from the physics education literature [90,91]: (i) a 20-item concept inventory covering basic topics such as force and motion, kinematics, units of measure, vectors, Cartesian graphs, in multiple-choice format; (ii) a brief task with five multiple-choice items and an open-ended exercise, for which the students had to write down the solution approach, the calculation performed and the numerical results obtained. The 20-item concept inventory was used as an entrance test (without evaluation) for the physics course, while the short task was used as the written midterm exam.

In total, the students completed the TAIP and the AAS in about 5 min. To complete the ATAIP, students took on average about 2 min.

C. Data analysis

Construct-related validation evidence of the TAIP (H1) was tested using confirmative factor analysis (CFA). Normality was checked by calculating each item’s skewness and kurtosis. The goodness of the data fit was assessed though the typical indices used in the literature (Table II). We evaluated the internal consistency by calculating McDonald’s ω , namely, the ratio of true-score variance and total variance. We chose this index instead of the traditional Cronbach’s α since it takes into account the loadings of the factor model and its use is recommended for multidimensional measures with well-defined dimensions. Average variance extracted (AVE) and composite reliability (CR) were also calculated to inspect whether the CFA extracted the dominant factors that explained most of the variance of the items (Table II).

TABLE II. Recommended fit indices range for confirmatory factor analysis.

Index	Range	Reference
Chi-square to degrees of freedom ratio ($\chi^2/\text{d.o.f.}$)	<5	[92–98]
Root mean squared error of approximation (RMSEA)	<0.08	[92–98]
Tucker-Lewis index (TLI)	>0.95	[92–98]
Comparative fit index (CFI)	>0.95	[92–98]
Average variance extracted (AVE)	≥ 0.5	[99]
Composite reliability (CR)	>0.7	[99]

TABLE III. Recommended fit indices range for Rasch analysis.

Index	Acceptable range	Reference
Principal component analysis of residuals (PCAR) loading	<0.4	[100,101]
Person reliability	>0.5	[102]
Item separation	>3	[102]
Person separation	>2	[102]
Infit and outfit mean square (MNSQ)	0.7–1.4	[102]
Differential item functioning (DIF)	p of a DIF contrast <0.05 and DIF contrast < 0.43 = negligible 0.43 < DIF < 0.64 = moderate DIF > 0.64 = large	[100]

A value of AVE greater than 0.5 indicates that more than half of the variance of the latent construct is explained by the measurement model. In addition, if the square root of the AVE of each factor is lower than all the correlations with the other factor, the scale also has adequate discriminant validity.

Then, we used a 1D Rasch model to gather further construct-related validation evidence for the full TAIP scale (see Table III for an overview of the fit indices calculated for each subscale). We first assessed multidimensionality using principal component analysis of residuals (PCAR) [100,101]. A residual is the unexplained, i.e., not predicted by the Rasch model, variance in the data and is divided into components, called contrasts, that do not depend on item difficulty or student ability. The aim of PCA of the residuals is to falsify the hypothesis that the unexplained variance is at the noise level so that the eigenvalue of contrast can be interpreted as the number of items that share a common characteristic. Since at least two items should share a common feature to form a second dimension, a contrast must have an eigenvalue of at least two to be above the noise level. To check whether the possible secondary dimensions are related to the expected subscales of the TAIP, we checked the residuals plot, in which the items are identified by two coordinates, difficulty and the loading of the item with the first contrast, respectively. If the items have loadings greater than 0.4 (see Table III) and are separated from the others, then it is likely that there is a secondary dimension in the test.

Finally, probability curves were examined for each scale to investigate whether all rating steps (1, 2,..., 5) were actually used by respondents [102,103].

Convergent validation evidence (H2) was obtained by examining the correlations between Rasch person measures in each of the subscales of the TAIP and person measures in the AAS.

Criterion-related validation evidence of the TAIP (H3) was obtained by examining the correlation between Rasch person measures in each of the subscales and the Rasch measures in the 20-item concept inventory, which was used as an entrance test.

Finally, to examine the gender invariance of the measurement model of the TAIP (H4), we conducted

hierarchically nested CFAs using a multigroup analysis [104].

Measurement invariance is the ability of an instrument to measure a given construct consistently across different groups of respondents. We assessed invariance by examining whether variations in the χ^2 and in the comparative fit index (CFI) were significant when forcing measurement weights and structural covariance [105].

Gender invariance was also tested by Rasch analysis using differential item functioning (DIF) for each of the subscales of the TAIP [102]. DIF indicates when a group of respondents performs differently than expected on an item, given the overall ability of the groups and the difficulty of the items. To investigate DIF, we examined the contrast for each item, comparing the observed and expected abilities of the two groups (female students and male students). See Table III for the adopted ranges of DIF.

Similarly, construct-related validation evidence for the ATAIP (H5) was examined using CFA and Rasch analysis. Criterion-related validation evidence (H6) was collected by examining the correlation between Rasch person measures in the ATAIP and Rasch measures in the brief physics task used as a midterm exam. We also examined the gender invariance of the measurement model of the ATAIP (H7) using hierarchically nested CFAs and DIF analysis.

All the statistical analyses were carried out through the software SPSS v. 29. Confirmatory factor analyses were carried out through AMOS v. 29. Rasch analysis was carried out through Winsteps 3.91.

V. RESULTS

A. CFA of the TAIP

The initial analysis confirmed the four-factor structure for the TAIP. Fit indices were satisfactory: $\chi^2/\text{d.o.f.} = 1.582 (p < 0.001)$; Tucker-Lewis index (TLI) = 0.954; comparative fit index (CFI) = 0.962; root mean squared error of approximation (RMSEA) = 0.049. However, one item in the Worry factor (“*I think about how important the...*”) had a weak standardized regression weight (0.43),

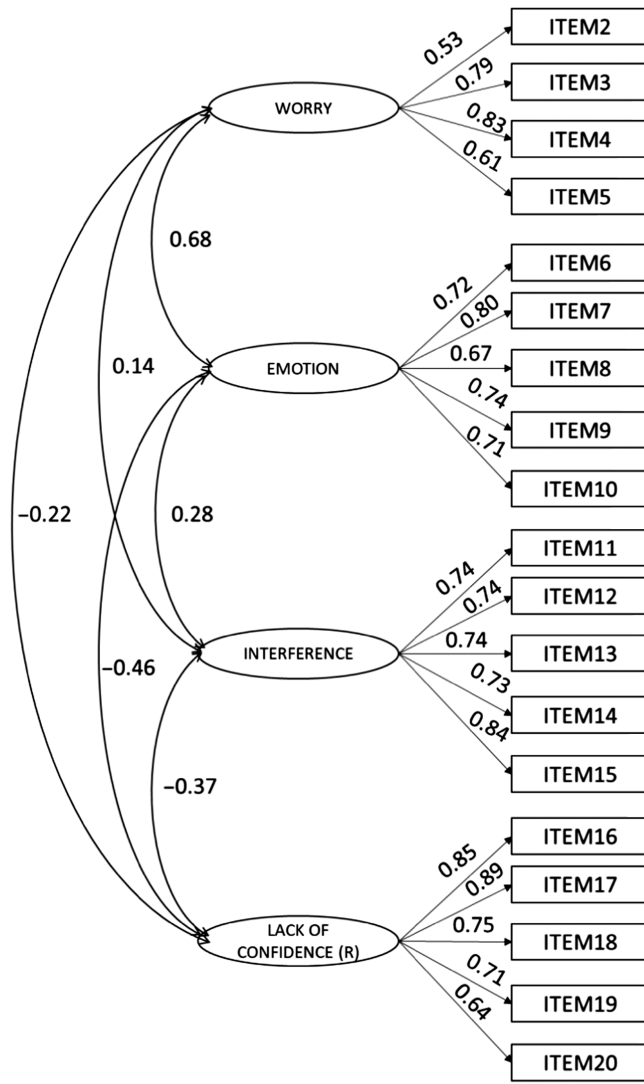


FIG. 1. CFA of the 19-item TAIP instrument.

so we decided to drop it, obtaining a slightly better fit, $\chi^2/\text{d.o.f.} = 1.492 (p < 0.001)$; TLI = 0.963; CFI = 0.971; RMSEA = 0.045. Figure 1 shows the regression weights and correlations between the four factors. Descriptive

statistics and correlations between the four factors, as well as McDonald’s ω , AVE, and CR for each of the four factors are shown in Table IV.

Note that the items on the Lack of Confidence scale are reversed, as in the Italian version of the PAF, in that they are formulated to measure subjects’ confidence in performance. Accordingly, all correlations between the Lack of Confidence factor and the other three factors are expected to be negative. All the values obtained are acceptable, confirming the validity of the four-factor structure. In particular, the square root of the AVE of each factor is lower than the correlation with the other factors, which also supports the discriminant validity of the scale.

The resulting final 19-item scale of TAIP is reported in the Appendix.

B. Rasch analysis of the TAIP

Rasch analysis was performed after the CFA on the remaining 19 items. PCA of the residuals confirmed the multidimensionality of the scale, as the eigenvalues of the first and second contrasts were 5.6792 and 3.8311, respectively. Looking at the residuals graph [Fig. 2(a)], we notice that five items are well separated from the others along the contrast loadings axis, with a coordinate much greater than 0.4, therefore, it is very likely that these five items form a secondary dimension. These items correspond to the Lack of Confidence scale, which therefore identify a secondary dimension for the TAIP. In order to check whether the remaining 14 items form different dimensions of the instrument, we then repeated the PCA of the residuals. The eigenvalues of the first and second contrasts were 4.2112 and 2.1281, respectively, confirming the multidimensionality of the 14-item scale. Looking at the graph of residuals [Fig. 2(b)], there are five items that are well separated from the others and all have a contrast loading of around 0.7. These items correspond to the Interference scale, which is therefore another secondary dimension of the TAIP. Finally, we repeated the PCA of residuals for the remaining nine items, obtaining in this case that only the eigenvalue of the first contrast is greater

TABLE IV. Descriptive statistics and correlations among the four TAIP-20 factors ($N = 244$). Pearson correlations between factorial scores: * $p < 0.05$; ** $p < 0.01$.

Factor	1	2	3	4
1. Worry (four Items)	0.78 ^a ; 0.49 ^b ; 0.79 ^c			
2. Emotionality (five items)	0.678*	0.86 ^a ; 0.54 ^b ; 0.85 ^c		
3. Interference (five items)	0.136	0.277*	0.87 ^a ; 0.58 ^b ; 0.87 ^c	
4. Lack of Confidence (R) (five items)	-0.216*	-0.460*	-0.374**	0.87 ^a ; 0.60 ^b ; 0.88 ^c
Mean (SD)	3.02 (0.60)	2.05 (0.72)	2.12 (0.70)	2.28 (0.64)
Kurtosis	-0.246	0.603	-0.093	-0.578
Asymmetry	-0.230	-0.272	0.624	0.081

^aMcDonald’s ω .

^bAverage variance extracted (AVE).

^cComposite reliability.

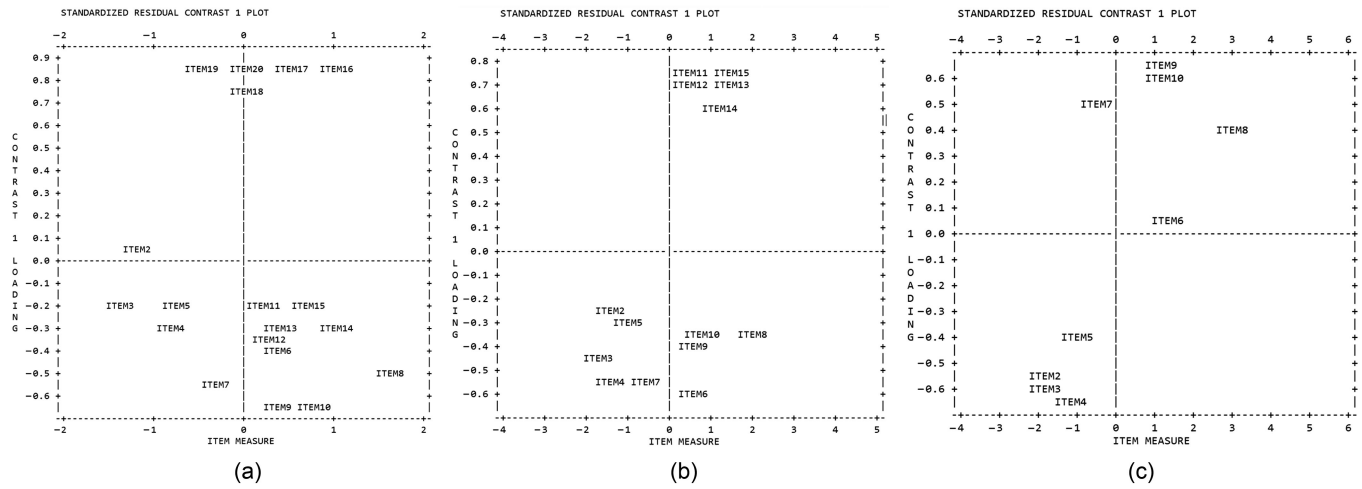


FIG. 2. PCA of residual plots for the TAIP subscales: (a) whole instrument; (b) Worry, Emotion, and Interference factors; (c) Worry and Emotion factors.

than 2 (2.4290). Looking at the residuals graph [Fig. 2(c)], we see two clearly identifiable clusters of items that roughly correspond to the Worry (bottom of the graph) and Emotionality (top of the graph) scales, with the sole exception of item 6 in the middle of the graph. However, as

one item cannot form a cluster, we can safely say that Worry and Emotionality form two distinct scales.

We then carried out a Rasch analysis for each of the four subscales identified. The results are shown in Tables V and VI. The probability curves for each of the scales are

TABLE V. Rasch measures for each scale of the TAIP ($N = 244$).

Scale	Average person measure (logit)	Eigenvalue of first contrast	Person reliability	Item separation	Person separation
Worry	1.93	1.7136	0.78	4.00	1.87
Emotionality	-1.50	1.9110	0.84	10.67	2.31
Interference	-1.28	1.7070	0.86	4.93	2.45
Lack of Confidence (R)	-0.96	1.8675	0.88	3.93	2.75

TABLE VI. Fit statistics for the adapted TAIP ($N = 244$).

Dimension	Item	Estimate	Standard error	Point-measure correlation	Infit MNSQ	ZSTD	Outfit MNSQ	ZSTD
Worry	Item 2	-0.45	0.13	0.72	1.06	0.7	1.07	0.7
	Item 3	-0.61	0.13	0.80	0.81	-2.2	0.78	-2.4
	Item 4	0.41	0.13	0.83	0.77	-2.8	0.77	-2.7
	Item 5	0.66	0.13	0.75	1.33	3.3	1.33	3.3
	Item 6	-0.27	0.12	0.77	1.25	2.6	1.21	2.1
Emotionality	Item 7	-1.79	0.12	0.84	0.85	-1.6	0.82	-1.7
	Item 8	2.40	0.15	0.70	1.12	1.1	0.87	-0.6
	Item 9	-0.10	0.12	0.83	0.93	-0.7	0.89	-1.2
	Item 10	-0.24	0.12	0.81	0.91	-0.9	0.89	-1.1
Interference	Item 11	-0.63	0.13	0.81	1.07	0.7	1.03	0.40
	Item 12	-0.49	0.13	0.83	0.89	-1.2	0.86	-1.5
	Item 13	0.04	0.13	0.79	0.98	-0.2	0.95	-0.6
	Item 14	1.18	0.13	0.74	1.40	3.9	1.49	4.1
	Item 15	-0.10	0.13	0.85	0.67	-4.1	0.66	-4.1
Lack of Confidence (R)	Item 16	0.89	0.14	0.80	0.92	-0.8	0.92	-0.7
	Item 17	0.35	0.14	0.84	0.87	-1.5	0.85	-1.5
	Item 18	-0.33	0.14	0.76	1.43	4.2	1.39	3.4
	Item 19	-0.78	0.14	0.86	0.75	-2.9	0.73	-2.9
	Item 20	-0.24	0.14	0.80	0.95	-0.6	0.93	-0.6

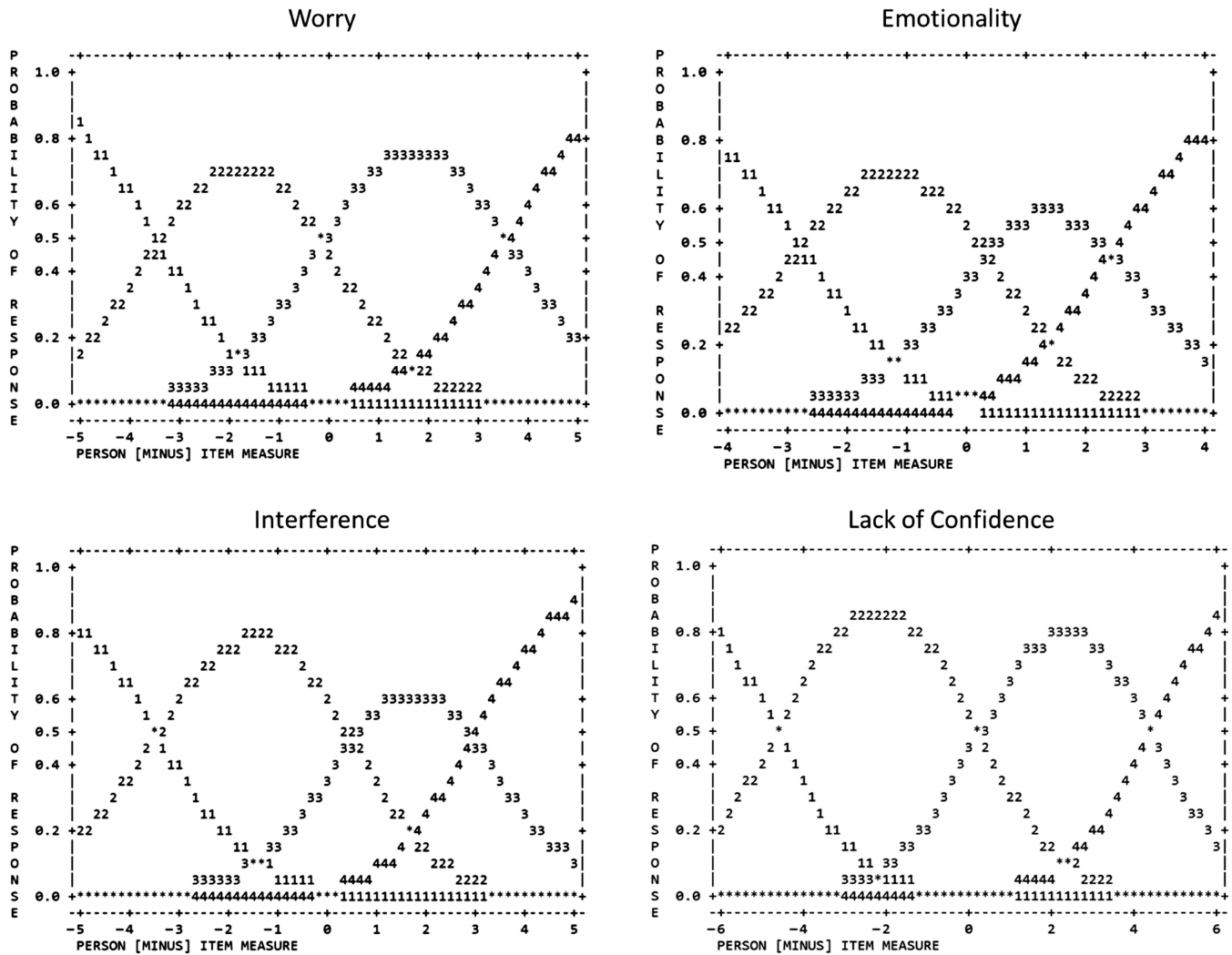


FIG. 3. Probability plots for the four scales of the TAIP.

shown in Fig. 3. All four scales have acceptable values for person reliability, item separation, and person separation, with the exception of the Worry subscale, which has a person separation value slightly lower than 2. However, the scale has acceptable values for person reliability and item separation. Regarding the item fit statistics (Table VI), all four subscales show acceptable values of expected variability, thus confirming the four-dimensional construct validity of the adapted PAF. The adapted four-point rating scale is also consistent with the item difficulty and person measures, as the step estimates increase monotonically, and each peak has a probability greater than 0.5.

C. Evidence for convergent and criterion-related validity of the TAIP

To collect convergent validation evidence, we first performed a one-factor CFA of the AAS. We found that the one-dimensional model was a mediocre fit for the data: $\chi^2/d.o.f. = 2.867(p = 0.090)$; TLI = 0.967; CFI = 0.994;

RMSEA = 0.088. When inspecting scale reliability, we found that the McDonald’s ω was 0.78. However, removing the fourth item (“Being asked an oral question from the seat”) would have raised the value to 0.82. So, we removed this item and performed a Rasch analysis with the three remaining items of the instrument. The analysis confirmed the unidimensionality of this three-item instrument as the eigenvalue of the first contrast was 1.9184. Person reliability was 0.83, item separation was 5.61, and person separation was 2.22. The average person measure was -0.35 logit. The three items had infit and outfit mean square (MNSQ) greater than 0.70 and lower than 1.50, with a point-measure correlation between 0.78 and 0.88. Then, we calculated Pearson correlations between the Rasch measure of the four scales of the TAIP and the three-item AAS. All four correlations were significant at $p < 0.01$: 0.368 (Worry); 0.497 (Emotionality); 0.318 (Interference); -0.413 (Lack of Confidence reversed).

To collect criterion-related validation evidence, we first performed a Rasch analysis of the 20-item concept

TABLE VII. Multigroup analysis of the TAIP instrument for the gender variable.

Model	$\Delta\chi^2$	d.o.f.	p	ΔCFI
Measurement weights	11.036	15	0.750	0.002
Structural covariances	16.150	25	0.910	0.004

inventory. PCA of residuals showed that the instrument is unidimensional (eigenvalue of the first contrast = 1.8588). Person reliability was 0.70, item separation was 8.95, and person separation was 1.52, which can be considered acceptable values for a concept inventory. The average person measure was -0.54 logit, which means that the questionnaire was slightly difficult for the students. Only one item (Q9) had an outfit MNSQ greater than 1.50 (2.48) but acceptable infit MNSQ (1.05). However, this item was also very difficult (measure = $+2.23$ logit), so some unexpected variations can be considered acceptable. For such reason, we decided to keep the item. Pearson correlations between the Rasch measures of the concept inventory and those of the four TAIP scales are all negative as expected and significant at $p < 0.05$, except for the reversed Lack of Confidence scale: -0.164 (Worry);

-0.133 (Emotionality); -0.151 (Interference); 0.106 (Lack of Confidence reversed).

D. Gender invariance of the TAIP

Table VII reports the χ^2 and CFI variations when constraining the TAIP items' factor loadings (measurement weights model) and when constraining factors' covariances (structural covariances), respectively. We obtained nonsignificant $\Delta\chi^2$ and differences in CFI values less than 0.01. Such evidence supports the measurement invariance of the TAIP instrument for the gender variable. The results of the DIF analysis (Fig. 4) confirmed gender invariance for all four subscales. Only two items in the Emotionality scale show statistically significant DIF ($p < 0.05$): “*I feel somewhat overwhelmed*” and “*I feel upset.*” However, being the DIF contrast less than 0.64, these items show moderate DIF, so we decided to keep these items in the final scale.

Once established the measurement invariance of the instrument for the gender variable, we calculated the difference between the average Rasch measures of female and male students for the four TAIP scales using the t statistics. Results are reported in Table VIII. Female students are significantly more likely than male students to agree on the Worry, Emotionality, and Interference items.

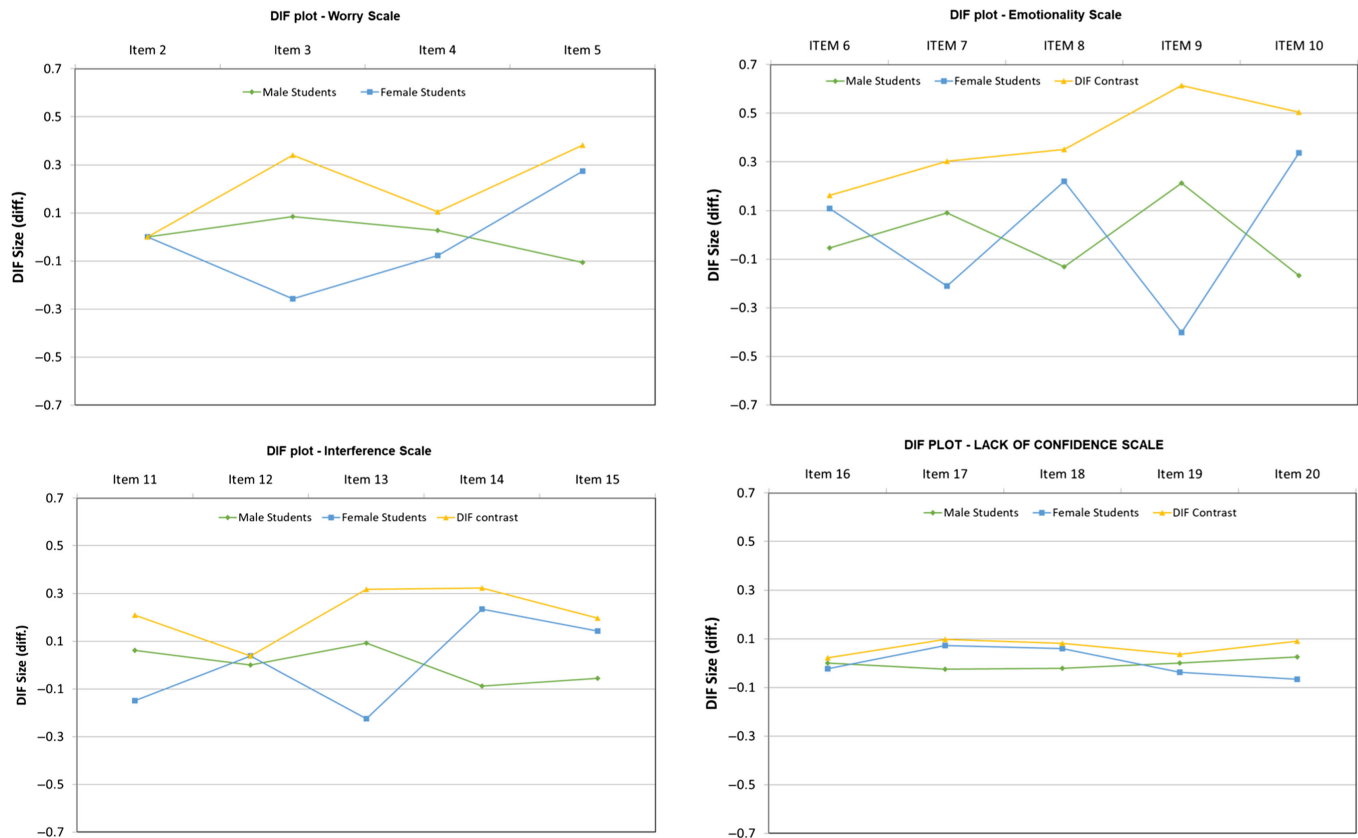


FIG. 4. DIF analysis of the four subscales of the TAIP. DIF contrast ranges: <0.43 = negligible; $(0.43; 0.64)$ = moderate; >0.64 = large.

TABLE VIII. Average Rasch measures (logit) for each of the four TAIP subscales and for gender.

Scale	Females (<i>N</i> = 72)	Males (<i>N</i> = 172)	<i>p</i>	Cohen's <i>d</i>
Worry	2.77	1.58	<0.001	0.52
Emotionality	-0.26	-2.02	<0.001	0.73
Interference	-1.15	-1.33	0.611	0.07
Lack of Confidence (R)	-1.80	-0.61	<0.01	0.40

Conversely, for the Lack of Confidence scale, which is in reversed form, female students are less likely to agree, namely they feel less confident than male peers in their performance.

E. CFA of the ATAIP

The initial fit of the five adapted items was not adequate: $\chi^2/\text{d.o.f.} = 3.352 (p < 0.05)$; TLI = 0.950; CFI = 0.975; RMSEA = 0.108. Analysis suggested adding residual covariates between item 1 and item 5, so we decided to parcel these two items. The new fit was acceptable: $\chi^2/\text{d.o.f.} = 1.655 (p = 0.157)$; TLI = 0.958; CFI = 0.986; RMSEA = 0.075. The average raw score of the brief scale was 2.49 (SD = 0.66), with a good reliability, McDonald's $\omega = 0.84$.

F. Rasch analysis of The ATAIP

Rasch's analysis confirmed the construct-related validity of the four-item ATAIP. Person reliability was 0.83, item separation was 5.22, and person separation was 2.24. PCA of residuals confirmed the unidimensionality of the scale since the eigenvalue of the first contrast was 1.6266. The average person measure was -0.01 logit (SD = 2.55 logit), indicating that on average, the students in this sample show significantly different levels of agreement on the four items of the scale (see Table IX). We note also that all four items of the ATAIP had acceptable fit statistics (see Table IX). Finally, the adopted four-point rating scale is also consistent with the item difficulty and person measures, as shown by the probability curves reported in Fig. 5.

TABLE IX. Fit statistics for the four ATAIP items (*N* = 117).

Dimension	Item	Estimate	Standard error	Point-measure				
				correlation	INFIT MNSQ	ZSTD	OUTFIT MNSQ	ZSTD
Test anxiety	Item 2	0.73	0.18	.71	1.29	2.1	1.29	2.0
	Item 3	-0.43	0.18	.85	0.90	-0.8	0.92	-0.6
	Item 4	1.07	0.18	.82	0.90	-0.7	0.85	-1.0
	Parcel items 1/5	-1.37	0.18	.80	0.85	-1.1	0.94	-0.3

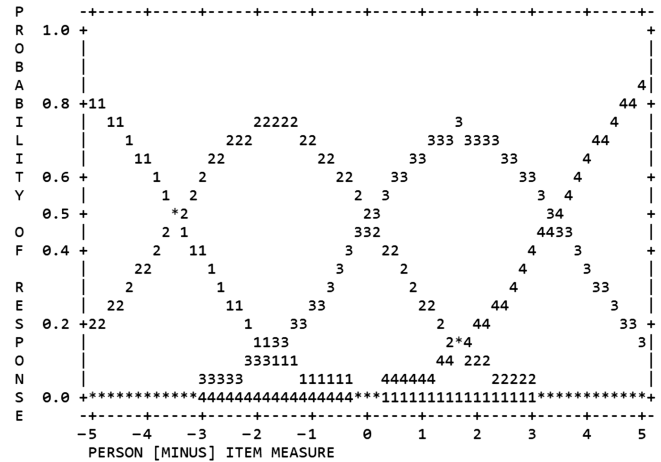


FIG. 5. Probability plot for the ATAIP.

G. Evidence for criterion-validity of ATAIP

First, we conducted a Rasch analysis of students' responses to the short task of the midterm exam, which consisted of five multiple-choice items and an open-ended question. The multiple-choice items were scored dichotomously, while the open question was scored on a partial credit scale from 1 to 5. Using Rasch analysis, we could reliably combine the scores of the two parts of the midterm exam task. Overall, this task was unidimensional (eigenvalue of the first contrast was 1.8556). Person reliability was 0.53, item separation was 7.99, and person separation was 1.06, which can be considered acceptable given that the midterm exam task consisted of a few items. Note that the average person measure was -3.13 logit, which means that the task was very difficult for the students. The correlation between the Rasch measures of the ATAIP and those of the short task was negative, -0.128, but not significant ($p = 0.169$).

H. Gender invariance of ATAIP

Gender invariance was tested for the four-item ATAIP with the parcel between items 1 and 5. We report in Table X the χ^2 and CFI variations when constraining items' loadings (measurement weights model) and items residuals (measurement residuals), since for this model, we had only one factor. We obtained nonsignificant $\Delta\chi^2$ and differences in CFI values less than or equal to 0.01. Such evidence

TABLE X. Multigroup analysis of the ATAIP for the gender variable.

Model	$\Delta\chi^2$	d.o.f.	p	ΔCFI
Measurement weights	4.981	4	0.173	0.010
Measurement residuals	7.485	8	0.485	0.004

supports the gender invariance of the instrument. Results of DIF analysis (Fig. 6) show that only one item (item 2: “*I wish the examination did not bother me so much*”) has substantial DIF, contrast = 1.22, $p < 0.01$, namely male (female) students agree more (less) than expected. Interestingly, this item was not assigned to any of the two scales in the original TAI. Hence, taking also into account that this was the only item to show DIF in the abbreviated scale, we decided to maintain it for this study.

We also note that, although female students on average agree more than male students on the four ATAIP items, namely, females have higher Rasch scores than males (+0.5762 vs -0.2910 logit), such difference is not significant, $t = 1.716$, d.o.f. = 115, $p = 0.089$, and the effect size is moderate (Cohen’s $d = 0.34$).

Given these results and the detected DIF for item 2, we conducted a differential group functioning (DGF) analysis for each ATAIP item using Rasch person and item measures.

Briefly, DGF combines DIF and differential person functioning (DPF), which provides the difficulty measure of a single item independently of the measures of the other items for two or more groups of subjects. In our case, we have two groups of students, namely female and male students. The results are shown in Table XI, where we report the item measure for female and male students, as well as the detected contrast, which can be interpreted similarly to the DIF contrast, and the associated probability. The DGF analysis shows that, as expected, item 2 has a large and significant DGF contrast, while item 4 and the

TABLE XI. Differential group functioning analysis of the ATAIP.

Item	Females ($N = 37$)	Males ($N = 80$)	Contrast	p
Item 2 ^a	−0.21	0.10	−1.18	0.03
Item 3	−0.03	0.01	−0.17	0.65
Item 4	0.12	−0.05	0.67	0.09
Parcel items 1/5	0.13	−0.06	0.77	0.06

^aItem showing differential item functioning.

parcel of items 1 and 5 have moderate but not statistically significant DGF contrasts.

VI. DISCUSSION

In this study, we investigated the psychometric properties of two instruments to measure physics-related test anxiety, which has been recognized as a factor that can negatively affect students’ performance in an examination, with potential negative effects on long-term learning processes [31,52,58,106]. We focused on test anxiety in the context of an undergraduate physics course to fill the gap in the literature related to the lack of a psychometrically sound instrument for use in physics examinations. To this end, we adapted the Italian translation of the German Test Anxiety Inventory (PAF, [69]), a self-report instrument that measures four dimensions of anxiety: worry, negative emotions, interference, namely intrusive thoughts during test situations, and lack of confidence in performance.

We also validated an abbreviated scale, based on the Text Anxiety Inventory, to be used with students in test situations where there is limited time to complete surveys outside of the examination task. We used both confirmatory factor analysis and Rasch analysis to better support the psychometric validity of the two scales. To our knowledge, this is the first study that validates two test anxiety scales using such a robust statistical approach. The results obtained for both scales are discussed below.

A. Validation of the TAIP

Our results suggest that the final scale adapted from the original PAF has good construct-related validity, namely a robust 19-item, four-factor structure with good internal consistency and adequate discriminant validity (H1). This is consistent with the results of the Italian validation study [72]. In line with our previous studies, we also used Rasch analysis to further strengthen the psychometric validity of the instrument. The Rasch analysis confirmed the multidimensionality of the 19-item instrument and the construct validity of each factor. In addition, we found that the rating scale used was appropriate, namely all four steps were likely to be used by the respondents.

We also collected sufficient convergent validation evidence (H2), namely we found that the Rasch scores in the

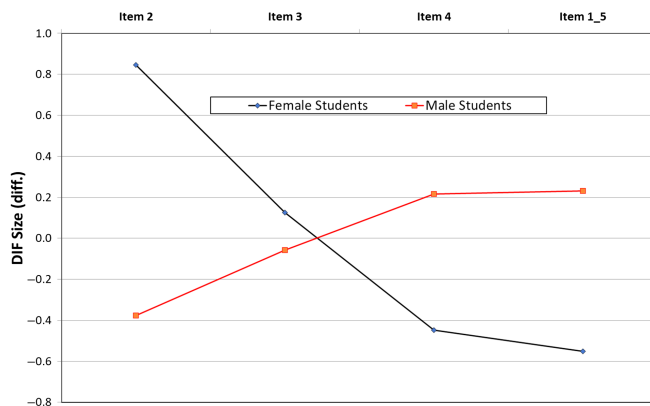


FIG. 6. DIF plot of the ATAIP items. DIF contrast ranges: <0.43 = negligible; (0.43; 0.64) = moderate; >0.64 = large.

four factors of the TAIP significantly correlated with the Rasch scores of the Abbreviated Anxiety Scale, an instrument we developed for this study by adapting four items of the Italian version of the Math Anxiety Inventory [35,36]. Such evidence confirms the associations between different forms of test anxiety [56,107]. Furthermore, the Rasch scores in three of the dimensions (Worry, Emotionality, and Interference) of the TAIP were negatively correlated with the Rasch person measures in the 20-item conceptual questionnaire developed in this study and used as an entrance test to the physics course.

This finding is consistent with studies that have shown the relationship between math anxiety and math performance [71,72]. However, contrary to the findings of the study of the Italian validation of the original PAF, the Lack of Confidence scale was not significantly correlated with the performance in a physics test.

In this regard, however, it should be noted that the Lack of Confidence items measured students' belief in their ability to perform well in a physics exam. Thus, the Lack of Confidence measured by the TAIP is a construct that is closely related to self-efficacy. Recent studies [39] report that the test anxiety dimension of emotivity and self-efficacy both affect performance in the high-stakes exam, while the relationship with performance in the low-stakes exam, such as the entrance questionnaire used in this study, becomes weaker and less clear. Therefore, while hypothesis H3 is substantially supported for the first three dimensions, further research is warranted to examine the role of the fourth dimension of the TAIP instrument on student's performance. Finally, we collected evidence of adequate gender measurement invariance of the TAIP instrument.

While the previous study in the Italian context [72] reported measurement invariance using only variations in χ^2 and CFI index, we also measured invariance using a different method, namely examining the DIF using Rasch measures for each subscale of the TAIP.

Overall, the results of our analyses confirm the gender invariance of the TAIP, in line with previous studies [72], in particular, that the four-factor structure is invariant for gender. Our results therefore confirm that the instrument can be reliably used to compare male and female students' perceptions of anxiety in a physics context. To this end, we found significant gender differences in the Rasch measures on the four dimensions of TAIP except Interference, with large effect sizes on the Worry, Emotionality, and Lack of Confidence scales, confirming that female students feel more anxious about exams than male peers [33].

This finding is also consistent with studies that have focused on examination anxiety in younger students [108]. The results obtained on the Interference scale confirm those obtained with Italian students [72], while the results on the Lack of Confidence scale confirm the gender differences in self-efficacy in physics [109,110]. Thus, also hypothesis H4 is confirmed.

B. Validation of the ATAIP

Our findings suggest that the abbreviated form of the TAI adapted for physics (ATAIP), after parceling two items, has an acceptable internal consistency, very similar to the one found in the literature [67]. The Rasch analysis also supports the construct validity of the final four-item scale (H5). It is important to note that establishing the construct validity of the unidimensional ATAIP does not contradict the establishment of the construct validity of the four-dimensional TAIP. The TAIP and ATAIP are based on different conceptualizations of test anxiety, with the TAIP being based on the PAF and the ATAIP being based on the TAI. As discussed in the review section, the PAF scale expands the original TAI scales (Worry and Emotion) by incorporating the Interference and Lack of Confidence scales. Consistent with the literature, we found in the present study that the Worry and Emotion scales are highly correlated (refer to Fig. 1). Therefore, it is acceptable to collapse the two scales into a more parsimonious model. However, as is often the case, using a more parsimonious model may result in some measurement weaknesses. In particular, in the present study, the criterion-related validity was not fully established. As anticipated, we found a negative correlation between the Rasch measures of the ATAIP and the Rasch measures of the short task used in the midterm exam. However, this correlation was not statistically significant. Although this result may sound somewhat unexpected (and thus hypothesis H6 is not entirely supported), it is consistent with the evidence reported by Lowe [62,63], who also found a nonsignificant correlation between the scores on the college student test anxiety measure and students' self-reported grades. Such a finding may be explained by considering that low levels of anxiety may be beneficial in some cases, as it may contribute to motivation to perform well on a given task [111]. Alternatively, since the brief ATAIP was administered at the end of a midterm exam that was very difficult for students, it is very likely that students with low levels of anxiety also did not perform well in the exam. Future studies are therefore warranted to verify the accuracy of our interpretation.

Finally, our results also support a substantial gender invariance of the measurement model of the ATAIP (H7). However, using Rasch analysis, we found one item (Item 2) that showed a significant DIF contrast. A possible reason for the detected DIF could be that Item 2 elicited a retrospective judgment about the just finished exam ("*I wish the examination did not bother me so much*"), which might differ between female and male students. Actually, such a retrospective judgment may be influenced by the students' accuracy of self-evaluation, defined as the difference between the confidence score in the performance and the actual performance score. Previous studies have found gender differences in the accuracy of self-evaluation in physics, namely girls are more likely to be

underconfident, while boys are more likely to be overconfident [112]. Being underconfident would in turn lead to an underestimation of one's own performance and further increase anxiety. In terms of the person measures, unlike the TAIP, the average ATAIP measures are not significantly different between female and male students. This finding is consistent with previous results reported for physiological hyperarousal from the Test Anxiety Measure [62,63]. An example item from this subscale reads: "*My heart is pounding in my chest during the test*". This subscale measures physical discomfort and worry about the consequences of failing tests [113], so it can be assumed that it measures the same construct as the ATAIP.

There may be three reasons for the different behavior of the two scales with respect to gender. First, we note that 2 out of 19 items in TAIP and 2 out of 4 items in ATAIP also showed moderate DIF and DGF for the gender variable, although the corresponding contrasts were not statistically significant at the $p < 0.05$ level. The existence of a DGF may signal a possible bias in the measures of female and male students. Second, we administered the TAIP and the ATAIP in two different moments of the physics course, at the beginning, after the entrance test that has no evaluation (TAIP), and in the middle after the midterm exam (ATAIP). Another possibility is that the ATAIP is not sensitive enough to measure such differences due to the small number of items. Given that a recent study with 13- to 14-year-old students also found that the brief TAI did not show gender measurement invariance [114], future studies could address such inconsistencies with a larger sample that spans different student ages.

VII. LIMITATIONS

This study has several limitations. First, we acknowledge that having treated the gender variable as binary may have caused a limitation to the study since it excluded from the analysis students who did not fit into the gender or sexual binary. While the percentage of respondents who preferred not to indicate their gender was less than 1%, we plan to involve in subsequent studies larger samples in order to make these data meaningful for quantitative analysis. Second, we involved a convenience sample, which limits the generalizability of the findings. In particular, it would have been important to include in our sample also students enrolled in medicine or life sciences undergraduate courses. In these courses, introductory physics is also mandatory, and anxiety related to the physics exams could be higher due to lower levels of perceived physics self-efficacy. Second, and related to the previous point, the sample consisted of a larger proportion of male students compared to female students. Although the ratio of female to male students in our sample is in line with the national ratio for undergraduate engineering and physics courses, future validation studies of the two proposed scales should include larger and more gender-balanced samples in order to

improve the measurement invariance of the two instruments. Furthermore, the measures were administered in the first semester. A more heterogeneous sample, including students attending courses in the second semester, could be useful in determining the extent to which test anxiety may vary during the university experience. Third and finally, for the TAIP, we were only able to use an adapted version of a math anxiety inventory and a concept inventory as convergent and criterion validity measures, respectively. We are planning to use a wider range of measures, including constructs such as motivation, self-efficacy, accuracy of self-assessment, and previous school grades, in order to obtain stronger evidence of convergent and criterion validity. Similarly, we were only able to collect criterion validation evidence for the ATAIP due to the time constraints of the midterm exam. Therefore, future studies should at least include another concurring instrument to measure test anxiety.

VIII. CONCLUSIONS

In this paper, we have presented psychometric evidence based on factorial and Rasch analysis for two test anxiety scales to be used in introductory physics courses at the undergraduate level.

The validation of both a long and brief scale to measure test anxiety can be useful for physics education researchers and instructors interested in the role of affective variables, such as anxiety, in the learning process. The TAIP scale can be particularly useful for researchers investigating the correlations or predictive relationships of each subscale on relevant outcome variables, such as scores on conceptual tests. However, researchers could also use the ATAIP to measure test anxiety reliably in combination with longer surveys, such as a self-efficacy scale or an attitude toward physics scale. Physics instructors may prefer the ATAIP because it is easier to administer during exams, and completing long questionnaires in that situation may interfere with students' performance. However, physics instructors can also utilize the TAIP to evaluate various aspects of test anxiety. This can aid in planning appropriate educational interventions to reduce maladaptive responses that may hinder performance in physics examinations. For a review of interventions to reduce test anxiety, refer to Ref. [115]. Among the strategies that can be implemented to reduce test anxiety, previous research studies suggest strengthening learning and coping strategies while targeting dysfunctional beliefs about one's own performance and redirecting them toward more adaptive thoughts [116]. Such activities could be delivered in an extracurricular lecture setting by psychologically and counsellor-trained lecturers. Furthermore, both TAIP and ATAIP scales can be used to assess test anxiety under different educational conditions (e.g., different textbooks and lecture approaches) and with students with different types of social conditions (e.g., different socioeconomic status). Similarly, both scales can be used with younger students at the

secondary school level as companions to math anxiety inventories [35,36] to investigate whether test anxiety is dependent on the specific discipline.

A follow-up study is being conducted to investigate how test anxiety changes over a semester in an introductory physics course for undergraduate students. Previous studies with secondary school students [117] suggest that test anxiety may not change significantly over time. This suggests that test anxiety, as measured by typical test anxiety scales such as those presented in this paper, may be a stable individual trait. The study also explores whether test anxiety differs depending on the type of test, including precourse tests, midterm exams, and final exams. Additionally, future research will examine the structural relationships between the dimensions of test anxiety identified in this study and antecedents suggested in the literature, such as social anxiety or avoidance motivation [118]. Further research is necessary to investigate which dimensions better explain the inverse relationship between test anxiety and performance, and whether other dimensions may be suitable indicators of physics test anxiety.

IX. ETHICAL STATEMENT

This research was approved by the Ethical Committee for Research on Human Subjects in Non-Biomedical Field of the corresponding author's institution.

ACKNOWLEDGMENTS

We acknowledge the kind collaboration of all the students who responded to the survey questions and the help of the colleagues in our departments who submitted the survey. We also kindly thank the authors of the Italian validation of the original PAF and TAI for allowing us to use the validated versions of these instruments. This work was supported by authors' institution through the project "drOpout pReventIon and EngagemeNT At The unIversity Of Naples (ORIENTATION)" [Rectorial Decree n. 3429].

APPENDIX

Hereafter, we report the complete TAIP and ATAIP scales.

1. Test anxiety inventory for physics (TAIP) scale

Read the following statements and think about your experience with a physics test. Please indicate the number

between 1 = hardly ever and 4 = almost always that corresponds to how you feel

(Worry subscale)

1. I think about how important the examination is for me.¹
2. I think about how important it is for me to receive a good result.
3. I worry about my results.
4. I am concerned about my grades.
5. I think about what will happen if I don't do well.

(Emotionality subscale)

6. I feel my heart beating fast.
7. I feel anxious.
8. I tremble with fear.
9. I feel somewhat overwhelmed.
10. I feel upset.

(Interference subscale)

11. Distracting thoughts keep "popping" into my head.
12. I am preoccupied by other thoughts that distract me.
13. I easily lose my train of thoughts.
14. I forget things because I am too preoccupied with my personal problems.
15. My concentration is interrupted by interfering thoughts.

(Lack of Confidence subscale)

16. I am confident about my performance.
17. I have faith in my own performance.
18. I am satisfied with myself.
19. I think that I will succeed.
20. I am convinced that I will do well.

2. Abbreviated test anxiety inventory physics (ATAIP) scale

Read the following statements and think about the physics exam that you have just taken. Please indicate the number between 1 = hardly ever and 4 = almost always that corresponds to how you feel

1. During tests I feel very tense.²
2. I wish examinations did not bother me so much.
3. I seem to defeat myself while working on important tests.
4. I feel very panicky when I take an important test.
5. During examinations I get so nervous that I forget facts I really know².

¹This item was deleted after the first confirmatory factor analysis.

²These items were parceled after the first confirmatory factor analysis.

- [1] R. Pekrun, K. R. Muis, A. C. Frenzel, and T. Goetz, *Emotions at School* (Taylor & Francis/Routledge, New York, 2017).
- [2] M. Laukenmann, M. Bleicher, S. Fuß, M. Gläser-Zikuda, P. Mayring, and C. von Rhöneck, An investigation of the influence of emotional factors on learning in physics instruction, *Int. J. Sci. Educ.* **25**, 489 (2003).
- [3] G. Sinatra, S. Broughton, and D. Lombardi, *Emotions in science education*, in *International Handbook of Emotions in Education*, edited by R. Pekrun and L. Linnembrink-García (Routledge, New York, 2014), pp. 415–436.
- [4] F. M. Teixeira dos Santos and E. F. Mortimer, How emotions share the relationship between a chemistry teachers and her high school students, *Int. J. Sci. Educ.* **25**, 1095 (2003).
- [5] R. Pekrun, A. J. Elliot, and M. A. Maier, Achievement goals and achievement emotions: Testing a model of their joint relations with academic performance, *J. Educ. Psychol.* **101**, 115 (2009).
- [6] D. Feldman and M. Kubota, Hope, self-efficacy, optimism, and academic achievement: Distinguishing constructs and levels of specificity in predicting college grade-point average, *Learn. Individ. Diff.* **37**, 210 (2015).
- [7] L. Day, K. Hanson, J. Maltby, C. Proctor, and A. Wood, Hope uniquely predicts objective academic achievement above intelligence, personality, and previous academic achievement. *J. Res. Pers.* **44**, 550 (2010).
- [8] N. Gillet, R. J. Vallerand, M.-A. K. Lafrenière, and J. S. Bureau, The mediating role of positive and negative affect in the situational motivation-performance relationship, *Motiv. Emot.* **37**, 465 (2013).
- [9] X. Oriol-Granado, M. Mendoza-Lira, C.-G. Covarrubias-Apablaza, and V.-M. Molina-López, Positive emotions, autonomy support and academic performance of university students: The mediating role of academic engagement and self-efficacy, *Rev. de Psicodid (English Ed.)*, **22**, 45 (2017).
- [10] A. Ben-Eliyahu, Academic emotional learning: A critical component of self-regulated learning in the emotional learning cycle, *Educ. Psychol.* **54**, 84 (2019).
- [11] R. Bruffaerts, P. Mortier, G. Kiekens, R. P. Auerbach, P. Cuijpers, K. Demyttenaere, and R. C. Kessler, Mental health problems in college freshmen: Prevalence and academic functioning, *J. Affect. Disord.* **225**, 97 (2018).
- [12] R. P. Auerbach, J. Alonso, W. G. Axinn, P. Cuijpers, D. D. Ebert, J. G. Green, and R. Bruffaerts, Mental disorders among college students in the World Health Organization world mental health surveys, *Psychol. Med.* **46**, 2955 (2016).
- [13] L. Dörrenbächer and F. Perels, Self-regulated learning profiles in college students: Their relationship to achievement, personality, and the effectiveness of an intervention to foster self-regulated learning, *Learn. Individ. Diff.* **51**, 229 (2016).
- [14] B. J. Fraser and D. L. Fisher, Effects of anxiety on science related attitudes, *Eur. J. Sci. Educ.* **4**, 441 (1982).
- [15] R. A. Hansen, Anxiety, in *Motivation in Education*, edited by S. Ball (Academic Press, New York, 1977).
- [16] A. Drapeau, A. Marchand, and D. Beaulieu-Prévost, Epidemiology of psychological distress, *Mental Illness: Understanding, Prediction and Control* (InTech Open Access Publisher, Croatia, 2011), pp. 105–134.
- [17] S. H. Ridner, Psychological distress: Concept analysis, *J. Adv. Nurs.* **45**, 536 (2004).
- [18] R. Beiter, R. Nash, M. McCrady, D. Rhoades, M. Linscomb, M. Clarahan, and S. Sammut, The prevalence and correlates of depression, anxiety, and stress in a sample of college students, *J. Affect. Disord.* **173**, 90 (2015).
- [19] E. M. Adlaf, L. Gliksman, A. Demers, and B. Newton-Taylor, The prevalence of elevated psychological distress among Canadian undergraduates: Findings from the 1998 Canadian Campus Survey, *J. Am. Coll. Health* **50**, 67 (2001).
- [20] J. L. Burris, E. H. Brechting, J. Salsman, and C. R. Carlson, Factors associated with the psychological well-being and distress of university students, *J. Am. Coll. Health* **57**, 536 (2009).
- [21] E. A. Maloney and S. L. Beilock, Math anxiety: Who has it, why it develops, and how to guard against it, *Trends Cognit. Sci.* **16**, 404 (2012).
- [22] I. C. Mammarella, F. Hill, A. Devine, S. Caviola, and D. Szűcs, Math anxiety and developmental dyscalculia: A study on working memory processes, *J. Clin. Exp. Neurol.* **37**, 878 (2015).
- [23] N. Casali, M. Ghisi, and R. Rizzato, Validation of the “Study-Anxiety” Questionnaire: A scale for the initial assessment of university students seeking psychological help, *J. Psychopathol. Behav. Assess.* **44**, 1158 (2022).
- [24] M. K. Udo, G. P. Ramsey, and J. V. Mallow, Science anxiety and gender in students taking general education science courses, *J. Sci. Educ. Technol.* **13**, 435 (2004).
- [25] S. Henschel, Antecedents of science anxiety in elementary school, *J. Educ. Res.* **114**, 263 (2021).
- [26] C. S. Ugwuanyi, M. O. Ede, C. N. Onyishi, O. V. Ossai, E. N. Nwokenna *et al.*, Effect of cognitive-behavioral therapy with music therapy in reducing physics test anxiety among students as measured by generalized test anxiety scale, *Medicine (Baltimore)* **99**, e16406 (2020).
- [27] J. V. Mallow, *Science Anxiety: Fear of Science and How to Overcome It* (H&H Publishing Company, Clearwater, FL, 1986).
- [28] R. Pekrun, The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice, *Educ. Psychol. Rev.* **18**, 315 (2006).
- [29] S. L. Britner, Motivation in high school science students: A comparison of gender differences in life, physical, and earth science classes, *J. Res. Sci. Teach.* **45**, 955 (2008).
- [30] A. González, M.-V C. Fernández, and P.-V. Paoloni, Hope and anxiety in physics class: Exploring their motivational antecedents and influence on metacognition and performance, *J. Res. Sci. Teach.* **54**, 558 (2017).
- [31] M. S. Chapell, Z. Blanding, M. E. Silverstein, M. Takahashi, B. Newman, A. Gubi, and N. McCann, Test anxiety and academic performance in undergraduate and graduate students, *J. Educ. Psychol.* **97**, 268 (2005).

- [32] B. Bewick, G. Koutsopoulou, J. Miles, E. Slaa, and M. Barkham, Changes in undergraduate students' psychological well-being as they progress through university, *Stud. Higher Educ.* **35**, 633 (2010).
- [33] N. C. Hall, J. G. Chipperfield, R. P. Perry, J. C. Ruthig, and T. Goetz, Primary and secondary control in academic development: Gender-specific implications for stress and health in college students, *Anxiety Stress Coping* **19**, 189 (2006).
- [34] S. Conneely and B. M. Hughes, Test anxiety and sensitivity to social support among college students: Effects on salivary cortisol. *Cognit. Brain. Behav.* **14**, 295 (2010), <https://www.cbbjournal.ro/index.php/en/test-anxiety/445-test-anxiety-and-sensitivity-to-social-support-among-college-students-effects-on-salivary-cortisol>.
- [35] S. Caviola, C. Primi, F. Chiesi, and I. C. Mammarella, Psychometric properties of the Abbreviated Math Anxiety Scale (AMAS) in Italian primary school children. *Learn. Individ. Diff.* **55**, 174 (2017).
- [36] C. Primi, C. Busdraghi, C. Tomasetto, K. Morsanyi, and F. Chiesi, Measuring math anxiety in Italian college and high school students: Validity, reliability and gender invariance of the Abbreviated Math Anxiety Scale (AMAS), *Learn. Individ. Diff.* **34**, 51 (2014).
- [37] M. S. Griggs, S. E. Rimm-Kaufman, E. G. Merritt, and C. L. Patton, The responsive classroom approach and fifth grade students' math and science anxiety and self-efficacy, *Sch. Psychol. Q.* **28**, 360 (2013).
- [38] J. V. Mallow, H. Kastrup, F. B. Bryant, N. Hislop, R. Shefner, and M. Udo, Science anxiety, science attitudes, and gender: Interviews from a binational study, *J. Sci. Educ. Technol.* **19**, 356 (2010).
- [39] A. Malespina and C. Singh, Gender differences in test anxiety and self-efficacy: Why instructors should emphasize low-stakes formative assessments in physics courses, *Eur. J. Phys.* **43**, 035701 (2022).
- [40] J. B. Stang, E. Altiere, J. Ives, and P. J. Dubois, Exploring the contributions of self-efficacy and test anxiety to 1338 gender differences in assessments, presented at PER Conf. 2020, virtual conference, [10.1119/perc.2020.pr.Stang](https://doi.org/10.1119/perc.2020.pr.Stang).
- [41] L. A. Brown, E. M. Forman, J. D. Herbert, K. L. Hoffman, E. K. Yuen, and E. M. Goetter, A randomized controlled trial of acceptance-based behavior therapy and cognitive therapy for test anxiety: A pilot study, *Behav. Change* **35**, 31 (2011).
- [42] D. D. Szafranski, T. L. Barrera, and P. J. Norton, Test anxiety inventory: 30 years later, *Anxiety Stress Coping* **25**, 667 (2012).
- [43] M. Zeidner and G. Matthews, *Anxiety 101* (Springer, New York, NY, 2011).
- [44] J. E. Sieber, H. F. O'Neil, Jr., and S. Tobias, *Anxiety, Learning, and Instruction* (Wiley, New York, NY, 1977).
- [45] M. Zeidner, Test anxiety in educational contexts: Concepts, findings, and future directions, in *Emotion in Education*, edited by P. A. Schutz and R. Pekrun (Academic Press, San Diego, CA, 2007), pp. 165–184.
- [46] D. W. Putwain, Deconstructing test anxiety, *Emot. Behav. Differ.* **13**, 141 (2008).
- [47] I. A. Friedman and O. Bendas-Jacob, Measuring perceived test anxiety in adolescents: A self-report scale. *Educ. Psychol. Meas.* **57**, 1035 (1997).
- [48] C. D. Spielberger and P. R. Vagg, *Test Anxiety: Theory, Assessment, and Treatment* (Taylor & Francis, Palo Alto, CA, 1995).
- [49] A. M. Sylvia, P. K. Shear, K. E. Jastrowski Mano, J. M. Guerin, and R. Quintino Mano, Test anxiety and reading comprehension: The key role of fluid reasoning, *Anxiety Stress Coping* **36**, 781 (2023).
- [50] M. Zeidner and G. Matthews, Test anxiety, in *Encyclopedia of Psychological Assessment*, edited by R. Fernández-ballesteros (SAGE Publications Ltd., London, 2003), pp. 965–969.
- [51] N. K. Segool, J. Carlson, A. Goforth, N. von der Embse, and J. Barterian, Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing, *Psychologie : Schweizerische Zeitschrift für Psychologie und ihre Anwendungen / herausgegeben von der Schweizerischen Gesellschaft für Psychologie und ihre Anwendungen* **50**, 489 (2013).
- [52] N. von der Embse, D. Jester, D. Roy, and J. Post, Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review, *J. Affect. Disord.* **227**, 483 (2018).
- [53] S. Duchesne and C. F. Ratelle, Patterns of anxiety symptoms during adolescence: Gender differences and sociomotivational factors, *J. Appl. Dev. Psychol.* **46**, 41 (2016).
- [54] R. M. Liebert and L. W. Morris, Cognitive and emotional components of test anxiety: A distinction and some initial data, *Psychol. Rep.* **20**, 975 (1967).
- [55] N. P. von der Embse, A. D. Mata, N. Segool, and E.-C. Scott, Latent profile analyses of test anxiety: A pilot study, *J. Psychoeduc. Assess.* **32**, 165 (2014).
- [56] K. Schnell, T. Ringeisen, D. Raufelder, and S. Rohrmann, The impact of adolescents' self-efficacy and self-regulated goal attainment processes on school performance—Do gender and test anxiety matter?, *Learn. Individ. Diff.* **38**, 90 (2015).
- [57] D. J. Burns, Anxiety at the time of the final exam: Relationships with expectations and performance, *J. Educ. Bus.* **80**, 119 (2004).
- [58] J. C. Cassady and R. E. Johnson, Cognitive test and academic performance, *Contemp. Educ. Psychol.* **27**, 270 (2002).
- [59] N. Karjanto and S.-T. Yong, Test anxiety in mathematics among early undergraduate students in a British university in Malaysia, *Eur. J. Eng. Educ.* **38**, 11 (2013).
- [60] V. Hodapp, Das Prüfungsangstlichkeitsinventar TAI-G: Eine erweiterte und modifizierte Version mit vier Komponenten [The Test Anxiety Inventory: An expanded and modified version with four components], *Zeit. Padag. Psychol.* **5**, 121 (1991).
- [61] M. Zeidner, *Test Anxiety: The State of the Art* (Plenum Press, New York, NY, 1998).
- [62] P. A. Lowe, An investigation into the psychometric properties of the Test Anxiety Measure for college students, *J. Psychoeduc. Assess.* **36**, 322 (2018).
- [63] P. A. Lowe, The Test Anxiety Measure for College Students: Examination of its psychometric properties

- using an online survey with a Canadian sample, *Can. J. Sch. Psychol.* **33**, 279 (2018).
- [64] R. M. Suinn, The STABS, a measure of test anxiety for behaviour therapy: Normative data, *Behav. Res. Ther.* **7**, 335 (1969).
- [65] R. M. Suinn, The desensitization of test-anxiety by group and individual treatment, *Behav. Res. Ther.* **6**, 385 (1968).
- [66] C. D. Spielberger, *Test anxiety inventory: Preliminary professional manual* (Consulting Psychologist Press, Palo Alto, CA, 1980).
- [67] J. Taylor and F. P. Deane, Development of a short form of the Test Anxiety Inventory (TAI), *J. Gen. Psychol.* **129**, 127 (2002).
- [68] I. G. Sarason, Stress, anxiety, and cognitive interference: Reactions to Tests, *J. Pers. Soc. Psychol.* **46**, 929 (1984).
- [69] V. Hodapp, S. Rohrmann, and T. Ringeisen, *Prüfung-sangstfragebogen (PAF) [Test-Anxiety Questionnaire]* (Hogrefe, Göttingen, the Netherlands, 2011).
- [70] K. Schnell, A. N. Tibubos, S. Rohrmann, and V. Hodapp, Test and math anxiety: A validation of the German Test Anxiety Questionnaire, *Polish Psychol. Bull.* **44**, 193 (2013).
- [71] F. Hoferichter, D. Raufelder, T. Ringeisen, S. Rohrmann, and W. M. Bukowski, Assessing the multi-faceted nature of test anxiety among secondary school students: An English version of the German Test Anxiety Questionnaire: PAF-E, *J. Psychol.* **150**, 450 (2016).
- [72] M. A. Donati, V. A. Izzo, A. Scabia, J. Boncompagni, and C. Primi, Measuring test anxiety with an invariant measure across genders: The case of the German test anxiety inventory, *Psychol. Rep.* **123**, 1382 (2020).
- [73] N. Mascret, S. Danthony, and F. Cury, Anxiety during tests and regulatory dimension of anxiety: A five-factor French version of the Revised Test Anxiety scale, *Curr. Psychol.* **40**, 5322 (2021).
- [74] M. Sahin, S. Çalişkan, and U. Dilek, Development and validation of the physics anxiety rating scale, *Int. J. Environ. Sci. Educ.* **10**, 183 (2015), <https://files.eric.ed.gov/fulltext/EJ1062997.pdf>.
- [75] F. C. Richardson and R. M. Suinn, The mathematics anxiety rating scale: Psychometric data, *J. Counsel. Psychol.* **19**, 551 (1972).
- [76] S. Tobias, *Overcoming Math Anxiety* (W.W. Norton & Company, New York, 1993).
- [77] R. Hembree, The nature, effects, and relief of mathematics anxiety, *J. Res. Math. Educ.* **21**, 33 (1990).
- [78] X. Ma, A meta-analysis of the relationship between anxiety toward mathematics, and achievement in mathematics, *J. Res. Math. Educ.* **30**, 520 (1999).
- [79] S. Henschel and T. Roick, The multidimensional structure of math anxiety revisited. Incorporating psychological dimensions and setting factors, *Eur. J. Psychol. Assess.* **36**, 123 (2018).
- [80] T. Ringeisen, D. Raufelder, K. Schnell, and S. Rohrmann, Validating the proposed structure of the relationships among test anxiety and its predictors based on control-value theory: Evidence for gender-specific patterns, *Educ. Psychol.* **36**, 1826 (2016).
- [81] E. Fennema and J. A. Sherman, Fennema-Sherman mathematics attitudes scales: Instruments designed to measure attitudes toward the learning of mathematics by females and males, *J. Res. Math. Educ.* **7**, 324 (1976).
- [82] D. R. Hopko, R. Mahadevan, R. L. Bare, and M. K. Hunt, The abbreviated math anxiety scale (AMAS): Construction, validity, and reliability, *Assessment* **10**, 178 (2003).
- [83] P. R. Pintrich, D. A. F. Smith, T. Garcia, and W. J. Mckeachie, Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ), *Educ. Psychol. Meas.* **53**, 801 (1993).
- [84] M. Crede and L. Phillips, A meta-analytic review of the motivated strategies for learning questionnaire, *Learn. Individ. Diff.* **21**, 337 (2011).
- [85] M. R. Stoeckel and G. H. Roehrig, Gender differences in classroom experiences impacting self-efficacy in an AP Physics 1 classroom, *Phys. Rev. Phys. Educ. Res.* **17**, 020102 (2021).
- [86] M. H. Ashcraft and A. M. Moore, Mathematics anxiety and the affective drop in performance, *J. Psychoeduc. Assess.* **27**, 197 (2009).
- [87] J. M. Ortiz-Lozano, A. Rua-Vieites, P. Bilbao-Calabuig, and M. Casadesús-Fa, University student retention: Best time and data to identify undergraduate students at risk of dropout, *Innov. Educ. Teach. Int.* **57**, 74 (2018).
- [88] T. Espinosa, K. Miller, I. Araujo, and E. Mazur, Reducing the gender gap in students' physics self-efficacy in a team and project-based introductory physics class, *Phys. Rev. Phys. Educ. Res.* **15**, 010132 (2019).
- [89] J. M. Nissen and J. T. Shemwell, Gender, experience, and self-efficacy in introductory physics, *Phys. Rev. Phys. Educ. Res.* **12**, 020105 (2016).
- [90] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, *Phys. Teach.* **30**, 141 (1992).
- [91] R. Thornton and D. Sokoloff, Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula, *Am. J. Phys.* **66**, 338 (1998).
- [92] J. B. Schreiber, F. K. Stage, J. King, A. Nora, and E. A. Barlow, Reporting structural equation modeling and confirmatory factor analysis results: A review, *J. Educ. Res.* **99**, 323 (2006).
- [93] L. T. Hu and P. M. Bentler, Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, *Struct. Eq. Model.* **6**, 1 (1999).
- [94] L. R. Tucker and C. Lewis, A reliability coefficient for maximum likelihood factor analysis, *Psychometrika* **38**, 1 (1973).
- [95] P. M. Bentler, D. Bonett, and G. Douglas, Significance tests and goodness of fit in the analysis of covariance structures, *Psychol. Bull.* **88**, 588 (1980).
- [96] P. M. Bentler, Comparative fit indexes in structural models, *Psych. Bull.* **107**, 238 (1990).
- [97] J. H. Steiger and J. C. Lind, Statistically-based tests for the number of common factors, in *Proceedings of the annual Spring Meeting of the Psychometric Society Annual Meeting, Iowa City, IA* (1980).
- [98] B. Wheaton, B. Muthen, D. F. Alwin, and G. Summers, Assessing reliability and stability in panel models, *Soc. Methodol.* **8**, 84 (1977).

- [99] F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis* (Pearson Education Limited, Essex, 2014).
- [100] J. M. Linacre, Winsteps® Rasch Tutorial 4. Available at <http://www.winsteps.com/a/winsteps-tutorial-4.pdf> (2012).
- [101] J. M. Linacre, WINSTEPS (Version 3.81.0) [Computer Software], Chicago, IL: Winsteps.com. (2014).
- [102] W. J. Boone, J. S. Staver, and M. S. Yale, *Rasch Analysis in Human Sciences* (Springer, Dordrecht, 2014).
- [103] W. J. Boone and J. S. Staver, *Advances in Rasch Analyses in the Human Sciences* (Springer, Dordrecht, 2020).
- [104] B. M. Byrne, Testing for multigroup invariance using AMOS graphics: A road less traveled, *Struct. Eq. Model.* **11**, 272 (2004).
- [105] A. W. Meade, E. C. Johnson, and P. W. Braddy, Power and sensitivity of alternative fit indices in tests of measurement invariance, *J. Appl. Psychol.* **93**, 568 (2008).
- [106] G. Cizek and S. Burg, *Addressing Test Anxiety in a High Stakes Environment* (Corwin Press, Thousand Oaks, CA, 2006).
- [107] K. Morsanyi, C. Busdraghi, and C. Primi, Mathematical anxiety is linked to reduced cognitive reflection: A potential road from discomfort in the mathematics classroom to susceptibility to biases, *Behav. Brain Funct.* **10**, 31 (2014).
- [108] P. A. Lowe and R. P. Ang, Cross-cultural examination of test anxiety among U.S. and Singapore elementary students on the Test Anxiety Scale for Elementary Students (TAS-E), *Educ. Psychol.* **32**, 107 (2012).
- [109] Y. Z. Kalender, E. Marshman, C. D. Schunn, T. J. Nokes-Malach, and C. Singh, Gendered patterns in the construction of physics identity from motivational factors, *Phys. Rev. Phys. Educ. Res.* **15**, 020119 (2019).
- [110] E. Bottomley, A. Kohnle, K. Mavor, P. Miles, and V. Wild, The relationship between gender and academic performance in undergraduate physics students: The role of physics identity, perceived recognition, and self-efficacy, *Eur. J. Phys.* **44**, 025701 (2022).
- [111] R. Alpert and R. N. Haber, Anxiety in academic achievement situations, *J. Abnorm. Soc. Psychol.* **61**, 207 (1960).
- [112] I. Testa, S. Galano, and O. Tarallo, The relationships between freshmen's accuracy of self-evaluation and the likelihood of succeeding in chemistry and physics exams in two STEM undergraduate courses, *Int. J. Sci. Educ.* **45**, 358 (2023).
- [113] T. E. Joiner, R. A. Steer, A. T. Beck, N. B. Schmidt, M. D. Rudd, and S. J. Catanzaro, Physiological hyperarousal: Construct validity of a central aspect of the tripartite model of depression and anxiety, *J. Abnorm. Psychol.* **108**, 290 (1999).
- [114] M. P. Kösters, L. H. Klaufus, and M. F. van der Wal, Validity and reliability of the short Test Anxiety Inventory (TAI-5) in Dutch adolescents, *J. Gen. Psychol.* **151**, 76 (2023).
- [115] T. Ergene, Effective interventions on test anxiety reduction: A meta-analysis, *Sch. Psychol. Int.* **24**, 313 (2003).
- [116] N. Reiss, I. Warnecke, T. Tolgou, D. Krampen, U. Luka-Krausgrill, and S. Rohrmann, Effects of cognitive behavioral therapy with relaxation vs. imagery rescripting on test anxiety: A randomized controlled trial, *J. Affect. Disord.* **208**, 483 (2017).
- [117] C. Fréchette-Simard, I. Plante, S. Duchesne, and K. E. Chaffee, A latent growth analysis of individual factors predicting test anxiety during the transition from elementary to secondary school, *J. Early Adolesc.* **43**, 265 (2023).
- [118] D. W. Putwain, N. P. von der Embse, E. C. Rainbird, and G. West, The development and validation of a new Multidimensional Test Anxiety Scale (MTAS), *Eur. J. Psychol. Assess.* **37**, 236 (2020).