

Article

Satisfaction and Frustration of Basic Psychological Needs in Classroom Assessment

Lia M. Daniels ^{1,*}, Kendra Wells ¹, Marlit Annalena Lindner ^{2,3}, Adam M. Beeby ¹ and Vijay J. Daniels ⁴

¹ Educational Psychology, Faculty of Education, College of Social Sciences and Humanities, University of Alberta, 6-123F Education North, Edmonton, AB T6G 2G5, Canada

² Institute of Educational Sciences, Europa-Universität Flensburg (EUF), 24943 Flensburg, Germany

³ Educational Research and Educational Psychology Department, Leibniz-Institute for Science and Mathematics Education (IPN), 24118 Kiel, Germany

⁴ Department of Medicine, Faculty of Medicine and Dentistry, College of Health Sciences, University of Alberta, Edmonton, AB T6G 2G3, Canada

* Correspondence: lia1@ualberta.ca

Abstract

Examinations are central to higher education, yet students consistently describe them as detrimental to well-being. Drawing on self-determination theory (SDT), we conducted three studies to examine whether multiple-choice examinations could be redesigned to satisfy students' basic psychological needs (BPNs) and support well-being. In Study 1 ($n = 400$), we developed and validated the Basic Psychological Need Satisfaction and Frustration Scale for Classroom Assessment (BPNSF-CA). Using bifactor exploratory structural equation modeling (bifactor ESEM), results supported a well-defined single global need fulfillment factor (G-factor) alongside six specific factors (autonomy support/frustration, competence support/frustration, relatedness support/frustration) as well as evidence of validity. In Study 2 ($n = 387$), we conducted a randomized experiment with three versions of a multiple-choice exam serving as the independent variable (flawed items, high-quality items, and high-quality + need-supportive features). Results showed that high-quality items improved performance, while only the addition of need-supportive features satisfied BPNs with differential patterns for the single G-factor and S-factors. In Study 3 ($n = 101$), we applied the intervention in a real classroom and tested the mediational role of BPN satisfaction. Results showed that redesigned exams (high-quality + need-supportive features) significantly enhanced perceptions of fairness and success via BPNs. We conclude with a discussion of all three studies, including implications and limitations.

Keywords: classroom assessment; basic psychological needs; well-being; intervention; higher education



Received: 1 December 2025

Revised: 30 December 2025

Accepted: 20 January 2026

Published: 2 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Classroom assessment is one of the most consequential features of post-secondary education, shaping students' grades, progression, and self-perceptions of competence. Despite its importance, students frequently characterize examinations as stressful, controlling, and undermining of their well-being [1–4]. The research literature on assessment and student well-being reveals a tension when it comes to solving this problem. On the one hand, psychometricians and measurement specialists have emphasized the importance of item quality, validity, and reliability [5]. From this perspective, the solution to compromised assessment experiences is to eliminate item-writing flaws and strengthen the validity of

inferences. On the other hand, educational researchers and reformers often argue that summative assessment is so flawed that it should be abandoned in favor of formative or authentic approaches [6,7]. This latter argument, however, has not displaced the dominance of multiple-choice examinations in higher education [8]. Thus, we propose another path forward, rooted in self-determination theory (SDT) [9,10]. SDT explains that when the three basic psychological needs (BPNs) of autonomy, competence, and relatedness are satisfied, student well-being is enhanced, and when they are frustrated, well-being is undermined [11,12]. Research has documented that teachers can and do influence students' well-being through need-supportive or -thwarting instructional practices [13,14]. However, the potential of applying need-supportive features to assessments, like multiple-choice exams, is underexplored [15].

We present three studies designed to advance both the measurement and application of BPNs in the domain of post-secondary classroom assessment. In Study 1, we developed and validated the Basic Psychological Need Satisfaction and Frustration Scale for Classroom Assessment (BPNSF-CA). Using bifactor exploratory structural equation modeling (bifactor ESEM), we demonstrated that BPNs in the domain of assessment are best conceptualized as a combination of a single global need fulfillment factor (G-factor) and six specific factors (S-factors) corresponding to autonomy, competence, and relatedness by satisfaction and frustration. In Study 2, we retained this measurement precision while experimentally manipulating multiple-choice (MCQ) exam design to test whether item quality and need-supportive features causally influenced BPN fulfillment and student performance. In Study 3, we tested the MCQ exam intervention in an authentic classroom, focusing on the mediational role of BPN fulfillment in linking exam design to indicators of student well-being. The progression from psychometric validation to randomized controlled experiment to ecological quasi-experimental design testing mediational mechanisms provides a fulsome initial exploration of how the redesign of MCQ exams according to psychological principles may benefit student well-being.

1.1. Psychological Need Satisfaction and Frustration in Self-Determination Theory

According to SDT, when the three BPNs, autonomy, competence, and relatedness, are satisfied at school, students have enhanced well-being and functioning, and when they are frustrated, students feel psychological distress and a reduced sense of well-being [10,12,14]. Importantly, SDT defines BPN as distinct while also noting that “socializing agents capable of nurturing one need often simultaneously support other needs” [16], p.16, allowing for a holistic appraisal of a context as generally need supporting or frustrating. As such, throughout this paper, we will draw on the literature that addresses both specific and holistic approaches to BPN.

According to Vansteenkiste et al. [10], autonomy refers to the degree to which an individual feels in control of their own behaviors and decisions. When autonomy is satisfied, a person feels that they are making decisions using their own free will, evoking a sense of volition and authenticity. When autonomy is frustrated, volition is replaced by pressure that creates an inner tension. The externally imposed nature of much of classroom assessment likely has consequences for autonomy. Competence refers to how capable or efficacious an individual feels in reaching their own goals. When the need for competence is satisfied, an individual feels capable of carrying out an activity in that domain. When the need for competence is frustrated, an individual feels helpless, which reduces their capacity to learn new skills and achieve their goals. In the domain of assessment, satisfaction or frustration regarding competence is at the forefront of the mind as assessments turn into grades that convey the mastery of learner outcomes. Relatedness refers to social connections that exist within the domain. When relatedness is satisfied, the relationships that an individual

has with others are meaningful, caring, and warm. When relatedness is frustrated, an individual feels excluded or isolated, or they may not perceive any relationship with others. Even the most caring instructors often depersonalize assessment in the name of objectivity. From the student perspective, Daniels and colleagues [15] showed that undergraduate students indeed report feeling neutral or disagreeing that assessment satisfies autonomy, competence, and relatedness when measured separately.

Usually, researchers measure students' BPNs through self-report surveys. A common measure of BPN satisfaction at school is Tian, Han, and colleagues' [17] Basic Psychological Needs at School. The scale was validated in four samples of Chinese students with adequate model fit via confirmatory factor analysis for satisfaction of BPN and invariance across students' age and gender. Other researchers have adapted scales to aspects of the school context without conducting a full validation study (e.g., Ref. [18] for engineering and Ref. [19] for calculus). BPN surveys have also been adapted to fit discrete pedagogical approaches such as gamification [20] and flipped classrooms [21]. Across all surveys, there is noticeable variation in the agent of relatedness, which sometimes focuses on the teacher, but other times involves peers or people generally.

Although there is some debate in the literature on the structure of satisfaction and frustration [10], contemporary measures tend to address both, thereby following Chen et al.'s 2015 [11] BPN Satisfaction and Frustration (BPN-SF) scale. Chen and colleagues used an iterative process involving selecting items from pre-existing scales, generating additional items, and translating and back-translating between English, Dutch, and Chinese. A total of 42 items were then tested through exploratory and confirmatory factor analyses and multi-group comparisons between versions until 24 balanced items were finalized as not only fitting a six-solution model but also demonstrating measurement invariance across countries. According to the manual, the BPN-SF [22] has been translated into several languages and used in many contexts, including schools. Overall, self-report surveys are a common way for researchers to collect data on students' BPNs, which they then analyze in several ways depending on their theoretical stance.

Specifying Global and Specific Factors in BPN Measurement

Across the surveys measuring BPN, researchers choose from a variety of methods to represent the underlying theoretical propositions and constructs [12]. On the one hand, it is common for researchers to use summed or latent subscales that treat autonomy, competence, and relatedness as distinct constructs that uniquely explain variance in outcomes [12]. On the other hand, multicollinearity between the individual BPN factors suggests that BPNs are often experienced as functionally intertwined in students' lived educational experiences, leading researchers to use higher-order factors to represent a unified factor of need satisfaction or frustration.

As a third option, researchers suggest that bifactor exploratory structural equation modeling (bifactor ESEM) is appropriate for modeling BPN because it simultaneously models a global need fulfillment factor (G-factor) alongside specific factors (S-factors) for autonomy, competence, and relatedness [12]. Evidence for this approach is quickly accumulating. For instance, Sánchez-Oliva et al. [23] demonstrated in a work setting that the G-factor was strongly related to efficacy and negatively associated with exhaustion, while S-factors showed differentiated patterns (competence positively predicted efficacy, relatedness reduced exhaustion, and autonomy was non-significant). Bifactor-ESEM solutions have also been reported in body image research [24] and in second-language learning [25]. These studies, however, only involved need satisfaction, not frustration, which requires an additional theoretically guided measurement decision, namely whether satisfaction and frustration are best represented as two separate G-factors or a single G-factor defined

by negative loadings from frustration items and positive loadings from satisfaction items. Tóth-Király et al. [26] tested this question directly by comparing 16 competing models in two separate samples and concluding that “a single overarching global factor reflecting a global continuum of need fulfillment” (p. 281) provided the best fit.

In highly structured contexts such as classroom assessment, we suggest that students are unlikely to parse specific design elements into discrete need categories; instead, assessment practices, collectively, are likely to exert their greatest effects on a global sense of need fulfillment. As such, much like how instructional practices can simultaneously impact multiple BPNs [16], design features of assessment will likely afford multiple need-relevant experiences as well as some specific emphases. We interpret the global factor as capturing the phenomenological sense of being supported versus pressured during assessment, with specific needs representing differentiated residual experiences that may be affected by design features that are more attuned to autonomy, competence, or relatedness, respectively. This perspective provides a theoretical basis for expecting assessment interventions to (a) consist of multiple elements and (b) exert their strongest effects on global need fulfillment (i.e., the G-factor) while producing more modest and tailored effects at the level of specific needs.

1.2. Classroom Assessment and Well-Being in Higher Education

Assessments, particularly MCQ examinations, are central to the higher education experience, yet they are among the most prominent sources of stress reported by students [2,3]. Unlike the diffuse, ongoing nature of instruction, assessments are time-constrained, high-consequence tasks where outcomes directly determine grades and future opportunities [27]. Unsurprisingly, students consistently describe assessments as anxiety-provoking, uncaring, and frustrating [1,4]. Empirical evidence corroborates these perceptions: test anxiety undermines performance at all grades [3] and exam-related stress is associated with depression, burnout, eating disorders, and suicidal ideation [28,29]. One way to unite this constellation of outcomes is as indicators of well-being.

According to Collie and Hascher [30], researchers tend to take either a model-determined approach or a model-informed approach to studying student well-being. In the former, researchers adhere to a pre-established model of well-being, such as Diener’s [31] subjective well-being or Ryff’s [32] psychological well-being. In the latter, researchers make conceptually informed decisions about variables that are relevant to well-being in a particular context. Dodd and colleagues [33] draw a similar distinction in their scoping review, noting that in higher education, researchers used a combination of “direct” measures of well-being, such as the Flourishing Scale [34], and “indirect” measures of well-being, such as self-esteem, hope, or resilience.

Considering indirect indicators at school, Hossain and colleagues [35] undertook a scoping review of 33 studies published between 1989 and 2020 and identified 91 separately labeled constructs used to measure student well-being. To organize this variation, they grouped the measures into eight overarching categories: positive emotions, absence of negative emotions, engagement, relationships, accomplishment, purpose in school, internal factors, and external factors. Looking for measurable elements of well-being in assessment specifically, Daniels and Wells [1] argued that “there is a consistent constellation of indicators of student well-being that have clear counterparts in the domain of assessment, including but not limited to test anxiety, emotions, efficacy beliefs, conceptions of fairness, effort, and perceptions of success” (p. 4). Conceptually, we agree with this formulation, and thus, in the forthcoming empirical studies, we use anxiety, stress, and fairness as elements of well-being related to assessment that are also sensitive enough to capture change in response to an MCQ exam intervention. These indicators meet Collie and Hascher’s [30]

recommendations as being conceptually sound, distinct from each other, and positively and negatively valenced.

1.2.1. Item-Writing Guidelines for Test Quality

A critical starting point in MCQ exam intervention is item quality. Haladyna and Downing's [36] taxonomy of MCQ item-writing guidelines synthesized decades of expert consensus into principles for eliminating flaws that compromise validity. These guidelines [5] include writing clear stems, ensuring only one correct answer, avoiding implausible distractors, and eliminating "all-of-the-above" or "none-of-the-above" options. Despite their wide availability, item-writing flaws remain pervasive [37,38], with some studies reporting fewer than 10% of items fully error-free [37]. Such flaws introduce construct-irrelevant variance and reduce the reliability and validity of the test [39]. Poor item quality may also exacerbate stress and undermine fairness, confounding the commonly observed associations between exams and negative well-being outcomes. In other words, the widespread distress linked to exams may be partly attributable to flawed multiple-choice items.

1.2.2. Need-Supportive Features of Test Design

High-quality multiple-choice items, while necessary, are unlikely to be sufficient to improve students' well-being in assessment. Based on decades of research, there is agreement that instructional practices such as taking the student's perspective, allowing input, providing choice, offering explanatory rationales, providing optimal challenge, giving specific feedback, being clear and consistent, and showing care and positive regard support students' BPNs [13,14,40]. In contrast, common practices that frustrate students' BPNs include ignoring student perspectives, exerting too much control, using pressuring language, creating chaotic course designs, changing standards, and being distant or cold [13,14,40]. Over 50 studies have found that teachers can adjust their instructional practices to better satisfy students' BPNs and support learning and well-being (see [13] for a review). While some studies identify specific connections between a practice and a specific BPN [40], other research has recognized that the correlations between practices and needs can be quite high, suggesting the associations may be holistic [12]. We should note, however, that the research on instructional practices is rarely tested with bifactor-ESEM formulations of BPNs.

Ahmadi and colleagues [40] recently created a classification system for instructional practices underpinning BPN interventions. Although the system is not specific to assessment, it is tailored to the domain of education and offers guidance in determining the type of features that could be relevant to an MCQ intervention. Using a Delphi panel, Ahmadi and colleagues identified 57 behaviors across the six BPNs that were agreed to be related to SDT. Drawing from this list, we translated instructional principles into four features that can be applied to the design of MCQ tests while ensuring each feature does not contradict high-quality item writing guidelines. First, students may experience more BPN satisfaction and less frustration in the context of an MCQ test when instructors create and share a test blueprint with students [41]. A test blueprint maps exam questions to course topics or learner outcomes, also perhaps indicating the level of cognitive complexity of the question. As a need-supportive practice, blueprints offer students transparency about what will be on the test, an indication that instructors are being honest, and reassurance that the content is appropriately linked to course topics. Second, instructors can organize the test according to the blueprint, thereby supporting students' clarity of thinking and showing they upheld the stated test parameters. When items are grouped, students may experience the test as less chaotic and more predictable. Third, instructors can show regard for their students by offering a supportive message at the start and end of the exam, thereby

satisfying relatedness even in an exam setting. This can be quite a simple and factual message that reminds students that the content is drawn from their textbook, lectures, or something more relational, depending on the overall style of the instructor. Fourth, to take students' perspectives and acknowledge emotions in assessment, practices regularly cited as satisfying BPN, instructors can create opportunities for student feedback on the test. This can be accomplished through error flagging, a comment box, or even opportunities to email or meet with the instructor. Each of these features is low-burden to the instructor, does not interfere with adherence to item-writing guidelines, and can be linked to the classification created by Ahmadi et al. in at least one way [13,14,40]. See Table 1.

Table 1. Classification of need-supportive features.

Design Feature	Description	Functional Description	Alignment with Ahmadi et al. 2023 [40]
Test blueprint	Mapping the number and cognitive complexity of questions to course outcomes/topics and providing students with the blueprint as a resource	Reduces uncertainty; increases perceived predictability of exam content; transparent alignment of learning and assessment	AS5 invitational language; AS11 provide resources; CS11 clarify expectations; CS14 self-monitoring
Item grouping	Organizing exam items into coherent sections aligned with course topics or learning objectives.	Reduces cognitive load; enhances clarity and predictability of exam; contributes to a sense of order and fairness in the assessment experience	CS12 displays explicit guidance, CT4 chaos
Supportive message	Providing a supportive message in keeping with the teaching style	Reduces evaluative threat; supports internalization of task value; signals instructor care	AS3 rationales; CS8 display hope; RS1 positive regard
Request for feedback	Ungraded question at the end of the exam for students to flag or highlight concerns	Affords voice; communicates respect; reinforces fairness perceptions	AS6 asks about student experiences; RS2 asks about progress; RT1 ignores students

Rather than targeting specific BPNs in isolation, we designed the four features to serve as a coordinated set of design features that can collectively shape how students experience assessments in terms of clarity, predictability, fairness, perceived support, and other mechanisms that consistently relate to BPN satisfaction and frustration [40]. From this perspective, features such as blueprinting and item grouping may simultaneously reduce uncertainty, enhance perceived controllability, and signal order rather than chaos, thereby contributing to students' global sense of being supported during assessment while also exerting effects specifically on competence satisfaction, for example. Similar logic applies to supportive messaging and requests for feedback, which we expect to have a global effect and draw on psychological functions that satisfy relatedness. Consistent with SDT and with bifactor models of need fulfillment, the theoretical expectation is not a one-to-one correspondence between features and needs, but an additive influence on students' holistic appraisal of need support, with residual differentiation at the level of specific needs [42].

1.3. Conceptual Framework and Study Overviews

We advance a conceptual model of assessment design that integrates psychometric rigor through high-quality items [5] with the priority of student well-being through SDT [10]. In this model, multiple-choice tests can be deliberately constructed to be both technically sound and BPN supportive. From a technical perspective, item-writing quality is necessary to ensure that inferences about student learning are valid. From a psychological

perspective, the addition of need-supportive features to test design collectively signals to students that they are in control, cared for, and positioned for success. When both aspects are considered, assessments may promote rather than undermine students' BPNs as a source of well-being. Indeed, this mediational mechanism is well articulated in SDT, which specifies that the satisfaction or frustration of BPNs serves as the psychological mechanism linking contextual features, in this case conveyed through MCQ exams, to downstream outcomes indicative of well-being [10,13]. Thus, in our model, the effects of MCQ exam design on student well-being are mediated by BPN fulfillment.

We conducted three studies extending from this model. Each study builds on the previous in scope and conceptual contribution. Materials, data, and code for all three studies are openly accessible at [Study 1 <https://doi.org/10.5683/SP3/KGC5J3> Study 2 <https://doi.org/10.5683/SP3/728IVW> Study 3 <https://doi.org/10.5683/SP3/UYU2YM>]. Study 1 focused on measurement, developing and validating the BPNSF-CA with bifactor ESEM to capture both global and specific need fulfillment in assessment contexts. A measurement tool was necessary to be sensitive to the domain, as well as to consider the role of the G-factor in the way students experience assessment. Study 2 provided causal evidence by experimentally manipulating MCQ test design, comparing tests with flawed items vs. high-quality items vs. high-quality items + need-supportive features. This study isolated the independent and combined effects of psychometric rigor and psychological support on both performance and BPN fulfillment, even though it lacked ecological validity. Study 3 tested the full process model in an authentic classroom setting. By embedding redesigned exams into a real course, we examined whether a high-quality, need-supportive MCQ test design improved students' stress, anxiety, perceived fairness, and success, and whether these effects operated through BPN fulfillment. Collectively, this progression from measurement precision to experimental causality to ecological mediation provides a preliminary test of the conceptual model and a way forward for assessment.

2. Study 1: Validation of the BPNSF-CA

The purpose of Study 1 was to develop and validate a questionnaire measuring students' need satisfaction and frustration in the domain of classroom assessment (BPNSF-CA). We designed the BPNSF-CA survey by modifying Chen and colleagues' [11] Basic Psychological Need Satisfaction and Frustration scale. Then we collected validation evidence through a self-report correlational study.

2.1. Method

2.1.1. Participants

We recruited and retained 400 North American undergraduate post-secondary students in any field of study from a pool of 2949 eligible participants in Prolific (4 participants accessed the survey but did not complete it). We limited the geographical location of students to minimize the variability of assessment practices and consequences. Participants ranged in age from 18 to 67 years ($M = 27.12$, $SD = 9.08$). Based on Prolific demographics, 50% identified as White, 24% as Asian, 10% as Black, 10% as mixed ethnicities, and 5% as other; 86% reported English as a first language. In terms of gender, 50% identified as women, 46% as men, and 4% as non-binary or preferred not to report.

2.1.2. Procedures

Ethics approval was granted by the university research ethics board (Pro00137385). Through Prolific, participants anonymously completed one online survey hosted on SurveyMonkey© and were remunerated £9/h, which is considered a "good" rate for a simple survey task. After the consent letter, participants were provided with the following prompt:

“Think about ONE course you are taking right now. Focusing on just that class, answer the following questions.” The survey contained the newly developed BPNSF-CA along with single-item measures of assessment-related stress, anxiety, and fairness as indicators of well-being.

2.1.3. Materials

To describe the course about which participants were reporting, we asked them to indicate if it was required or elective, its size, its broad discipline, and the types of graded assessments used. We also asked participants to confirm their status as an undergraduate student. The BPNSF-CA contains 24 items adapted from Chen et al. [11]: 4 items each for autonomy satisfaction (AS), competence satisfaction (CS), relatedness satisfaction (RS), autonomy frustration (AF), competence frustration (CF), and relatedness frustration (RF). Items avoided negatives and double-barreled phrasing [43], and frustration items used strong verbs to reflect active thwarting [44]. Responses used a 5-point Likert scale (1 = strongly disagree and 5 = strongly agree). The full scale is in Appendix A. We selected the criterion variables to be indirect, model-informed indicators of well-being [30] relevant to the domain of assessment. Specifically, we used single-item measures of stress (“The assessments in this class add to the stress in my life.”; $M = 3.59$ and $SD = 1.10$), anxiety (“I feel panicky when completing assessments in this course.”; $M = 3.02$ and $SD = 1.29$), and fairness (“So far, I experience the assessments in this class as fair.”; $M = 3.67$ and $SD = 0.97$).

2.1.4. Power Analysis

We implemented a Monte Carlo simulation for a bifactor model using Bader et al.’s [45] syntax in R. Expected factor loadings were set at 0.50 for the global factor (G) and 0.40 for the specific factors (S) based on estimates discerned from the existing literature. Simulations indicated acceptable convergence at $n = 250$ and excellent convergence at $n = 300$. Given the added complexity of the exploratory portion of bifactor ESEM and following guidance that successful convergence together with simulation can be taken as evidence of sufficient power, we targeted $n = 400$.

2.1.5. Plan for Analyses and Hypotheses

We inspected data for response patterns and tested for outliers, choosing to retain all participants. We calculated frequencies of the types of courses and assessments that participants were reflecting on as they responded to the BPNSF-CA. Main analyses were run in MPLUS 8.3 [46] with the maximum likelihood estimator to control for non-normality in the data. We tested four measurement models following Alamer’s [25] decision tree, contrasting confirmatory factor analyses (CFA), exploratory structural equation modeling (ESEM), bifactor CFA, and bifactor ESEM (Figure 1).

We defined a good model fit at CFI and TLI ≥ 0.95 and RMSEA and SRMR ≤ 0.05 , and an adequate model fit at CFI and TLI ≥ 0.90 and RMSEA and SRMR ≤ 0.07 [25]. We compared models using a sequential strategy articulated by Morin and colleagues [47] in which ESEM is deemed better than CFA when model fit is improved, latent factor correlations are reduced, the cross-loadings are small to moderate and/or easy to explain, and the target items loading on S-factors are well-defined at ideally >0.50 . Next, if the model fit for a bifactor model improves over the original CFA or ESEM and the G-factor and S-factors are well-defined by examination of the factor loadings, then the corresponding bifactor model is considered the most appropriate solution. We report standardized factor loadings (λ) for each item, uniquenesses (δ , i.e., residual variances), and the model-based omega coefficients (ω) as an indicator of reliability using Dueber’s [48] calculator.

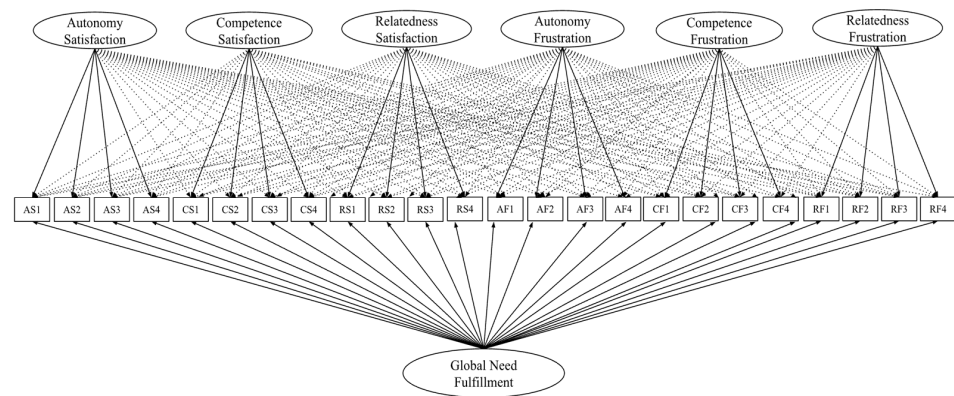


Figure 1. Representation of the Bifactor ESEM tested for the BPNSF-CA. Dotted lines represent cross loadings.

H1. A bifactor ESEM will provide the best fit of the data, including one G-factor, three specific satisfaction factors (autonomy support, AS; competence support, CS; and relatedness support, RS), and three specific frustration factors (autonomy frustration, AF; competence frustration, CF; and relatedness frustration, RF).

To test nomological validity, we regressed the best-fitting BPNSF-CA model on the three manifest items of stress, anxiety, and fairness in a single structural equation model. We examine the same goodness-of-fit indicators and report the associations between BPNSF-CA and the indicators of well-being.

H2. The G-factor will have the largest associations with indicators of student well-being. Specifically, we expect negative correlations with stress and anxiety and a positive correlation with fairness.

H3. The S-factors of autonomy AS, competence CS, and relatedness support RS will be positively associated with fairness and negatively associated with stress and anxiety, and the inverse will be true for the S-factors of autonomy AF, competence CF, and relatedness frustration RF. These correlations may be reduced in magnitude and/or differentiated because of the disaggregated effects of the G-factor [24].

2.2. Results

All participants answered the self-report undergraduate student status through the self-report item, confirming Prolific's screening. Participants were equally divided between natural and social sciences and were most often thinking about final exams when they completed the BPNSF-CA (Supplementary Materials, Table S1).

Model comparisons indicated that the CFA did not achieve adequate fit (Table 2). The ESEM model demonstrated excellent fit and a substantial improvement over CFA ($\Delta\text{CFI} = +0.08$, $\Delta\text{TLI} = +0.10$, and $\Delta\text{RMSEA} = -0.056$). Item loadings showed well-defined factors with all but five items loading ≥ 0.50 on their targets, and most cross-loadings were small to moderate (Table 3: mean $|\lambda| = 0.09$; range $|0|$ to $|0.48|$). Composite reliabilities for ESEM factors ranged from $\omega = 0.56$ to 0.71 .

The bifactor ESEM also achieved excellent fit, with no measurable improvement in global fit indices over ESEM but yielding a clearly defined and highly reliable G-factor (item loadings $|\lambda| = 0.47\text{--}0.72$; $M|\lambda| = 0.59$; $\omega = 0.95$). All six S-factors retained meaningful residual specificity with reliabilities from $\omega = 0.79$ to 0.88 . Cross-loadings remained small to moderate (mean $|\lambda| = 0.11$; range $|0|$ to $|0.36|$). Given the strong G-factor and interpretable S-factors, we retained the bifactor-ESEM solution (see Supplementary Materials Tables S2–S4).

Table 2. Model fit indices for four measurement models of the BPNSF-CA.

Model	χ^2	<i>p</i>	<i>df</i>	SRMR	RMSEA (CIs)	CFI	TLI
CFA	554.05	<0.001	237	0.06	0.06 (0.05, 0.06)	0.92	0.90
Bifactor CFA	628.84	0.00	228	0.07	0.07 (0.06, 0.07)	0.90	0.87
ESEM	147.79	0.47	147	0.016	0.004 (0.000, 0.024)	1.00	1.00
Bifactor ESEM	129.67	0.47	129	0.013	0.004 (0.000, 0.025)	1.00	1.00

Notes. CFA = confirmatory factor analysis; ESEM = exploratory structural equation modeling; *df* = degrees of freedom; CFI = Comparative Fit Index; TLI = Tucker–Lewis index; RMSEA = Root Mean Square Error of Approximation; CI = confidence interval; and SRMR = Standardized Root Mean Square Residual.

Table 3. Factor loadings of the ESEM and bifactor ESEM of the BPNSF-CA.

Item	Exploratory Structural Equation Model Factor Loadings								Bifactor Exploratory Structural Equation Model Factor Loadings							
	AS λ	CS λ	RS λ	AF λ	CF	RF λ	δ		AS λ	CS λ	RS λ	AF λ	CF λ	RF λ	G- λ	δ
AS1	0.51	−0.06	0.34	−0.11	−0.02	0.12	0.46	0.37	−0.15	0.13	−0.01	0.20	0.24	0.24	0.53	0.45
AS2	0.18	0.000	0.44	−0.01	−0.15	−0.18	0.50	0.09	−0.08	0.23	0.07	0.03	−0.01	0.66	0.66	0.50
AS3	0.73	0.08	−0.001	−0.09	−0.004	0.09	0.39	0.67	0.02	0.08	−0.07	0.04	0.06	0.45	0.33	0.33
AS4	0.73	0.11	−0.08	−0.04	0.02	−0.09	0.38	0.57	0.04	−0.02	−0.02	0.07	−0.02	0.52	0.40	0.40
CS1	0.09	0.53	0.13	−0.04	0.11	−0.18	0.50	0.05	0.36	0.08	0.02	0.05	−0.12	0.58	0.50	0.50
CS2	−0.02	0.79	−0.002	−0.06	0.18	−0.02	0.44	−0.04	0.61	−0.04	−0.01	0.06	−0.05	0.47	0.41	0.41
CS3	0.05	0.59	0.11	−0.004	−0.31	0.16	0.43	0.003	0.42	0.03	0.03	−0.27	0.09	0.55	0.43	0.43
CS4	0.01	0.51	0.03	0.09	−0.40	−0.003	0.42	−0.06	0.38	−0.03	0.09	−0.36	−0.03	0.53	0.42	0.42
RS1	0.11	0.11	0.48	−0.28	0.03	0.06	0.47	0.11	−0.02	0.31	−0.13	0.14	0.13	0.61	0.46	0.46
RS2	−0.02	0.13	0.39	−0.14	0.08	−0.17	0.64	0.05	0.05	0.41	−0.08	0.05	−0.14	0.48	0.56	0.56
RS3	0.13	0.14	0.64	0.01	0.02	−0.11	0.35	0.08	0.000	0.40	0.09	0.17	0.03	0.67	0.34	0.34
RS4	0.28	0.07	0.45	0.02	0.01	−0.17	0.46	0.13	−0.05	0.16	0.12	0.22	0.07	0.67	0.44	0.44
AF1	0.01	0.02	−0.02	0.55	0.30	−0.01	0.47	−0.01	0.04	−0.03	0.36	0.27	−0.01	−0.58	0.46	0.46
AF2	−0.48	0.14	0.04	0.25	0.10	0.20	0.49	−0.28	0.18	0.23	0.10	−0.10	−0.04	−0.65	0.38	0.38
AF3	0.01	−0.09	−0.04	0.71	−0.09	−0.001	0.47	−0.05	−0.02	−0.07	0.46	−0.03	−0.01	−0.56	0.47	0.47
AF4	0.04	0.01	0.003	0.72	0.004	0.06	0.46	−0.003	0.05	−0.01	0.46	0.03	0.02	−0.56	0.47	0.47
CF1	0.03	−0.11	−0.02	0.10	0.64	0.13	0.31	0.15	−0.08	0.11	0.02	0.49	0.03	−0.65	0.30	0.30
CF2	−0.03	−0.23	−0.03	0.03	0.54	0.19	0.26	0.09	−0.19	0.08	−0.03	0.43	0.09	−0.71	0.25	0.25
CF3	−0.003	−0.15	0.06	0.11	0.57	0.20	0.31	0.05	−0.13	0.04	0.06	0.52	0.17	−0.62	0.29	0.29
CF4	−0.07	−0.02	0.08	0.18	0.70	−0.002	0.35	0.01	−0.02	0.10	0.11	0.57	−0.002	−0.55	0.35	0.35
RF1	−0.03	−0.14	−0.05	0.06	0.06	0.63	0.33	0.07	−0.09	0.04	−0.02	0.04	0.36	−0.72	0.33	0.33
RF2	0.04	0.15	−0.33	0.13	0.11	0.53	0.39	0.12	0.18	−0.13	0.01	−0.03	0.22	−0.70	0.39	0.39
RF3	−0.02	−0.07	0.15	0.001	0.02	0.74	0.44	0.04	−0.09	0.07	−0.02	0.10	0.52	−0.53	0.43	0.43
RF4	0.12	−0.01	−0.04	0.08	0.20	0.52	0.53	0.14	−0.02	−0.05	0.03	0.21	0.36	−0.54	0.51	0.51
ω	0.64	0.70	0.56	0.66	0.71	0.70	-	0.80	0.79	0.79	0.78	0.81	0.88	0.95	-	-

Notes. AS = autonomy support; CS = competence support; RS = relatedness support; AF = autonomy frustration; CF = competence frustration; RF = relatedness frustration; and G- λ = G-factor.

We tested nomological associations in a structural model that regressed stress, anxiety, and fairness on the G- and S-factors. This model fit adequately, χ^2 (180) = 216.48, p = 0.033, CFI = 0.99, TLI = 0.985, SRMR = 0.015, RMSEA = 0.023, 90% CI [0.007, 0.033]. As hypothesized, the G-factor was significantly associated in the expected directions with all three indicators of well-being: negatively with stress and anxiety, and positively with fairness (Table 4). Specific factors showed a differentiated pattern: CS and RS were positively related to fairness; AF and CF were positively related to stress and anxiety; RF was negatively related to fairness; remaining associations were non-significant once G was accounted for (Supplementary Materials Table S5).

Table 4. Associations with Indicators of Student Well-being.

Factors	Stress		Anxiety		Fairness	
	β	CI	β	CI	β	CI
G-factor	−0.63	−0.51, −0.75	−0.67	−0.52, −0.82	0.64	0.54, 0.74
Specific factors						
AS	−0.10	−0.24, 0.04	−0.07	−0.27, 0.13	−0.01	−0.14, 0.12
CS	0.08	−0.03, 0.19	0.02	−0.12, 0.16	0.29	0.19, 0.40
RS	−0.12	−0.35, 0.12	0.12	−0.47, 0.24	0.15	0.01, 0.29
AF	0.44	0.31, 0.58	0.34	0.17, 0.52	0.07	−0.03, 0.16
CF	0.29	0.16, 0.42	0.53	0.36, 0.71	−0.11	−0.23, 0.01
RF	−0.13	−0.32, 0.07	−0.06	−0.32, 0.20	−0.16	−0.23, −0.001

Notes. G-factor = global need fulfillment; AS = autonomy support; CS = competence support; RS = relatedness support; AF = autonomy frustration; CF = competence frustration; RF = relatedness frustration; β = standardized weight; CI = confidence interval. **Bold** indicates statistical significance.

2.3. Brief Discussion and Limitations

The results of Study 1 show that the BPNSF-CA is an adequate domain-specific measure of BPN satisfaction and frustration. The results affirm current theorizing and empirical studies showing that basic psychological needs are best described through a single G-factor and S-factors (e.g., [23,25]). While we contribute to the proliferation of measurement tools cautiously, we believe the BPNSF-CA is an important addition because assessment can be viewed as distinct from instructional classroom experiences because of its bounded and consequential nature. Particularly for theoretically grounded well-being research to become more infused in the design of assessment, a tool like the BPNSF-CA is critical.

One limitation of this study is that students responded to the BPNSF-CA for a recalled course. This means students' responses reflected a diversity of content areas and assessments. While a limitation, this also means that the BPNSF-CA functioned well under conditions of diversity, thereby suggesting robustness to the items. In Study 2, we sought to expand on Study 1 by using the BPNSF-CA as a dependent variable, measuring the effects of four theorized need-supportive features on MCQ exams specifically.

3. Study 2: Experimental Test of Need-Supportive Features

The purpose of Study 2 was to test how multiple-choice tests of varying levels of item quality and need-supportive features predicted BPN fulfillment and exam performance. To achieve this, we used an experimental 1×3 between-subject design with random assignment of participants to one of three different test conditions. The BPNSF-CA and test score served as the main outcome variables.

3.1. Method

3.1.1. Participants

We recruited 400 psychology students (M age = 29.00, SD = 8.57; 73% women) via Prolific, restricting eligibility to current students in the United States, Canada, the United Kingdom, or Australia with $\geq 95\%$ approval ratings (total available N = 1108). We chose these additional inclusion criteria because completion of a test requires attention and is somewhat higher risk than a basic survey. Following data-quality checks (attention checks, response time, patterns, etc.; see Supplementary Materials on data adequacy), the final sample consisted of 387 university students (73% women; M age = 29, SD = 8.57) who were pursuing either a certificate (n = 16), or a bachelor's (n = 201), master's (n = 102), or PhD degree (n = 67) in psychology.

3.1.2. Procedures

Ethics approval was obtained from the university research ethics board (Pro00141097). Participants completed a pre-test motivation survey and were then randomly assigned to one of three experimental conditions: Test A (flawed items), Test B (high-quality items), or Test C (high-quality items + need-supportive features). After completing the 20 MCQ test, participants completed two manipulation checks, responded to the full 24-item BPNSF-CA, and provided demographic information. Due to the fact that students had to take a test, remuneration was set at GBP 17.96/h, which is considered excellent.

3.1.3. Materials

The pre-test motivation survey (see Supplementary Materials) contained nine items written by the researchers to measure participants' level of motivation for completing the test in terms of effort, value, and expected performance (1 = strongly disagree to 5 = strongly agree). Although drawn from distinct theories, the nine items were summed

into a single motivation score ($\alpha = 0.80$, $M = 37.35$, and $SD = 3.98$), which was used to ensure no difference in effortful investment on the exam between groups.

We created three isomorphic 20-item psychology MCQ tests that served as the independent variable. Test A included common item-writing flaws (e.g., verbose stems, implausible distractors, ≥ 4 options, negative stems, all-of-the-above/complex combinations). Test B revised each item to adhere to established guidelines [5]. See Supplementary Materials Table S6 for samples of revised test items. Test C used the same high-quality items as Test B and added four need-supportive features: a test blueprint, item grouping matching the blueprint, brief supportive messages at start and end, and a feedback opportunity. Figure 2 shows the layout of the exam with these features. The questions dealt with basic psychology content in the domains of development, psychopathology, social, and methods, and were written at an introductory level.

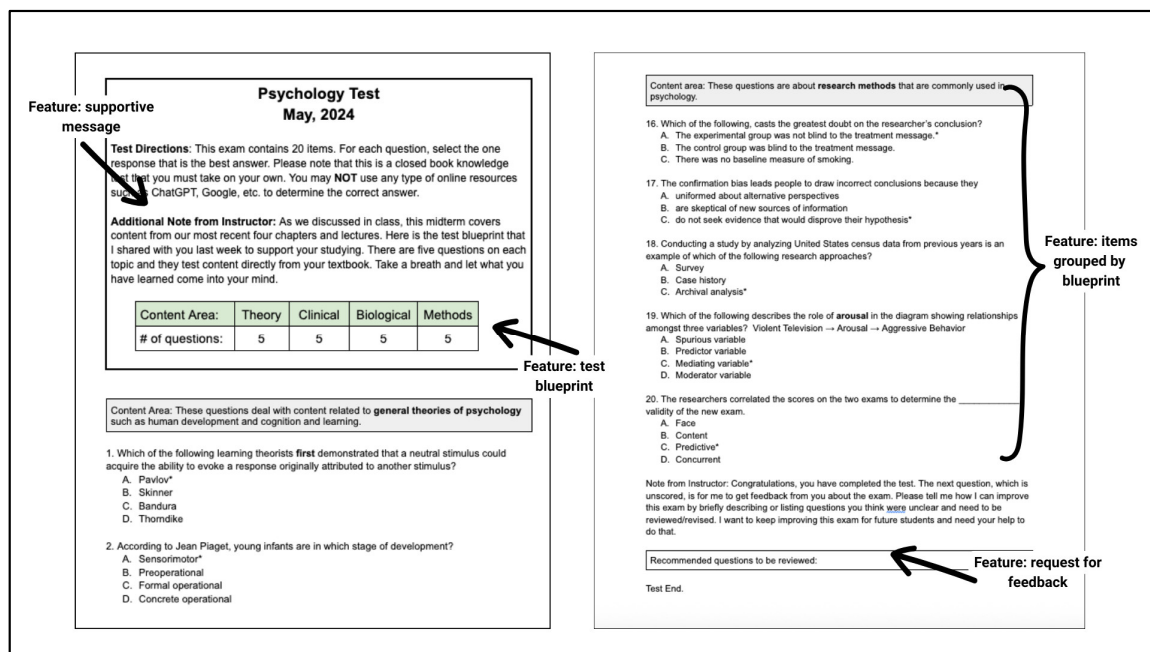


Figure 2. Sample of exam pages, including the four need-supportive features. Asterisks denote the correct answer to the item.

The post-test survey contained two questions that were designed as manipulation checks: an estimate of performance and perceived quality of test design (1 = very poor quality, 2 = poor, 3 = average, 4 = good, and 5 = excellent quality). The dependent variables were the BPNSF-CA and test performance calculated by summing the test items participants answered correctly (1 = correct; 0 = incorrect; and maximum = 20).

3.1.4. Power Analysis

An a priori power analysis using G*Power 3.1 (fixed-effects omnibus for three groups) indicated that $n = 390$ would be required to detect a small-to-medium effect ($f = 0.20$) with $\alpha = 0.05$ and power = 0.80. In addition, a Monte Carlo simulation for the bifactor measurement model, using Study 1 estimates (G loadings ≈ 0.59 ; S loadings ≈ 0.40), suggested that $n \approx 130$ per condition (total $N \approx 390$) would ensure adequate convergence and precision. We therefore targeted $n = 400$ to mitigate attrition.

3.1.5. Plan for Analyses and Hypotheses

As preliminary analyses, we tested for group equivalence on pre-test motivation using analyses of variance (ANOVA). We also confirmed the structure of the BPNSF-CA by

testing four competing models following the same steps as Study 1. Finally, we examined participants' responses to the manipulation check questions.

The hypotheses and main analysis plan were pre-registered at AsPredicted (https://aspredicted.org/GPJ_Z2L on 31 May 2024). We used multiple indicators, multiple causes (MIMIC) models to compare latent means across conditions in a sequence of null, saturated, and invariant models. MIMIC approaches are more appropriate than multi-group approaches to measurement invariance when testing multiple contrast variables, such as experimental/control groups, and with relatively small sample sizes [49–51]. Interpretation of MIMIC models follows common standards of good model fit with CFI and TLI ≥ 0.95 and RMSEA and SRMR ≤ 0.05 , and an adequate model fit at CFI and TLI ≥ 0.90 and RMSEA and SRMR ≤ 0.07 [49].

Following [50], we used R to run three increasingly restrictive models that determine the effect of the independent variable on the G-factor and six S-factors of the BPNSF-CA. First, we ran a null model in which the independent variable is treated as having no effect on the latent means and items' intercepts. If this model has a poor fit, it means there is an effect of the test condition on the outcome that has not yet been modeled. Second, we ran a saturated model in which the independent variable is allowed to influence all items' intercepts, but not the latent means. If this model has a poor fit, it means the test condition variable has an effect that may extend to the latent means as not yet modeled. Third, we ran an invariant model that allows the independent variable to influence all latent means but not items' intercepts. Evidence that the test condition has a significant effect on the latent means is found when the invariant model provides a better fit to the data than the null or saturated models [50]. To determine the nature of that association, the individual path weights are interpreted. We repeated this process three times to make all comparisons between conditions: Test A compared to B; Test A compared to C; and Test B compared to C.

H4. *Test A will have a negative effect on students' global and specific need satisfaction and a positive effect on specific need frustration relative to Test B and Test C (H4a). Test B will have a negative effect on students' global and specific need satisfaction and a positive effect on specific need frustration relative to Test C (H4b). In both instances, we recognize that the effects on S-factors may be diminished because of the G-factor.*

To compare students' item-level performance across the three levels of the independent variable, we employed generalized linear mixed-effects models (GLMMs) [52]. GLMMs enhance the precision of standard errors by taking both fixed-effects (i.e., representing the effects of experimental manipulations) and random-effects of students and items (i.e., random intercepts) into account. The control condition (i.e., Test A) was used as the reference level (i.e., intercept). All models were run in R, using the lme4 package [53].

H5a. *Students will perform better on Test B and Test C (together and individually) compared to Test A.*

H5b. *Students will perform better on Test C compared to Test B.*

3.2. Results

According to ANOVA, there were no significant differences in participants' motivation for completing the test at baseline: $F(2, 383) = 0.79, p = 0.46$. We tested the fit of four competing measurement models, CFA, bifactor, ESEM, and bifactor ESEM, using the same process as described in Study 1 (Supplementary Materials Tables S7 and S8). We retained the bifactor ESEM as the best-fitting and most meaningful model. On the post-test

manipulation check questions, participants did not differ significantly between conditions in the number of questions they believed they answered correctly, $F(2, 381) = 1.27$, $p = 0.28$. However, participants who completed Test C did perceive the test as significantly higher quality than participants who completed Test A or B, $F(2, 383) = 3.27$, $p < 0.05$. The goodness-of-fit results from the MIMIC models are reported in Table 5. We were unable to meaningfully consider the effects of Test A vs. Test B on students' need fulfillment as they pertain to H4a, because although the model successfully converged, it was uninterpretable in terms of fit indices.

Table 5. Goodness-of-fit indices for MIMIC models.

Model	χ^2	p	df	CFI	TLI	SRMR	RMSEA (CIs)	AIC	BIC	ABIC
Model 1: Test A vs. Test B on BPNSF-CA ^a										
Null	217.139	1.00	331	1.00	1.053	0.052	0.00 (0.000, 0.000)	16,380.88	16,441.28	16,387.389
Saturated	213.509	1.00	324	1.00	1.052	0.054	0.00 (0.000, 0.000)	16,391.95	16,477.22	16,401.134
Invariant	220.591	1.00	341	1.00	1.054	0.055	0.00 (0.000, 0.000)	16,364.91	16,389.79	16,367.592
Model 2: Test A vs. Test C on BPNSF-CA										
Null	262.577	<0.001	153	0.951	0.905	0.049	0.053 (0.042, 0.063)	16,390.52	17,082.59	16,464.377
Saturated	208.156	<0.001	129	0.965	0.918	0.021	0.049 (0.036, 0.061)	16,386.12	17,163.36	16,469.066
Invariant	250.08	<0.001	146	0.954	0.905	0.024	0.053 (0.041, 0.064)	16,377.84	17,094.76	16,454.354
Part.Invar. ^b	220.068	<0.001	142	0.965	0.927	0.022	0.046 (0.034, 0.058)	16,368.42	17,099.52	16,446.446
Model 3: Test B vs. Test C on BPNSF-CA										
Null	202.554	0.005	153	0.984	0.970	0.039	0.029 (0.017, 0.039)	24,566.78	25,338.17	24,719.455
Saturated	150.244	0.097	129	0.993	0.985	0.016	0.021 (0.000, 0.034)	24,562.33	25,428.66	24,733.795
Invariant	172.564	0.066	146	0.992	0.983	0.018	0.022 (0.000, 0.034)	24,549.89	25,348.97	24,708.044

Notes. ^a BPNSF-CA is included in the model as a bifactor ESEM with six S-factors and one global G-factor.

^b Part.Invar. = partial invariant model.

For Test A vs. Test C, the saturated model marginally improved fit over the invariant model (Table 6), indicating monotonic differential item functioning concentrated on AS1, AS2, RF1, and RF2. Allowing direct effects for these items yielded a partial-invariant model. In this model, the grouping variable (1 = Test A, 2 = Test C) positively predicted the G-factor (Table 7; $\beta = 0.223$, $p = 0.004$), indicating higher global need fulfillment in response to Test C; S-factor differences were non-significant. For Test B vs. Test C, the invariant and saturated models were equivalent; constraining effects to latent means was justified. Test C showed a higher G-factor ($\beta = 0.180$ and $p = 0.006$) and autonomy satisfaction ($\beta = 0.145$ and $p = 0.049$), and lower relatedness frustration ($\beta = -0.162$ and $p = 0.023$) than Test B.

Across all participants, descriptive performance was $M = 12.04$ and $SD = 3.26$. Collapsing across high-quality conditions, students in Tests B and C with high-quality items answered about three more items correctly ($M \approx 13$, $SD = 2.85$) than those in Test A ($M \approx 10$, $SD = 3.27$). Item-level GLMMs with random intercepts for students and items confirmed these differences (Table 8). Using Test A as the reference, the fixed effect for Quality (Test B) was $\gamma_1 = 0.684$, $SE = 0.09$, and $z = 7.285$, where $p < 0.001$; the fixed effect for Quality + BPN (Test C) was $\gamma_2 = 0.719$, $SE = 0.09$, and $z = 7.662$, where $p < 0.001$. In a contrast model comparing Test B vs. Test C, the fixed effect for Quality + BPN was $\gamma = 0.038$, $SE = 0.10$, $z = 0.390$, $p = 0.697$, indicating no additional performance benefit of need-supportive features beyond item quality. Random-effect variances were 0.315 ($SD = 0.561$) for students and 0.751 ($SD = 0.866$) for items in the omnibus model, reflecting meaningful clustering at both levels. These results align with the pre-registered hypotheses.

Table 6. Change in goodness-of-fit indices for MIMIC model comparisons.

Model Comparisons	ΔCFI	ΔTLI	$\Delta RMSEA$	ΔAIC	ΔBIC
Model 2: Test A vs. Test C on BPNSF-CA ^a					
Null-Saturated ^b	0.014	0.013	−0.004	−4.401	80.776
Null-Invariant	0.003	0	0	−12.675	12.169
Saturated-Invariant	−0.011	−0.013	0.004	−8.274	−68.607
Saturated-Partial	0	0.009	−0.003	−17.696	−63.839
Model 3: Test B vs. Test C on BPNSF-CA					
Null-Saturated	0.009	0.015	−0.008	−4.451	90.489
Null-Invariant	0.008	0.013	−0.007	−16.891	10.799
Saturated-Invariant	−0.001	−0.002	0.001	−12.44	−79.69

Notes. ^a BPNSF-CA is included in the model as a bifactor ESEM with six S-factors, AS, CS, RS, AF, CF, and RF, and one global G-factor. ^b Null model has no estimates between the test condition and BPNSF-CA. Saturated model tests effects on intercepts. Partial = partial invariant model in which some paths have been freed to achieve an acceptable model fit.

Table 7. Effect of experimental conditions on BPNSF-CA.

Outcome	Model 2 Effects of Test A vs. Test C			Model 3 Effects of Test B vs. Test C		
	β	p	R^2	β	p	R^2
G-factor	0.223	0.004	0.05	0.180	0.006	0.032
S-factors:						
AS	0.090	0.338	0.008	0.145	0.049	0.021
RS	0.107	0.193	0.012	0.105	0.090	0.011
CS	−0.004	0.98	0.000	−0.005	0.951	0.000
AF	0.036	0.667	0.001	0.051	0.455	0.003
RF	−0.158	0.181	0.025	−0.162	0.023	0.026
CF	−0.055	0.652	0.003	−0.059	0.524	0.004

Notes. G-factor = global need fulfillment; S-factors AS = autonomy support; CS = competence support; RS = relatedness support; AF = autonomy frustration; CF = competence frustration; RF = relatedness frustration; β = standardized weight. **Bold** indicates statistical significance.

3.3. Brief Discussion and Limitations

The results of the current study show that only Test C, which included both high-quality items and need-supportive features, significantly enhanced students' global need fulfillment. Additional specific effects emerged where students reported higher specific autonomy satisfaction and lower relatedness frustration in Test C than Test B above and beyond increases in global need fulfillment. In contrast, comparisons between Test A and Test C revealed differences only at the global level, perhaps suggesting that the magnitude of difference between a flawed exam and a redesigned exam was so pronounced that only the G-factor captured the shift. Importantly, these findings suggest that the need-supportive features provide a coordinated set of affordances that shape students' overall experience of the assessment in ways that are most readily captured by global need fulfillment rather than by isolated, feature-specific effects.

Participants completing high-quality tests (B or C) answered, on average, three more items correctly ($M \approx 13$) than those completing flawed items ($M \approx 10$), aligning with current psychometric stances that quality items always improve test scores [39]. Yet there were no statistically significant differences in performance between Tests B and C. This pattern of results could mean that intentionally need-supportive features do not further enhance performance, which might suggest to some instructors that these features are unnecessary. However, it also supports an interpretation that need-supportive features, as introduced in Test C, do not hinder performance, meaning instructors can safely add them without fear of distracting students from test content. Both possibilities need to recognize

an important limitation: regarding the observation of improved performance in Tests B and C compared to Test A, the findings may also be driven to some extent by higher chances of a priori guessing due to a reduction in implausible distractors. This interpretation needs to be tested in future research.

Table 8. Model parameters of the generalized linear mixed-effects models predicting student performance across the experimental conditions.

Model 1 (Tests A vs. B vs. C)				
Fixed Effects	Est.	SE	z	p
Intercept (Control) γ_0	0.067	0.20	0.329	0.742
Quality γ_1	0.684	0.09	7.285	<0.001
Quality + BPN γ_2	0.719	0.09	7.662	<0.001
Random Effects	VAR (<i>SD</i>)			
Students (Random Intercept)	0.315 (0.561)			
Items (Random Intercept)	0.751 (0.866)			
Number of Observations	7740			
Number of Students (Items)	387 (20)			
Model 2 (Tests A vs. B and C)				
Fixed Effects	Est.	SE	z	p
Intercept (Control) γ_0	0.067	0.20	0.329	0.742
Quality + Quality + BPN γ_1	0.701	0.08	8.643	<0.001
Random Effects	VAR (<i>SD</i>)			
Students (Random Intercept)	0.315 (0.561)			
Items (Random Intercept)	0.751 (0.866)			
Number of Observations	7740			
Number of Students (Items)	387 (20)			
Model 3 (Tests B vs. C)				
Fixed Effects	Est.	SE	z	p
Intercept (Quality) γ_0	0.808	0.25	3.287	<0.001
Quality + BPN γ_1	0.038	0.10	0.390	0.697
Random Effects	VAR (<i>SD</i>)			
Students (Random Intercept)	0.339 (0.582)			
Items (Random Intercept)	1.110 (1.053)			
Number of Observations	5160			
Number of Students (Items)	258 (20)			

Notes. Est. = estimate; SE = standard error; Intercept = Test A (control) as reference (Model 1 and 2); Model 3 does not include the control and refers to Test B (Quality) as the intercept; Control = Test A; Quality = Test B; Quality + BPN = Test C.

The most pressing limitation of Study 2 was that the setting lacked ecological validity. Indeed, it could be argued that because the study involved psychology students outside of a real classroom, the exam experience was relatively inconsequential. Readers may worry that, given the artificial setting, where students were paid to take an exam, the results are artificial. We would have a similar concern if not for the rigor of the true experimental design and appropriate powering, which create the controlled conditions to have confidence in the findings. Nonetheless, to overcome this limitation in Study 3, we embedded the intervention into an ecologically valid classroom where the exams have meaningful consequences for students.

4. Study 3: An Ecological Classroom-Based Quasi-Experiment

The purpose of Study 3 was to embed the MCQ design intervention in an ecological context and test the hypothesized sequence that improvements in indicators of well-being were mediated by BPN fulfillment. To achieve this, we used a two-group post-test only quasi-experimental design in which students in two sections of actual psychology classes completed either unmodified MCQ tests or high-quality, need-fulfilling exams. Each semester, the course involved two midterm MCQ tests and a cumulative final exam.

4.1. Method

4.1.1. Participants

Two cohorts of an upper-level undergraduate psychology course (enrollment per cohort = 300) taught by the same instructor served as control and experimental groups. From the September cohort (control), 53 students participated, and from the January cohort (experimental), 48 participated. Participation was voluntary, and students were entered into a draw for a CND 50 gift card. All participants provided informed consent.

4.1.2. Procedures

All procedures were approved by the institutional ethics board (Pro00142406). Students in the September cohort completed the instructor's original MCQ exams, while students in the January cohort completed redesigned MCQ exams consisting of high-quality items and the four need-supportive features that we applied in Study 2. Students wrote three course exams (two midterms and a final), indicating their responses on a scannable bubble sheet. After each of the three exams, students had an opportunity to complete a survey assessing BPN satisfaction, stress, anxiety, perceptions of fairness, and perceived success. To maximize response rates, we used the first survey response from each student regardless of the examination as data.

4.1.3. Materials

The independent variable consisted of two versions of actual course exams. In the control group, we used the instructor's original MCQ exams as treatment as usual. In the experimental group, the redesigned MCQ exams consisted of high-quality items [5], and the four need-supportive features tested in Study 2 are as follows: blueprint, grouping, supportive messaging, and feedback. The exams tested in Study 3 essentially map onto Test A and Test C from Study 2. Students were provided with the blueprint before the exam as a study guide. See Supplementary Material for information on the tests.

Dependent variables were measured after each exam. The post-test survey contained a reduced six-item BPNSF-CA (one item per AS, CS, RS, AF, CF, and RF), which was summed into a single indicator of need fulfillment after reverse scoring the frustration items ($\alpha = 0.65$). For well-being, we used the same single-item measures of stress, anxiety, and fairness as in Study 1. We asked participants to indicate their perceived success by responding to the item: "I felt successful on this assessment." on a 1 Strongly Disagree to 5 Strongly Agree Likert scale.

4.1.4. Power Analysis

We conducted an a priori power analysis for mean differences between two independent groups using G*Power3.1. It indicated that at least 45 participants per group were needed to achieve 95% power to detect $d = 0.70$ ($\alpha = 0.05$; Faul et al. [54]). As such, the number of participants was adequate for the analyses even if it represented a small portion of the full number of students in the course.

4.1.5. Plan for Analyses and Hypotheses

As preliminary analyses, we examined descriptive statistics and intercorrelations. As main analyses, we conducted five independent-samples *t*-tests comparing the control and experimental groups on BPN satisfaction, stress, anxiety, fairness, and perceived success. Although our analyses were not pre-registered, we had a priori hypotheses and thus did not use a correction procedure.

H6. *The experimental group will report significantly more need fulfillment, well-being, and perceived success than the control group.*

Then, we tested four mediation models in JASP [55] using the structural equation modeling function and maximum likelihood estimator. JASP-SEM estimates direct, indirect, and total effects within a path-analytic framework and therefore does not require fit indices for interpretation. Total BPNSF-CA was specified as the mediator between the predictor (group 0 = control; 1 = experimental) and well-being outcomes, and indirect effects were evaluated using bootstrap 95% confidence intervals.

H7. *The effect of group on indicators of well-being will be partially mediated by need fulfillment.*

4.2. Results

As evidence of the validity of the single-item measures, perceptions of fairness and success were significantly positively correlated, as were stress and anxiety (Table 9). BPN fulfillment was also significantly associated with all outcomes in the expected directions.

Table 9. Zero-order correlations for Study 3 variables.

Variable	1	2	3	4	5
1. BPN Satisfaction	—				
2. Stress	−0.41 ***	—			
3. Anxiety	−0.20	0.55 ***	—		
4. Fairness	0.67 ***	−0.26 *	−0.06	—	
5. Perceived Success	0.55 ***	−0.22 *	−0.11	0.54 ***	—

Notes. Pearson's *r*. Two-tailed significance: * $p < 0.05$, and *** $p < 0.001$.

As shown in Table 10, results of the independent samples *t*-tests indicated that the experimental group reported significantly higher BPN satisfaction, fairness, and perceived success than the control group. Differences in stress were marginal ($p = 0.068$), and anxiety did not differ significantly ($p = 0.727$) between groups.

Direct, indirect, and total effects are in Table 11. Mediation analyses in SEM showed significant indirect effects of group through BPN satisfaction on fairness and perceived success. The indirect effect on stress approached significance, with confidence intervals just crossing zero. The anxiety pathway was non-significant. These results indicate that redesigned exams primarily enhance positive indicators of well-being via need satisfaction rather than reducing negative indicators.

Table 10. Descriptive statistics and *t*-tests comparing control and experimental groups.

Variable	Control <i>M</i> (<i>SD</i>)	Experimental <i>M</i> (<i>SD</i>)	<i>t</i> (<i>df</i>)	<i>p</i>	<i>d</i>
BPN Satisfaction	3.28 (0.63)	3.57 (0.70)	2.19 (94) *	0.031	0.45
Stress	3.21 (1.08)	2.77 (1.24)	1.85 (95)	0.068	−0.38

Table 10. Cont.

Variable	Control M (SD)	Experimental M (SD)	t (df)	p	d
Anxiety	3.85 (1.10)	3.77 (1.03)	0.35 (95)	0.727	−0.07
Fairness	3.81 (1.13)	4.36 (0.92)	2.61 (95) *	0.011	0.53
Perceived Success	3.11 (0.87)	3.52 (1.02)	2.13 (95) *	0.036	0.44

Notes. Control $n = 43$ and experimental $n = 53$. Cohen's d values represent standardized mean differences.

* $p < 0.05$.

Table 11. Direct, indirect, and total effects of group on well-being through BPN fulfillment.

Outcome Variable	Effect Type	Estimate	SE	z	p	95% CI LL	95% CI UL
Stress	Direct	−0.20	0.19	−1.08	0.28	−0.167	0.573
	Indirect	−0.17	0.09	−1.94	0.05	−0.002	0.340
	Total	−0.37	0.20	−1.87	0.06	−0.018	0.763
Anxiety	Direct	0.02	0.20	0.08	0.93	−0.416	0.383
	Indirect	−0.09	0.06	−1.47	0.14	−0.029	0.206
	Total	−0.07	0.20	−0.35	0.72	−0.326	0.469
Fairness	Direct	0.25	0.16	1.55	0.12	−0.574	0.068
	Indirect	0.30	0.14	2.12	0.03	−0.575	−0.023
	Total	0.55	0.21	2.64	0.01	−0.963	−0.142
Success	Direct	0.21	0.17	1.19	0.23	−0.542	0.133
	Indirect	0.22	0.11	2.00	0.05	−0.440	−0.004
	Total	0.43	0.20	2.15	0.03	−0.815	−0.038

Notes. Group 0 = control and 1 = experimental. Indirect effects represent mediation through BPN satisfaction. SE = standard error; CI = confidence interval; LL = lower limit; and UL = upper limit.

4.3. Brief Discussion and Limitations

The purpose of Study 3 was to address the lack of ecological validity in Studies 1 and 2. Replicating the findings of Study 2, the addition of a coordinated set of need-supportive features to MCQ tests with high-quality items resulted in a significant increase in BPN fulfillment in the experimental group relative to the control group. This finding, however, used a summed scale of BPNs rather than a bifactor ESEM with G-factor, a clear reduction in measurement sophistication [56] relative to Study 2. Although it is a limitation of Study 3, it was necessary to shorten the post-test survey for use with students in a classroom setting with multiple administration sessions.

In terms of the indicators of well-being, we also detected significant differences between the two groups for perceptions of fairness and success, but not for the negative indicators. In turn, it was perceptions of fairness and success that were also mediated by BPN fulfillment. Overall, this suggests that the collective addition of need-supportive features to high-quality exams may have uneven effects on positive or negative indicators of well-being. One explanation for this is that the group of need-supportive features was insufficient in reducing need frustration. Future research will want to bring greater precision to understanding how the need-supportive features function. Moreover, this imbalance highlights the need for future research to continue to use both positively and negatively valenced indicators of student well-being, as recommended by Collie and Hascher [30].

In addition to compromising on measurement specification, two other limitations of Study 3 must be addressed. First, Study 3 was not pre-registered. Second, we did not have a baseline measure of equivalence in the two-group quasi-experimental design, thereby reducing causal confidence. In future research, a composite and generic motivation scale such as the one we used in Study 2 could be added to classroom-based research without interacting with the intervention itself. While these limitations are valid, the consistency in results across the three studies is reassuring, as we discuss next.

5. General Discussion

Across three studies, we tested the premise that applying a coordinated set of need-supportive features to multiple-choice tests can bring about gains in students' well-being and performance via the fulfillment of general and specific BPN. In this general discussion, we focus on how the progression from measurement to causal inference to ecological mediation cumulatively offers emerging evidence for starting to rethink how assessments can be designed to simultaneously ensure rigor and support well-being. We discuss implications for theory, research, and practice before dealing with limitations.

5.1. Domain-Specific Measurement and Benefits of Bifactor

Both Study 1 and Study 2 provided validity evidence for the creation of a domain-specific measure of need satisfaction and frustration in classroom assessment. The evidence of reliability and validity of the BPNSF-CA should allow researchers to confidently move forward with domain-specific investigations connecting assessment with students' need fulfillment. Importantly, Study 3 showed that even a reduced version of the BPNSF-CA retained predictive validity in authentic classroom contexts, underscoring its flexibility for applied research. As a caveat, the reliability of the six-item BPNSF-CA was less than ideal, suggesting that additional psychometric testing should be carried out to confirm the best items to form a shortened version [57].

For Studies 1 and 2, we retained the bifactor model for two primary reasons: (a) factor loadings revealed a clearly defined single G-factor that captured shared variance in BPNs, while S-factors retained meaningful residual specificity; and (b) the G-factor demonstrated the expected associations with indicators of well-being, allowing for a differentiated pattern of S-factor correlations to emerge. This dual evidence of clear global structure plus differentiated specific effects strengthens the case for bifactor modeling [47]. By retaining the bifactor ESEM, we found associations that likely would have been obscured in traditional CFA or ESEM solutions. Beyond its psychometric advantages, the bifactor structure also carries theoretical implications for intervention research. In contexts such as classroom assessment, where students' experiences are shaped by multiple, simultaneous design features, the G-factor may best represent students' holistic appraisal of being supported versus pressured. From this perspective, assessment interventions may be expected to exert their strongest effects at the level of global need fulfillment, with specific needs being further satisfied or frustrated by assessment features that are particularly attuned to the psychological mechanisms of the specific need.

5.2. Redesigning Multiple-Choice Exams for Well-Being

Studies 2 and 3 extended the measurement work into an intervention. The two studies offset each other in terms of strengths and limitations: Study 2 used sophisticated pre-registered analyses with a large sample and a randomized true experiment that lacked ecological validity, whereas Study 3 made compromises in terms of method and measurement to test the intervention with real students taking consequential exams. What they shared, however, was a high-fidelity implementation of the intervention aligned with our theorizing, namely that high-quality items [5] paired with four need-supportive features would result in measurable improvements in students' BPNs and indicators of well-being. The results from the two studies provide preliminary evidence that this theorizing is appropriate and that the effects are cautiously promising.

The work of improving the quality of MCQ items is tedious, but the standards and processes are clear in the literature [5]. Thus, although it will be important to find effective ways to support instructors in improving items, the guidance on how to do so is readily available. It is plausible that recent advances in artificial intelligence and training of

models may help reduce the burden of improving items, perhaps even automating this step of the intervention [58]. We cannot make the same comments for need-supportive features because there is essentially no body of empirical literature on BPN-supportive and -frustrating practices in assessment, making this a particularly novel contribution of the current research.

For this intervention, we translated common BPN-supportive instructional principles [40] into four assessment features: blueprint, grouping, supportive messaging, and a chance to provide feedback. Thus, the effects of the intervention represent these four features as a unit, meaning we were unable to discern how each feature individually would align with either the G-factor or S-factors. It is common for BPN interventions to contain multiple components that can be classified according to specific BPN while exerting broader effects [40,59]. While from a theoretical perspective there may be a desire to test each feature on its own, continuing with a coordinated set of design features is more in keeping with the way in which instructors design assessment (and students experience assessment) [42]. By extension, the results from Studies 2 and 3 suggest that while the four features exerted important effects, they did not adequately reduce S-factors of frustration or improve negative indicators of well-being. To tackle this, we suggest future research follows the model of Ahmadi et al. [40] and uses a Delphi study or other similar approach to create a large list of features that are agreed on by experts as likely to increase satisfaction and decrease frustration, and then test those in systematic ways. Because assessment has such a high-stakes setting, we recognize that some level of ill-being may always exist, but additional features may help minimize it.

5.3. Needs Satisfaction as a Mechanism to Well-Being

In addition to its ecological embedding in a real psychology course, Study 3 provided us with a chance to test the theorized mediational mechanism between the coordinated set of need-supportive features and indicators of well-being through BPNs. BPN fulfillment significantly mediated the intervention's effects on fairness and perceived success, but indirect effects for stress and anxiety were weaker or non-significant. This pattern is like the asymmetry noted in the earlier studies by which the BPN S-factors were differentially associated with positively valenced indicators of well-being than negative ones. The mediational design allows us to consider additional explanations for these earlier results. For example, it is plausible that fairness perceptions and feelings of success are proximal cognitive appraisals of the immediate assessment context. Students may more readily recognize and internalize fairness and accomplishment when a specific exam seems to support their autonomy, competence, and relatedness. Stress and anxiety, by contrast, are diffuse affective states shaped by multiple contextual and personal factors (e.g., workload, trait dispositions, and external stressors) and may require additional or more direct supports (e.g., counseling and workload adjustments) to make a meaningful shift [60]. This nuanced pattern underscores an important possible implication: at the very least, as tested in these studies, the need-supportive design of MCQ exams appears to be better suited to enhancing the pleasantness of the test in terms of fairness and success than reducing stress or anxiety. The finding aligns with prior SDT evidence suggesting that need-supportive practices primarily bolster adaptive motivational states, whereas reductions in maladaptive states often require complementary interventions [13].

5.4. Implications for Theory, Research, and Practice

Theoretically, this program of research demonstrates the value of integrating psychometric views on assessment and robust psychological theories. It is rare for test developers and motivation researchers to collaborate because their programs of research tend to focus

on different outcomes, methods, and dissemination outlets. The results of these studies show that combining these approaches may help effectively deal with well-being in the domain of assessment. We encourage continued theorizing that involves both quality assessment design and psychological frameworks [15].

For research, the progression from scale validation to causal inference to ecological mediation illustrates the importance of triangulating methods to build cumulative evidence [61]. The evidence from the three studies shows that the BPNSF-CA provides a robust foundation for testing assessment-related need fulfillment. We encourage researchers to think deeply about the structure of measurement tools and to consider the potential of G- and S-factors whenever possible, while also recognizing that research in ecological contexts sometimes requires simpler measurement approaches [57]. Future research can build on our openly archived materials and code to identify and test the effect of need-supportive features on other forms of assessment, such as essays, labs, rubrics, presentations, etc. Moreover, longitudinal research is required to explore whether repeated exposure to high-quality need-supportive assessments within and across courses compounds their effects.

For practice, the findings point to four actionable, low-cost modifications instructors can apply to the design of MCQ examinations without compromising rigor: sharing a test blueprint, aligning items with that blueprint, providing supportive messaging, and inviting student feedback. The results of these studies, even with their limitations, suggest that there is no obvious drawback to these need-supportive features in terms of student well-being or performance. The simplicity of these modifications is important because motivation interventions have been shown to be more effective if they are perceived as easy to do [62].

5.5. Limitations and Future Directions

While the results of these three studies represent an important advancement, they must be considered in light of three key limitations. First, although bifactor ESEM provided a powerful and theoretically appropriate framework for modeling students' basic psychological needs in the assessment domain, this approach also introduces methodological burdens. Bifactor models require large samples to achieve stable parameter estimates [45] and are often paired with relatively lengthy instruments to ensure sufficient item coverage across global and specific factors [26,47]. Moreover, continuing with a bifactor solution in subsequent analyses, such as testing intervention effectiveness, is equally demanding. These requirements may pose challenges for ecological classroom studies, where instructional time is precious, sample sizes are smaller, and survey fatigue is a concern. Future research needs to continue exploring ways to balance methodological rigor with ecological feasibility in terms of measurement and design.

Second, our reliance on single items measuring stress, anxiety, fairness, and perceived success under the umbrella of well-being is also problematic. Although single-item measures can be pragmatic and have shown adequate validity in previous research [57], they inevitably prevent any measurement of reliability and restrict the conceptual richness of well-being. While our selected single items map onto several common aspects of well-being, follow Collie and Hascher's [30] recommendations, and showed appropriate correlations, they still limit the fullness of well-being. Future research may want to use qualitative methodologies to allow for richer descriptions.

Finally, in Studies 2 and 3, our intervention focused exclusively on multiple-choice tests. This focus was deliberate: given the prevalence of MCQ exams in higher education [8] and the well-documented links between examinations and compromised student well-being [3,28], we viewed exams as a particularly difficult space in which to test the intervention. In other words, we were stacking the odds against the intervention. As

such, the emerging success builds confidence in the possibilities for revising other formats of assessment. Nonetheless, all inferences from these three studies must be restricted to MCQ tests. Essays, presentations, lab reports, and other performance assessments also shape students' need satisfaction and well-being in the domain of assessment [63,64]. Our theory is that the technical quality of assessment, when paired with need-supportive features, can be applied to and tested with other assessment formats. Expanding beyond MCQ exams will be crucial for understanding the broader potential of need-supportive assessment design across the higher education landscape.

6. Conclusions

Taken together, these three studies demonstrate that the longstanding belief that multiple-choice tests are uniformly detrimental to student well-being may be somewhat simplistic. By applying a rich theory of psychological well-being to the design of multiple-choice tests, we found preliminary evidence of the contrary. Although there are important limitations to each study that require tempered conclusions, the studies suggest that the BPNSF-CA is a valid and useful tool for the domain of classroom assessment, that need-supportive features extracted from SDT can be effectively applied to exams, and that enhanced BPNs likely serve as a nutriment for well-being in the domain of assessment, much like instruction. Although more research is necessary on need-supportive features, forms of assessment, and indicators of well-being, our results provide preliminary evidence that when instructors combine item-writing quality with need-supportive features, they can create exams that are both rigorous and humane, thereby meeting two pressing concerns for higher education [65].

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/higheredu5010015/s1>. Table S1: Types of Graded Assessments Used in the Course Students were thinking About; Table S2: Descriptive Statistics and Correlations for the BPNSF-CA items from the Retained Bifactor ESEM Model; Table S3: Study 1 Standardized Factor Loadings for CFA and bifactor CFA; Table S4: Study 1 Correlations for CFA and ESEM Models; Table S5: Study 1 Latent Zero-order Correlations for the ESEM Model; Table S6: Sample Revisions to MCQ Items between Test A and Test B and C to meet Quality Requirements; Table S7: Study 2 Fit Indices for Four Measurement Models of the BPNSF-CA; Table S8: Study 2 Factor Loadings of the b-ESEM. Reference [66] is cited in the supplementary materials.

Author Contributions: L.M.D. was responsible for conception, design, analysis, interpretation, drafting, revising, and funding for all three studies. K.W. was responsible for conception of Studies 1–3, design, analysis, interpretation, and drafting. M.A.L. was responsible for conception, design, analysis, interpretation, and drafting of Study 2. A.M.B. was responsible for designing and drafting Studies 2–3. V.J.D. was responsible for conception and design of Studies 1–2. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Social Sciences and Humanities Research Council of Canada, grant number 435-2022-1075.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of The University of Alberta (codes Pro00137385, 6 December 2023; Pro00141097, 15 April 2024; and Pro00142406, 28 May 2024).

Informed Consent Statement: Informed consent was obtained from all participants involved in the studies.

Data Availability Statement: Materials, data, and code for all three studies are openly accessible at Study 1 <https://doi.org/10.5683/SP3/KGC5J3> Study 2 <https://doi.org/10.5683/SP3/728IVW> Study 3 <https://doi.org/10.5683/SP3/UYU2YM>.

Acknowledgments: We would like to thank graduate students in the Alberta Consortium for Motivation and Emotion for their input on the intervention materials. During the preparation of this manuscript/study, the author(s) used ChatGPT v4 for the purposes of editing. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SDT	Self-determination theory
BPNs	Basic psychological needs
BPNSF-CA	Basic Psychosocial Need Satisfaction and Frustration—Classroom Assessment Scale
CFA	Confirmatory factor analysis
ESEM	Exploratory structural equation modeling
AS	Autonomy satisfaction
AF	Autonomy frustration
CS	Competence satisfaction
CF	Competence frustration
RS	Relatedness satisfaction
RF	Relatedness frustration

Appendix A

Basic Psychological Needs Satisfaction and Frustration—Classroom Assessment Scale (BPNSF-CA)

Instructions: Think about ONE course you are taking right now. Focusing on just that class, answer the following questions. There are no right or wrong answers, we are simply interested in your own perspective. Response scale: 1 = strongly disagree; 2 = disagree; 3 = neither agree nor disagree; 4 = agree; 5 = strongly agree.

Autonomy Satisfaction

I feel that I have a lot of input in the assessments used in this class. (AS1)

I feel free to express my opinions about the assessments in this class. (AS2)

I feel I can make decisions about the assessments in this course. (AS3)

I feel able to make choices related to the assessments in this class. (AS4)

Autonomy Frustration

I feel like there are no opportunities to make choices about assessments in this class. (AF2)

I feel forced to do assessments that I wouldn't choose to do if it was up to me. (AF3)

I feel pressured by the assessments in this class. (AF1)

Assessments for this class feel like a chain of obligations. (AF4)

Relatedness Satisfaction

I feel that my instructor tries to understand how assessments affect me. (RS3)

My instructor designed assessments in a way that makes me feel that they care about me. (RS1)

I feel that my instructor takes my perspectives into consideration when it comes to assessment. (RS4)

I feel like my instructor tries to prevent me from feeling overwhelmed by assessments in this class. (RS2)

Relatedness Frustration

Assessment is a barrier to feeling supported by my instructor in this class. (RF4)

I feel disconnected from my instructor because of the assessments in this class. (RF1)

It seems like my instructor is indifferent about the stress that assessment creates for me. (RF2)

I feel my connection with my instructor is hurt by assessment in this class. (RF3)

Competence Satisfaction

I feel that the types of assessments in this class allow me to show my learning. (CS1)

I feel capable of completing the assessments in this class. (CS4)

I feel competent completing assessments in this class. (CS3)

I feel a sense of accomplishment completing the assessments in this class. (CS2)

Competence Frustration

I feel doubtful about whether or not I can do the assessments in this class well. (CF4)

I feel a sense of incompetence as I work on the assessments in this class. (CF1)

I feel ineffective in completing assessments in this class. (CF3)

The assessments in this class make me feel like a failure. (CF2).

References

1. Daniels, L.M.; Wells, K. Connecting Students' Descriptions of Classroom Assessment in Higher Education with Wellness. *Assess. Eval. High. Educ.* **2025**, *50*, 366–380. [\[CrossRef\]](#)
2. Linden, B.; Stuart, H.; Ecclestone, A. Trends in Post-Secondary Student Stress: A Pan-Canadian Study. *Can. J. Psychiatry* **2023**, *68*, 521–530. [\[CrossRef\]](#) [\[PubMed\]](#)
3. von der Embse, N.; Jester, D.; Roy, D.; Post, J. Test Anxiety Effects, Predictors, and Correlates: A 30-Year Meta-Analytic Review. *J. Affect. Disord.* **2018**, *227*, 483–493. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Wass, R.; Timmermans, J.; Harland, T.; McLean, A. Annoyance and Frustration: Emotional Responses to Being Assessed in Higher Education. *Act. Learn. High. Educ.* **2020**, *21*, 189–201. [\[CrossRef\]](#)
5. Haladyna, T.M.; Downing, S.M.; Rodriguez, M.C. A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Appl. Meas. Educ.* **2002**, *15*, 309–333. [\[CrossRef\]](#)
6. Guo, P.; Saab, N.; Post, L.S.; Admiraal, W. A Review of Project-Based Learning in Higher Education: Student Outcomes and Measures. *Int. J. Educ. Res.* **2020**, *102*, 101586. [\[CrossRef\]](#)
7. Suhonen, S. Automatically Scored, Multiple-Attempt, Recurring Weekly Exams in a Physics Course: Can They Improve Student Wellbeing and Learning Outcomes? In Proceedings of the 51st Annual Conference of the European Society for Engineering Education (SEFI), Dublin, Ireland, 11–14 September 2023.
8. Rawlusk, P.E. Assessment in Higher Education and Student Learning. *J. Instr. Pedagog.* **2018**, *21*, 1–34.
9. Ryan, R.M.; Deci, E.L. Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. *Am. Psychol.* **2000**, *55*, 68–78. [\[CrossRef\]](#)
10. Vansteenkiste, M.; Soenens, B.; Ryan, R.M. Basic Psychological Needs Theory: A Conceptual and Empirical Review of Key Criteria. In *The Oxford Handbook of Self-Determination Theory*; Oxford University Press: Oxford, UK, 2023; pp. 84–123.
11. Chen, B.; Vansteenkiste, M.; Beyers, W.; Boone, L.; Deci, E.L.; Van der Kaap-Deeder, J.; Duriez, B.; Lens, W.; Matos, L.; Mouratidis, A.; et al. Basic Psychological Need Satisfaction, Need Frustration, and Need Strength across Four Cultures. *Motiv. Emot.* **2015**, *39*, 216–236. [\[CrossRef\]](#)
12. Howard, J.L.; Slemp, G.R.; Wang, X. Need Support and Need Thwarting: A Meta-Analysis of Autonomy, Competence, and Relatedness Supportive and Thwarting Behaviors in Student Populations. *Personal. Soc. Psychol. Bull.* **2025**, *51*, 1552–1573. [\[CrossRef\]](#)
13. Reeve, J.; Cheon, S.H. Autonomy-Supportive Teaching: Its Malleability, Benefits, and Potential to Improve Educational Practice. *Educ. Psychol.* **2021**, *56*, 54–77. [\[CrossRef\]](#)
14. Gilbert, W.; Bureau, J.S.; Poellhuber, B.; Guay, F. Predicting College Students' Psychological Distress through Basic Psychological Need-Relevant Practices by Teachers, Peers, and the Academic Program. *Motiv. Emot.* **2021**, *45*, 436–455. [\[CrossRef\]](#)
15. Daniels, L.M.; Pelletier, G.; Radil, A.I.; Goegan, L.D. Motivating Assessment: How to Leverage Summative Assessments for the Good of Intrinsic Motivation. In *Teaching on Assessment; Theory to practice: Educational psychology for teachers and teaching*; Information Age Publishing, Inc.: Waxhaw, NC, USA, 2021; pp. 107–128, ISBN 978-1-64802-428-3.
16. Vansteenkiste, M.; Ryan, R.M.; Soenens, B. Basic Psychological Need Theory: Advancements, Critical Themes, and Future Directions. *Motiv. Emot.* **2020**, *44*, 1–31. [\[CrossRef\]](#)
17. Tian, L.; Han, M.; Huebner, E.S. Preliminary Development of the Adolescent Students' Basic Psychological Needs at School Scale. *J. Adolesc.* **2014**, *37*, 257–267. [\[CrossRef\]](#)
18. Oh, H.; Patrick, H.; Kilday, J.; Ryan, A. The Need for Relatedness in College Engineering: A Self-Determination Lens on Academic Help Seeking. *J. Educ. Psychol.* **2024**, *116*, 426–447. [\[CrossRef\]](#)

19. Wang, C.; Cho, H.J.; Wiles, B.; Moss, J.D.; Bonem, E.M.; Li, Q.; Lu, Y.; Levesque-Bristol, C. Competence and Autonomous Motivation as Motivational Predictors of College Students' Mathematics Achievement: From the Perspective of Self-Determination Theory. *Int. J. STEM Educ.* **2022**, *9*, 41. [\[CrossRef\]](#)
20. Sailer, M.; Hense, J.U.; Mayr, S.K.; Mandl, H. How Gamification Motivates: An Experimental Study of the Effects of Specific Game Design Elements on Psychological Need Satisfaction. *Comput. Hum. Behav.* **2017**, *69*, 371–380. [\[CrossRef\]](#)
21. Zainuddin, Z.; Perera, C.J. Exploring Students' Competence, Autonomy and Relatedness in the Flipped Classroom Pedagogical Model. *J. Furth. High. Educ.* **2019**, *43*, 115–126. [\[CrossRef\]](#)
22. Van der Kaap-Deederr, J.; Soenens, B.; Ryan, R.M.; Vansteenkiste, M. *Manual of the Basic Psychological Need Satisfaction and Frustration Scale (BPNSFS)*; Ghent University: Ghent, Belgium, 2020.
23. Sánchez-Oliva, D.; Pulido-González, J.J.; Leo, F.M.; González-Ponce, I.; García-Calvo, T. Effects of an Intervention with Teachers in the Physical Education Context: A Self-Determination Theory Approach. *PLoS ONE* **2017**, *12*, e0189986. [\[CrossRef\]](#)
24. Swami, V.; Maïano, C.; Morin, A.J.S. A Guide to Exploratory Structural Equation Modeling (ESEM) and Bifactor-ESEM in Body Image Research. *Body Image* **2023**, *47*, 101641. [\[CrossRef\]](#)
25. Alamer, A. Exploratory Structural Equation Modeling (ESEM) and Bifactor ESEM for Construct Validation Purposes: Guidelines and Applied Example. *Res. Methods Appl. Linguist.* **2022**, *1*, 100005. [\[CrossRef\]](#)
26. Tóth-Király, I.; Morin, A.J.S.; Bőthe, B.; Orosz, G.; Rigó, A. Investigating the Multidimensionality of Need Fulfillment: A Bifactor Exploratory Structural Equation Modeling Representation. *Struct. Equ. Model. A Multidiscip. J.* **2018**, *25*, 267–286. [\[CrossRef\]](#)
27. Yukhymenko-Lescroart, M.; Sharma, G. Sense of Life Purpose Is Related to Grades of High School Students via Academic Identity. *Heliyon* **2022**, *8*, e11494. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Pascoe, M.C.; Hetrick, S.E.; Parker, A.G. The Impact of Stress on Students in Higher Education. *Int. J. Stress Manag.* **2019**, *27*, 162–174. [\[CrossRef\]](#)
29. Ribeiro, Í.J.S.; Pereira, R.; Freire, I.V.; de Oliveira, B.G.; Casotti, C.A.; Boery, E. Stress and Quality of Life among University Students: A Systematic Literature Review. *Health Prof. Educ.* **2018**, *4*, 70–77. [\[CrossRef\]](#)
30. Collie, R.J.; Hascher, T. Student Well-Being: Advancing Knowledge of the Construct and the Role of Learning and Teaching Factors. *Learn. Instr.* **2024**, *94*, 102002. [\[CrossRef\]](#)
31. Diener, E.; Suh, E.M.; Lucas, R.E.; Smith, H.L. Subjective Well-Being: Three Decades of Progress. *Psychol. Bull.* **1999**, *125*, 276–302. [\[CrossRef\]](#)
32. Ryff, C.D. Happiness Is Everything, or Is It? Explorations on the Meaning of Psychological Well-Being. *J. Personal. Soc. Psychol.* **1989**, *57*, 1069–1081. [\[CrossRef\]](#)
33. Dodd, A.L.; Priestley, M.; Tyrrell, K.; Cygan, S.; Newell, C.; Byrom, N.C. University Student Well-Being in the United Kingdom: A Scoping Review of Its Conceptualisation and Measurement. *J. Ment. Health* **2021**, *30*, 375–387. [\[CrossRef\]](#)
34. Diener, E.; Wirtz, D.; Tov, W.; Kim-Prieto, C.; Choi, D.; Oishi, S.; Biswas-Diener, R. New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Soc. Indic. Res.* **2010**, *97*, 143–156. [\[CrossRef\]](#)
35. Hossain, S.; O'Neill, S.; Strnadová, I. What Constitutes Student Well-Being: A Scoping Review Of Students' Perspectives. *Child Indic. Res.* **2023**, *16*, 447–483. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Haladyna, T.M.; Downing, S.M. A Taxonomy of Multiple-Choice Item-Writing Rules. *Appl. Meas. Educ.* **1989**, *2*, 37–50. [\[CrossRef\]](#)
37. Breakall, J.; Randles, C.; Tasker, R. Development and Use of a Multiple-Choice Item Writing Flaws Evaluation Instrument in the Context of General Chemistry. *Chem. Educ. Res. Pract.* **2019**, *20*, 369–382. [\[CrossRef\]](#)
38. Tarrant, M.; Ware, J. Impact of Item-Writing Flaws in Multiple-Choice Questions on Student Achievement in High-Stakes Nursing Assessments. *Med. Educ.* **2008**, *42*, 198–206. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Downing, S.M. Construct-Irrelevant Variance and Flawed Test Questions: Do Multiple-Choice Item-Writing Principles Make Any Difference? *Acad. Med.* **2002**, *77*, S103. [\[CrossRef\]](#)
40. Ahmadi, A.; Noetel, M.; Parker, P.; Ryan, R.M.; Ntoumanis, N.; Reeve, J.; Beauchamp, M.; Dicke, T.; Yeung, A.; Ahmadi, M.; et al. A Classification System for Teachers' Motivational Behaviors Recommended in Self-Determination Theory Interventions. *J. Educ. Psychol.* **2023**, *115*, 1158–1176. [\[CrossRef\]](#)
41. Young, K.J.; Lashley, S.; Murray, S. Influence of Exam Blueprint Distribution on Student Perceptions and Performance in an Inorganic Chemistry Course. *J. Chem. Educ.* **2019**, *96*, 2141–2148. [\[CrossRef\]](#)
42. Brunet, J.; Gunnell, K.E.; Teixeira, P.; Sabiston, C.M.; Bélanger, M. Should We Be Looking at the Forest or the Trees? Overall Psychological Need Satisfaction and Individual Needs as Predictors of Physical Activity. *J. Sport Exerc. Psychol.* **2016**, *38*, 317–330. [\[CrossRef\]](#)
43. Gideon, L. *Handbook of Survey Methodology for the Social Sciences*; Gideon, L., Ed.; Springer: New York, NY, USA, 2012; ISBN 978-1-4614-3875-5.
44. Murphy, B.A.; Watts, A.L.; Baker, Z.G.; Don, B.P.; Jolink, T.A.; Algoe, S.B. The Basic Psychological Need Satisfaction and Frustration Scales Probably Do Not Validly Measure Need Frustration. *Psychol. Assess.* **2023**, *35*, 127–139. [\[CrossRef\]](#)

45. Bader, M.; Jobst, L.J.; Moshagen, M. Sample Size Requirements for Bifactor Models. *Struct. Equ. Model. A Multidiscip. J.* **2022**, *29*, 772–783. [CrossRef]
46. Muthén, L.K.; Muthén, B.O. *Mplus: Statistical Analysis with Latent Variables: User's Guide (Version 8)*; Muthen & Muthen: Los Angeles, CA, USA, 2017.
47. Morin, A.J.; Myers, N.D.; Lee, S. Modern Factor Analytic Techniques: Bifactor Models, Exploratory Structural Equation Modeling (ESEM), and Bifactor-ESEM. In *Handbook of Sport Psychology*, 4th ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2020; pp. 1044–1073.
48. Dueber, D. Bifactor Indices Calculator: A Microsoft Excel-Based Tool to Calculate Various Indices Relevant to Bifactor CFA Models. 2017. Available online: <https://doi.org/10.13023/edp.tool.01> (accessed on 19 January 2026).
49. Marsh, H.W. Application of Confirmatory Factor Analysis and Structural Equation Modeling in Sport and Exercise Psychology. In *Handbook of Sport Psychology*, 3rd ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2007; pp. 774–798.
50. Morin, A.J.; Arens, A.K.; Marsh, H.W. A Bifactor Exploratory Structural Equation Modeling Framework for the Identification of Distinct Sources of Construct-Relevant Psychometric Multidimensionality. *Struct. Equ. Model. A Multidiscip. J.* **2015**, *23*, 116–139. [CrossRef]
51. Muthén, B. A Method for Studying the Homogeneity of Test Items with Respect to Other Relevant Variables. *J. Educ. Stat.* **1985**, *10*, 121–132. [CrossRef]
52. Snijders, T.A.B.; Bosker, R.J. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, 2nd ed.; SAGE: Thousand Oaks, CA, USA, 2012.
53. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting Linear Mixed-Effects Models Using Lme4. *J. Stat. Softw.* **2015**, *67*, 1–48. [CrossRef]
54. Faul, F.; Erdfelder, E.; Buchner, A.; Lang, A.-G. Statistical Power Analyses Using G*Power 3.1: Tests for Correlation and Regression Analyses. *Behav. Res. Methods* **2009**, *41*, 1149–1160. [CrossRef] [PubMed]
55. JASP Team. JASP (Version 0.18.3). 2025. Available online: <https://jasp-stats.org/> (accessed on 19 January 2026).
56. Howard, J.L. Psychometric Approaches in Self-Determination Theory: Meaning and Measurement. In *The Oxford Handbook of Self-Determination Theory*; Oxford University Press: Oxford, UK, 2023; pp. 438–454. [CrossRef]
57. Gogol, K.; Brunner, M.; Goetz, T.; Martin, R.; Ugen, S.; Keller, U.; Fischbach, A.; Preckel, F. “My Questionnaire Is Too Long!” The Assessments of Motivational-Affective Constructs with Three-Item and Single-Item Measures. *Contemp. Educ. Psychol.* **2014**, *39*, 188–205. [CrossRef]
58. Firoozi, T.; Daniels, L.; Daniels, V.; Gierl, M. An Augmented Intelligence System for Automated Quality Control and Feedback Generation of Multiple Choice Test Items. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium, Blue Sky, and WideAIED*; Cristea, A.I., Walker, E., Lu, Y., Santos, O.C., Isotani, S., Eds.; Springer Nature: Cham, Switzerland, 2025; pp. 86–93.
59. Teixeira, P.J.; Marques, M.M.; Silva, M.N.; Brunet, J.; Duda, J.; Guardia, J.L.; Lindwall, M.; Lonsdale, C.; Markland, D.; Moller, A.C.; et al. A Classification of Motivation and Behavior Change Techniques Used in Self-Determination Theory-Based Interventions in Health Contexts. *Mot. Sci.* **2020**, *6*, 438–455. [CrossRef]
60. Keptner, K.M.; Fitzgibbon, C.; O'Sullivan, J. Effectiveness of Anxiety Reduction Interventions on Test Anxiety: A Comparison of Four Techniques Incorporating Sensory Modulation. *Br. J. Occup. Ther.* **2021**, *84*, 289–297. [CrossRef]
61. Braver, S.L.; Thoemmes, F.J.; Rosenthal, R. Continuously Cumulating Meta-Analysis and Replicability. *Perspect. Psychol. Sci.* **2014**, *9*, 333–342. [CrossRef]
62. Reeve, J.; Cheon, S.H. Teachers Become More Autonomy Supportive after They Believe It Is Easy to Do. *Psychol. Sport Exerc.* **2016**, *22*, 178–189. [CrossRef]
63. Janisch, C.; Liu, X.; Akrofi, A. Implementing Alternative Assessment: Opportunities and Obstacles. *Educ. Forum* **2007**, *71*, 221–230. [CrossRef]
64. Jopp, R.; Cohen, J. Choose Your Own Assessment-Assessment Choice for Students in Online Higher Education. *Teach. High. Educ.* **2020**, *27*, 738–755. [CrossRef]
65. Horowitch, R. *The Perverse Consequences of the Easy A*; The Atlantic: Washington, DC, USA, 2025.
66. Rodriguez, M.C. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educ. Meas. Issues Pract.* **2005**, *24*, 3–13. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.