Psicothema

Volumen 38, no. 1
ISSN 0214–9915 • eISSN 1886-144X

Psicothema

Editor-in-Chief: Laura E. Gómez. Univ. de Oviedo, Spain
Deputy Editor: Rebeca Cerezo. Univ. de Oviedo, Spain
Associate Editors: Susana Al-Halabí. Univ. de Oviedo, Spain
Isabel Benítez, Univ. de Granada, Spain
Jorge F. del Valle, Univ. de Oviedo, Spain
Eduardo Fonseca, Univ. de La Rioja, Spain
José Carlos Núñez, Univ. de Oviedo, Spain
Javier Suárez-Álvarez. Univ. of Massachusetts Amherst, USA
Paz Suárez-Coalla, Univ. de Oviedo, Spain

Managing Editor: Gloria García-Fernández, Univ. de Oviedo, Spain

Editorial Office: Leticia García, COPPA, Spain
M.ª Ángeles Gómez, COPPA, Spain

Honorary Editor: José Muñiz, Univ. Nebrija, Spain

ADVISORY BOARD

Olivia Afonso. Oxford Brookes Univ., UK
Leandro Almeida. Univ. do Minho, Portugal
David Álvarez. Univ. de Oviedo, Spain
Rui Alves. Univ. of Porto, Portugal
Antonio M. Amor González. Univ. de Salamanca, Spain
Diego Ardura. UNED, Spain
Natalia Arias. Univ. Nebrija, Spain
Ignacia Arruabarrena. Univ. del País Vasco, Spain
Roger Azevedo. Univ. of Central Florida, USA
Giulia Balboni. Univ. de Perugia, Italy
Mónica Bernaldo de Quirós. Univ. Complutense de Madrid, Spain
Ana Bernardo. Univ. de Oviedo, Spain
Verner P. Bingman. Univ. Browling Green, USA
Teresa Bobes Bascarán. Univ. de Oviedo, Spain
Roser Bono. Univ. de Barcelona, Spain
Amaia Bravo. Univ. de Oviedo, Spain
José Luis Carballo, Univ. Miguel Hernández, Spain
Thomas J. Carew. New York Univ., USA
José Pedro Espada. Univ. Miguel Hernández, Spain
Martin Debbané. Univ. de Ginebra, Switzerland
Paula Elosua. Univ. del País Vasco, Spain
Joyce L. Epstein. Univ. John Hopkins, USA
Rubén Fernández Alonso, Univ. de Oviedo, Spain
Sergio Fernández Artamendi, Univ. de Sevilla, Spain
Mª Estrella Fernández Alba. Univ. de Oviedo, Spain
María Fernández Sánchez. Univ. de Salamanca, Spain
Pere J. Ferrando. Univ. Rovira i Virgili, Spain
Victoria A. Ferrer. Univ. de las Islas Baleares, Spain
José M. García Montes. Univ. de Almería, Spain
Alba González de la Roz. Univ. de Oviedo, Spain
Francisco González-Lima. Univ. of Texas, USA
Ana González Menéndez. Univ. de Oviedo, Spain
Steve Graham. Arizona State Univ., USA
Mattias Grünke. Univ. of Cologne, Germany
Ana Hernández. Univ. de Valencia, Spain
M.ª Dolores Hidalgo. Univ. de Murcia, Spain
Stephen T. Higgins. Univ. Vermont, USA
Cándido Inglés. Univ. Miguel Hernández, Spain
Christa Labouliere. Columbia University, USA
Alfonso Lara Torralbo. Univ. de Córdoba, Spain
Susana Lázaro Visa. Univ. de Cantabria, Spain
Pablo Livacic Rojas. Univ. de Santiago de Chile, Chile
Mónica López. Univ. Groningen, The Netherlands
José Antonio López Pina. Univ. de Murcia, Spain
Verónica Marina Guillén. Univ. de Cantabria, Spain
Eduardo Martín Cabrera. Univ. de La Laguna, Spain
Víctor Martínez Loredo, Univ. de Oviedo, Spain
Ana Miranda. Univ. de Valencia, Spain
Fabia Morales Vives. Univ. Rovira i Virgili, Spain
Mª Lucía Morán. Univ. de Cantabria, Spain
Patricia Navas Macho, Univ. de Salamanca, Spain
Javier Ortuño-Sierra, Univ. de La Rioja, Spain
José Luis Padilla. Univ. de Granada, Spain
Mercedes Paíno. Univ. de Oviedo, Spain
Alicia Pérez de Albéniz. Univ. de La Rioja, Spain
Celestino Rodríguez. Univ. de Oviedo, Spain
Mª Fe Rodríguez Muñoz. UNED, Spain
Sonia Romero. Univ. Aut. de Madrid, Spain
Pedro Rosário. Univ. do Minho, Portugal
David Scanlon. Boston College, USA
Roberto Secades-Villa, Univ. de Oviedo, Spain
Albert Sesé. Univ. de las Islas Baleares, Spain
Giorgios Sideridis. Harvard Medical School, USA
Steve Sirecci, Univ. of Massachusetts, USA
Miguel Ángel Sorrel, Univ. Autónoma de Madrid, Spain
Mark Torrance. Nottingham Trent Univ., UK
Luis Valero Aguayo. Univ. de Málaga, Spain
Antonio Valle. Univ. de A Coruña, Spain
Wouter Vanderplasschen. Ghent Univ., Belgium
Antonio Verdejo-García. Monash Univ., AUS
Eva Vicente. Univ. de Zaragoza, Spain
Jianzhong Xu. Mississippi State Univ., USA
Jin H. Yoon. Univ. of Texas, USA

Psicothema is indexed by Social Sciences Citation Index (WOS), Scopus, Google Scholar, SciELO, Dialnet, EBSCO Essentials Academic Search Premier, PubMed, PsycINFO, IBECS, Redinet, Psicodoc, Pubpsych, Fuente Académica Plus, IBZ Online, Periodicals Index Online, MEDLINE, EMBASE, ERIH PLUS, Latindex, MIAR, CARHUS Plus+ 2018, Rebiun, DOAJ, & Crossref.

D.L. AS 3779-1989 ISSN: 0214 - 9915 CODEN PSOTEG

∞ This paper meets the requirements of ISO 9706:1994, Information & documentation – Paper for documents – Requirements for permanence, effective with Volume 7, Issue 2, 1995.

Publisher address:
► Colegio Oficial de Psicología del Principado de Asturias
Ildefonso Sánchez del Río, 4 - 1º B
33001 Oviedo (Spain)
Tel.: +34 985 28 57 78 • Fax: +34 985 28 13 74
E-mail: psicothema@cop.es • http://www.psicothema.com

Psicothema

PUBLICATION GUIDELINES

Psicothema publishes empirical work in English which is done with methodological rigor and which contributes to the progress of any field of scientific psychology. As an exception, the Editorial Board may accept publication of work in Spanish if the content justifies such a decision. Theoretical work may also be accepted, if requested by the Editorial Board, with preference given to articles that engage with critical research issues or which discuss controversial approaches.

Submission of articles

1. Articles should be submitted via the journal's web page: www.psicothema.com (Authors section – submission of articles): http://www.psicothema.es/submit

2. Submissions must comply with the rules for preparation and publication of articles, as well as the ethical standards specified below.

3. Studies must be unpublished. Articles which have been fully or partially published elsewhere will not be accepted, nor will articles that are in the process of publication or which have been submitted to other journals for review. It will be assumed that all those who appear as authors have agreed to do so, and all those cited for personal correspondence have consented.

4. The activities described in the published articles will comply with generally accepted ethical standards and criteria, both in terms of work with human beings and animal experimentation, as well as all aspects of professional and publishing ethics.

5. The original work may be submitted in Spanish initially and receipt will be acknowledged immediately. If so, and if it is accepted, the authors will be responsible for translating it into English for publication.

6. Authors may only submit one article for consideration by Psicothema per year.

7. Names and surnames should be entered on the platform in the form they will be cited (a single surname, two separate surnames, hyphenated surnames, etc.). The affiliation of all authors must be indicated. **A maximum of two affiliations per author may be indicated. Affiliations must follow the format "entity or university (country, in English)"**. Do not include information about research groups or departments. Only one person may appear as corresponding author, who will be responsible for ensuring that the author names, order, and affiliations are correct.

8. Authors should suggest three people who they believe would be suitable reviewers for the article, clearly indicating their institutional affiliation and email address. Authors may also indicate people who, for whatever reason, they do not wish to be involved in the review process for their work. Pleas bear in mind the recommendations from the Committee on Publication Ethics (COPE) when suggesting the three reviewers https://publicationethics.org/files/Ethical_guidelines_for_peer_reviewers_0.pdf

9. Manuscripts are screened by the Editorial Board to assess relevance and interest for the journal and whether it follows the rules. Articles must faithfully conform to the editorial rules and fall within the editorial scope of the journal. It is a necessary, though not sufficient, condition that articles must comply with the rules for publication. Articles which do not follow Psicothema's rules will be rejected. In general, within around 10 days the Editorial Board will communicate a decision of interest to begin the review process.

10. Psicothema is only able to publish about 10% of the manuscripts it receives, which is why we apply a very rigorous screening and selection system. Many submissions are considered **non-priorities** by the Editorial Board without being sent for review.

11. If an article passes the Editorial Board screening, it will be sent to a minimum of two reviewers to evaluate its scientific quality. The journal has a **policy of "double blind" reviews**, meaning that both authors and reviewers are anonymous during the review process. To that end, manuscripts must not contain information that would allow the authors to be identified. Most reviewers report back within the agreed three week period. The review process, from receiving an article to the decision to modify it or reject it, usually takes around two months.

12. If, after receiving the reviewers' reports, the Editorial Board decides that the article needs "modifications" to be published, the authors should send the modifications in the requested format together with a point-by-point response to all the comments made by the reviewers and the Editorial Board. Failure to respond in the required format within the set timescale will lead to the article being rejected and removed from the management platform, with no possibility of re-submission.

13. The Editorial Board is responsible for the final decision to accept the article for publication or not. The editors usually make their decisions as quickly as possible once they have received all the necessary reports.

14. After an article has been accepted, and before publication, the authors must sign a copyright agreement. Printing rights and rights of reproduction in any format or medium belong to Psicothema, who will not reject any reasonable request from authors for permission to reproduce their contributions.

15. It is the authors' responsibility to obtain relevant permissions to reproduce copyright-protected material. They are also responsible for disclosing possible conflicts of interest, declaring sources of funding and their participation in the research, and providing access, where necessary, to databases, procedure manuals, scores, and other experimental material that may be relevant. These aspects must be declared in the articles, as described below.

For any questions or clarifications, the journal can be contacted via the email address psicothema@cop.es

Manuscript preparation

1. File format: Articles must be sent in DOC or DOCX format. Microsoft Word documents must not be locked or password-protected, they should not have comments in the margins or information that might reveal the authors' identities. The file should be anonymised in "file properties" so that author information does not appear.

2. Length: The maximum length for articles is **6,000 words** (including the title, abstracts, key words, in-text references, acknowledgements, figures, and tables). The 6,000 word limit **does not include the list of references**. If authors wish to provide supplementary material, the article should include a unique, persistent web link (see point 18 about supplementary material).

3. Format: The articles must be in Microsoft Word format, using **12-point Times New Roman**, in a single column with 3 cm margins, paragraphs left-aligned and double spaced (except for tables and figures which may use single spacing). Page numbers must be included in the lower right corner. Limit sections and subsections to three levels of headings and follow the recommendations in the APA 7th edition about "Sentence case" in the list of references. Psicothema does not allow footnotes, annexes, or appendices. Any such content should be incorporated appropriately into the text (see point 18 about supplementary material).

4. Language: Although articles may be submitted and reviewed in Spanish, accepted articles are usually published in English. Once articles are accepted, the authors must provide an English translation of the reviewed article, within the indicated timeframe, for publication. Psicothema accepts American and British English, but not a mix of the two. Any text in English must be of appropriate professional quality, which will be reviewed by a professional native-speaking translator. Following that review, Psicothema may suggest changes, or if necessary, request a new translation or revision of the translation, the costs of which will be borne by the article's authors.

5. Title page: The first page of the article contains the article title in English and in Spanish, the running title (in English), the total number of words

in the article (not counting references) and a **declaration of authorship, originality and the fact that the work is previously unpublished**. This declaration is obligatory as one of the measures the journal takes to avoid plagiarism. The submitted text must be anonymized, avoiding use of the authors names or anonymizing other possible references that may identify them. Follow the APA 7th edition rules for capitalization of titles and subtitles (i.e., "Title case"). Use upper case for the first letter of all nouns, verbs, adjectives, adverbs, pronouns, and any word longer than three letters.

6. **Title:** The title should be short, descriptive, clear, accurate, and easy to read. It should engage the reader's interest and name variables or topics addressed. Ensure that the main key phrase of the topic is in the article title and avoid superfluous words. Remember that searches normally use key phrases rather than individual words (for example, "mental health in people with disability" not just "health"). Try to include the topic at the start of the title. If the title is "creative", add a more descriptive subtitle after a colon. A descriptive title will help the article to be found in databases. The Editorial Board reserves the right to change titles and abstracts of articles accepted for publication in order to follow the above rules and enhance the article's impact and dissemination.

7. **Abstracts and key words:** the second page of the article contains the abstracts (in Spanish and English) and 3-5 key words or terms. Abstracts must be no more than 200 words and **structured** in four sections: Background, Method, Results, and Conclusions. The abstract should be a single paragraph with these titles in bold, followed by colons and upper case. The key words cover essential elements of the paper such as the research topic, population, method, or application of the results. Avoid general terms and empty words (pronouns, adverbs etc.), or redundant words such as analysis, description, research, etc. Nouns are preferred. Pay particular attention to selection of key words as they are used to index the article.

8. **Article:** The article introduction begins on the third page. The introductory section should not include the article title, or the subtitle "Introduction", or subsections. Following that, the "Method" section should contain the following subsections "Participants", "Instruments", "Procedure", and "Data Analysis", and no others, in no other order, and with no other titles. Where appropriate, in the procedure section it is obligatory to provide information about ethical aspects of the study, the ethics committee that approved the study and the reference code (anonymized during the review process). For research with children, express mention must be made about obtaining informed consent. Pay particular attention to the APA rules about the presentation of statistical and mathematical results in the text, as well as tables and figures. At the end, there should be a single "Discussion" section which should include both discussion along with limitations and conclusions of the study. The discussion section should not have any subsections.

9. **Declaration of author contributions:** Where there is more than one author, there must be a declaration of responsibilities at the end of the article, before the references, specifying what contribution each of the authors made. To specify each author's contribution, use the criteria established by the CRediT taxonomy (Contributor Roles Taxonomy; https://credit.niso.org). Please use the full name of each author as it appears in the manuscript to declare their contributions, followed by the CRediT roles performed. Follow this example: **John White:** Conceptualization, Methodology, Software. **Nuria García-Fernández:** Data curation, Writing - Original draft. **Lucinda Jackson:** Visualization, Investigation. **Laura Gayo:** Supervision, Software, Validation. **Michael Gutiérrez:** Writing - Review and Editing.

   If a group of authors made equal contributions, please also use the CRediT taxonomy to specify their contributions: **John White:** Conceptualization, Writing – Original draft, Writing - Review and Editing. **Lucinda Jackson:** Conceptualization, Writing – Original draft, Writing review and Editing.

   Psicothema does not permit the use of other formulas to indicate equal contributions, such as 'contributed equally to this work ', co-first authors, co-last authors, or co-senior authors.

10. **Corresponding author:** Psicothema allows only **one corresponding author**, who will take primary responsibility for communication with the journal during the manuscript submission, peer review, and publication process, as well as for ensuring providing correct details of authorship (including the names of co-authors, addresses and affiliations), ethics, acknowledgements, sources of funding, conflict of interests, and declarations. The corresponding author is responsible for having ensured that all authors have agreed to be so listed, and have approved the manuscript submission to the journal. After publication, the corresponding author is the point of contact for queries about the published paper. It is their responsibility to inform all co-authors of any matters arising in relation to the published paper and to ensure such matters are dealt with promptly.

11. **Acknowledgements:** any acknowledgements should be included at the end of the text, before the references, in a separate section titled "Acknowledgements".

12. **Sources of Funding:** Priority will be given to work supported by competitive national and international projects. A section titled "Funding" must be included following the "Acknowledgements" section (if one is included) and before the list of references. The "Funding" section must clearly specify the funding body with the assigned code in brackets. It must also be clearly indicated whether the source of funding had any kind of participation in the study. If there was no participation, include the following sentence, "The source of funding did not participate in the design of the study, the data collection, analysis, or interpretation, the writing of the article, or in the decision to submit it for publication". If no funding was received, add the following, "This study did not receive any specific assistance from the public sector, the commercial sector, or non-profit organizations".

13. **Conflict of interests:** Authors must report any economic or personal relationship with other people or organizations that may inappropriately influence their work. If there are none, following the funding section, in a section titled "Conflict of Interest", authors should state: "The author(s) declare(s) that there are no conflicts of interest".

14. **Declaration of availability of data:** The authors should state, in a section titled "Data Availability Statement", whether the research data associated with the article is available and where or under what conditions it may be accessed. They may also include links (where appropriate) to the dataset.

15. **Reference style:** Articles must be written following the guidelines in the 7th edition of the Publication Manual of the American Psychological Association. Articles that do not comply with these rules will be rejected. Some of the requirements are summarized below.

Bibliographical references in the text should include the author's surname and year of publication (in brackets, separated by a comma). If the author's name forms part of the narrative, it should be followed by the year in brackets. If there are more than two authors, only the first author's surname is given, followed by "et al." and the year; if there is confusion, add subsequent authors until the work is clearly identified. In every case, the references in the bibliography must be complete (up to 20 authors). When citing different articles in the same brackets, order them alphabetically. To cite more than one study from the same author or authors from the same year, add the letters a, b, c, as necessary, repeating the year (e.g., 2021a, 2021b).

The list of references at the end of the article must be alphabetical and comply with the following rules:

a) **Books:** Author (surname, comma, initials of first name(s) and a full stop); if there are various authors, separate them with a comma; before the final author use a comma and "&"; year (in brackets) and full stop, The full title in italics and full stop; finally, the publisher. For example:

   Lezak, M., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological assessment* (4th ed.). Oxford University Press.

b) **Chapters of books with various authors, reports from conferences or similar:** Author(s); year; title of the work being cited, followed by "In", the director(s), editor(s), or compiler(s) and in brackets Ed., adding an s if plural; the title of the book in italics and in brackets the page numbers of the cited chapter; the publisher. For example:

   de Wit, H., & Mitchell, S. H. (2009). Drug effects on delay discounting. In G. J. Madden & W. K. Bickel (Eds.), *Impulsivity: The behavioral and neurological science of discounting* (pp. 213-241). American Psychological Association.

c) Journal articles: Author(s); year; article title; full name of the journal in italics; volume number in italics; issue number in brackets with no space between it and the volume number; first and last page number. The doi should be included in URL format. For example:

Muñiz, J., & Fonseca-Pedrero, E. (2019). Diez pasos para la construcción de un test. *Psicothema, 31*(1), 7-16. https://doi.org/10.7334/psicothema2018.291

For documents that do not have a doi, it is no longer necessary to use "Retrieved from", instead give the URL directly. For example:

Walker, A. (2019, November 14). *Germany avoids recession but growth remains weak*. BBC News. https://www.bbc.com/news/business-50419127

d) Pay particular attention to the rules in the 7th edition of the APA manual for citing work presented in **conferences, doctoral theses, and software,** as well as the rules for the **use of acronyms in text** and in the references section.

e) When the original version of the cited work (book, chapter, or article) **is not in English**, cite the original title and give the English translation in square brackets (with no separation from the original, without using italics).

For further information and other cases, consult the 7th edition of the APA publication manual or the following page: https://apastyle.apa.org/style-grammar-guidelines/references/examples

16. Figures and tables should be included at the end of the manuscript, one per page. They should also follow the APA 7th edition guidelines, be appropriately numbered and cited in the text, indicating approximately where they should be placed. They must have a short, descriptive title that helps understand the content, and follow the APA recommendations about title case, with no full stop. They should be 7 or 14 cm wide and have clear, legible lettering and symbols. Avoid wasted space and make best use of the space available. Figures must be submitted in editable formats, consistent with the format of the rest of the article. If that is not possible, they must have a minimum resolution of 300ppp.

17. Pre-registration of studies and plans of analysis: as a general rule, Psicothema recommends pre-registering submitted studies. If authors have pre-registered studies or plans of analysis, links to that pre-registration should be provided in the article.

18. Supplementary material. Psicothema recommends sharing the data that has been used in the research and supplementary material in institutional or thematic open-access repositories, federated in the European Open Science Cloud (EOSC). Provide a web link if access is to be provided to databases or any other supplementary material, using unique, persistent identifiers.

19. We encourage authors to consult the following standard guidelines when preparing their manuscripts (although due to the multidisciplinary nature of the journal, this is not obligatory):

Case Reports - **CARE**

Diagnostic accuracy - **STARD**

Observational studies - **STROBE** (von Elm et al., 2008), **MQCOM** (Chacón et al., 2019) o **GREOM** (Portell et al., 2015)

Randomized controlled trial - **CONSORT** and **SPIRIT** (Hopewell et al., 2022)

Systematic reviews, meta-analyses – **PRISMA** (Page et al., 2020)

Test adaptation - **International Test Commission Guidelines** (Hernández et al., 2020)

Test development - **Ten steps for test development** (Muñiz & Fonseca, 2019)

Publication of articles

1. Publication rates: Psicothema is an **"open access"** journal. All of the articles will always be free to those who want to read or download them. In order to provide this open access, Psicothema charges a publication fee which the authors or their funders must pay. The price depends on the length of the manuscript. In general, the average price per article is between €180 and €210, based on a mean of 6-7 pages per article, at €30 per laid-out page.

2. Print Proofs: Once an article has been accepted for publication, the contact person will receive an email with the print proofs in PDF format to check and correct spelling-typographical errors. Only minimal corrections can be made to the content of the article once it has been accepted. **Substantial modifications and changes will not be accepted** other that correcting printing or translation errors, possible errors detected during the review process, or incorporating suggestions made by the Editorial Board. No changes will be accepted in this phase to authorship, addition of new affiliations, or details such as including research groups or departments. Galley proofs should be checked carefully, following the instructions provided with them, to confirm that they match the accepted original. Corrected proofs should be returned within the requested timeframe (48-72 hours). Corrections must be made in the PDF file itself, no other means of correction will be accepted. It is vital to check that names, surnames, ORCID codes, and affiliations are all correct in this stage. The corresponding author is responsible for gaining approval from all co-authors for the corrected print proofs. If the proof article is not reviewed within the timeframe or manner specified, that version of the article will be published and subsequent changes or corrections will not be possible.

3. Published version: Once the edition of Psicothema containing the article is published, the author will receive a copy of their article in PDF format. The final version typeset by Psicothema will be available online via DOI. We strongly recommend sharing the final version published by Psicothema on social networks, (Facebook, Twitter, LinkedIn…), university and public repositories (Mendeley, Cosis…), scientific social networks (ResearchGate, Academia.edu, Kudos ...), personal and institutional websites, blogs, Google Scholar, ORCID, Web of Science ResarcherID, ScopusID...

Ethical standards

Psicothema is committed to the scientific community to ensure the ethical and quality standards of published articles. Its references are the "Core practices" defined by the **Committee on Publication Ethics (COPE)** for journal editors, the **American Psychological Association (APA)** Code of Conduct, and the Code of Ethics for Psychology from the **Spanish General Council of Psychology**.

Use of inclusive, non-sexist language. At Psicothema, we are firmly committed to equality and respect for all, recognizing and appreciating diversity. For this reason, authors should ensure that they use bias-free language, avoid stereotypes, and engage with inclusive, non-sexist language, albeit prioritizing grammatical correctness, economy of language, and accuracy, given the limitations of space. Pay particular attention to the presentation of data, so that participants' characteristics are described and analysed properly, without presenting information that is irrelevant to testing hypotheses, achieving objectives, or presenting results of the study. Avoid condescending, obsolete, or inappropriate language, as well as the use of labels related to stereotypes. We recommend reporting where potential gender differences are found in the results.

Responsible authorship. Psicothema promotes transparency via the declaration of authors' contributions. All signatories must have made substantial contributions in each of the following aspects: (1) conception and design of the study, or data acquisition, or analysis and interpretation of data, (2) drafting the article or critical review of the intellectual content, and (3) final approval of the submitted version. The list and order of authors should be carefully reviewed before the initial submission of the article. Any addition, removal, or re-ordering must be done before the article is accepted, with the approval of the Psicothema Editorial Board and the consent of all named authors. A form for this is available on request.

Open science. To facilitate the reproducibility of research and reuse of data, code, types of software, models, algorithms, protocols, methods, and any other useful material related to the project should be shared.

We recommend that authors publish the original study data in public open-access repositories online, such as FigShare (http://figshare.com), Mendeley Data (https://data.mendeley.com/), Zenodo (http://zenodo.org/), DataHub (http://datahub.io) and DANS (http://www.dans.knaw.nl/). Where data or supplementary material is shared, a corresponding reference should be included in the manuscript and the list of references, using unique, persistent identifiers.

Funding sources. In the acknowledgements section, authors should include data on the organizations that provided economic funding for the study or preparation of the article, and briefly describe the role any funding body played in designing the study, data collection, analysis, and interpretation, writing the article, or the decision to submit it for publication. If there was no participation from the funding body, this should be indicated as suggested in the "Preparation of Articles" section. The author responsible for submitting the article should include this metadata at the time of submission in the corresponding section.

San Francisco declaration on research assessment (DORA). As part of its commitment to open knowledge, Psicothema follows this initiative because it shares the need to address the quality assessment of scientific articles (not only the journals in which they are published), to consider the value and impact of all research outputs (including data and software), and to consider the societal impact of research from a broader perspective (including qualitative indicators, such as the influence on scientific policies and practices, together with a responsible use of quantitative indicators). To this end, it is committed to remove restrictions on the number of references that can be included in the bibliography, not counting them as part of the maximum number of words, to encourage responsible authorship practices and to provide information about the specific contributions of each author (CRediT), to mandate the citation of primary literature in favor of reviews in order to give credit to the group(s) who first reported a finding, and to make available a variety of journal-based metrics and article-level metrics (PlumX).

Good publishing practice in gender equality. Psicothema is committed to gender policies that lead to real equality between men and women in society through various actions: (1) pursuing equal proportions of women and men in the editorial team, as well as in those who review the articles; (2) recommending the use of inclusive language in scientific articles; (3) recommending that articles report whether the original study data considered sex or gender in order to identify possible differences; and (4) including the full names of the authors of published articles. To that end, authors must include their full names (not just first initials) in the metadata, which will appear in the published articles.

Authors' rights

Acknowledgement of receipt. Receipt of the article will be immediately communicated to the authors by email.

Screening. Articles will be reviewed by the Editor-in-Chief, Executive Editor, Managing Editor, and the Associate Editors. The editorial team may directly reject studies if, in their opinion, they do not follow the journal's publication rules, do not meet the minimum requirements, or do not fit the journal's objectives or priorities.

Review. Once past the Editorial Board screening, the articles will be reviewed by external reviewers and by the Associate Editor responsible for managing the article. The Associate Editor and the Managing Editor will consider the external reviewers' reports and will make the final decision on publication.

Reasoned reply. Except in cases of articles considered to be "non-priority" in the initial screening phase, authors will be given a reasoned response about the Editorial Board's final decision when that involves rejection (i.e., articles rejected after the peer review phase).

Confidentiality. Authorship of articles received will be kept anonymous and the evaluation process will be confidential, we commit to not disseminating the article more than necessary for the evaluation process and until the article is accepted for publication.

Use of data. The members of the Editorial Board will not use the results of unpublished work without the express consent of the authors.

Declaration of privacy. The names and email addresses provided to Psicothema will only be used for purposes established in the journal, they will not be given to third parties or used for commercial purposes.

Complaints and claims. Efforts will be made to respond to and resolve complaints and claims quickly and constructively. Complaints or claims should be send by email to psicothema@cop.es, clearly and accurately specifying the nature of the complaint, the contact details of the person making it, and sufficient data to demonstrate any possible violation of the journal's declaration of ethics. Complaints about published content must be made as soon as possible after publication, and after having first contacted the corresponding authors to try and find a direct resolution. Psicothema may be contacted where it is not appropriate to contact the authors, if the authors do not respond, or if they do not resolve the issue. If possible, documentation must be included as evidence of the situation. Psicothema will acknowledge receipt of the complaint by email, and may request additional information or documentation for clarification. Depending on the nature or complexity of the issue, if the content is reviewed and sufficiently documented, the Editorial Board will study the case and make any decision in accordance with the directives of the Committee of Publishing Ethics (COPE). The Editor-in-chief will make the final decision and a response will be sent by email. Other people and institutions will be consulted as necessary, including university authorities or subject-matter experts, and legal advice may be sought if the complaint has legal implications. Complainants will have to expressly request that a complaint be treated confidentially and the Editorial Board will do so as far as appropriate and in line with our management processes. It is possible that complainants will not receive any information about the state of any investigation until a final decision is reached, and it is important to bear in mind that investigations may take some time. Complaints that are outside Psicothema scope or that are presented in an offensive, threatening, or defamatory manner will be dismissed. Personal criticism or comments are not acceptable. Communication will be terminated if it is not cordial and respectful, or if there is persistent vague or unfounded complaint. Psicothema reserves the right to take appropriate legal measures if a complainant insists on a complaint that is unfounded, false or malicious.

Authors' responsibilities

Editorial rules. Authors should read and accept the editorial rules and journal's instructions before sending a manuscript. While the article is undergoing the evaluation process at Psicothema, it must not be in any evaluation process at other journals.

Ethical rules. Authors must comply with the ethical standards specified in the Psicothema rules for authors.

License for public communication. The authors cede to Psicothema the public communication rights of their article for free dissemination through the internet, portals and electronic devices, through its free provision to users for online consultation, printing, download and archive, guaranteeing free, open access to the publication.

Publication licence. The authors accept the Psicothema copyright policy and cede it the right of publication. Psicothema publishes its articles under CC-BY-NC-ND license.

Reviewers' responsibilities

Editorial rules. Reviewers must read and accept the journal's editorial rules and instructions before reviewing an article. They must also follow the COPE ethical directives for reviewers.

Professional responsibility. Reviewers must only accept articles for review for which they have sufficient knowledge to perform a proper review.

Conflict of interests. Reviewers will constructively and impartially review articles for which they consider themselves qualified, abstaining from reviewing articles in which there might be a conflict of interest.

Confidentiality. Reviewers will respect the confidentiality of the review process and will not use information obtained during the peer review process for personal gain or to others' advantage, or to discredit or disadvantage others. They will not involve other people in the review process without the authorization of Psicothema.

Suspicion of a breach of ethics. Reviewers will inform the Editorial Board if they detect poor practice, fraud, plagiarism, or self-plagiarism, as well as any other irregularity related to research or publication ethics.

Deadline for reviews. Reviewers will commit to meeting the review timeframes set by Psicothema, informing the Editorial Board if they need additional time or are unable to send a report after having accepted a review request.

Preparation of the report. The format of Psicothema's review report is open, but reviewers must use a short scoring rubric. Reviewers must be objective and constructive in their reviews, offering feedback that will help authors improve their articles. Reviewers must make fair, impartial, constructive assessments of the article's strengths and weaknesses, and avoid disparaging personal comments or baseless accusations. They must not suggest that authors add references to the reviewer's own work (or that of colleagues) just to increase the number of citations or raise the visibility of their work or the work of associated; suggestions must only be based on valid academic reasons.

The journal's responsibilities

The Editorial Board is not responsible for the ideas or opinions expressed by the authors in the journal articles or the reviewers in their reports. The opinions and facts expressed in the articles are solely and exclusively the authors' responsibility and do not represent the journal's opinions or scientific policies. The editorial organization is not responsible in any case for the credibility or authenticity of the articles.

Psicothema will strive to avoid scientific fraud, which includes fabrication, falsification, or omission of data; plagiarism; duplicate publications; and authorial conflicts. Particular attention in plagiarism is paid to avoiding passing others' work off as one's own, co-opting others' ideas without recognition, giving incorrect information about the source of a reference, and paraphrasing a source without mentioning it. Detection of **fraud or plagiarism** will lead to the rejection of the submitted or published article.

The Psicothema Editorial Board undertakes to ensure that everyone involved (authors, reviewers, editors, and journal management) comply with the expected ethical standards in every phase of the publishing process, from reception to publication of an article, basing this on the recommendations from **The Committee on Publication Ethics (COPE)** to resolve possible conflicts.

The readers' rights

Readers have the right to read all articles published in Psicothema for free immediately after their publication.

Updating published articles

Psicothema is committed to correcting important scientific errors or ethical issues in published articles. In order to be transparent about any change, the following criteria and procedures have been established for updating our published articles.

Minor errors. Minor errors that do not affect the readability or meaning, such as spelling, grammar, or layout mistakes are not sufficient for and do not justify an update, regardless of the source of the error.

Metadata errors. Requests to correct errors in an article's metadata (for example, title, author name, abstract) must be made during the galley correction process. Once an article has been published, corrections can only be made if the Editorial Board believes the request to be reasonable and important. Once approved, the article will be updated and republished on the Psicothema website, with notification to relevant databases.

Author name and affiliation. Authors must make any desired changes to author names, surnames and affiliations during the galley correction process. Once the article has been published, no changes will be made without valid, convincing reasons, especially if the ORCID code has been supplied correctly. Changes to names after publication will only be in exceptional cases where the authors adopt a new name (such as for marriage or after gender transition) and want it updated. In such cases the Committee on Publication Ethics (COPE) recommendations will be followed.

Corrections. Requests may be submitted to correct errors that affect scientific interpretation. Once a request is approved, the article will be updated and re-published on the Psicothema website, together with a notice of correction. This notice will be a separate publication with a link to the updated article in the most recent edition of the journal, in order to notify readers that there has been a significant chance to the article and that the revised version is available on the website. Relevant databases will then be notified about the update.

Retractions. If an article needs to be retracted from the research literature due to inadvertent errors during the review process, serious ethical violations, fabrication of data, plagiarism, or other reasons that threaten the integrity of the publication, Psicothema will follow the recommendations from the Committee on Publication Ethics (COPE) for retractions. In this case, the original publication will be amended with a "RETRACTED" mark but will remain available on the Psicothema website for future reference. Retractions will be published with the same authorship and affiliation as the retracted article, so that the notice and the original retracted article may be properly found in indexing databases. The retraction notice will be published in the most current edition of the journal. Partial retractions may be published in cases where results are partly incorrect. An article will only be removed completely from the Psicothema website and indexing databases in very exceptional circumstances, where leaving it online would constitute an illegal act or could cause significant harm.

Expression of concern. Psicothema may publish such a concern if the investigation about supposed bad conduct in research is inconclusive, complex, or very prolonged. In this case, a Psicothema editor may choose this option, detailing their concerns point by point, and any actions ongoing.

Comments and responses. Psicothema will only exceptionally publish comments on or responses to articles published in the journal, when the comments (i.e., short letters to the editors from readers who wish to publicly question a specific article) affect the editorial content of an article published in the journal, contain evidence of a claim, and the result of an investigation by the editorial team does not lead to rejecting the criticism, or correction or retraction of the article. In these exceptional cases, once the comment has been approved for peer review, the editorial team will contact the authors of the article in question and invite them to reply to the comment. The reply will allow the authors to publicly respond to the concerns raised. If the authors do not provide a reply within the set timescale or decide not to reply, the comment will be published along with a note explaining the absence of a response. Both comments and replies will be reviewed to ensure that the comment addresses significant aspects of the original article without becoming essentially a new article, the response directly addresses the concerns without evasion, and the tone of both is appropriate for a scientific journal. Only one round of comments and responses will be facilitated about any single article. Nonetheless, Psicothema recommends that readers direct their comments directly to the authors involved and use alternative forums for additional public discussion.

Archive and Digital Preservation Policies

Conditions for self-archiving preprints. Preprint versions of articles (the version initially submitted to the journal) may be shared at any time anywhere. Sharing an article (without review) on a preprint server, for example, is not considered prior publication. Because of that, before final publication, we recommend that authors self-archive the preprint

version on personal or institutional websites, scientific social networks, repositories, reference managers… Once published, if the preprint remains available in a repository, it must be specified that, "This is an electronic version of an article published on Psicothema (year). The final version is available on the official web page", also highlighting the full reference to the published article, including its DOI.

Psicothema's preservation policy. Psicothema publications are available on the journal website and in international open-access repositories online such as SciELO and Dialnet. Psicothema focuses on disseminating content and making it accessible through indexing services. Online access is free, while the printed version is subscription access. In addition, Psicothema allows self-archiving of preprint, post-print, and editorial version (immediately after publication). Exploitation rights and Psicothema's self-archiving permissions may be found at https://dulcinea.opensciencespain.org/ficha1034.

Conditions for self-archiving postprint articles. Authors are encouraged to share the final version of their accepted articles (the authors' version including changes suggested by reviewers and editors) on social networks and repositories until the editorial version is published in an edition of the journal. It must be expressly indicated that this is an article "in press" in the Psicothema journal. Once the editorial version is published, if the postprint is still available in a repository, it must be specified that, "This is an electronic version of an article published on Psicothema (year). The final version is available on the official web page", also highlighting the full reference to the published article, including its DOI.

Archive. The journal undertakes to provide XML metadata or in other specific formats, immediately after its publication and within three months in order to promote its dissemination in databases. Psicothema uses various national and international repositories: Clarivate Analytics, Scopus, Google Scholar, S2ciELO, Dialnet, EBSCO Essentials Academic Search Premier, PubMed, PsycINFO, IBECS, Redinet, Psicodoc, Pubpsych, Fuente Académica Plus, IBZ Online, Periodicals Index Online, MEDLINE, EMBASE, ERIH PLUS, Latindex, MIAR, CARHUS Plus+ 2018, Rebiun, DOAJ, & Crossref.

Digital preservation. In order to preserve permanent access to digital objects hosted on its own servers, Psicothema makes backups, monitors the technological environment to foresee possible migrations of obsolete formats or software, preserves digital metadata, uses DOI Digital Object Identifier) and ORCID. All articles published in Psicothema are also hosted and available in the institutional repository of the Universidad de Oviedo (REUNIDO). The articles published on Psicothema's website are available in easily reproducible format (PDF).

Anti-Plagiarism Policy

In compliance with our code of ethics and in order to guarantee the originality of the manuscripts submitted for evaluation, Psicothema applies anti-plagiarism software to all manuscripts that meet the minimum criteria of the preliminary review and are to be subjected to the review process.

The submission will be rejected in case of detecting practices of plagiarism or scientific fraud, either during the preliminary review by the editorial committee or once the peer review has started.

Articles

Using artificial intelligence in test construction: A practical guide
Javier Suárez-Álvarez, Qiwei He,
Nigel Guenole & Damiano D'Urso..... 1-12

Waiting times in clinical psychology in public mental health units: Predictors of attendance at the first appointment and early dropout
María del Mar Miras-Aguilar, Jose Ruiz-Gutiérrez, Sandra Martínez-Gómez,
Saioa Pérez-García-Abad, Carmen Ramos-Barrón, Emilio Pariente-Rodrigo,
Lourdes Piñán-Setién, Noelia Otero-Cabanillas, María Isabel Priede & César González-Blanch..... 13-22

Assessing positive digital experiences: A Spanish validation of the Digital Flourishing Scale for Adolescents
Alfredo Zarco-Alpuente, Víctor Ciudad Fernández, Jasmina Rosic,
Sophie Janicke-Bowles, Tamara Escrivà-Martínez, & Paula Samper-García 23-35

Assessing impulsivity in adolescents: Psychometric properties of the Spanish short S-UPPS-P
Esteve Montasell-Jordana, Eva Penelo, Laura Blanco-Hinojo,
Beatriz Lanceta, Laura Gomàriz-Camacho, Mar Gràcia,
Anna Soler, Jesús Pujol, & Joan Deus 36-45

Psychometric properties of the Teachers´ Responses to Bullying Questionnaire (TRBQ) in Spanish students
Laura Rodríguez-Pérez, Rosario Del Rey, Noemí García-Sanjuán & Noelia Muñoz-Fernández 46-57

GRUPO EDITORIAL DE PSICOFUNDACIÓN

Psicothema (2026) 38(1) 1-12

Psicothema

https://www.psicothema.com/es • ISSN 0214–9915 • eISSN 1886-144X

Colegio Oficial de Psicología del Principado de Asturias

Psicothema

Article

Using Artificial Intelligence in Test Construction: A Practical Guide

Javier Suárez-Álvarez[1], Qiwei He[2], Nigel Guenole[3] and Damiano D'Urso[4]

[1] University of Massachusetts Amherst (USA)
[2] Georgetown University (USA)
[3] University of London (United Kingdom)
[4] Independent Researcher (Netherlands)

ARTICLE INFO

Received: 01/08/2025
Accepted: 09/10/2025

Keywords:
Artificial intelligence
Test construction
Automated item generation
Validity
Fairness

ABSTRACT

Background: Artificial Intelligence (AI) is increasingly used to enhance traditional assessment practices by improving efficiency, reducing costs, and enabling greater scalability. However, its use has largely been confined to large corporations, with limited uptake by researchers and practitioners. This study aims to critically review current AI-based applications in test construction and propose practical guidelines to help maximize their benefits while addressing potential risks. **Method:** A comprehensive literature review was conducted to examine recent advances in AI-based test construction, focusing on item development and calibration, with real-world examples to demonstrate practical implementation. **Results:** Best practices for AI in test development are evolving, but responsible use requires ongoing human oversight. Effective AI-based item generation depends on quality training data, alignment with intended use, model comparison, and output validation. For calibration, essential steps include defining construct validity, applying prompt engineering, checking semantic alignment, conducting pseudo factor analysis, and evaluating model fit with exploratory methods. **Conclusions:** We propose a practical guide for using generative AI in test development and calibration, targeting challenges related to validity, reliability, and fairness by linking each issue to specific guidelines that promote responsible, effective implementation.

Uso de la Inteligencia Artificial en la Construcción de Pruebas: Una Guía Práctica

Palabras clave:
Inteligencia artificial
Construcción de pruebas
Generación automatizada de ítems
Validez
Equidad

RESUMEN

Antecedentes: La inteligencia artificial (IA) se utiliza crecientemente para mejorar las prácticas tradicionales de evaluación, aumentando la eficiencia, reduciendo costos y facilitando la escalabilidad. Sin embargo, su uso se ha limitado a grandes corporaciones, con escasa adopción por parte de investigadores y profesionales. Este estudio revisa críticamente las aplicaciones de la IA en la construcción de pruebas y propone guias prácticas para maximizar sus beneficios y abordar posibles riesgos. **Método:** Se realizó una revisión exhaustiva de la literatura para examinar los avances en aplicaciones basadas en IA en la construcción de pruebas, con énfasis en el desarrollo y calibración de ítems, y se incluyeron ejemplos del mundo real para mostrar su implementación práctica. **Resultados:** Las mejores prácticas para el uso de IA en el desarrollo de pruebas están en evolución, pero requieren supervisión humana. Para generar ítems se necesitan datos de calidad, alineación con el uso previsto, comparación de modelos y validación. Para calibrar, hay que definir el constructo, optimizar las instrucciones (prompts), verificar la alineación semántica, realizar análisis factoriales pseudoexploratorios y evaluar el ajuste del modelo. **Conclusiones:** Se propone una guía práctica que vincula los desafíos de validez, fiabilidad y equidad con recomendaciones para una implementación responsable y eficaz.

Cite as: Suárez-Álvarez, J., He, Q., Guenole, N., & D´Urso, D. (2026). Using artificial intelligence in test construction: A practical guide. *Psicothema, 38*(1), 1-12. https://doi.org/10.70478/psicothema.2026.38.01
Corresponding author: Javier Suárez-Álvarez, suarezj@umass.edu
This article is published under Creative Commons License 4.0 CC-BY-NC-ND

Artificial Intelligence (AI) is being adopted globally at an unprecedented pace. ChatGPT alone reached 800 million weekly users by April 2025, achieving 90% of its current global user base in just three years. In comparison, the Internet took over 23 years to reach the same level of global adoption (Meeker et al., 2025). Most importantly, its capabilities are still evolving. The Organisation for Economic Co-operation and Development (OECD, 2025) established an independent committee of experts, which estimated that it has reached only about half of its full potential (OECD, 2025). As AI continues to grow, finding ways to use it effectively while reducing potential risks is a major focus for governments, researchers, and practitioners. Educational and psychological assessments are no exception as AI is transforming how tests are designed, delivered, and interpreted.

Educational and psychological assessments are crucial for both individual and societal progress, as they support the identification of needs and the monitoring of progress over time. However, as emphasized in the *Standards for Educational and Psychological Testing* jointly developed by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), assessments must be relevant, valid, and fair to be effective (AERA, APA, & NCME, 2014). Historically, the improvement of these assessments has progressed alongside advances in methodology and technology. For example, in the 20th century, standardized testing provided a systematic method for evaluating the skills and knowledge of large populations (Sireci et al., 2025). Optical scanners later automated the scoring process, enhancing efficiency and reducing errors. Computer-adaptive testing (CAT) advanced the measurement field by adjusting test difficulty based on individual performance, optimizing the accuracy and relevance of assessments for each test-taker (Zenisky & Sireci, 2002).

Traditional test development followed a rigorous process that typically began with defining the assessment purpose and construct to be measured, manually crafting assessment items, and refining them based on pilot studies and psychometric analysis (AERA et al., 2014; Downing & Haladyna, 2006; Lane et al., 2016; Muñiz & Fonseca-Pedrero, 2019). While this systematic approach is still considered the gold standard for creating relevant, valid, fair measurement tools, it does have its drawbacks. Crafting assessment items manually is time-consuming and often expensive, particularly when done by experienced subject-matter experts (SMEs). Additionally, if the assessments' purpose and construct are innovative and groundbreaking such as AI literacy or prompt engineering, finding the appropriate SMEs can be challenging, which limits accessibility for the broader research community (European Commission, OECD, & Code.org., 2025). Another common challenge is generating a sufficiently large pool of items from which to create parallel versions of tests to counteract item content becoming public online (Bißantz et al., 2024). Designing assessments that reflect test takers' funds of knowledge and cultural backgrounds to enhance engagement, and performance is particularly challenging in traditionally developed assessments, due to rigid blueprints, administration conditions, and high development costs (Walker et al., 2023). Traditional test development is also at an increasing risk of assessing skills that humans routinely use machines to perform (Swiecki et al., 2022).

To address these limitations, researchers have long proposed the use of Automated Item Generation (AIG) and predicting item parameters based on item attributes. AIG enables the creation of diverse item versions based on item templates, reducing item reuse and improving cost efficiency (Bejar et al., 2002; Luecht, 2025). Similarly, statistical modeling approaches have been recommended for decades to estimate item complexity by assigning a difficulty score based on item attributes, allowing developers to systematically predict item performance without relying on extensive field testing (Embretson, 1983, 1999; Sheehan & Mislevy, 1994; Sheehan et al., 2006). These analytical methods offer the potential to streamline development by replacing large-scale pilot studies with model-based predictions. However, it is only with recent technological advancements in generative and representational AI using embeddings that these approaches are beginning to realize their full operational potential (see Table 1 for key operational definitions).

In recent years, the automation of test content generation has significantly streamlined the traditionally manual and costly development processes (Attali et al., 2022; Gierl & Haladyna, 2012; von Davier et al., 2024). Automated scoring systems are now routinely used for evaluating constructed responses - a task that previously required human judgment (von Davier et al., 2022; Yamamoto et al., 2019). When well-design prompts are used, large language models (LLM) can enhance efficiency and quality over traditional automated item generation methods (Bezirhan & von Davier, 2023). LLMs can also be used to obtain item parameters estimates prior to collecting empirical data (Feng et al., 2025; Guenole et al., 2024, 2025). AI technologies are helping to define and refine new constructs, like AI literacy, computational thinking, and prompt engineering, that are becoming increasingly important in digital learning environments (European Commission, OECD, & Code.org., 2025). The use of AI enables the development of innovative item formats such as interactive simulations, scenario-based assessments, and chat-based dialogues (Foster & Piacentini, 2023). AI algorithms can be used to map assessment items to learning standards or curriculum frameworks, thereby assisting with instructional alignment and reducing the burden on subject-matter experts (Butterfuss & Doran, 2025). AI supports adaptive testing and personalized learning paths that respond to individual learner characteristics (Arslan et al., 2024; Sireci et al., 2024; Suárez-Álvarez

Table 1
Key Definitions of AI-Driven Methods in Educational and Psychological Assessment

Name	Description	Example
Generative AI (GenAI)	A class of AI models that can generate new content, such as text, images, or code, based on learned patterns from data.	ChatGPT (OpenAI, 2023)
Machine Learning (ML)	A subset of AI that enables systems to learn from data and improve performance on tasks without being explicitly programmed.	Neural Networks (von Davier, 2018).
Natural Language Processing (NLP)	A field of AI focused on enabling machines to understand, interpret, and respond to human language.	Analyzing students' written responses to assess problem-solving strategies (Yaneva von & Davier, 2023).
Large Language Model (LLM)	A type of NLP model trained on massive text to generate and understand human-like language.	GPT-4 or Claude 3 Opus (OpenAI, 2023; Anthropic, 2024)

et al., 2024; Yan et al., 2024). Digital assessments also capture log (process) data, providing invaluable insights into test takers' cognitive processes and engagement with tasks (He et al., 2021, 2023; Ulitzsch et al., 2023; Suárez-Álvarez et., 2022). Although log (process) data has primarily been used to refine estimates of test takers' proficiencies (Pohl et al., 2021; Wise et al., 2021), it can also be employed to identify item attributes and predict item performance.

The goal of this paper is to summarize current best practices in the applications of Generative AI in modern educational and psychological test construction, specifically focusing on item generation and item calibration. These applications are emphasized because they offer significant benefits in terms of cost efficiency and scalability within educational and psychological assessments, and they also present potential threats to reliability, validity, and fairness. Although these applications have been predominantly utilized by large corporations like Duolingo (von Davier et al., 2024), their adoption among the wider research and practitioner community remains limited. The mission of this paper is to disseminate the latest technological advancements to a broader audience, ensuring that these innovations benefit a diverse group and contribute to the development of a wide range of groundbreaking assessments. Finally, a cautionary commentary is included, outlining strategies to maximize the benefits of AI-driven methods in test construction while minimizing potential risks.

Generative AI in Educational Assessment

Generative AI (GenAI hereafter) has emerged as an innovative tool rapidly adopted across various professional fields, efficiently managing repetitive and time-consuming tasks. Education assessment has been significantly transformed by these advancements, with GenAI becoming a contemporary trend in education. AI facilitates interactive and authentic assessment formats, including simulations, virtual reality (VR) integration, and gamified learning experiences. Automated grading and instant feedback reduce teachers' workloads while enabling personalized learning experiences (Mao et al., 2024). Educational chatbots, also known as educational conversational agents (ECAs), are designed to assist teachers, enhance students' learning processes, and evaluate their performance (Chang et al., 2023). Some chatbots are student-oriented, serving as personalized learning assistants that guide students to answers, evaluate their responses, and foster engagement (Kuhail et al., 2023). Others are tailored to support teachers by preparing class materials, managing course schedules, and tracking deadlines (Ramandanis et al., 2023). The applications of GenAI are widely utilized across various subjects, adapting to different educational formats and needs. In this section we describe emerging methods in educational assessments that leverage GenAI for Automated Item Generation (AIG) and summarize current best practices for implementing them.

Automated Item Generation (AIG)

Automated item generation (AIG) has long been a subject of study in employment and educational assessments (Bejar et al., 2002). Creating test questions—especially for medical licensing and certification—requires significant time and financial resources because it depends on expert input for writing scenarios and crafting credible answer choices. Technologies like machine learning or AI that could help lower these development costs are of great interest to test creators. Traditionally, AIG has focused either on non-verbal formats like visual matrix puzzles (Embretson, 1999), or on techniques resembling fill-in-the-blank exercises similar to MadLibs. Since then, GenAI has significantly transformed both reading and language assessment.

In Maas's (2024) recent research, the team applied a fine-tuned Conditional Transformer Language (CTRL) model to generate English reading comprehension questions for educational purposes, with a focus on controllability and alignment to classroom needs. The model was trained on the Reading Comprehension dataset from Examinations (RACE) and clustered latent traits to allow educators to specify desired question types, for example, cloze-style, title-related, or general questions. The training helped improve the generation of questions tailored to specific reasoning skills. The research found that while the fine-tuned model demonstrated promising results in generating relevant and contextual reading questions, challenges such as overfitting and maintaining consistency in generated outputs remain. This required further refinement for practical classroom adoption (Maas, 2024). Another study compared human-designed and AI-generated English reading comprehension materials, using tools like Twee and Kimi to generate multiple-choice questions based on middle school materials. This research used mixed methods by using both quantitative data and qualitative data to explore the human-AI collaboration in comprehension questions generation. The results of the study showed that the AI tool was significantly more time-efficient, requiring only a fraction of the time needed by the human teacher to complete the task, while generating material of comparable quality, although the human was superior in terms of clarity, relevance, and consistency of the questions with the educational objectives. The study also proved that AI tools can effectively complement teachers in content creation, enhancing efficiency while requiring human guidance to ensure pedagogical depth and appropriateness for classroom contexts (Jen et al., 2024).

In addition to the Generative Pre-trained Transformer (GPT) model, widely used for text generation through applications like ChatGPT, the BERT model, which underlies Google's search engine capabilities, has also been widely discussed. For example, Kumar's study combined GPT and BERT in a two-stage architecture to improve the coherence and contextual accuracy of automated text generation. Before training, the team preselected models from GPT, Large Scale Decision-Making (LSDM), and Gated Recurrent Units (GRU) and finally selected GPT as the text generation model. After fine-tuning the model with metrics like Bilingual Evaluation Understudy (BLEU) Score and perplexity to gauge the model's performance, the combined model outperformed the single model across various tasks like question-answering and summarization. The research indicated the potential of combining several models for better AI-driven content creation for future diverse applications (Kumar et al., 2024). GenAI chatbots were also powerful tools for language learning and adaptive questions generation during the learning process. Yang et al. (2022) implemented Ellie, a task-based AI voice chatbot, to support Korean EFL students in practicing English speaking. The chatbot fostered meaningful conversations and achieved high task success rates, with students positively perceiving it as a fun and effective learning tool despite some technical and comprehension challenges. The results highlight the potential of AI chatbots to enhance language education while

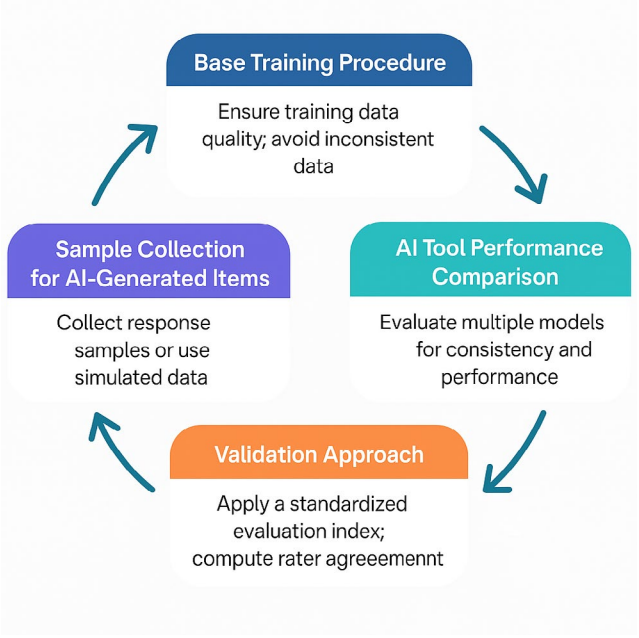
recommending further development to address usability issues and expand application scope (Yang et al., 2022). Von Davier (2018) used a recurrent neural network (RNN) trained on 3,000 test items from the International Personality Item Pool (IPIP) database (Goldberg, 1999), which shows the initial framework of modern test design with a collaboration between human and AI.

Earlier studies noticed that due to limitations in models and data, practical AI-driven AIG was still far off, though the models have been well developed with machine learning techniques. However, as previously noted, the field advanced rapidly when researchers replaced recurrent networks with self-attention-based architectures (Vaswani et al., 2017), enabling simpler designs that support parallel training and allow models to be pre-trained on broad text data before being adapted to specific tasks.

Real-World Example: NAEP Reading Passage Generation

To illustrate how GenAI can support item development, we present an example from the U.S. National Assessment of Educational Progress (NAEP) focused on the generation of reading passages. This process includes ensuring high-quality and consistent training data, evaluating multiple AI models for performance and reliability, applying standardized validation metrics, and collecting response samples to test and refine newly generated items (Figure 1).

Figure 1
Cyclical Framework for Generative AI-Based Test Development



A recent analysis of NAEP reading tasks revealed inconsistencies in readability scores across the training data. We curated reading passages from NAEP-released items spanning Grades 4 and 8, covering the years 1992 to 2020. To maintain consistency in item design, we focused exclusively on text-based passages paired with multiple-choice questions, deliberately excluding content that incorporated tables or figures. This process yielded 24 passages for Grade 4 and 23 passages for Grade 8. To assess the difficulty level and establish a robust base sample, we applied four widely accepted readability indices: Average Reading Level Consensus, Automated Readability Index (Smith & Senter, 1967), Flesch-Kincaid Grade Level (Kincaid et al., 1975), and SMOG Index (McLaughlin, 1969). Contrary to expectations, the results revealed minimal distinction between grades—approximately 75% of the passages exhibited similar readability scores, making them indistinguishable in terms of grade-level appropriateness.

Inconsistencies such as these can introduce substantial variability in model performance. Moreover, training on biased or misaligned data risks reinforcing and amplifying those biases in model outputs. This is especially concerning when employing general-purpose pre-trained models, where human oversight becomes essential to ensure cultural relevance, fairness, and appropriateness.

To address these challenges and construct a clearly defined, representative training set, we collaborated closely with item developers. Together, we identified and selected six prototypical passages for each grade to serve as the foundation for model training. Figure S1 (Supplementary Material) shows the results from four readability metrics before and after the selection process. It apparently shows a smaller variance after the careful selection for training data. This more accurate training set significantly contributes to the accuracy of AI generation results. It is noted that AI generated results kept at the comparable level as the training set index results. The Fleisch Kincaid Grade Level index consistently showed the lowest value of readability compared with their peers.

NAEP reading passage generation findings indicate that AI-generated nonfiction passages demonstrate a significantly higher difficulty level than fiction passages. This discrepancy likely stems from the inherent variability and creative divergence of fiction writing, which contrasts with the more structured nature of nonfiction texts. Figure S2 (Supplementary Material) presents AI-generated fiction and nonfiction passages for Grade 4. While the nonfiction passages exhibit relatively higher readability scores across all indices—suggesting a level above Grade 4—the fiction passages more closely match the required difficulty range.

To improve the performance of AI in generating fiction content, augmenting the input prompts has shown promise. For example, including explicit labels such as "fiction" or "nonfiction" during training, and emphasizing genre-specific textual features in the prompts, can help guide the AI towards producing passages more consistent with training expectations. These refinements contribute to marginal improvements in readability scores and better alignment with task design.

In this example, we trained AI models using LLMs implemented in ChatGPT, Meta AI, and Claude to generate 40 new passages for Grade 4 and Grade 8 respectively. The readability of these AI-generated passages was reassessed to determine whether they matched the target grade levels. To enhance generation quality, we employed an iterative approach to prompt engineering. Initially, we provided a general description of key differences between Grade 4 and Grade 8 reading levels, including vocabulary complexity, sentence structure, and word count. Our preliminary prompts led to AI-generated passages that mimicked these linguistic features but did not consistently align with expected readability index score ranges. To refine the process, we revised our prompts by explicitly quantifying readability standards, detailing the significance of readability indices, and explaining how they are calculated. This structured approach improved alignment with actual readability levels. Among the three AI tools, ChatGPT demonstrated the most

effective performance in passage generation, particularly when utilizing customized GPT functions. The language in the reading passage generated from ChatGPT shows richer descriptions and is highly consistent with the grade level indexes.

As pointed out earlier, we used the consistent evaluation method by using the four readability indicators. This evaluation standard is unchanged between human and AI generated items. As Figure S2 (c) shows (Supplementary Material), the language in the reading passage generated from ChatGPT shows richer descriptions and is highly consistent with the grade level indexes.

Finally, we invited human item developers to help validate the generated items by giving multiple dimensions and calculated the consistency. Though there was no real data collected to validate the items, the experienced human developers give a relatively objective evaluation. In the future study, it is highly recommended to consider using simulated data and/or new sample data collection to make a further validation on the passages.

Practical Guide for Generative AI-Based Test Development

This section provides a practical guide (Table 2) for developing tests using GenAI, aimed at maximizing relevance, validity, and fairness throughout the test construction process.

1. Ensure Consistency and Quality in Training Data

Ensuring the quality of the training dataset is essential for conveying accurate information during the learning process. All materials must undergo rigorous review to confirm the inclusion of high-quality items before they are used for AI training (AERA et al., 2014; Downing & Haladyna, 2006; Lane, Raymond, & Haladyna, 2016; Muñiz & Fonseca-Pedrero, 2019). This step is vital to support critical learning and clear representation of labels in the model.

2. Align AI Use with Intended Uses and Task Type

When using AI for item generation, it is essential to consider both the intended use and the nature of the task. AI models tend to excel at rule-based or logic-driven tasks, yet they often struggle with fiction and emotionally nuanced content. Tasks that require complex human emotion or creativity typically demand additional validation to ensure quality and appropriateness.

3. Compare Multiple AI Models for Reliability

To ensure consistent and reliable outcomes, it is highly recommended to employ at least two AI models and carefully evaluate their performance. Comparing outputs, such as those from ChatGPT and the Claude model, can help identify discrepancies, assess robustness, and improve the overall quality of generated items.

4. Apply a Standardized Validation Approach

Use a consistent evaluation index to assess both training and AI-generated outputs. This ensures alignment with baseline standards and allows for meaningful performance comparisons. Treat AI-generated responses as those from a "human" rater to calculate inter-rater agreement. For example, by verifying whether passages fall within the same readability grade level. This guideline aligns with and extends general guidance on evidence for test validation (Sireci & Benítez, 2023) specifically to AI-based assessments.

5. Verify and Validate AI-Generated Items

While collecting new human response data to evaluate freshly generated items is the most rigorous validation approach, it may not always be feasible due to cost and time restriction. In AI contexts, "verification" often denotes confirming that AI systems are working correctly internally before submitting them for validation scrutiny. This involves checking that AI algorithms generate items as intended, free from technical errors, bias, or unintended patterns, which creates an additional layer addressing the "black box" nature of AI compared to traditional assessment development. For example, consider using AI-simulated data to calibrate item parameters and compare them with the training set (e.g., through Differential Item Functioning analysis), or apply NLP techniques to measure semantic distance between AI-generated items and the original dataset to ensure content alignment and diversity.

Generative AI in Psychological Assessment

GenAI is increasingly applied in psychological assessment and practice, with examples ranging from enhancing diagnostic accuracy and therapeutic interventions in clinical psychology (De la Fuente & Armayones, 2025) to using ChatGPT as a simulated patient to support interactive training and skill development (Sanz et al., 2025). Recent advances in Representational AI using embeddings and GenAI have led to novel approaches in psychological assessment, offering alternatives to traditional self-report methods and enhancing item development, and validation. Generative models (decoders) help create text, such as test items, while representational models (encoders) convert text into numerical formats (embeddings) for analysis. This approach offers a promising way to modernize and improve measurement in psychology (Wulff & Mata, 2025). These embeddings can be used in methods like Pseudo Factor Analysis (PFA) to explore psychological constructs and address issues such as overlap between scales (Guenole et al., 2025). On the other hand, Large Language Models (LLMs) such as GPT-4o and Claude 3 can be used to predict correlations between personality items more accurately than human experts (Schoenegger et al., 2025). Another application comes from Fan et al. (2023), who examined the psychometric properties of personality scores inferred by AI chatbots. These scores, derived from users' free-text input during conversational interactions, showed acceptable reliability and convergent validity but limited discriminant and criterion-related validity. Yuan et al. (2024) examined how users perceive personality scores generated by AI chatbots compared to traditional self-report questionnaires. While users found both methods similarly satisfying and accurate, they tended to view the survey-based results as more trustworthy, likely due to their greater familiarity and simplicity. Sun et al. (2024) presented a framework for developing and validating an AI chatbot based on the Big Five personality model. They emphasize the chatbot's ability to elicit rich, narrative responses aligned with psychological constructs and report improved validity outcomes compared to existing tools. In this section we describe emerging methods in psychological assessment that leverage LLMs for scale

construction. We discuss item generation, how to check semantic item alignment, and PFA.

Item Generation, Semantic Item Alignment, and Pseudo Factor Analysis (PFA)

When designing a new assessment, conceptual clarification of how the construct is similar to and different from related constructs is an important step. This can occur qualitatively using subject matter experts before data are collected, but LLMs present the possibility to approach this task analytically with sentence encoders. A sentence encoder is a transformer-based model trained on text to produce highly dense numerical representations of sentences in vector form. These representations are commonly known as embeddings. Association measures such as cosine similarity can be used to compare the similarities of embeddings created from construct definitions. This allows practitioners to determine the constructs' semantic positions in a nomological network, in turn allowing us to move to item generation.

One of the most important requirements is designing effective instructions for the AI, known as prompt engineering, to ensure the output aligns with your goals while minimizing hallucinations and misinterpretations. Prompt engineering with few constraints on instructions leads to direct item generation, where we instruct the LLM to generate items measuring the focal construct without restrictions. We can also use guided item generation methods, where we provide detailed instructions about item requirements, such as construct definitions, item templates, and other constraints necessary such as item polarity (Ferrando et al., 2025). Whether direct or guided item generation is used, we can provide or omit example items in the LLM prompt. If no item examples are given, the approach is zero-shot prompting, giving less control over the items that are created. If we do give examples, we refer to the method as few-shot prompting, which grounds the model in the task context.

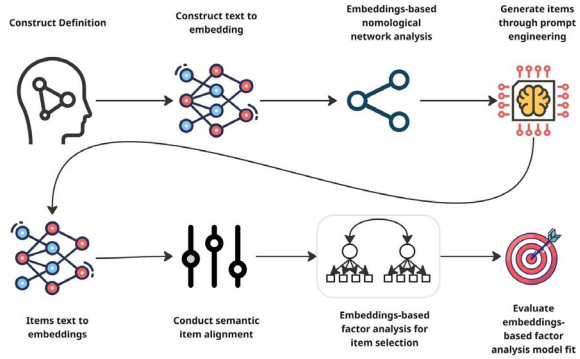
Despite giving instructions regarding item features, generated items might not always match our criteria. Quality checks can be implemented as constraints during the item generation process itself. Alternatively, items might be checked with a prompting approach post generation. If the number of items is small (e.g. several hundred or fewer) it is feasible to check these manually and ultimately all items should be human reviewed. As suggested in the educational assessment section, LLMs can also be used to check semantic item alignment with construct definitions. To check semantic item alignment, encodings are generated between the items and the construct definitions, and the cosine similarities are calculated. Items should have high similarities with their parent constructs and low similarities with non-parent constructs. High and low here do not have fixed values, item parent similarities and item non-parent similarities need to be interpreted relative to one another.

With items generated and pre-screened via semantic item analysis, the factor structure of the items can be examined before responses data are collected with PFA. Similar to traditional factor analysis, PFA allows for different degrees of prior expectations through the use of target rotation. This flexibility enables both fully exploratory analyses, with no prior assumptions, and semi-confirmatory approaches to examine how items group and cluster. At the heart of PFA is the "substitutability assumption", or the idea that the embedding vector for an item statement can stand in for an empirical response vector. This involves forming a cosine similarity matrix between the item embeddings from the previous step, and factor analyzing the matrix in essentially the same way that a correlation matrix of real item responses is analyzed.

Real-World Example: Moral Foundations Scale Calibration

As in the previous section, we use a real-world example to illustrate how GenAI can support AI-based item calibration. This section focuses on the design of a measure targeting executive moral foundations (Graham et al., 2009). Moral foundations are important for senior executives because they make decisions that affect many workers, and these decisions are frequently evaluated in moral terms. Moral foundations are conceptually distinct from familiar industrial psychology constructs, yet they are infrequently included in executive assessment processes. We propose a new moral foundations scale using AI. We show that when our proposed pipeline is followed (Figure 2), PFA can be an effective data-less method for obtaining item pre-knowledge in scale development. We also discuss the challenges relevant to PFA including assessing model fit without sample sizes using raw residuals. We begin our analysis pipeline after we have generated items. More details on the item generation process itself are available in Guenole (2025).

Figure 2
Analytical Pipeline for Generative AI-Based Item Calibration



To prepare items for analysis, we first prepared a file of our moral foundations' items (Supplementary Material: factor.csv). We used the MiniLM sentence encoder to generate embeddings of these items in a Jupyter notebook (matrix_generation.ipynb). The notebook uses MiniLM to convert each item into a numerical representation called an embedding, which captures the semantic meaning of the item. Each embedding has hundreds of numbers (dimensions), and the notebook organizes these into columns (one column per dimension). The notebook calculates how similar each item embedding row is to every other item, creating a similarity matrix, much like how you'd calculate correlations between item responses. The output matrix (matrix.csv) can then be prepared for factor analysis by setting any diagonals that are less than one due to rounding errors to 1, as they are in a correlation matrix (matrix.csv). Early theorizing about why this approach works

rests on a substitutability assumption (Guenole et al. 2025). This is the notion that a numerical item embedding can substitute for an empirical item response vector under certain conditions.

Next, a factor analysis can be performed on the similarity matrix in R (pfa.R) using any extraction and rotation method. Maximum likelihood estimation with oblique rotation, which allows the resulting factors to be related to each other, have been shown to work well in earlier work. The output includes familiar results from traditional factor analysis, such as eigenvalues, a scree plot, and a pattern loading matrix showing which items load onto which factors. While we present the factor analysis for the final item set, we intentionally included about twice as many items as we intended to keep. This gave us the flexibility to run several rounds of analysis, removing items that didn't load well on any factor or that cross-loaded on multiple factors. After each round of removal, we updated the matrix and repeated the analysis to refine the item set. The items, embedding code, and R code to produce the final factor model are included in Supplemental Materials.

Most methods conventionally used to decide on item retention in the context of EFA can be used with PFA. In the current example we soon discuss, we proposed ensuring that items have their highest loading on their parent factor; that this loading is higher than its loading on any other factor; that this loading is higher than its average loading across all other factors; and that its loading is higher than the average of all other item loadings on that factor. From the pattern matrix in Table S1 (Supplementary Materials) we see that this is the case for most items of the newly developed executive moral foundation scale. From the scree plot in Figure S3 (Supplementary Materials), we see that six factors are plausible, which in fact was the expectation at the outset.

One important point about this approach is that the factor analysis is based on the embedding similarities rather than human responses and therefore there is no sample size. Sample sizes are required for many model-based fit tests and indexes. It is not recommended to simply assume an arbitrarily large sample size, because model fit statistics are influenced by sample size and the correct sample size is required. In this case, we recommend using model free and exploratory approaches to checking model fit based on interpreting the raw residuals. There are several exploratory approaches that might be useful depending on the goal and we describe these here now.

We first plot a heat map of the residual correlations. What we hope to see is that most residual correlations are white indicating they are near zero. We do not want to see any obvious patterns with blocks of blue or red indicating systematically low or high residual correlations between the items after conditioning on the latent factors. In Figure S4 (Supplementary Materials) we see this is mostly the case. We might also plot the distribution of off-diagonal elements of the residual correlation matrix, expecting to see relatively small residuals with few outliers. Again, this appears mostly the case in Figure S5 (Supplementary Materials). Finally, we may choose to plot the original versus the residual correlations. Ideally, we would see a horizontal band of residuals clustered around zero, which is broadly what we see in Figure S5 (Supplementary Materials). We also calculated the Root Mean Square Residual (.037) and the Common Part Accounted for (CAF, Lorenzo-Seva et al., 2011) (.87) which are both indicative of good fit.

Critically, we do not yet present empirical relations with actual factor loadings from participant responses, and this is always an important step. Earlier work by Guenole et al. (2025) shows that pseudo factor loadings are related to empirical loadings, but this is an important next step for the executive moral foundations assessment. We also note while the pseudo and empirical loadings themselves have been shown to be highly correlated. The pseudo factor loadings do not yet differentiate reverse keyed items in the way conventional items do, because cosine similarities between embeddings tend to be positive. Nonetheless, it is still critical to compare pseudo factor structures derived from embeddings with empirical factor structures based on human responses. Ultimately, the empirical factor structure remains the gold standard. Once empirical data are available, alignment between models can be assessed using quantitative metrics such as Tucker's congruence coefficient (values > .85 indicate fair similarity; > .95 indicate strong alignment) and correlation coefficients between corresponding factors (Guenole et al., 2025). Readers may also wish to explore alternative approaches to assessing item dimensionality and discrimination through embedding-based network models (Russell-Lasalandra et al., 2024).

Practical Guide for Generative AI-Based Item Calibration

This section provides a practical guide (Table 2) for item calibration using GenAI, aimed at maximizing relevance, validity, and fairness throughout the test construction process.

6. Use Sentence Encoders to Establish Semantic Construct Validity

Before item generation, clarify how the target construct is similar to or distinct from related constructs. By comparing the semantic similarity of construct definitions within a nomological network, developers can validate construct boundaries early in the design process, improving alignment and focus on subsequent item development.

7. Apply Prompt Engineering Strategies for LLM-Based Item Generation

When generating non-cognitive assessment items with LLMs, use prompt engineering strategies that match the desired level of control. Guided prompts with examples (few-shot) offer greater precision, while minimal prompts without examples (zero-shot) allow more creativity but less control. The choice should reflect the specificity and psychometric standards required for the assessment.

8. Conduct Semantic Item Alignment to Ensure Construct Relevance

To ensure AI-generated items align with the intended construct, apply semantic alignment checks either during or after item generation. This can involve manual review or LLM-based methods, such as calculating cosine similarity between item and construct embeddings. Items should show relatively higher similarity to their target construct than to unrelated ones, guiding item selection and refinement.

9. Use Embedding-Based Factor Analysis with Iterative Refinement for Item Selection

To evaluate AI-generated items, convert item text into embeddings using an LLM and analyze the resulting similarity matrix with factor analysis. Begin with a large item pool to allow for iterative refinement, removing items with weak or cross-loadings. Assign items to factors using systematic criteria based on loading strength and distinctiveness. Ensure the process is transparent and reproducible using shared data and code.

10. Use Model-Free Exploratory Techniques to Evaluate Fit in Embedding-Based Factor Analysis

When factor analyzing item embeddings without response data, traditional fit indices can't be used due to the lack of a sample size. Instead, apply model-free exploratory methods such as heatmaps of residual correlations, distributions of off-diagonal residuals, and plots comparing original to residual correlations to assess whether the latent structure fits the data well.

Table 2
Practical Guide to Generative AI–Based Test Development and Calibration

Generative AI-Based Application	Guidelines
Test Development	*1. Ensure Consistency and Quality in Training Data*
	2. Align AI Use with Intended Uses and Task Type
	3. Compare Multiple AI Models for Reliability
	4. Apply a Standardized Validation Approach
	5. Verify and Validate AI-Generated Items
	6. Use Sentence Encoders to Establish Semantic Construct Validity
	7. Apply Prompt Engineering Strategies for LLM-Based Item Generation
Item Calibration	*8. Conduct Semantic Item Alignment to Ensure Construct Relevance*
	9. Use Embedding-Based Factor Analysis with Iterative Refinement for Item Selection
	10. Use Model-Free Exploratory Techniques to Evaluate Fit in Embedding-Based Factor Analysis

Maximizing Benefits While Reducing Risks

As public trust and engagement in standardized testing declines (Borgonovi & Suárez-Álvarez, 2025; Suárez-Álvarez et al., 2024), AI-driven methods, such ML, NLP, and LLM (see Table 1 for definitions), are being increasingly applied to optimize traditional measurement approaches (Hao et al, 2024; Yaneva & von Davier, 2023). While these innovations offer important gains in efficiency, cost, and scalability, there is a risk that, without also addressing broader concerns of trust, equity, and relevance, educational and psychological measurement may become increasingly disconnected from evolving scientific standards, societal needs, and ethical principles (Burstein et al., 2025; Johnson et al., 2025; Walker et al., 2023). Therefore, to fully harness the benefits of technological innovations like AI in promoting individual and societal progress, it is essential to understand their limitations (Bulut et al., 2024; Dixon-Roman, 2024; Dumas, Greiff, & Wetzel, 2025; Hao et al., 2024; Ho, 2024; Yan, Greiff et al., 2024; Swiecki et al., 2022).

The following section summarizes current limitations of AI-based methods for test construction, organized into four key areas: validity (explainability), reliability (consistency, and generalizability), fairness (training data quality), and data security and privacy. Each issue is linked to specific guidelines to support implementation. However, given the conceptual and practical overlap among these issues and the guidelines to address them, some level of interaction between them is to be expected.

Validity and the "Black Box" Problem

One of the most pressing validity concerns is the lack of transparency in how large AI models make predictions, a challenge often referred to as the *black box problem*. Unlike theory-driven methods grounded in Karl Popper's falsifiability principle, where a scientific theory must be testable and subject to empirical disconfirmation, data-driven AI models do not typically allow for such scrutiny. While these models can serve valuable roles in educational and psychological measurement, the absence of a clear theoretical foundation increases the risk of speculative or spurious conclusions. Rather than discarding theory when confronted with data inconsistencies, we argue for refining theoretical frameworks using advanced methodologies. Empirical inquiry should be guided, and at minimum verified, by theory, not divorced from it.

Furthermore, Explainable Artificial Intelligence (XAI) aims to make AI models more transparent and interpretable, addressing concerns related to model opacity and validity (Samek et al., 2017). By providing clear and understandable explanations of how decisions are made, XAI helps build trust and facilitates validation, particularly in high-stakes domains. This approach has shown promising results in healthcare, improving both clinician understanding and patient outcomes (Doshi-Velez & Kim, 2017; Holzinger et al., 2019). Given these successes, there is growing interest in applying XAI techniques to the educational (Khosravi et al., 2022) and psychological fields (Joyce et al., 2023) to enhance the interpretability and acceptance of AI-driven assessment tools. Our current efforts focus on adapting XAI methods to support transparent and valid test development processes.

Guideline 4 directly addresses the validity concern by establishing systematic methods for evaluating whether AI-generated outputs align with intended constructs. It helps make the AI's decision-making process more interpretable and transparent, reducing the "black box" nature of the model. *Guideline 5* supports construct validity by ensuring that the generated items are actually measuring what they are intended to measure. Through expert review, semantic alignment, or empirical validation, this step helps mitigate the opacity of the model's outputs. *Guideline 6* helps clarify how constructs are defined and differentiated prior to item generation, enhancing conceptual transparency. *Guideline 8* ensures that generated items align with the intended construct, providing a data-driven check on construct representation. Finally, *Guideline 9* offers a framework for analyzing the dimensionality of AI-generated items, thereby supporting construct validity through empirical evidence.

Reliability and the "Hallucination" Problem

Another major threat is (un)reliability. AI models can produce errors, respond inconsistently to identical prompts, and struggle with abstract reasoning, logical inference, or unfamiliar content, issues commonly referred to as *hallucinations*. Although *Guidelines 2* and *3* are intended to mitigate these risks by encouraging task-model alignment and multi-model comparisons, consistent human verification remains essential (see also *Guidelines 4* and *5*).

Guideline 7 recommends using prompt engineering strategies that align with the intended purpose to structure, and guide prompts effectively. This approach reduces variability, increases the consistency of AI-generated items, and is also expected to enhance validity. *Guideline 9* advises applying embedding-based factor analysis iteratively to identify and remove items with weak or inconsistent loadings, thereby enhancing item stability and internal consistency. Finally, *Guideline 10* encourages the use of model-free exploratory techniques to empirically assess internal consistency and dimensional coherence. These methods help identify unreliable or poorly fitting items and support improvements to both internal consistency and the underlying structure of the scale.

Fairness and the "Alignment Gap"

Fairness is compromised when pre-trained models, such as those behind ChatGPT, are used without scrutiny of the cultural responsiveness of their training data. This *alignment gap* reflects a disconnect between model training and intended test use. When sufficient task-specific data are available, *Guideline 1* recommends training models directly on curated, high-quality content. However, when relying on general-purpose pre-trained models, extreme caution is warranted. Human oversight and review are essential to ensure cultural relevance and appropriateness (see *Guidelines 4 and 5*). Our approach maintains a clear boundary between AI-based assessments and the ultimate decision-making responsibilities of psychologists and educators, reinforcing that AI serves as an aid rather than a substitute.

Guideline 6 also aims to ensure that constructs are clearly defined and culturally grounded, helping to reduce the risk of biased construct representation. *Guideline 8* recommends systematically evaluating whether items accurately reflect the target construct across diverse populations. Additionally, *Guideline 7* supports greater control over content generation by incorporating constraints that promote inclusivity and cultural responsiveness.

Data Security and Privacy

Although not directly related to validity, reliability, and fairness, data privacy and security are crucial ethical considerations. Consumer-facing tools like ChatGPT may use submitted prompts and generated responses to further train their models. This poses risks when test content or sensitive data are entered into such platforms. Also, the legal and ethical aspects of content ownership generated by AI warrant future discussion to inform policy and practice.

This issue is addressed through strong data governance practices that ensure sensitive information used in AI-assisted test construction is protected throughout the development process. This includes establishing clear protocols for data access, ensuring compliance with privacy regulations, avoiding the use of open-access consumer AI tools that may reuse input data (such as ChatGPT's free version), and using secure environments for storing and processing both training data and AI-generated content. Effective governance also involves transparency in how data are handled and ensuring that personal or confidential educational data are not inadvertently exposed or misused.

Concluding Remarks

GenAI holds great promise for transforming assessments by enabling faster, more adaptive, and scalable test development. Techniques like embedding-based item evaluation can streamline early test design and reduce costs, helping bridge the gap between semantic AI models and traditional psychometric practices (Guenole et al., 2025; Russell-Lasalandra et al., 2024). However, these innovations must be implemented with caution. Risks such as academic misconduct, technical vulnerabilities, and disciplinary skepticism highlight the need for thoughtful integration (Alasadi et al., 2023; Dolenc et al., 2024; Farrelly et al., 2023; Wang et al., 2023). Crucially, the effectiveness of AI-based tools depends on their alignment with core psychometric principles. Without clear evidence of reliability, validity, and fairness, even the most advanced systems remain superficial. Moving forward, assessment professionals must balance innovation with rigorous empirical standards and ethical safeguards to ensure responsible use of GenAI.

Authors Contributions

Javier Suárez-Álvarez: Conceptualization, Writing - Original draft. **Qiwei He**: Conceptualization, Writing - Original draft. **Nigel Guenole**: Conceptualization, Software, Writing - Original draft. **Damiano D'Urso**: Software, Writing - Review & editing

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors

Declaration of Interests

The author(s) declare(s) that there is no conflict of interest

Data Availability Statement

Supplementary material for this article is available online in the following link: https://osf.io/fvzyx/?view_only=27238879597b42f984ec7e7b2c721041

References

Alasadi, E. A., & Baiz, C. R. (2023). Generative AI in education and research: Opportunities, concerns, and solutions. *Journal of Chemical Education, 100*(8), 2965–2971. https://doi.org/10.1021/acs.jchemed.3c00323

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Anthropic. (2024). Claude 3 Opus [Large language model]. https://www.anthropic.com

Arslan, B., Lehman, B., Tenison, C., Sparks, J. R., López, A. A., Gu, L., & Zapata-Rivera, D. (2024). Opportunities and challenges of using generative AI to personalize educational assessment. *Frontiers in Artificial Intelligence, 7,* 1460651. https://doi.org/10.3389/frai.2024.1460651

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence, 5,* 903077. https://doi.org/10.3389/frai.2022.903077

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2002). A feasibility study of on-the-fly item generation in adaptive testing. *ETS Research Report Series*, i-44.

Bezırhan, U., & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence, 5*, 100161. https://doi.org/10.1016/j.caeai.2023.100161

Bißantz, S., Frick, S., Melinscak, F., Iliescu, D., & Wetzel, E. (2024). The potential of machine learning methods in psychological assessment and test construction. *European Journal of Psychological Assessment, 40*(1), 1–4. https://doi.org/10.1027/1015-5759/a000817

Borgonovi, F. & Suárez-Álvarez, J (2025). *How can adult skills assessments best meet the demands of the 21st century?*. OECD Social, Employment and Migration Working Papers, No. 319. OECD Publishing. https://doi.org/10.1787/853db37b-en

Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., Ormerod, C., Fabiyi, D. G., Ivan, R., Walsh, C., Rios, O., Wilson, J., Yildirim-Erbasli, S. N., Wongvorachan, T., Liu, J. X., Tan, B., & Morilova, P. (2024). The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges. *Chinese/English Journal of Educational Measurement and Evaluation, 5*(3), 3. https://doi.org/10.59863/MIQL7785

Burstein, J. (2025, April 17). *The Duolingo English Test responsible AI standards (Duolingo Research Report No. DRR-25-05)*. Duolingo. https://englishtest.duolingo.com/research

Butterfuss, R., & Doran, H. (2025). An application of text embeddings to support alignment of educational content standards. *Educational Measurement: Issues and Practice, 44*(1), 73–83. https://doi.org/10.1111/emip.12581

Chang, D. H., Lin, M. P.-C., Hajian, S., & Wang, Q. Q. (2023). Educational design principles of using AI chatbot that supports self-regulated learning in education: Goal setting, feedback, and personalization. *Sustainability, 15*(17), 12921.

De la Fuente, D., & Armayones, M. (2025). AI in psychological practice: What tools are available and how can they help in clinical psychology? *Psychologist Papers, 46*(1), 18-24. https://doi.org/10.70478/pap.psicol.2025.46.03

Dixon-Román, E. (2024). AI and psychometrics: Epistemology, process, and politics. *Journal of Educational and Behavioral Statistics, 49*(5), 709–714. https://doi.org/10.3102/10769986241280623

Dolenc, K., & Brumen, M. (2024). Exploring social and computer science students' perceptions of AI integration in (foreign) language instruction. *Computers and Education: Artificial Intelligence, 7*, 100285.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv.* https://doi.org/10.48550/arXiv.1702.08608

Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Lawrence Erlbaum Associates Publishers.

Dumas, D., Greiff, S., & Wetzel, E. (2025). Ten guidelines for scoring psychological assessments using artificial intelligence [Editorial]. *European Journal of Psychological Assessment, 41*(3), 169–173. https://doi.org/10.1027/1015-5759/a000904

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179–197. https://doi.org/10.1037/0033-2909.93.1.179

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64*(4), 407-433.

European Commission, OECD, & Code.org. (2025, May). *Empowering learners for the age of AI: An AI literacy framework for primary and secondary education* (Review draft). https://www.oecd.org/digital/empowering-learners-ai-literacy-framework

Fan, J., Sun, T., Liu, J., Zhao, T., Zhang, B., Chen, Z., … Hack, E. (2023, January 5). How well can an AI chatbot infer personality? Examining psychometric properties of machine-inferred personality scores. *PsyArXiv.* https://doi.org/10.31234/osf.io/pk2b7

Farrelly, T., & Baker, N. (2023). Generative artificial intelligence: Implications and considerations for higher education practice. *Education Sciences, 13*(11), 1109.

Feng, W., Tran, P., Sireci, S., & Lan, A. S. (2025). *Reasoning and sampling-augmented MCQ difficulty prediction via LLMs*. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), *Artificial intelligence in education. AIED 2025* (Lecture Notes in Computer Science, Vol. 15880). Springer, Cham. https://doi.org/10.1007/978-3-031-98459-4_3

Ferrando, P. J., Morales-Vives, F., Casas, J. M., & Muñiz, J. (2025). Likert scales: A practical guide to their design, construction and use. *Psicothema, 37*(4), 1–15. https://doi.org/10.70478/psicothema.2025.37.24

Foster, N. & Piacentini, M (2023). *Innovating assessments to measure and support complex skills.* OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Gierl, M. J., & Haladyna, T. M. (2012). *Automatic item generation*. Routledge. https://doi.org/10.4324/9780203803912

Goldberg, L. R. (1999). *A broad-bandwidth, public domain personality inventory measuring the lower-level facets of several five-factor models*. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), Personality Psychology in Europe (Vol. 7, pp. 7-28). Tilburg University Press

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5), 1029.

Guenole, N., D'Urso, E. D., Samo, A., Sun, T., & Haslbeck, J. (2025).Enhancing scale development: Pseudo factor analysis of language embedding similarity matrices. *PsyArXiv*. https://osf.io/preprints/psyarxiv/vf3se_v2

Guenole, N., Samo, A., Sun, T. (2024). Pseudo-Discrimination Parameters from Language Embeddings. *OSF*. https://osf.io/9a4qx_v1

Guenole, N. (2025). *Psychometrics.ai: Transforming Behavioral Science with Machine Learning*. https://psychometrics.ai

Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice, 43*(2), 16-29. https://doi.org/10.1111/emip.12602

He, Q., Borgonovi, F., Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Identifying generalized behavioral patterns with sequence mining. *Computers and Education, 166*, 104170. https://doi.org/10.1016/j.compedu.2021.104170

He, Q., Borgonovi, F., & Suárez-Álvarez, J. (2023). Clustering sequential navigation patterns in multiple-source reading tasks with dynamic time warping method. *Journal of Computer Assisted Learning, 39*, 719–736. https://doi.org/10.1111/jcal.12748

Ho, A. D. (2024). Artificial intelligence and educational measurement: Opportunities and threats. *Journal of Educational and Behavioral Statistics, 49*(5), 715-722. https://doi.org/10.3102/10769986241248771

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(4), e1312. https://doi.org/10.1002/widm.1312

Jen, F.-L., Huang, X., Liu, X., & Jiao, J. (2024). *Can generative AI really empower teachers' professional practices? Comparative study on human-tailored and GenAI-designed reading comprehension learning materials*. In L. K. Lee, P. Poulova, K. T. Chui, M. Černá, F. L. Wang, & S. K. S. Cheung (Eds.), *Technology in Education. Digital and Intelligent Education. ICTE 2024. Communications in Computer and Information Science, vol. 2330* (pp. 112–123). Springer.

Johnson, M. S. (2025, April). *Responsible AI for measurement and learning: Principles and practices* (ETS Research Report No. RR-25-03). ETS Research Institute.

Joyce, D. W., Kormilitzin, A., Smith, K. A., & Cipriani, A. (2023). Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *NPJ Digital Medicine, 6*(6). https://doi.org/10.1038/s41746-023-00751-9

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence, 3*, 100074. https://doi.org/10.1016/j.caeai.2022.100074

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.* (Research Report No. 56). Institute for Simulation and Training. https://stars.library.ucf.edu/istlibrary/56

Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies, 28*, 973–1018. https://doi.org/10.1007/s10639-022-11177-3

Kumar, P., Manikandan, S., & Kishore, R. (2024). *AI-driven text generation: A novel GPT-based approach for automated content creation*. 2024 2nd International Conference on Networking and Communications (ICNWC). IEEE.

Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.). (2016). *Handbook of test development (2nd ed.)*. Routledge.

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research, 46*(2), 340–364. https://doi.org/10.1080/00273171.2011.564527

Luecht, R. M. (2025). *Assessment engineering in test design: Methods and applications* (1st ed.). Routledge. https://doi.org/10.4324/9781003449464

Maas, A. C. (2024). *An empirical study on training generative AI to create appropriate questions for English reading comprehension* [Doctoral dissertation, Tohoku University]. Tohoku University Repository.

Mao, J., Chen, B., & Liu, J. C. (2024). Generative artificial intelligence in education and its implications for assessment. TechTrends, *68*(1), 58-66.

Meeker, M., Simons, J., Chae, D., & Krey, A. (2025). *Trends – artificial intelligence (AI).* BOND. https://www.bondcap.com/report/tai/

McLaughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of Reading, 12*(8), 639-646.

Muñiz, J., & Fonseca-Pedrero, E. (2019). Ten steps for test development. *Psicothema, 31*(1), 7–16. https://doi.org/10.7334/psicothema2018.291

OECD (2025). *Introducing the OECD AI capability indicators.* OECD Publishing. https://doi.org/10.1787/be745f04-en

OpenAI. (2023). GPT-4 technical report. https://arxiv.org/abs/2303.08774

Pohl, S., Ulitzsch, E., & von Davier, M. (2021). Reframing rankings in educational assessments. *Science, 372*(6540), 338–340. https://doi.org/10.1126/science.abd3300

Ramandanis, D., & Xinogalos, S. (2023). Designing a chatbot for contemporary education: A systematic literature review. *Information, 14*(9), 503.

Russell-Lasalandra, L. L., Christensen, A. P., & Golino, H. (2024, September 12). Generative psychometrics via AI-GENIE: Automatic item generation and validation via network-integrated evaluation. *PsyArXiv.* https://doi.org/10.31234/osf.io/fgbj4

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv.* https://doi.org/10.48550/arXiv.1708.08296

Sanz, A., Tapia, J. L., García-Carpintero, E., Rocabado, J. F., & Pedrajas, L. M. (2025). ChatGPT simulated patient: Use in clinical training in Psychology. *Psicothema, 37*(3), 23-32. https://doi.org/10.70478/psicothema.2025.37.21

Schoenegger, P., Greenberg, S., Grishin, A., Lewis, J., & Caviola, L. (2025). AI can outperform humans in predicting correlations between personality items. *Communications Psychology, 3*, 23. https://doi.org/10.1038/s44203-025-00123-1

Sheehan, K. M., Kostin, I., & Persky, H. (2006, April). *Predicting item difficulty as a function of inferential processing requirements: An examination of the reading skills underlying performance on the NAEP Grade 8 Reading Assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), San Francisco, CA. Educational Testing Service.

Sheehan, K., & Mislevy, R. J. (1994). *A tree-based analysis of items from an assessment of basic mathematics skills (ETS RR-94-14).* Educational Testing Service.

Sireci, S., & Benítez, I. (2023). Evidence for test validation: A guide for practitioners. *Psicothema, 35*(3), 217-26. https://doi.org/10.7334/psicothema2022.477

Sireci, S. G., Crespo Cruz, E, Suárez-Álvarez, J, & Rodríguez Matos, G. (2025). *Understanding UNDERSTANDardization research*. In R. Bennett, R., L. Darling-Hammond & A. Barinarayan (Eds.), Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy, Routledge. https://doi.org/10.4324/9781003435105

Sireci, S. G., Suárez-Álvarez, J., Zenisky, A. L., & Oliveri, M. E. (2024). Evolving educational testing to meet students' needs: Design-in-real-time assessment. *Educational Measurement: Issues and Practice, 43*(4), 112–118. https://doi.org/10.1111/emip.12653

Smith, E. A., & Senter, R. J. (1967). *Automated readability index* (Vol. 66, No. 220). Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command.

Suárez-Álvarez, J., Fernández-Alonso, R., García-Crespo, F. J., & Muñiz, J. (2022). The use of new technologies in educational assessments: Reading in a digital world. *Psychologist Papers, 43*(1), 36–47. https://doi.org/10.23923/pap.psicol.2986

Suárez-Ávarez, J., Oliveri, M. E., Zenisky, A., & Sireci, S. G (2024). Five key actions for redesigning adult skills assessments from learners, employees, and educators. *Journal for Research on Adult Education, 47*, 321–343. https://doi.org/10.1007/s40955-024-00288-8

Sun, T, B., Drasgow, F., & Zhou, M. X. (2024, May 1). Developmente and validation of an artificial chatbot to assess personality. *PsyArXiv*. https://doi.org/10.131234/osf.io/ahtr9

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence, 3*, 100075. https://doi.org/10.1016/j.caeai.2022.100075

Ulitzsch, E., Shin, H. J., & Lüdtke, O. (2023). Accounting for careless and insufficient effort responding in large-scale survey data—Development, evaluation, and application of a screen-time-based weighting procedure. *Behavior Research Methods, 56*(2), 804–825. https://doi.org/10.3758/s13428-022-02053-6

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008. https://arxiv.org/abs/1706.03762

von Davier, A. A., Runge, A., Park, Y., Attali, Y., Church, J., & LaFlair, G. (2024). The item factory: Intelligent automation in support of test development at scale. In H. Jiao & R. W. Lissitz (Eds.), *Machine learning, natural language processing, and psychometrics* (Marces Book Series) (pp. 1– 25). Information Age Publishing Inc.

von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika, 83*(4), 847–857. https://doi.org/10.1007/s11336-018-9608-y

von Davier, M., Tyack, L., & Khorramdel, L. (2022). Scoring graphical responses in TIMSS 2019 using artificial neural networks. *Educational and Psychological Measurement, 83*(3), 556–585. https://doi.org/10.1177/00131644221098021

Walker, M. E., Olivera-Aguilar, M., Lehman, B., Laitusis, C., Guzman-Orth, D., & Gholson, M. (2023). *Culturally responsive assessment: Provisional principles* (ETS RR-23-11). Educational Testing Service. https://doi.org/10.1002/ets2.12374

Wang, Y., Pan, Y., Yan, M., Su, Z., & Luan, T. H. (2023). A survey on ChatGPT: AI-generated contents, challenges, and solutions. *Open Journal of Computer Science, 4*, 280–286. https://doi.org/10.48550/arXiv.2305.18339

Wise, S. L., Im, S., & Lee, J. (2021). The impact of disengaged test taking on a state's accountability test results. *Educational Assessment, 26*(3), 163–174. https://doi.org/10.1080/10627197.2021.1956897

Wulff, D. U., & Mata, R. (2025). Semantic embeddings reveal and address taxonomic incommensurability in psychological measurement. *Nature Human Behaviour*, 1-11. https://doi.org/10.1038/s41562-024-02089-y

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2019). *Introduction of multistage adaptive testing design in PISA 2018.* OECD Education Working Papers, No. 209, OECD Publishing, https://doi.org/10.1787/b9435d4b-en

Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024). Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour, 8*, 1839–1850. https://doi.org/10.1038/s41562-024-02004-5

Yaneva, V., & von Davier, M. (Eds.). (2023). *Advancing natural language processing in educational assessment* (1st ed.). Routledge. https://doi.org/10.4324/9781003278658

Yang, H., Kim, H., Lee, J. H., & Shin, D. (2022). Implementation of an AI chatbot as an English conversation partner in EFL speaking classes. *ReCALL, 34*(3), 327–343. https://doi.org/10.1017/S0958344022000039

Yuan, L. (I.), Sun, T., Dennis, A. R., & Zhou, M. (2024). Perception is reality? Understanding user perceptions of chatbot-inferred versus self-reported personality traits. *Computers in Human Behavior: Artificial Humans, 2*, 100057. https://doi.org/10.1016/j.chbah.2024.100057

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*(4), 337–362. https://doi.org/10.1207/S15324818AME1504_02

GRUPO EDITORIAL DE PSICOFUNDACIÓN

Psicothema (2026) 38(1) 13-22

Psicothema

https://www.psicothema.com/es • ISSN 0214–9915 • eISSN 1886-144X

Colegio Oficial de Psicología del Principado de Asturias

Psicothema

Article

Waiting Times in Clinical Psychology in Public Mental Health Units: Predictors of Attendance at the First Appointment and Early Dropout

María del Mar Miras-Aguilar[1], Jose Ruiz-Gutiérrez[1], Sandra Martínez-Gómez[1], Saioa Pérez-García-Abad[2], Carmen Ramos-Barrón[3], Emilio Pariente-Rodrigo[3], Lourdes Piñán-Setién[3], Noelia Otero-Cabanillas[3], María Isabel Priede[3] and César González-Blanch[1,4]

[1] Mental Health Unit, University Hospital Marqués de Valdecilla - IDIVAL (Spain)
[2] Mental Health Service, Navarre Healthcare Service – Osasunbidea (Spain)
[3] Primary Care, Cantabrian Healthcare Service (Spain)
[4] Department of Psychology, International University of La Rioja (UNIR) (Spain)

ARTICLE INFO

Received: 10/05/2025
Accepted: 28/07/2025

Keywords:
Clinical psychology
Primary care
Waiting lists
Attendance
Dropout

ABSTRACT

Background: Waiting lists in mental health are a growing problem. This study analyzes their impact on attendance and early dropout from treatment in the Santander health area of the Spanish National Health System. **Method:** A retrospective observational study was conducted with 2,765 patients referred from Primary Care to four Mental Health Units during 2021. Logistic regressions were applied to analyze the influence of waiting times on attendance at the first appointment and early dropout, and ROC curves were used to identify optimal cut-off points. **Results:** The median waiting time was 51 days for the first appointment and 35 between the first and second. A total of 84.6% attended their first session, with higher attendance in women, older individuals, those with work-related disability, and shorter waiting times. Early dropout (15.8%) was associated with longer time between appointments, being male, and being younger. The discriminative power of the cut-off points was poor. **Conclusions:** Waiting times exceed recommended standards and negatively affect treatment continuity. Structural reforms and more human resources are needed to improve access to and the effectiveness of psychological care.

Tiempos de Espera en Psicología Clínica de las Unidades de Salud Mental Públicas: Predictores de Asistencia a Primera Consulta y Abandono Temprano

Palabras clave:
Psicología clínica
Atención primaria
Listas de espera
Asistencia
Abandono

RESUMEN

Antecedentes: Las listas de espera en salud mental son un problema creciente. Este estudio analiza su impacto en la asistencia y el abandono temprano del tratamiento en el área de salud de Santander del Sistema Nacional de Salud español. **Método:** Se realizó un estudio observacional retrospectivo con 2.765 pacientes derivados desde Atención Primaria a cuatro Unidades de Salud Mental durante 2021. Se aplicaron regresiones logísticas para analizar la influencia de los tiempos de espera en la asistencia a la primera cita y el abandono temprano, y curvas ROC para identificar puntos de corte óptimos. **Resultados:** La mediana del tiempo de espera fue de 51 días para la primera cita y 35 entre la primera y segunda. Asistieron a la primera cita el 84,6%, siendo más probable en mujeres, personas de mayor edad, con incapacidad laboral y menor tiempo de espera. El abandono temprano (15,8%) se asoció con mayor tiempo entre consultas, ser hombre y menor edad. El poder discriminativo de los puntos de corte fue pobre. **Conclusiones:** Los tiempos de espera exceden las recomendaciones y afectan la continuidad del tratamiento. Se requieren reformas estructurales y más recursos humanos para mejorar el acceso y la efectividad de la atención psicológica.

Cite as: Miras-Aguilar, M. del M., Ruiz-Gutiérrez, J., Martínez-Gómez, S., Pérez-García-Abad, S., Ramos-Barrón, C., Pariente-Rodrigo, E., Piñán-Setién, L., Otero-Cabanillas, N., Priede, M. I., & González-Blanch, C. (2026). Waiting times in clinical psychology in public mental health units: Predictors of attendance at the first appointment and early dropout . *Psicothema, 38*(1), 13-22. https://doi.org/10.70478/psicothema.2026.38.02
Corresponding author: César González-Blanch, cesar.gonzalezblanch@scsalud.es
This article is published under Creative Commons License 4.0 CC-BY-NC-ND

Waiting lists in healthcare services represent a major global challenge, significantly impacting both accessibility and quality of care. This issue is particularly critical in mental healthcare, where the high and growing prevalence of mental disorders continues to overburden healthcare systems worldwide. In 2019, one in every eight people—around 970 million individuals globally—were living with a mental disorder, with anxiety and depressive disorders being the most prevalent (Institute for Health Metrics and Evaluation, 2022). The situation worsened with the onset of the COVID-19 pandemic, which led to an estimated 26% increase in anxiety disorders and a 28% increase in major depressive disorders in just one year (World Health Organization [WHO], 2022). In 2020, 53.2 million additional cases of major depression and 76.2 million new cases of anxiety disorders were recorded worldwide (Santomauro et al., 2021). By 2021, the number of global cases of mental disorders exceeded 440 million (Fan et al., 2025).

In Spain, recent data reflects a worsening trend. According to the National Statistics Institute (INE, 2025), 14.6% of the population over 15 years old experienced depressive symptoms in 2023. Moreover, the Ministry of Health (2024) indicates that 34% of the population reported experiencing some type of mental health problem, with anxiety disorders (10%), sleep disorders, and depressive disorders being the most commonly reported conditions.

Access to public mental healthcare services is essential for the timely detection and treatment of mental health problems. In this context, within the National Health System (NHS) of Spain, Primary Care (PC) serves as the first point of contact with the healthcare system, where around 20 to 55% of total appointments address mental health problems (Alonso-Gómez et al., 2019). However, the strain for the treatment of these problems largely falls on Mental Health Units (MHU), consisting of healthcare teams including clinical psychologists, psychiatrists, mental health nurses, as well as social workers in some cases. Therefore, coordination between PC and MHU is essential to provide high-quality thorough healthcare.

Despite the fact that the first recommended treatment approach for most mental disorders is psychological treatment (Gaudiano & Miller, 2013), it is necessary for patients to access these services within a reasonable time. Previous studies revealed an average waiting time for a first appointment in Clinical Psychology between 32 and 74 days in different Spanish cities, such as Pamplona (Goñi-Sarries et al., 2008), Burgos (Martín-Jurado et al., 2012), Madrid (Díaz et al., 2017), Badalona (Tejedo-García, 2018), and even clinical psychologists themselves have reported an average of 120 days for access to psychological care in Community of Madrid (Cuellar-Flores et al., 2022). The data on subsequent appointments is not encouraging either, as an average of 50 days has been recorded (Cuellar-Flores et al., 2022), which significantly hinders the implementation of formal psychological treatments. These studies highlight the significant variability and long waiting lists in the different regions of Spain, and are far from what the evidence recommends regarding the frequency of psychological treatment sessions. The study by Clark et al. (2018) found that interventions which started within the first six weeks from referral yielded better therapeutic outcomes, highlighting the urgent need to reduce waiting times to improve clinical results, as well as a weekly frequency to increase the effectiveness and efficiency of psychological treatments (Erekson, et al., 2015; 2022).

Long waiting lists in mental healthcare have significant repercussions, affecting both the care provided and the mental health of patients (Peipert et al., 2022). Delayed care may increase the chronicity of disorders and worsen the severity of symptoms (Cuijpers et al., 2021; Patel et al., 2015; Reichert & Jacobs, 2018; Wang, 2004). Furthermore, prolonged waiting times may demotivate patients, reducing their resilience and treatment response, and producing feelings of hopelessness regarding future interventions (Punton et al., 2022; Van Dijk et al., 2023). Additionally, limited and slow access to psychological therapies has led to a predominantly psychopharmacological first approach in PC, contrary to the recommendations of clinical guidelines from the National Institute of Health and Care Excellence (NICE, 2011). Previous studies in Spain found that 47% of patients referred to Clinical Psychology were already receiving psychopharmacological treatment (Díaz et al., 2017; Martín-Jurado et al., 2012). The situation not only goes against best practice, but also contributes to the chronicity of mental disorders and increased long-term costs (González-Blanch et al., 2023).

Following this line, prolonged waiting time is considered as one of the most determining factors in the attendance of clinical psychologist appointments (Gallucci et al., 2005; Loumidis & Shropshire, 1997; Miranda-Chueca et al., 2003; Vellisca et al., 2014). The negative impact of long waiting lists is reflected in lower attendance at the first appointment and higher early dropout (Steinert et al., 2017; Swift et al., 2012). Early dropout refers to the premature termination of the treatment without fulfilment of the therapeutic goals or possible benefits that may have been possible with normal termination of the therapy (Swift & Greenberg, 2012). Although attendance rates at the first appointment in Spain have been reported to range from 80% to 90% (García-Pedrajas et al., 2018; Miranda-Chueca et al., 2003; Tejedo-García, 2018; Vellisca et al., 2014), early dropout rates in psychological treatments are commonly observed to range from 20% to 25%; (García-Pedrajas et al., 2018; Hanevik et al., 2023; Swift & Greenberg, 2012).

Several sociodemographic and clinical variables have been examined in an attempt to explain attendance rates, although the results remain contradictory. For example, the study by Vellisca et al., (2014) found no significant association between attendance at the first appointment and various sociodemographic variables (i.e. sex, age and population type). However, other studies have found a significant relationship between attendance at the first appointment and older age (especially over 25 years old), having a temporary work disability or previous history of mental health treatment (Fenger et al., 2011; Loumidis & Shropshire, 1997; Moratalla & Lobo, 2002). Additionally, predictors of early dropout from psychological treatment have been found to include being under 45 years old, unemployed, lower educational level and poor social support, although severity of symptoms was not a predictor (Fenger et al., 2011; Hanevik et al., 2023; Swift & Greenberg, 2012).

Despite advancements in mental healthcare research, several gaps remain in the literature. First, previous studies have focused on specific centres within a region, hindering the capacity to capture the variability and representativeness of an entire healthcare area. Second, the lack of studies conducted after the COVID-19 pandemic limits the understanding of the evolution of healthcare demands and the population needs following the impact of the pandemic on public mental healthcare services. Finally, although previous

studies have found inconsistent results in the relationship between sociodemographic variables and attendance at the first appointment and early dropout, waiting times are considered central factors for both variables. These discrepancies highlight the need to focus our analysis on the impact of waiting times, since it is the index most influenced by the different Healthcare Services in Spain. Furthermore, studies that control for other variables potentially influencing attendance and early dropout are very limited.

The objectives of this study, conducted in the healthcare area of Santander, Cantabria (Spain), are threefold: (i) to examine waiting times for a first and second appointment, (ii) to analyze the influence of waiting times in the attendance at the first appointment and early dropout from psychological treatment, while controlling the effect of several sociodemographic and clinical variables, across all referral received throughout an entire year in every MHU within a healthcare area, and (iii) to determine an optimal cut-off for waiting times at the first and second appointments which maximises attendance and minimises early dropout.

Method

Participants

The sample study included all patients aged 18 years and older referred by a general practitioner for a first treatment appointment with a clinical psychologist of the four MHUs belonging to the Healthcare Area of Santander between 1st January to 31st December 2021. Patients were selected during a whole year to remove any seasonal effect from the sample recruitment. A first treatment appointment was considered as those patients attending a clinical psychologist appointment for the first time in the Cantabrian Healthcare Service or, in cases with a history of prior psychological care, when more than one year had passed since their last appointment at the MHU. Patients were excluded if (i) they were referred from other mental health professionals from the same MHU, such as a psychiatrist or from other healthcare services different from PC, (ii) they had notified the MHU in advance to cancel the appointment before attending, and (iii) the reason for referral should be addressed in other healthcare facilities more appropriate or in specialised programs.

Instruments

An ad-hoc protocol for data collection was elaborated, based exclusively on information retrieved from electronic health records (EHRs). The protocol included the following variables:

Sociodemographic Variables

Sex, age, civil status, maximum level of education attained, and current employment status.

Clinical Variables

History of psychological care (defined as an appointment in any mental healthcare resource in the Cantabrian Healthcare Service prior to referral), reason for the appointment recorded by the general practitioner according to the International Classification of Primary Care (ICPC-2), which was recoded in accordance with the International Classification of Diseases-10 (ICD-10) diagnoses to improve categorization, prescription and type of psychopharmacological treatment at the time of referral, and the existence of a temporary work disability at the time of referral.

Healthcare Variables

MHU handling the demand, waiting time (defined as the number of days between the referral of the general practitioner and the first appointment with the clinical psychologist), attendance at the first appointment, and clinical discharge at the first appointment. Finally, for patients who were offered a second appointment, the time between appointments was recorded (defined as the number of days between the first and second appointments). Early dropout was registered in patients who were not clinically discharged in the first appointment, but did not attend the second appointment nor resume follow-up within a year from the first appointment.

Procedure

A single-group retrospective observational cohort design was conducted in the Healthcare Area I of Cantabria, corresponding to the city of Santander, during the year 2021. This Healthcare Area includes 20 health centres and 40 clinics that refer patients to four MHUs (Puertochico, López Albo I and II and Nueva Montaña), assisting a predominantly urban population of over 315,000 habitants in the year 2021. The characteristics of the different MHUs are displayed in Table 1.

When a general practitioner identifies a mental health problem in a patient and considers that the patient may benefit from psychological treatment, an electronically recorded referral is made to the corresponding MHU assigned to their PC centre. Subsequently, the patient is scheduled for a first in-person appointment with the

Table 1
Characteristics, Population and Resources of the Mental Health Units of Santander in 2021

Variables	MHU López Albo I	MHU López Albo II	MHU Nueva Montaña	MHU Puertochico	Healthcare Area I (Santander)
Population[a]	75,320	100,073	76,115	63,908	315,416
Population above 14 years old[a]	66,104	87,406	66,030	57,028	276,568
Number of Health Centres	5	5	4	6	20
Number of CP per MHU	2	3	2	2	9
CP of MHU per 100.000 habitants	2.66	2.99	2.63	3.13	2.85

Note. CP = clinical psychologist; PC = primary care; MHU = mental health unit.
[a]Number of healthcare cards in the year 2021 obtained through internal correspondence with Primary Care Management of the Cantabrian Healthcare Service.

clinical psychologist who has the earliest availability. To maximize attendance, the Cantabrian Healthcare Service contacts the patient via phone to inform the date of their appointment, and a mobile message is sent to remind them two days before. In this study, data collection was conducted by retrieving EHRs from the Cantabrian Healthcare Service using specific software programs (VisorCorp for PC and Altamira for specialised care). Due to the retrospective nature of the study, general practitioners were not informed about the study nor its objectives, ensuring that their referral and treatment criteria were not influenced.

We took measures to ensure the privacy and confidentiality of the data throughout the study. Given the de-identified nature of the data and the practical challenges of obtaining informed consent from every individual whose data was included in the study, we did not request informed consent from participants. We believe that the absence of identifiable personal information in the EHRs and the impracticality of obtaining consent for large datasets justifies the exemption. Recognizing that the use of EHRs for research purposes involves ethical considerations, we followed best practices to minimize any potential risks to participants. This approach was reviewed and approved by the local Ethics Committee (2021.410).

Data Analysis

Descriptive analyses included the mean (*M*), standard deviation (*SD*), Median (*Mdn*) and interquartile range (*IQR*) for quantitative variables, while frequency (*n*) and percentage (%) were reported for categorical variables. Due to the violation of normality assumption in every continuous variable, the non-parametric Mann-Whitney *U* test was used to make comparisons with two different groups. Then, multiple logistic regression assumptions (linearity in the logit for continuous predictors, absence of multicollinearity, independence of errors, and absence of overly influential outliers) were confirmed and it was used to calculate the relationships between attendance at the first appointment and waiting time, as well as early dropout and time between first and second appointments, while statistically adjusting for the confounding effects of other sociodemographic, clinical and healthcare variables of relevance according to the literature. We used the adjusted odds ratio (*aOR*) as effect size for every variable included in the models. A $p < .05$ was considered as the minimum threshold for statistical significance. To assess the discriminative capacity of waiting times in predicting attendance at the first appointment and early dropout, receiver operating characteristic (ROC) curve analyses were conducted. The area under the curve (AUC), sensitivity and specificity was reported. The Youden Index (J = Sensitivity + Specificity – 1) was also calculated to determine optimal cut-off points. Every analysis was carried out using the statistical program Statistical Package for the Social Sciences (SPSS) version 25.0.

Results

Descriptive Analyses

The final sample of the study consisted of 2,765 patients. The sociodemographic and clinical characteristics of the sample are displayed in Tables 2 and Table 3, respectively.

Table 2
Sociodemographic Characteristics of the Study Sample

		n	%
Age		2,765	
	18-24 years	377	13.6
	25-39 years	773	28.0
	40-65 years	1,409	41.0
	> 65 years	206	7.4
Sex		2,765	
	Women	1,953	70.6
Civil Status		2,233	
	Single	518	23.2
	In a relationship	470	21.1
	Married	880	39.4
	Divorced	291	13.0
	Widowed	74	3.3
Level of Education		1,092	
	Primary education	89	8.1
	Secondary education	135	12.4
	Upper secondary education	165	15.1
	Vocational training	358	32.8
	College Diploma	345	31.6
Current employment status		2,319	
	Student	224	9.7
	Working	906	39.1
	Unemployed	290	12.5
	Temporary work Disability	529	22.8
	Permanent Work Disability	38	1.6
	Retired	171	7.4
	Homemaker	88	3.8
	Working and Studying	41	1.8
	Other	32	1.4

Note. The mean age of the study sample was 43.1 years old (*SD* = 14.9)

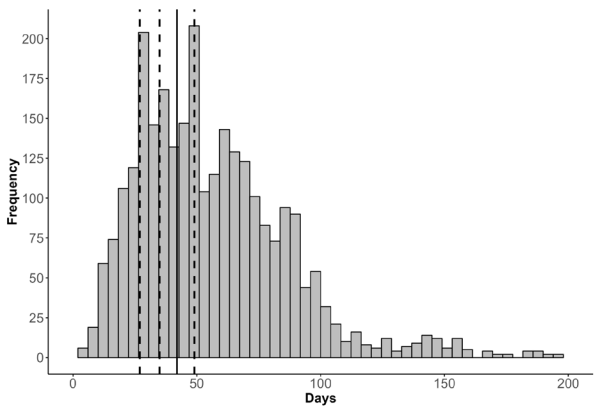
Table 3
Clinical Characteristics of the Study Sample

	n	%
History of psychological care	2,763	
Yes	953	34.5
Reason for appointment	2,765	
Anxiety disorders	1,440	52.1
Adjustment disorders	496	17.9
Depressive disorders	418	15.1
Other disorders	411	14.9
Psychopharmacological treatment at the time of referral medication	2,763	
Yes	1,744	63.1
Type of psychopharmacological treatment	1,744	
Anxiolytic	717	41.1
Anxiolytic and antidepressant	659	37.8
Antidepressant	317	18.2
Others	51	2.9

The distributions of the waiting time for the first and second appointment are presented in Figures 1 and 2, respectively. The average waiting time for the first appointment with a clinical psychologist was 58.2 days (*SD* = 35.5), with a median of 51 days (*IQR* = 40), a minimum of 2 days, and a maximum of 329 days. Notably, in 65.6% of the sample (n = 1,727) the waiting time for the first appointment exceeded the recommended clinical standard of 6weeks. The attendance rate for the first appointment was 84.6% and clinical discharge at the first appointment was provided to 21.3% of the patients. Additionally, the average waiting time for a second appointment was 40.9 days (*SD* = 23.4), with a median of 35 days (*IQR* = 22), a minimum of 3 days, and a maximum of 220 days. Among patients who were offered a second appointment, the

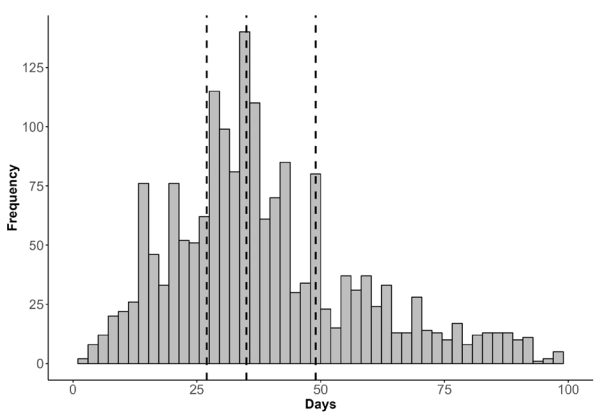
attendance rate was 84.2%, thus 15.8% did not attend the second appointment, nor resumed subsequent care within the 1-year follow-up period (i.e., early dropout).

Figure 1
Waiting Time Distribution for the First Appointment



Note. Straight line placed in 42 days to represent the recommended clinical standard for a first appointment with a clinical psychologist. Dashed lines indicate the 25th (34 days), 50th (51 days), and 75th (74 days) percentiles. Values exceeding 200 days were grouped into the 200 category to improve visualisation.

Figure 2
Waiting Time Distribution Between First and Second Appointment



Note. Dashed lines indicate the 25th (27 days), 50th (35 days), and 75th (49 days) percentiles. Values exceeding 100 days were grouped into the 100 category to improve visualisation.

Predictors for the Attendance at the First Appointment

The main variables associated with attendance at the first appointment were analysed. The Mann-Whitney test found statistically significant differences in the waiting time ($U = 531002.5$; $p = .022$; $r = .07$) between the group that did not attend the first appointment ($Mdn = 53$; $IQR = 37$) and the group that did attend ($Mdn = 51$; $IQR = 32$). A multiple logistic regression was performed to predict attendance at the first appointment based on waiting time, while statistically controlling for the variables of age, sex, history of psychological care, presence of a temporary work disability and psychopharmacological treatment. The model statistically predicted attendance at the first appointment ($\chi^2(2750) = 66.58$; $p < .001$; Nagelkerke $R^2 = .024$) and correctly classified 84.7% of the cases. The coefficients of the variables included in the model are presented in Table 4. The results indicate that attendance at the first appointment was significantly influenced by shorter waiting time, but also by being female, older age and the presence of a temporary work disability, with each of these variables making an independent contribution to the prediction.

To evaluate the discriminative ability of waiting time in predicting attendance at the first appointment, a ROC curve analysis was performed. The AUC was 0.535 (95% CI [0.506–0.564]), indicating a poor discriminative performance. Consistently, the Youden Index did not identify any clinically meaningful threshold, with the highest value observed at 44 days ($J = 0.082$). At this threshold, sensitivity was 0.682 and specificity was 0.399, further reflecting a limited ability of waiting time to distinguish between attendees and non-attendees.

Predictors for Early Dropout

Main predictors of early dropout at the second appointment were examined. Statistically significant differences were found in the waiting time for the second appointment ($U = 156993$; $p < .001$; $r = .27$) between individuals who dropped out ($Mdn = 42$; $IQR = 26$) and those who did not drop out ($Mdn = 35$; $IQR = 23.75$). A multiple logistic regression model was performed to predict dropout at the second appointment based on waiting time and time between appointments, while statistically controlling for age, sex, history of psychological care and presence of a temporary work disability. The results indicated that the model was statistically significant in predicting early dropout ($\chi^2(1812) = 53274$; $p < .001$; Nagelkerke $R^2 = .029$). The coefficients for the variables included in the model are presented in Table 3. Statistically significant predictors of early dropout were longer waiting time between appointments, but also younger age and being male, which played a significant predictive role in the likelihood of early dropout.

To complement these findings and further assess the discriminative utility of waiting time between appointments, a second ROC curve analysis was conducted. The AUC was 0.633 (95% CI [0.601–0.666]), suggesting a modest discriminative ability to distinguish individuals at risk of early dropout. The Youden Index identified 36 days as the optimal cut-off point ($J = 0.203$), corresponding to a sensitivity of 0.668 and a specificity of 0.536. This suggests that when the interval between appointments exceeds approximately one month, the risk of early dropout increases significantly.

Discussion

The aim of the study was to analyse waiting times for access to specialised psychological care from PC and its relationship with attendance at the first appointment and early dropout from psychological treatment while controlling for several sociodemographic and clinical variables. The study revealed that the median waiting time for specialised psychological care at MHU is 51 days for the first appointment and 35 days for the second. The attendance rate for the first appointment was 85%, which was influenced by shorter waiting time, being female, older age and the presence of a temporary work disability. On the other hand, an early

Table 4
Logistic Regression Models to Examine Potential Predictors of Attendance at the First Appointment and Early Dropout

	Attendance at the first appointment (n = 2,757)				Early dropout (n = 1,820)			
Variables	***aOR***	**95% CI**		***p***	***aOR***	**95% CI**		***p***
		LL	*UL*			*LL*	*UL*	
Age	1.018	1.011	1.026	<.001	0.986	0.977	0.995	.004
Sex	0.789	0.629	0.989	.040	1.404	1.058	1.863	.019
History of psychological care	0.840	0.675	1.045	.118	1.206	0.917	1.587	.180
Presence of a TWD	1.835	1.360	2.476	<.001	0.789	0.569	1.095	.156
Presence of any psychopharmacological treatment	1.116	0.891	1.397	.340	0.986	0.728	1.288	.825
Waiting time	0.996	0.993	0.998	.002	0.997	0.993	1.001	.107
Time between appointments		—			1.016	1.011	1.021	<.001

Note. aOR = adjusted odds ratio; CI = confidence interval; *LL* = lower limit; TWD = temporary work disability; *UL* = upper limit.

dropout rate of 16% was found after the first appointment, being mainly related to longer waiting time for a second appointment, being male and younger age.

These findings reflect a concerning reality in the field of public mental healthcare and highlight a significant structural problem regarding access to psychological care. The results indicate access difficulties, with waiting times reaching seven weeks for a first appointment and five weeks for a second. Although our data fall within an intermediate range compared to other national studies—where waiting times for the first consultation range from 30 to 120 days (Cuéllar-Flores et al., 2022; Díaz et al., 2017; Goñi-Sarries et al., 2008; Martín-Jurado, 2012; Tejedo-García, 2018)—they still exceed current recommendations. On the other hand, research on waiting times for a second appointment is scarce. Some recent studies, such as that by Cuéllar-Flores et al. (2022), report an average of seven weeks in the Community of Madrid, while Benítez-Ortega et al. (2021) report an eight-week interval in Andalucia. Although our study shows slightly shorter waiting times, they remain above the recommended thresholds and could negatively impact the therapeutic process and patient recovery (Reichert & Jacobs, 2018; van Dijk et al., 2023). Overall, patients experience significant delays, exceeding the recommended six-week timeframe for a first appointment (Clark et al., 2018), as well as the one-week interval for subsequent sessions (Erekson et al., 2015, 2022).

Regional heterogeneity in waiting times may stem from differences in healthcare resources, Clinical Psychology staffing, and the internal organization of each regional system. Social determinants such as socioeconomic status, education, and community context also shape mental healthcare demand and access, contributing to observed inequalities (Kirkbride et al., 2024). Although the number of clinical psychologists has increased since 2003—reaching 6,010 professionals under age 65 by 2021 (Ministry of Health, 2022)—only 2,615 are estimated to work in the public healthcare system, resulting in a ratio of 5.56 per 100,000 inhabitants (Duro-Martínez, 2021; Fernández-García, 2021). This shortage, combined with the growing prevalence of mental disorders, has led to longer waiting lists for both initial and follow-up appointments. While structural and social factors are essential to understanding these disparities, certain interpretations of them may conflict with the need to ensure access to psychological treatments in the public system, ultimately reinforcing existing inequalities (González-Blanch, 2025).

The lower waiting times reported in previous studies may be due to differences in the time periods during which they were conducted, as there has been a progressive increase in the prevalence of mental disorders (WHO, 2017). In this regard, the possible discrepancies with earlier research reflect pre-pandemic realities, whereas the COVID-19 pandemic led to a significant rise in the demand for mental health care (Pfefferbaum & North, 2020), thereby contributing to the prolonged waiting times observed in our study. Moreover, the organizational structure of the healthcare system may also play a role, particularly the tendency to prioritize the intake of new patients by increasing the number of weekly first appointments in an effort to reduce its waiting time. While this approach is understandable from an accessibility standpoint, it may have adverse effects on long-term treatment quality, as it limits the system's ability to provide continuous and structured subsequent care.

These structural limitations may also help explain the high proportion of patients who were already receiving psychopharmacological treatment —nearly two-thirds—with anxiolytics being the most frequently prescribed medications. Although our study does not establish a direct link between waiting times and the prescription of psychopharmacological treatments, prolonged delays in accessing psychological care—along with other limitations in PC—may contribute to the continued reliance on medication as a faster and more accessible solution (Marquina-Márquez et al., 2022). Clinical guidelines, such as those from NICE (2022), recommend psychological therapy as the first-line intervention for anxiety and depression. However, the high rates of psychopharmacological prescription observed in our sample—despite these guidelines—point to a persistent gap between recommended practice and actual clinical implementation.

The results of this study highlight that prolonged waiting times not only affect accessibility to psychological treatment but also compromise its continuity, increasing the risk of early dropout. In line with previous literature (Gallucci et al., 2005; Loumidis & Shropshire, 1997; Miranda-Chueca et al., 2003; Vellisca et al., 2014), the longer the delay for a first appointment, the higher the absenteeism rate. However, when examined more closely through ROC curves, waiting time showed a limited capacity to establish a clinically useful cut-off point for distinguishing between attendees and non-attendees. While the regression analysis confirmed that shorter waiting times were significantly associated with higher attendance, the ROC results indicate that no single cut-off point offers sufficient sensitivity and specificity to identify a critical threshold beyond which the risk of non-attendance increases markedly. The optimal threshold identified was 44 days, but

it presented very low discriminative capacity, suggesting that attendance at the first appointment is not determined solely by structural factors such as waiting times.

In this regard, sociodemographic and clinical characteristics appeared to play an important role. Being female, older age, and those in the situation of temporary work disability were more likely to attend the first appointment. These results are consistent with previous studies, such as those by Moratalla and Lobo (2002), Fenger et al., (2011) and Loumidis and Shropshire (1997). Specifically, in the case of temporary work disability, these patients may experience greater functional impairment, which could justify both the referral and the motivation to receive treatment (Lau et al., 2016). Additionally, they have more flexibility to attend since they are not subject to a work schedule that could interfere. However, the role of other external factors, such as institutional pressure to justify the temporary work disability, cannot be ruled out, as it may be related to a poorer response to psychological treatment (González-Blanch et al., 2021).

Similarly, a longer time interval between the first and second appointment is associated with a significant increase in the likelihood of early dropout. In this case, the ROC analysis showed a modest improvement in discriminative capacity, identifying a threshold of approximately 36 days beyond which the risk of early dropout increases notably, offering more informative guidance for service planning. This finding could be explained by a progressive loss of motivation, as well as feelings of frustration or distrust towards the healthcare system (Punton et al., 2022; van Dijk et al., 2023). Additionally, prolonged waiting time between appointments may create a sense of discontinuity, affecting the perception of treatment effectiveness (Swift & Greenberg, 2012). On the other hand, these delays, particularly between appointments, could interfere with the consolidation of a strong therapeutic alliance, which is especially important during the early clinical encounters. The absence or fragility of this alliance may negatively influence the progress of the psychotherapeutic process and increase the risk of dropout (Flückiger et al., 2018; Horvath et al., 2011; Roos & Werbart, 2013; Sharf et al., 2010). As a result, this could lead to the chronicity of disorders, the worsening of symptoms, and a growing sense of helplessness regarding future interventions (Cuijpers et al., 2021; Patel, 2015; Peipert et al., 2022; Reichert, 2018; Wang, 2004). Alternatively, it is also possible that during the waiting time, there could be spontaneous remission of symptoms, which may reduce the perception of the need for intervention and contribute to either not accessing treatment or dropping out once it has begun.

However, while waiting time appears to play a relevant role in early dropout, it is also important to consider individual factors. In this regard, early dropout was more common among men and younger individuals. The higher dropout observed in men could be explained by their lower tendency to seek professional help (Nam et al., 2010; Wang et al., 2007), which may hinder their commitment to treatment. Regarding age, it has been observed that younger patients have a lower adherence rate to psychological interventions, possibly due to higher levels of stigma towards mental disorders in this age group (Benjet et al., 2022; Clarkin et al., 2024).

Finally, it is important to highlight the strengths and limitations of the present study. One of its main strengths is, firstly, the extensive data collection period, which spans an entire year, allowing for a more robust and less biased representation of the healthcare reality. Additionally, direct access to information through the thorough review of all referrals via the EHR ensures precise and reliable data collection. Moreover, the fact that the study includes the entire healthcare area of an autonomous community broadens its applicability within the regional context and provides a more comprehensive and representative view of the functioning of a mental healthcare service. However, some limitations should be considered. Firstly, the sample is limited exclusively to referrals from PC, excluding other routes such as specific hospital programs or psychiatrists from the same MHU. Although these represent a small percentage of the total patients attended, their exclusion means that the results do not fully reflect all the entry pathways into the psychological care system. Secondly, although the study focused on waiting times, which are one of the most system-dependent factors, variables such as the patient's level of motivation, perceived need, or personal practical barriers (e.g., work schedule, family care, transportation, etc.) were not recorded and could enhance the analysis of predictors for attendance and dropout in future studies. Finally, it should be noted that, although the study encompasses an entire healthcare area within one autonomous community—specifically, Area I of Cantabria—, the findings regarding the impact of waiting times on adherence may not be generalizable to other regional healthcare contexts with different organizational structures or levels of resource allocation.

In conclusion, this study highlights the importance of addressing waiting times not only as an indicator of healthcare system efficiency but also as a clinically relevant factor that affects access to and adherence to psychological treatment. The situation described calls for a thorough review of the healthcare system, promoting structural reforms that enable more accessible, continuous, and effective psychological care.

In this regard, one of the key actions to achieve these goals involves increasing the number of clinical psychologists by expanding the availability of specialised training positions. This would help address growing demand and improve access to evidence-based psychological treatments. Additionally, it is essential to promote strategies that improve the efficient use of available resources, strengthen coordination across different levels of care, and support the development of quality assessment plans to evaluate the system's performance and identify service needs.

Among these approaches, stepped-care models are increasingly being implemented as a way to organise mental health services to maximise the effectiveness and efficiency of allocation of resources by ensuring that the intensity of intervention matches the individual's clinical needs (McGorry & Mei, 2021). The treatments following this model are structured along a continuum of intensity ranging from low-intensity (e.g. self-help or group therapy) to high intensity (e.g., specialised or multidisciplinary intervention) and have been shown to improve the treatment response and remission of depressive and anxiety disorders (Jeitani et al., 2024).

In Spain, the PsicAP project has demonstrated the effectiveness of brief psychological interventions in PC (Cano-Vindel et al., 2022). Based on this experience, Cantabria began integrating clinical psychologists into PC centres in 2023, which could represent a significant change in the structure and functioning of Mental Health Units. Future studies should evaluate the impact of these measures on reducing waiting times and improving care continuity.

Author Contributions

María del Mar Miras-Aguilar: Conceptualization, Methodology, Funding acquisition, Investigation, Data curation, Formal analysis, Visualization, Writing – Original draft, Writing – Review & editing. **Jose Ruiz-Gutiérrez:** Conceptualization, Methodology, Funding acquisition, Investigation, Data curation, Formal analysis, Visualization, Writing – Original draft, Writing – Review & editing. **Sandra Martínez-Gómez:** Conceptualization, Methodology, Funding acquisition, Investigation, Data curation, Writing – Review & editing. **Saioa Pérez-García-Abad:** Conceptualization, Methodology, Funding acquisition, Investigation, Data curation, Writing – Review & editing. **Carmen Ramos Barron:** Funding acquisition, Supervision, Writing – Review & editing. **Emilio Pariente Rodrigo:** Funding acquisition, Supervision, Writing – Review & editing. **Lourdes Piñán Setién:** Funding acquisition, Supervision, Writing – Review & editing. **Noelia Otero Cabanillas:** Funding acquisition, Supervision, Writing – Review & editing. **César González-Blanch:** Conceptualization, Methodology, Funding acquisition, Writing – Review & editing, Supervision, Project administration.

Acknowledgements

We thank Amador Priede for providing valuable information about the Cantabrian Health Service.

Funding

Valdecilla Healthcare Research Institute - IDIVAL, Grant/Award Number: PRIMVAL 22/02. The source of funding did not participate in the design of the study, the data collection, analysis, or interpretation, the writing of the article, or in the decision to submit it for publication.

Conflict of Interests

The authors declare that there are no conflicts of interest.

Data Availability Statement

Data available on request from the authors.

References

Alonso-Gómez, R. A., Reina, L. L., Méndez, I. F., García, J. M., & Briñol, L. G. (2019). El psicólogo clínico en los centros de salud. Un trabajo conjunto entre atención primaria y salud mental [The clinical psychologist in health centers. A joint work between primary care and mental health]. *Atención Primaria, 51*(5), 310–313. https://doi.org/10.1016/j.aprim.2018.08.012

Benítez-Ortega, J. L., Venceslá-Martínez, J. F., López-Pérez-Díaz, Á G., Rodríguez-Gómez, A., Gómez-Gómez, V., Martínez-Cervantes, R. J., Romero-Gamero, R., & Vázquez-Morejón, A. J. (2021). Calidad asistencial de la psicología clínica en el Servicio Andaluz de Salud evaluada por los facultativos [Quality of care of clinical psychology in the Andalusian Health Service as assessed by professionals]. *Apuntes de Psicología*, *39*(3), 143-158. https://doi.org/10.55414/ap.v39i3.910

Benjet, C., Borges, G., Orozco, R., Aguilar-Gaxiola, S., Andrade, L. H., Cia, A., Hwang, I., Kessler, R. C., Piazza, M., Posada-Villa, J., Sampson, N., Stagnaro, J. C., Torres, Y., Viana, M. C., Vigo, D., & Medina-Mora, M. (2022). Dropout from treatment for mental disorders in six countries of the Americas: A regional report from the World Mental Health Surveys. *Journal of Affective Disorders*, *303*, 168-179. https://doi.org/10.1016/j.jad.2022.02.019

Cano-Vindel, A., Muñoz-Navarro, R., Moriana, J. A., Ruiz-Rodríguez, P., Medrano, L. A., & González-Blanch, C. (2022). Transdiagnostic group cognitive behavioural therapy for emotional disorders in primary care: The results of the PsicAP randomized controlled trial. *Psychological Medicine*, *52*(15), 3336–3348. https://doi.org/10.1017/S0033291720005498

Clark, D. M., Canvin, L., Green, J., Layard, R., Pilling, S., & Janecka, M. (2018). Transparency about the outcomes of mental health services (IAPT approach): An analysis of public data. *The Lancet*, *391*(10121), 679-686. https://doi.org/10.1016/s0140-6736(17)32133-5

Clarkin, J., Heywood, C., & Robinson, L. J. (2024). Are younger people more accurate at identifying mental health disorders, recommending help appropriately, and do they show lower mental health stigma than older people? *Mental Health & Prevention*, *36*, 200361. https://doi.org/10.1016/j.mhp.2024.200361

Cuellar-Flores, I., Garzón, L. F., Félix-Alcántara, M. P., Olivares, B. M., De la Vega Rodríguez, I., González, M. F., Albarsanz, M. L. P., Rivera, S. V., & Belmonte, M. J. M. (2022). Indicadores asistenciales y estándares de calidad asistencial para la psicología clínica en los centros de salud mental del Sistema Madrileño de Salud evaluados por sus profesionales [Care indicators and care quality standards for clinical psychology in mental health centers assessed by clinical psychologists of Madrid Health System]. *Apuntes de Psicología, 40*(2), 71-86. https://doi.org/10.55414/ap.v40i2.1414

Cuijpers, P., Karyotaki, E., Ciharova, M., Miguel, C., Noma, H., & Furukawa, T. A. (2021). The effects of psychotherapies for depression on response, remission, reliable change, and deterioration: A meta-analysis. *Acta Psychiatrica Scandinavica*, *144*(3), 288-299. https://doi.org/10.1111/acps.13335

Díaz, J., Díaz-De-Neira, M., Jarabo, A., Roig, P., & Román, P. (2017). Estudio de derivaciones de atención primaria a centros de salud mental en pacientes adultos en la Comunidad de Madrid [Study of primary care referrals to mental health centers in adult patients in Madrid Region]. *Clínica y Salud*, *28*(2), 65-70. https://doi.org/10.1016/j.clysa.2017.03.001

Duro-Martínez, J. C. (2021). ¿Sabemos cuántos profesionales especialistas en Psicología Clínica trabajan en el Sistema Nacional de Salud español? [Do we know how many professionals specialists in clinical psychology work in the Spanish National Health System]. *Papeles del Psicólogo*, *42*(2), 81-93. https://doi.org/10.23923/pap.psicol.2955

Erekson, D. M., Bailey, R. J., Cattani, K., Klundt, J. S., Lynn, A. M., Jensen, D., Merrill, B. M., Schmuck, D., & Worthen, V. (2022). Psychotherapy session frequency: A naturalistic examination in a university counseling center. *Journal of Counseling Psychology*, *69*(4), 531-540. https://doi.org/10.1037/cou0000593

Erekson, D. M., Lambert, M. J., & Eggett, D. L. (2015). The relationship between session frequency and psychotherapy outcome in a naturalistic setting. *Journal of Consulting and Clinical Psychology*, *83*(6), 1097-1107. https://doi.org/10.1037/a0039774

Fan, Y., Fan, A., Yang, Z., & Fan, D. (2025). Global burden of mental disorders in 204 countries and territories, 1990–2021: results from the

global burden of disease study 2021. *BMC Psychiatry*, *25*(1). https://doi.org/10.1186/s12888-025-06932-y

Fenger, M., Mortensen, E. L., Poulsen, S., & Lau, M. (2011). No-shows, drop-outs and completers in psychotherapeutic treatment: Demographic and clinical predictors in a large sample of non-psychotic patients. *Nordic Journal of Psychiatry*, *65*(3), 183–191. https://doi.org/10.3109/08039488.2010.515687

Fernández-García, X. (2021). Situación de la psicología clínica en el Sistema Nacional de Salud (SNS) y perspectivas de crecimiento [Situation of clinical psychology in the Spanish National Health System and growth perspectives]. *Ansiedad y Estrés*, *27*(1), 31-40. https://doi.org/10.5093/anyes2021a5

Flückiger, C., Del Re, A. C., Wampold, B. E., & Horvath, A. O. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy*, *55*(4), 316-340. https://doi.org/10.1037/pst0000172

Gallucci, G., Swartz, W., & Hackerman, F. (2005). Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. *Psychiatric Services*, *56*(3), 344-346. https://doi.org/10.1176/appi.ps.56.3.344

Garcia-Pedrajas, C., Bayona, E. M., Ibáñez, O. P., Guerre, S. O., & Fornas, C. V. (2018). Longitudinal descriptive study of diagnostic concordance between primary care and psychology support program in primary care. *Anales de Psicología*, *34*(1), 23-29. https://doi.org/10.6018/analesps.34.1.251451

Gaudiano, B. A., & Miller, I. W. (2013). The evidence-based practice of psychotherapy: Facing the challenges that lie ahead. *Clinical Psychology Review*, *33*(7), 813-824. https://doi.org/10.1016/j.cpr.2013.04.004

Goñi-Sarriés, A., De Jalón Aramayo, E. G., González, N. L., & Urra, E. L. (2008). Análisis de las derivaciones desde atención primaria a salud mental [Analysis of referrals from primary care to mental health]. *Anales de Psiquiatría*, *24*(2), 83-88. https://dialnet.unirioja.es/servlet/articulo?codigo=2578508

González-Blanch, C., Muñoz-Navarro, R., Medrano, L. A., Moriana, J. A., Ruiz-Rodríguez, P., & Cano-Vindel, A. (2021). Moderators and predictors of treatment outcome in transdiagnostic group cognitive-behavioral therapy for primary care patients with emotional disorders. *Depression and Anxiety, 38*(7), 757-767. https://doi.org/10.1002/da.23164

González-Blanch, C., Barrio-Martínez, S., Priede, A., Martínez-Gómez, S., Pérez-García-Abad, S., Miras-Aguilar, M., Ruiz-Gutiérrez, J., Muñoz-Navarro, R., Ruiz-Rodríguez, P., Medrano, L. A., Prieto-Vila, M., Carpallo-González, M., Aguilera-Martín, Á., Gálvez-Lara, M., Cuadrado, F., Moreno, E., García-Torres, F., Venceslá, J. F., Corpas, J., . . . Cano-Vindel, A. (2023). Cost-effectiveness of transdiagnostic group cognitive behavioural therapy versus group relaxation therapy for emotional disorders in primary care (PsicAP-Costs2): Protocol for a multicentre randomised controlled trial. *PLoS ONE*, *18*(3), e0283104. https://doi.org/10.1371/journal.pone.0283104

González-Blanch, C. (2025). Algunas prevenciones más: una crítica sobre la prevención cuaternaria en salud mental [Some further considerations: A critique of quaternary prevention in mental health]. *Papeles del Psicólogo/Psychologist Papers*, *46*(2), 118–124. https://doi.org/10.70478/pap.psicol.2025.46.15

Hanevik, E., Røvik, F. M. G., Bøe, T., Knapstad, M., & Smith, O. R. F. (2023). Client predictors of therapy dropout in a primary care setting: A prospective cohort study. *BMC Psychiatry*, *23*(1), 358. https://doi.org/10.1186/s12888-023-04878-7

Horvath, A. O., Del Re, A. C., Flückiger, C., & Symonds, D. (2011). Alliance in individual psychotherapy. *Psychotherapy*, *48*(1), 9-16. https://doi.org/10.1037/a0022186

Instituto Nacional de Estadística. (2025). Spain Health Survey (SHS): Year 2023 [Press release]. https://www.ine.es/dyngs/Prensa/en/ESdE2023.htm

Institute for Health Metrics and Evaluation. (2022). Global Health Data Exchange (GHDx). https://vizhub.healthdata.org/gbd-results/

Jeitani, A., Fahey, P. P., Gascoigne, M., Darnal, A., & Lim, D. (2024). Effectiveness of stepped care for mental health disorders: An umbrella review of meta-analyses. *Personalized Medicine in Psychiatry*, *47*, 100140. https://doi.org/10.1016/j.pmip.2024.100140

Kirkbride, J. B., Anglin, D. M., Colman, I., Dykxhoorn, J., Jones, P. B., Patalay, P., Pitman, A., Soneson, E., Steare, T., Wright, T., & Griffiths, S. L. (2024). The social determinants of mental health and disorder: evidence, prevention and recommendations. *World Psychiatry, 23*(1), 58–90. https://doi.org/10.1002/wps.21160

Lau, B., Victor, M., & Ruud, T. (2016). Sickness absence and presence among employees in treatment for common mental disorders. *Scandinavian Journal of Public Health*, *44*(4), 338-346. https://doi.org/10.1177/1403494815621418

Loumidis, K. S., & Shropshire, J. M. (1997). Effects of waiting time on appointment attendance with clinical psychologists and length of treatment. *Irish Journal of Psychological Medicine*, *14*(2), 49-54. https://doi.org/10.1017/s0790966700002986

Marquina-Márquez, A., Olry-de-Labry-Lima, A., Bermúdez-Tamayo, C., Ferrer, L. I., & Marcos-Marcos, J. (2022). Identifying barriers and enablers for benzodiazepine (de)prescription: A qualitative study with patients and healthcare professionals. *Anales del Sistema Sanitario de Navarra, 45*(2), e1005. https://doi.org/10.23938/assn.1005

Martín-Jurado, A., De la Gándara Martín, J., Carbajo, S. C., Hernández, A. M., & Sánchez-Hernández, J. (2012). Análisis de concordancia de las derivaciones de atención primaria a salud mental [Concordance analysis of referrals from primary care to mental health]. *Medicina de Familia SEMERGEN*, *38*(6), 354-359. https://doi.org/10.1016/j.semerg.2011.12.005

McGorry, P. D., & Mei, C. (2021). Clinical staging for youth mental disorders: Progress in reforming diagnosis and clinical care. *Annual Review of Developmental Psychology, 3*(1), 15–39. https://doi.org/10.1146/annurev-devpsych-050620-030405

Ministerio de Sanidad. (2022). *Estrategia de Salud Mental del Sistema Nacional de Salud. Período 2022-2026* [Mental Health Strategy of the National Health System 2022 - 2026]. https://www.sanidad.gob.es/areas/calidadAsistencial/estrategias/saludMental/

Ministerio de Sanidad. (2024). *Aspectos relevantes del informe anual del Sistema Nacional de Salud 2023* [Relevant aspects of the annual report of the National Health System 2023]. https://www.sanidad.gob.es/estadEstudios/estadisticas/sisInfSanSNS/tablasEstadisticas/InfAnualSNS2023/ASPECTOS_RELEVANTES_2023.pdf

Miranda-Chueca I, Peñarrubia-María MT, García-Bayo I, Caramés-Durán E, Soler-Vila M, Serrano-Blanco A. (2003). ¿Cómo derivamos a salud mental desde atención primaria? [How do we refer to mental health from primary care?]. *Atención Primaria, 32*(9), 524-530. https://doi.org/10.1016/S0212-6567(03)70782-3

Moratalla, B. G., & Lobo, A. O. (2002). Ausencias en las primeras consultas de un centro de salud mental: Un estudio controlado [No attendance to initial appointments in a mental health centre: a controlled study]. *Revista de la Asociación Española de Neuropsiquiatría*, *22*(83), 27-36. https://doi.org/10.4321/s0211-57352002000300003

Nam, S. K., Chu, H. J., Lee, M. K., Lee, J. H., Kim, N., & Lee, S. M. (2010). A meta-analysis of gender differences in attitudes toward seeking professional psychological help. *Journal of American College Health*, *59*(2), 110-116. https://doi.org/10.1080/07448481.2010.483714

National Institute for Health and Care Excellence. (2011). *Generalised anxiety disorder and panic disorder in adults: Management* (NICE guideline No. CG113). https://www.nice.org.uk/guidance/cg113

National Institute for Health and Care Excellence. (2022). *Depression in adults:Treatment and management* (NICE guideline No. NG222). https://www.nice.org.uk/guidance/ng222

Patel, R., Shetty, H., Jackson, R., Broadbent, M., Stewart, R., Boydell, J., McGuire, P., & Taylor, M. (2015). Delays before diagnosis and initiation of treatment in patients presenting to mental health services with bipolar disorder. *PLoS ONE*, *10*(5), e0126530. https://doi.org/10.1371/journal.pone.0126530

Peipert, A., Krendl, A. C., & Lorenzo-Luaces, L. (2022). Waiting lists for psychotherapy and provider attitudes toward low-intensity treatments as potential interventions: Survey study. *JMIR Formative Research*, *6*(9), e39787. https://doi.org/10.2196/39787

Pfefferbaum, B., & North, C. S. (2020). Mental Health and the Covid-19 Pandemic. *New England Journal of Medicine*, *383*(6), 510-512. https://doi.org/10.1056/nejmp2008017

Punton, G., Dodd, A. L., & McNeill, A. (2022). 'You're on the waiting list': An interpretive phenomenological analysis of young adults' experiences of waiting lists within mental health services in the UK. *PLoS ONE*, *17*(3), e0265542. https://doi.org/10.1371/journal.pone.0265542

Reichert, A., & Jacobs, R. (2018). The impact of waiting time on patient outcomes: Evidence from early intervention in psychosis services in England. *Health Economics*, *27*(11), 1772-1787. https://doi.org/10.1002/hec.3800

Roos, J., & Werbart, A. (2013). Therapist and relationship factors influencing dropout from individual psychotherapy: A literature review. *Psychotherapy Research*, *23*(4), 394-418. https://doi.org/10.1080/10503307.2013.775528

Santomauro, D. F., Herrera, A. M. M., Shadid, J., Zheng, P., Ashbaugh, C., Pigott, D. M., Abbafati, C., Adolph, C., Amlag, J. O., Aravkin, A. Y., Bang-Jensen, B. L., Bertolacci, G. J., Bloom, S. S., Castellano, R., Castro, E., Chakrabarti, S., Chattopadhyay, J., Cogen, R. M., Collins, J. K., . . . Ferrari, A. J. (2021). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *The Lancet*, *398*(10312), 1700-1712. https://doi.org/10.1016/s0140-6736(21)02143-7

Sharf, J., Primavera, L. H., & Diener, M. J. (2010). Dropout and therapeutic alliance: A meta-analysis of adult individual psychotherapy. *Psychotherapy*, *47*(4), 637-645. https://doi.org/10.1037/a0021175

Steinert, C., Stadter, K., Stark, R., & Leichsenring, F. (2017). The Effects of waiting for treatment: A meta-analysis of waitlist control groups in randomized controlled trials for social anxiety disorder. *Clinical Psychology & Psychotherapy*, *24*(3), 649–660. https://doi.org/10.1002/cpp.2032

Swift, J. K., & Greenberg, R. P. (2012). Premature discontinuation in adult psychotherapy: A meta-analysis. *Journal of Consulting and Clinical Psychology*, *80*(4), 547–559. https://doi.org/10.1037/a0028226

Swift, J. K., Whipple, J. L., & Sandberg, P. (2012). A prediction of initial appointment attendance and initial outcome expectations. *Psychotherapy*, *49*(4), 549-556. https://doi.org/10.1037/a0029441

Tejedo-García, A. (2018). Análisis comparativo de las derivaciones desde atención primaria de salud a salud mental: Atención psicológica y atención psiquiátrica [Comparative analysis of referrals from primary health care to mental health: Psychological care and psychiatric care]. *Informaciones Psiquiátricas, 233*, 23-50. https://dialnet.unirioja.es/servlet/articulo?codigo=6983196

van Dijk, D., Meijer, R., Van Den Boogaard, T., Spijker, J., Ruhé, H., & Peeters, F. (2023). Worse off by waiting for treatment? The impact of waiting time on clinical course and treatment outcome for depression in routine care. *Journal of Affective Disorders*, *322*, 205-211. https://doi.org/10.1016/j.jad.2022.11.011

Vellisca, M. Y., Latorre, J. I., Orejudo, S., Gascón, S., Nolasco, A., & Villanueva, V. J. (2014). Patrón asociado a la inasistencia a la primera consulta de un centro de salud mental [Pattern associated with non-attendance to first appointment at a mental health center]. *Revista de Psicopatología y Psicología Clínica*, *19*(2), 141-146. https://doi.org/10.5944/rppc.vol.19.num.2.2014.13064

Wang, P. S., Aguilar-Gaxiola, S., Alonso, J., Angermeyer, M. C., Borges, G., Bromet, E. J., Bruffaerts, R., De Girolamo, G., De Graaf, R., Gureje, O., Haro, J. M., Karam, E. G., Kessler, R. C., Kovess, V., Lane, M. C., Lee, S., Levinson, D., Ono, Y., Petukhova, M., . . . Wells, J. E. (2007). Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. *The Lancet*, *370*(9590), 841-850. https://doi.org/10.1016/s0140-6736(07)61414-7

Wang, P. S., Berglund, P. A., Olfson, M., & Kessler, R. C. (2004). Delays in initial treatment contact after first onset of a mental disorder. *Health Services Research*, *39*(2), 393-416. https://doi.org/10.1111/j.1475-6773.2004.00234.x

World Health Organization (2017). *Depression and other common mental disorders: Global health estimates.* World Health Organization. https://iris.who.int/handle/10665/254610

World Health Organization. (2022). *Mental health and COVID-19: Early evidence of the pandemic's impact* [Scientific brief]. World Health Organization. https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci_Brief-Mental_health-2022.1-eng

GRUPO EDITORIAL DE PSICOFUNDACIÓN

Psicothema (2026) 38(1) 23-35

Psicothema

https://www.psicothema.com/es • ISSN 0214–9915 • eISSN 1886-144X

Colegio Oficial de Psicología del Principado de Asturias

Psicothema

Article

Assessing Positive Digital Experiences: A Spanish Validation of the Digital Flourishing Scale for Adolescents

Alfredo Zarco-Alpuente[1], Víctor Ciudad Fernández[1], Jasmina Rosič[2], Sophie Janicke-Bowles[3], Tamara Escrivà-Martínez[1,4] and Paula Samper-García[1]

[1] University of Valencia (Spain)
[2] Media Psychology Lab, Department of Communication Science, KU Leuven (Belgium)
[3] School of Communication Research, Chapman University, Orange (USA)
[4] Instituto de Salud Carlos III, Madrid (Spain)

ARTICLE INFO

Received: 25/04/2025
Accepted: 05/09/2025

Keywords:
Digital flourishing
Adolescents
Scale adaptation
Self-determination theory
Digital communication

ABSTRACT

Background: Adolescents are immersed in digital communication, which can benefit or harm their well-being. Digital flourishing captures positive perceptions of this communication—connectedness, authentic self-presentation, positive social comparison, civil participation, and self-control—and how it contributes to well-being. In Spain there is still no validated instrument for adolescents. **Method**: We adapted and validated the Digital Flourishing Scale for Adolescents (DFSA) for Spanish adolescents. Study 1 involved a pilot survey (n = 13) and cognitive interviews (n = 10) to improve clarity and cultural relevance. Study 2 used a cross-sectional survey (n = 1,786) to examine the DFSA's latent structure, measurement invariance by gender and age, internal reliability of scores, and validity evidence based on relationships to other variables. Study 3 assessed test-retest reliability of scores and longitudinal measurement invariance over six weeks (n = 289). **Results**: Study 1 improved item clarity and cultural relevance through linguistic adjustments. Study 2 confirmed a five-factor model, showing strict age invariance and metric gender invariance. All subscales correlated with well-being indicators. Study 3 showed poor to moderate temporal stability of scores but supported scalar longitudinal invariance. **Conclusions**: The Spanish DFSA is a promising tool for assessing adolescents' digital flourishing in the Spanish context.

Evaluando las Experiencias Digitales Positivas: Validación Española de la Escala de Florecimiento Digital para Adolescentes

Palabras clave:
Florecimiento digital
Adolescentes
Adaptación de escala
Teoría de la autodeterminación
Comunicación digital

RESUMEN

Antecedentes: Los adolescentes están inmersos en la comunicación digital, con efectos positivos y negativos en su bienestar. El florecimiento digital describe percepciones positivas de dicha comunicación—conectividad, autoexpresión auténtica, comparación social positiva, participación cívica y autocontrol—y su aporte al bienestar. En España no existe un instrumento validado para adolescentes. **Método:** Adaptamos y validamos la Escala de Florecimiento Digital para Adolescentes (DFSA) españoles. Estudio 1: incluyó encuesta piloto (n = 13) y entrevistas cognitivas (n = 10) para mejorar claridad y adecuación cultural. Estudio 2: encuesta transversal (n = 1.786) examinando estructura latente de DFSA, invarianza métrica por sexo y edad, fiabilidad interna de las puntuaciones y evidencia de validez basada en las relaciones con otras variables. Estudio 3 evaluó fiabilidad test-retest de las puntuaciones e invariancia longitudinal en seis semanas (n = 289). **Resultados:** Estudio 1: mejoró claridad y relevancia cultural. Estudio 2: confirmó un modelo de cinco factores, con invariancia estricta por edad e invariancia métrica por género. Todas las subescalas se correlacionaron con indicadores de bienestar. Estudio 3: mostró estabilidad temporal de las puntuaciones baja-moderada, confirmando invariancia longitudinal escalar. **Conclusiones:** La DFSA española es una herramienta prometedora para evaluar el florecimiento digital de los adolescentes en España.

Cite as: Zarco-Alpuente, A., Ciudad Fernández, V., Rosič, J., Janicke-Bowles, S., Escrivá-Martínez, T., & Samper-García, P. (2026). Assessing positive digital experiences: A Spanish validation of the Digital Flourishing Scale for Adolescents. *Psicothema, 38*(1), 23-35. https://doi.org/10.70478/psicothema.2026.38.03
Corresponding author: Paula Samper-García, paula.samper@uv.es
This article is published under Creative Commons License 4.0 CC-BY-NC-ND

Contemporary adolescents grow up fully immersed in digital communication technologies, significantly transforming how they spend their time and interact with their environment (Holly et al., 2023). While early research emphasized the potential risks of digital communication, recent scholarship has called for a more nuanced understanding that includes the positive aspects of digital communication (Vanden Abeele, 2021). One such approach is the emerging construct of digital flourishing, which emphasizes that beneficial use of digital communication can satisfy adolescents' developmental needs and promote both hedonic and eudaimonic well-being (Gudka et al., 2023; Janicke-Bowles et al., 2023).

Digital flourishing refers to positive perceptions of digital communication experiences and behaviours contributing to well-being and fulfilment (Janicke-Bowles et al., 2023). To operationalize this construct, Janicke-Bowles et al. (2023) developed the Digital Flourishing Scale (DFS) for adults, which was later adapted for adolescents (DFSA) (Rosič et al., 2022). This instrument captures five interrelated dimensions: connectedness (feeling socially connected online), authentic self-presentation (expressing one's true self online), positive social comparison (feeling inspired after socially comparing online), civil participation (engaging respectfully and constructively online), and self-control (managing time spent online).

The theoretical foundation of digital flourishing draws significantly from Self-Determination Theory (SDT; Deci & Ryan, 2000). According to SDT, the satisfaction of the basic psychological needs for relatedness, autonomy, and competence is essential for well-being. Digital flourishing builds on this framework by proposing that digital communication can support these needs. Empirical studies have consistently found that adolescents who report higher levels of digital flourishing also experience greater psychological need satisfaction and related well-being outcomes (Janicke-Bowles et al., 2023; Janicke-Bowles, 2024; Rosič et al., 2022).

To the best of our knowledge, the DFSA is currently the only validated instrument specifically designed to assess digital flourishing in adolescence. It is currently available in English, Slovenian (Rosič et al., 2022), Dutch (Schreurs & Vandenbosch, 2024), and Chinese (Yao et al., 2025). However, it has not yet been adapted to widely spoken languages such as Spanish. While other frameworks have assessed general flourishing in Spanish among adults (e.g. De la Fuente et al., 2017), the DFSA provides a unique tool to evaluate adolescents' positive digital communication. This study aims to adapt the DFSA for Spanish-speaking adolescents using a multimethod approach (i.e. cognitive interviewing, a cross-sectional and longitudinal study) to evaluate its psychometric properties, evidence of validity based on the relationship with other variables, measurement invariance, and temporal reliability.

Digital flourishing is theorized to support basic psychological needs, namely relatedness, competence, and autonomy (Janicke-Bowles et al., 2023). During developmental period of adolescence these needs become more salient and therefore, digital flourishing is especially relevant. Regarding relatedness, adolescents increasingly prioritize peer relationships for identity validation and emotional support, decreasing compliance with parents (Berk, 2022; Girelli et al., 2019). For competence, adolescents prefer independent decisions and complex tasks, seeking challenges that foster achievement and mastery (Berk, 2022). Autonomy needs manifest as adolescents actively pursue independence through self-determined decisions and activities (Girelli et al., 2019).

Moreover, adolescents are among the highest users of digital media (Boer et al., 2020). Digital media use plays a vital role during adolescence, providing platforms for socialization, learning and self-expression (Holly et al., 2023). The positive interactions adolescents have while using digital media are part of the context that can contribute to the satisfaction of basic psychological needs and shape their development (Holly et al., 2023).

Digital communication with peers may provide adolescents with a sense of belonging, satisfying their need for relatedness by making them feel connected and less lonely (Rosič et al., 2024). This virtual context offers flexibility in choosing what to share, who to interact with, and when, supporting the fulfilment of relational needs (O'Keeffe et al., 2011). When adolescents learn to interact responsibly online and navigate online communication challenges like presenting themselves authentically in spaces shaped by "positivity bias" and idealized portrayals, digital communication also contributes to the need for competence (Schreurs & Vandenbosch, 2024). Positive social comparisons online, especially in areas like academics, sports, and relationships, offer insights into their perceived competence and can evoke motivation, inspiration, and benign envy (Meier & Schäfer, 2018). Civil participation online is also relevant for competence, as adolescents' psychosocial and cognitive development fosters prosocial and civil engagement in online discussions (Lysenstøen et al., 2021). Finally, as their cognitive abilities mature, adolescents gain greater self-control over digital interactions, an important aspect of autonomy in a context of constant connectivity (Hoareau et al., 2021; Rosič et al., 2022). These dimensions of connectedness, civil participation, authentic self-presentation, positive social comparison, and self-control, form the core of digital flourishing and have been theorized and empirically proven to relate to the basic psychological needs' satisfaction (Janicke-Bowles et al., 2023).

Previous studies measuring digital flourishing using the DFSA have consistently supported a better fit for multidimensional model with five-factor structure than high-order structure among adolescent (Rosič et al., 2022) and adult samples (Janicke-Bowles et al., 2023), although both structures were acceptable. Therefore, digital flourishing can be investigated either through a composite score or by analysing its five dimensions separately, as each dimension captures distinct but complementary aspects of positive digital experiences. This study examines whether the five-factor structure replicates in a new cultural context, namely Spain, which presents a distinctive setting in terms of digital engagement. Spain represents a unique environment, ranking seventh worldwide in active social media use (83.6%), notably above the global average (62.3%) and higher than the United States (70.1%) and Slovenia (76.9%) (DataReportal, 2024), where the DFS(A) have previously been applied. Consequently, Spanish adolescents navigate unique demands from ubiquitous connectivity (Vanden Abeele, 2021).

From an SDT perspective, broader social systems shape the opportunities adolescents have to pursue and satisfy their basic psychological needs. In highly connected environments, digital communication may both enable and constrain these opportunities, depending on how access is regulated. For example, recent restrictions on smartphone use in Valencian schools (see resolution of 17 April 2024 DOGV - Generalitat Valenciana) may impact digital flourishing by creating tension between institutional regulations and widespread peer smartphone use. Thus, adapting an instrument assessing positive

digital communication perceptions among Spanish adolescents requires an understanding of their specific context.

In addition to contextual relevance, examining the DFSA's associations with theoretically and empirically grounded constructs allows for a more comprehensive validation of the instrument within the Spanish adolescent population.

First, previous research has shown that all five dimensions of digital flourishing are significantly associated with the satisfaction of basic psychological needs (i.e. relatedness, competence, autonomy) (Rosič et al., 2022). The connectedness subscale was significantly associated with all three needs, showing the strongest correlation with relatedness. The civil participation and self-control subscales were most significantly related to autonomy, while the positive social comparison and authentic self-presentation demonstrated the strongest associations with competence (Rosič et al., 2022). We expected positive correlations between DFSA dimensions and basic psychological needs satisfaction.

In terms of broader well-being, satisfaction with life is a personal evaluation of life quality based on the alignment between individual aspirations and actual circumstances (Kjell & Diener, 2021). The dimensions of digital flourishing have been associated with higher levels of overall well-being, including life satisfaction (Janicke-Bowles et al., 2023). Therefore, we expected that higher levels of digital flourishing will be positively correlated with greater satisfaction with life.

Conversely, loneliness is a subjective experience of distress from a lack of social connection or belonging (Beutel et al., 2017). Digital communication (i.e. texting, group chatting) can foster the development of social connections and a sense of belonging among adolescents (Vincent, 2016). However, many adolescents report feelings of loneliness and isolation when communicating on social media, which can harm their sense of belonging and subsequently diminish their well-being (Smith et al., 2021). Consequently, higher loneliness was expected to negatively correlate with connectedness.

Authenticity can be defined as perceiving one's actions as self-authored and is achieved by acting in accordance with one's values, preferences, and needs (Ryan & Ryan, 2019), is another construct related to digital flourishing. Digital communication provides new opportunities for authentic self-expression, such as spontaneously and informally sharing daily activities and thoughts (Manning et al., 2017), which many adolescents do through apps such as BeReal or Instagram. Being authentic has been linked to higher well-being (Smallenbroek et al., 2017). Higher authenticity on social media was expected to positively correlate with authentic self-presentation.

Although much research links online social comparison to lower well-being, recent studies suggest that positive (or upward) comparison, which evokes benign envy, can inspire and enhance well-being (Meier & Schaefer, 2018; Meier et al., 2020). This process of inspiration is also considered in relation to digital flourishing. Specifically, content that is either creative, transformative in nature or portrays human's moral nature, is especially powerful to elicit inspiration (Chang, 2022). In turn, the experience of inspiration from online content or interactions has been found to increase love and compassion over time (Janicke-Bowles et al., 2022). We hypothesised that higher social media-induced inspiration would be positively related to positive social comparison.

On the negative side of digital interactions, Internet aggression includes harmful behaviours toward others online such as cyberbullying (Ybarra & Mitchell, 2004). Although most adolescents experience positive social interactions online, a significant minority are affected by negative interactions, either as perpetrators, targets, or both (Werner et al., 2010). These aggressive behaviours can include rude, threatening, harassing comments, unwanted sexual remarks, and social exclusion (Ybarra & Mitchell, 2004). Adolescents who engage more frequently in respectful online discourse and civil participation are significantly less likely to engage in aggressive or harmful digital communication (Jones & Mitchell, 2015). We hypothesised that higher rates of Internet aggression would be negatively related to civil participation.

Finally, problematic social media use (PSMU) refers to users' perceptions that their social media use cannot be controlled and is overused, characterized by the presence of various symptoms: preoccupation, tolerance, withdrawal, relapse, mood modification, detrimental consequences in important life domains and displacement of activities due to social media use (Boer et al., 2020). Such problematic use has been associated with a range of mental health problems (Huang, 2020). Research highlights that individuals with lower self-control dispositions are more likely to present PSMU (Osatuyi & Turel, 2018). Thus, we expected higher PSMU to negatively correlate with self-control.

The present research adapted the DFSA (Rosič et al., 2022) to the digital communication experiences of Spanish adolescents, following standard scale development procedures (Carpenter, 2018). In Study 1, a pilot survey and cognitive interviews with adolescents were conducted to assess clarity of the scale translated to Spanish. In Study 2, a cross-sectional survey was conducted to replicate the latent structure of the DFSA, evaluate measurement invariance for gender and age, and assess the scale's validity evidence based on its relationships to other variables. In Study 3, a longitudinal survey was conducted with a subsample of the participants from Study 2 to explore the temporal reliability and longitudinal measurement invariance of the scale. For the final Spanish DFSA version with the adaptations made after the study, see the OSF document 'DFSA'.

This study received approval from the University of [blinded] ethics committee (2039883). Prior to participation, all individuals were fully briefed on the study's objectives and gave their informed consent. For participants > 14 years, parental consent was obtained. Those ≥ 14 years could choose to provide their birth date and initials for a follow-up conducted 6 weeks later, which was done to explore the temporal reliability and longitudinal measurement invariance of the scale in Study 3. The responses of participants under 14 remained entirely anonymous. The database has also been used in other articles [Blinded].

This study was preregistered in November 2023 before the data analysis on the Open Science Framework (OSF) at https://osf.io/be4wh/?view_only=bc0e99ccd6334f66aaf463ccd7b0403b. Data, scripts, supplementary materials, and other resources are available on the same OSF page.

Method

Study 1: Pilot Survey and Cognitive Interviews

Participants

A total of 20 adolescents were initially recruited through the researchers' personal networks to participate in a pilot survey. The final sample consisted of 13 adolescents (12-18 years, M_{age} = 15.62, SD_{age} = 2.04, 69.2% girls). For the cognitive interview phase, 10 adolescents participated across two group sessions: one conducted in person (n = 8) and another online (n = 2) due to logistical constraints.

Instruments

In the pilot survey, participants rated each item's clarity on a 3-point scale (*1 = I don't understand anything; 2 = I understand it well, but not completely; 3 = I understand it perfectly*) and answered an open-ended question about any comprehension issues or suggestions. These measures collected both quantitative and qualitative feedback on the clarity and cultural relevance of the translated DFSA items.

Procedure

The original English version of the DFSA was translated into Spanish using a forward-backward translation procedure by two bilingual researchers. The resulting versions were reviewed by native Spanish speakers, and discrepancies were resolved to ensure semantic, idiomatic, experiential, and conceptual equivalence, resulting in a preliminary Spanish version.

A pilot survey was then administered using Qualtrics between September 2023 and May 2023. Based on reported comprehension issues, semi-structured cognitive interviews were conducted to assess validity based on response processes (Ryan et al., 2012). Following a hybrid model, both think-aloud and verbal probing techniques were employed (Padilla & Benítez, 2014). Details on the sample and specific changes made to the DFSA can be found in the OSF folder *'Cognitive Interview'*.

To ensure the methodological rigor of the adaptation process, we evaluated the Spanish version of the DFSA against the International Test Commission (ITC) guidelines for test adaptation (Hernández et al., 2020). A checklist documenting compliance with each criterion is available in the OSF document 'ITC adaptation checklist'.

Data Analysis

For quantitative pilot survey data, the percentage of participants for the three response options was calculated for each item to assess item clarity. Items were flagged for revision if over 25% of participants indicated partial or no understanding. Open-ended responses were analysed thematically, and researcher notes and observations of cognitive interviews were examined to identify common interpretation issues and improvement suggestions.

Results

According to the OSF document 'Pilot Survey Comprehensibility', 14 of 21 items were well understood by over 75% of participants. However, four items raised concerns, with nearly half indicating limited understanding, prompting cognitive interviews.

Based on this feedback, a series of changes were implemented across the scale. The introductory text was revised using more familiar and age-appropriate terminology (e.g. replacing "online applications" with "online activities") and updated to reflect the platforms most used by Spanish adolescents (e.g. replacing Viber with Telegram and including BeReal, Twitter, and gaming chats). Wording across items was adjusted to enhance specificity and personal meaning. For instance, some item content was also rephrased to better align with adolescents' digital communication experiences. For example, in the civil participation dimension, the item referring to "politics" was reworded to "current affairs (such as sports, politics, or celebrities)," as the original formulation was perceived as abstract or detached from participants' online interactions. All changes are available in the OSF under the files 'DFSA Changes' and 'DFSA Comparative'.

Study 2: Cross-Sectional Study

Participants

Out of initial 3,464 participants, we removed participants who: (1) did not accept the informed consent (n = 82), (2) were not between 13 and 19 years old or did not answer age question (n = 511), (3) had no access or didn't use social media (n = 53), and (4) failed at least two out of the three attention check questions (e.g. "If you are reading this, select 'Agree'.") (Buchanan & Scofield, 2018) (n = 457). The final sample consisted of 1,786 participants (M_{age} = 15.22, SD_{age} = 1.20, 49.0% girls, 66% Compulsory Secondary Education, 87% Spanish nationality). For more detailed results see the OSF document "Sociodemographic Study 2".

Instruments

Demographic Variables. Adolescents reported their age and gender (*1 = boy, 2 = girl, 3 = non-binary, 4 = prefer not to say*). Responses for the option "non-binary" and "prefer not to say" were included in the analyses, except for the gender invariance testing. Adolescents' educational level was categorized as follows: compulsory secondary education (ages 12-16), post-compulsory secondary education (ages 16-18), and vocational training levels (ages 16-20). Additionally, participants indicated their nationality.

Digital Flourishing in Adolescence. The 21-item DFSA in Spanish with five factors using a scale from *1 (Not at all true of me) to 5 (Very true of me), with an option "Not applicable to me"* was used. Reliability indices: connectedness (α = .65, ω = .68), civil participation (α = .73, ω = .76), positive social comparison (α = .78, ω = .81), authentic self-presentation (α = .82, ω = .86), and self-control (α = .79, ω = .83).

The Satisfaction of Basic Psychological Needs. We used the 12-item Brief Scale Measuring Basic Psychological Needs Satisfaction (BPNS; Girelli et al., 2019) evaluated on a 5-point Likert-type scale ranging from *1 (Not true at all) to 5 (Very true)*. Since no validated Spanish version for adolescents was available, we conducted a confirmatory factor analysis (CFA) to examine the internal structure and support the validity of the interpretations derived from the scores in our sample. The analysis confirmed a three-factor structure: Relatedness (e.g., "I like the people I know") (α = .78, ω = .81), Competence (e.g., "I feel good at doing many things") (α = .84, ω = .86), and Autonomy (e.g., "I feel free to decide how to do my own things") (α = .83, ω = .87), in line with the original model. See the OSF documents "CFA BPNS" and "Construct Validity Evidence for the BPNS" for further information regarding its construct validity in this sample.

Satisfaction With Life. We used the 3-item Satisfaction with Life Scale (SWLS-3; Ortuño-Sierra et al., 2019; Kjell & Diener, 2021) (e.g., "The conditions of my life are excellent") evaluated on a 7-point Likert-type scale ranging from *1 (Strongly disagree) to 7 (Strongly agree)*. Following Kjell and Diener's (2021) recommendations, the last two items out of five were removed. Internal consistency for the scale was excellent (α = .87, ω = .87).

Loneliness. The Three-Item Loneliness Scale (TILS; Trucharte et al., 2023) was used on a 3-point Likert scale ranging from 1 (Hardly ever) to 3 (Often) (e.g. "How often do you feel that you lack company?"). Reliability indices: α = .88 and ω = .89.

Subjective Authenticity of Positive Self-Content on Social Media. One item from the Virtual Self subscale of the Psycho-Social Aspects of Facebook Use (Bodroža & Jovanović et al., 2016) was adapted ("When you posted messages on social media during the last month, did you have the impression that these messages showed who you really are?"). As this questionnaire was not available in Spanish, it was translated and adapted for the present study. Responses were given on a 5-point Likert scale ranging from *1 (Never) to 5 (Very often)*. This item obtained an association of .51 with the Authentic self-presentation factor from the DFSA (Rosič et al., 2022).

Social Media-Induced Inspiration. Two items of the Social Media-Induced Inspiration Scale (SMII; Meier & Schäfer, 2018) were used: "When I use social media, I am inspired by the posts of other users to do something [new]" and "When I use social media, I experience inspiration." The word "Instagram" was replaced with "social media". Answers ranged from *1 (Strongly disagree) to 5 (Strongly agree)* with the option "Not applicable to me". As this questionnaire was not available in Spanish, it was translated and adapted for the present study. The Spearman-Brown coefficient was .71.

Internet Aggression. The 4-item Internet Aggression Scale (IAS; Werner et al., 2010) was used (e.g. "I used the Internet to play a joke or annoy someone I was mad at.") with a scale ranging from *1 (Never) to 4 (5 or more times) with the option "Not applicable to me"* (α = .86 and ω = .87). As this questionnaire was not available in Spanish, it was translated and adapted for the present study.

Social Media Disorder. The 9-item Social Media Disorder Scale (SMD-S; Boer et al., 2020) was used (e.g. "How often have you felt bad when you have not been able to use social networks?"). We adapted an original dichotomous Yes/No response format to a 6-point Likert scale, following Savci et al. (2018). Reliability indices in this sample are excellent (α = .90; ω = .90).

Procedure

Data collection took place in educational institutions between September 2023 and May 2024 in person, using either paper or digital formats (e.g. Qualtrics via tablet, smartphone, computer). While no monetary compensation was offered, participation was encouraged by providing a personalised report of the results and an educational workshop. Participants were recruited from schools that had collaborated in previous research and the official directory of educational institutions by the Generalitat Valenciana (GVA). School staff (e.g. counsellors, head teachers, or psychology departments) agreed to explain the study's aims and coordinate data collection within class time.

Data Analysis

First, internal consistency of the test scores was assessed using Cronbach's α and McDonald's ω, with polychoric correlation matrices. For the two-item scale (i.e. the Social Media-Induced Inspiration Scale), Spearman-Brown coefficient was calculated (Eisinga et al., 2013).

Second, multiple confirmatory factor analysis (CFA) models were tested to confirm the theoretical latent structure for the DFSA: a one-factor model, an uncorrelated five-factor model, a correlated five-factor model, and a five-factor model with a second order factor. Model fit was evaluated using the Comparative Fit Index (CFI), Tucker-Lewis Index (TLI) (i.e. ≥ .95 = excellent and ≥ .90 = acceptable), Root Mean Square Error of Approximation (RMSEA) with confidence intervals, Standardized Root Mean Square Residual (SRMR) (i.e. ≤ .06 = excellent and ≤ .08 = acceptable) (Hu & Bentler, 1999), and the χ^2 statistic (Kyriazos, 2018).

Third, measurement invariance of the test scores was examined across gender and age groups (early adolescence [13-14 years], middle adolescence [15-16], and late adolescence [17-19 years]) using a stepwise approach: (1) a configural model was tested without any restrictions (i.e. configural invariance); next, models were tested with constrained (2) factor loadings (i.e. metric invariance); (3) item intercepts (i.e. scalar invariance); and (4) residual variances (i.e. strict invariance). Responses for the option "non-binary", "prefer not to say", and "other" were excluded for gender invariance testing due to the low number of cases, which made it unfeasible to analyse the factorial model exclusively for these groups. To assess if constraining the models resulted in a significant reduction in model fit (i.e., measurement invariance), the χ^2 test, *p*-values, changes in CFI (≤ .01) and RMSEA (≤ .015), were examined (Chen, 2007), with ΔCFI and ΔRMSEA prioritized over the χ^2 due to its sensitivity to significant differences even when they are negligible (Kyriazos, 2018). When full invariance was not supported, partial invariance was subsequently tested by freeing parameters exhibiting the largest statistically significant cross-group differences. All CFA and invariance models used Maximum Likelihood with robust correction (MLR), with missing data handled using Full Information Maximum Likelihood.

Lastly, to assess validity evidence based on relationships to other variables, a Spearman correlation matrix was computed. CFA were conducted for each measure with at least three items (McNeish, 2023). Factor scores were then computed for each subscale.

Analyses were conducted using R version 4.3.2, the packages psych (Revelle, 2023), lavaan (Rosseel, 2012), semTools (Jorgensen et al., 2021), and ggcorrplot (Kassambara, 2019).

Results

Table 1 presents descriptive statistics for the study variables and Table 2 indicates the descriptive statistics of the DFSA items.

Table 3 shows the statistical fit of the CFA models. Model 1 demonstrate poor fit according to the cut-off scores. Model 2 shows a better fit, with an acceptable RMSEA, but poor remaining fit indices. Model 3 shows the best fit, with excellent values for all fit indices and an acceptable TLI. Model 4, which considers a second-order factor encompassing the five factors, indicates an acceptable CFI and TLI and excellent RMSEA and SRMR but fits notably worse than Model 3. Therefore, Model 3 was retained in further analyses.

Figure 1 presents the measurement model from Model 3. Most factor loadings were above .50. All correlations between latent factors were statistically significant except for the correlation between Factor 3 (Positive social comparison) and Factor 5 (Self-Control), which was not significant.

Table 4 indicates gender (boys and girls) invariance models. The configural model indicates acceptable CFI and RMSEA. Although the metric model indicates a significant χ^2 change, CFI and RMSEA remain within cut-offs. However, the scalar model showed a significant reduction in goodness-of-fit exceeding the cut-off. This indicated

Table 1
Descriptive Statistics and Validity Evidence Based on Relationships with Other Variables

Variables	*n*	*M*	*SD*	*Mdn*	*MAD*	Min	Max	Skewness	Kurtosis
DFSA connectedness	1,725	2.50	0.92	2.67	0.99	0	5	-0.29	-0.05
DFSA authentic self-presentation	1,726	2.69	0.87	2.75	0.74	0	5	-0.30	0.06
DFSA positive social comparison	1,725	2.23	0.96	2.25	1.11	0	5	0.12	-0.01
DFSA civil participation	1,725	2.93	0.69	3.00	0.59	0	5	-0.42	1.01
DFSA self-control	1,786	2.58	0.84	2.75	0.74	0	5	-0.30	0.06
BSBP Relatedness	1,786	13.84	4.90	15	2.97	0	20	-1.57	2.32
BSBP Competence	1,786	13.59	4.99	15	2.97	0	20	-1.37	1.73
BSBP Autonomy	1,786	13.74	5.08	15	4.45	0	20	-1.34	1.59
Life satisfaction	1,786	5.06	1.39	5.33	1.48	1	7	-0.78	-0.01
Loneliness	1,786	4.26	2.09	4	1.48	0	9	0.10	0.09
Subjective authenticity of positive self-content on social media	1,595	3.79	1.45	4	1.48	1	6	-.40	-0.66
Social Media-Induced Inspiration Scale	1,786	5.64	2.85	6	2.97	0	12	-0.56	-0.22
Internet Aggression Scale	1,786	4.80	2.92	4	1.48	0	20	1.22	3.81
Social Media Disorder Scale	1,786	20.46	10.11	20	10.38	0	52	0.03	-0.02

Note. SD: Standard Deviation; *MAD:* Median absolute deviation.

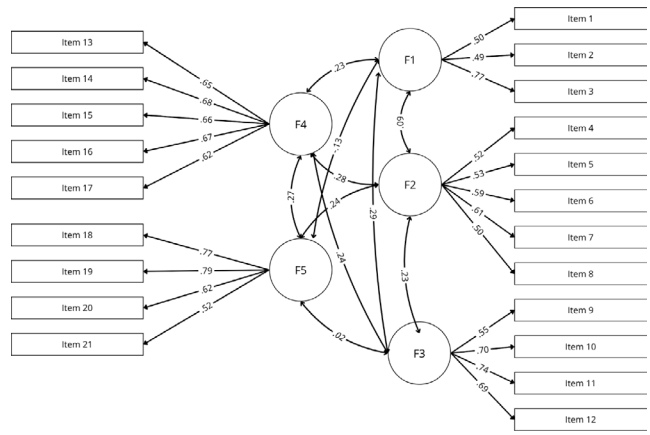
Table 2
Descriptive Statistics and Discrimination Indices for Individual Items of the Digital Flourishing Scale

Subscale	Item	*M*	*SD*	Skewness	Kurtosis	% Floor	% Ceiling	Item-total correlation
Connectedness	1	2.71	1.32	-0.18	-0.45	6.3	10.0	0.37
Connectedness	2	2.47	1.10	-0.25	-0.40	4.5	1.4	0.36
Connectedness	3	2.31	1.28	0.07	-0.51	8.6	5.1	0.49
Authentic self-presentation	1	2.54	1.14	0.05	-0.18	3.8	5.0	0.56
Authentic self-presentation	2	2.77	1.08	-0.30	0.22	3.3	5.0	0.58
Authentic self-presentation	3	2.69	1.30	0.02	-0.46	4.9	11.2	0.58
Authentic self-presentation	4	2.73	1.20	-0.23	-0.18	4.5	6.9	0.59
Authentic self-presentation	5	2.72	1.17	-0.25	-0.01	4.8	6.3	0.54
Positive social comparison	1	2.69	1.16	-0.33	-0.32	4.2	3.5	0.47
Positive social comparison	2	2.34	1.21	0.01	-0.41	7.2	3.8	0.61
Positive social comparison	3	2.02	1.27	0.34	-0.37	11.2	4.1	0.62
Positive social comparison	4	1.86	1.35	0.45	-0.49	17.2	4.4	0.56
Civil participation	1	3.02	1.00	-0.86	1.04	3.0	2.7	0.41
Civil participation	2	3.04	1.07	-0.91	0.56	3.2	2.3	0.43
Civil participation	3	3.06	0.95	-0.56	0.83	1.5	4.2	0.45
Civil participation	4	3.11	0.99	-0.54	0.87	1.9	6.3	0.49
Civil participation	5	2.41	1.20	0.03	-0.29	6.3	4.7	0.39
Self-control	1	2.54	1.10	-0.26	-0.29	4.1	2.1	0.62
Self-control	2	2.50	1.12	-0.21	-0.46	4.1	1.9	0.64
Self-control	3	2.39	1.10	-0.18	-0.43	4.7	1.5	0.56
Self-control	4	2.90	1.02	-0.78	0.41	2.8	1.0	0.48

Note. % Floor = Percentage of participants endorsing the lowest possible score on the item. % Ceiling = Percentage of participants endorsing the highest possible score on the item. Item-total correlation indicates the item's ability to discriminate between high and low scorers on the subscale.

a relevant loss in fit, suggesting that constraining item intercepts between men and women resulted in a non-negligible decrease in model fit to the data. Therefore, to continue comparing nested models, a partial invariance analysis was conducted. The intercept of DFSA Civil participation item 2 was identified as the most problematic. By freeing this intercept in the partial scalar invariance model, the changes in fit indices compared to the metric model were below the cut-off, achieving partial scalar invariance. Finally, strict invariance was assessed. The initial strict invariance model (with only DFSA Civil participation item 2 intercept freed) showed a ΔCFI violating the criterion. Further analysis identified the residual variance of DFSA Civil participation item 2 as the most problematic. By freeing both the intercept and the residual variance of DFSA Civil participation item 2, partial strict invariance was supported. In summary, complete metric invariance and partial strict invariance have been established. This means that men and women share the same latent structure and factor loadings. Furthermore, after freeing the intercept and residual variance of DFSA Civil participation item 2, strict invariance was achieved, which is crucial for comparing both latent factor means and variances between the groups.

Figure 1
Measurement Model of the DFSA.



Note. For the sake of clarity, unique variances and intercepts were omitted. Non-significant estimates are written in italics. Factor 1: Authentic self-presentation; Factor 2: Civil participation; Factor 3: Positive social comparison; Factor 4: Connectedness; Factor 5: Self-control.

Table 3
Confirmatory Factor Analyses Models

Model	χ^2	*df*	CFI	TLI	RMSEA	SRMR
One factor (Model 1)	4253.388*	189	.380	.311	.125 [.122, .128]	.115
Five uncorrelated factors (Model 2)	862.972*	189	.897	.886	.051 [.047, .054]	.083
Five correlated factors (Model 3)	519.960*	179	.948	.939	.037 [.033, .041]	.033
Five factors model with a second order factor (Model 4)	623.564*	184	.934	.924	.041 [.038, .045]	.047

Note. χ^2: Chi-Square; df: Degrees of Freedom; CFI: Comparative Fit Index; TLI: Tucker-Lewis Index; RMSEA: Root Mean Square Error of Approximation; SRMR: Standardized Root Mean Square Residual. *$p < .05$.

Table 4
Confirmatory Factor Models Assessing Gender Invariance

Model	χ^2	*df*	CFI	RMSEA	$\Delta\chi^2$	Δdf	*p*-value	ΔCFI	ΔRMSEA
Boys	378.644*	179	.940	.040	-	-	-	-	-
Girls	357.366*	179	.947	.038	-	-	-	-	-
Measurement Invariance Models									
Configural invariance	901.710	358	.943	.039	-	-	-	-	-
Metric invariance	945.500	374	.940	.039	33.980	16	.005	.003	.000
Scalar invariance	1069.100	390	.927	.042	125.310	16	<.001	.013	.003
Partial scalar invariance (DFSA Civil Participation - item 2 intercept freed)	1012.031	389	.934	.040	67.297	15	<.001	.007	.001
Partial strict invariance (DFSA Civil Participation - item 2 intercept freed)	1184.858	410	.916	.044	114.200	21	<.001	.017	.003
Partial strict invariance (DFSA Civil Participation - item 2 intercept and residual freed)	909.381	409	.924	.042	19.518	14	.146	.006	.005

Note. $\Delta\chi^2$: Chi-square difference across the previous and the current model; *Δdf:* Degrees of Freedom Difference across the previous and the current model; p-value: Probability Value; ΔCFI: Change in Comparative Fit Index across the previous and the current model; ΔRMSEA Change in Root Mean Square Error of Approximation across the previous and the current model

Table 5 reports age (early, middle and late adolescents) invariance models. The configural model obtained acceptable fit according to CFI and RMSEA. The metric model indicated non-significant χ^2 difference and minimal loss of fit in CFI and RMSEA. Similarly, the scalar and strict invariance models do not show a significant χ^2 difference, with CFI and RMSEA below the cut-off thresholds, supporting the assumption of age invariance across factor loadings, item intercepts, and unique variances. Therefore, the different adolescent age groups share a similar latent structure in the DFSA.

Figure 2 indicates the associations among variables. DFSA dimensions positively correlated with basic psychological needs satisfaction, and life satisfaction, except for positive social comparison, which was not significantly associated with life satisfaction. Loneliness was negatively associated with DFSA connectedness. Subjective authenticity of positive self-content on social media was associated with higher of DFSA authentic self-presentation. Social media-induced inspiration was positively correlated with DFSA positive social comparison. Internet aggression was negatively associated with DFSA civil participation. Finally, problematic social media use was negatively associated with DFSA self-control.

Study 3: A Longitudinal Survey

Participants

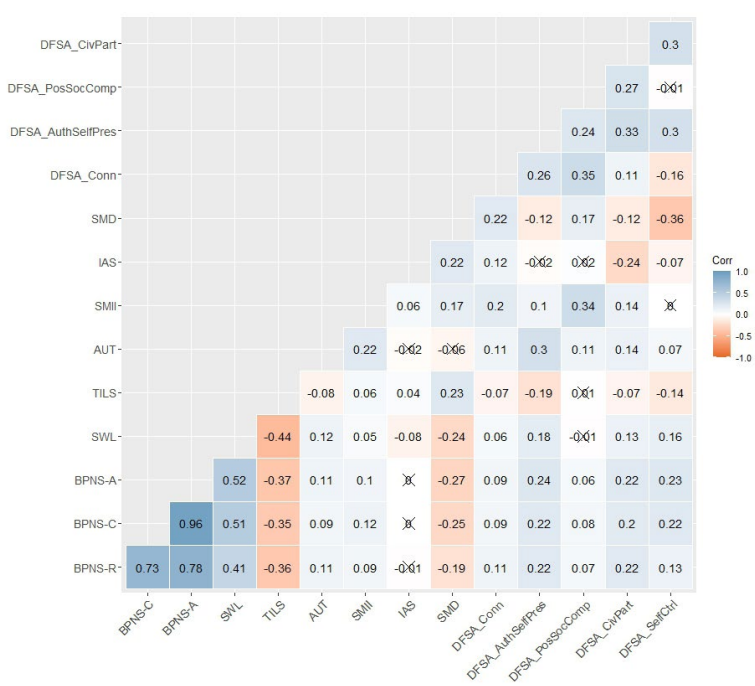
A subsample of 286 adolescents from the cross-sectional Study 2 participated in a follow-up assessment 6 weeks later (M_{age} = 15.71,

Table 5
Confirmatory Factor Models Assessing Age Invariance

Model	χ^2	*df*	CFI	RMSEA	$\Delta\chi^2$	Δdf	*p*-value	ΔCFI	ΔRMSEA
Early Adolescence (13-14)	253.726	179	0.953	0.035	-	-	-	-	-
Middle Adolescence (15-16)	365.989	179	0.941	0.039	-	-	-	-	-
Late Adolescence (17-19)	288.115	179	0.948	0.038	-	-	-	-	-
Measurement Invariance Models									
Configural	1105.0	537	0.942	0.035	-	-	-	-	-
Metric	1161.5	569	0.940	0.034	42.849	32	0.095	0.002	0.001
Scalar	1204.8	601	0.938	0.034	43.220	32	0.089	0.002	0.000
Strict	1281.1	643	0.938	0.033	50.372	42	0.176	0.001	0.001

Note. χ^2: Chi-square; *df:* Degrees of freedom; CFI: Comparative Fit Index; RMSEA: Root Mean Square Error of Approximation. $\Delta\chi^2$: Chi-square difference across the previous and the current model; Δdf: Degrees of freedom difference across the previous and the current model; ΔCFI: Change in CFI across the previous and the current model; ΔRMSEA: Change in RMSEA across the previous and the current model.

Figure 2
Spearman Correlation Matrix Among Variables



Note. BPNS-R: Basic Psychological Needs Satisfaction Relatedness; BPNS-C: Competence; BPNS-A: Autonomy; SWL: Satisfaction with Life; TILS: Loneliness; AUT: Subjective Authenticity of Positive Self-Content on Social Media; SMII: Social Media-Induced Inspiration Scale; IAS: Internet Aggression Scale; SMD: Social Media Disorder; DFSA_Conn: Connectedness; DFSA_AuthSelfPres: Authentic Self-Presentation; DFSA_PosSocComp: Positive Social Comparison; DFSA_CivPart: Civil Participation; DFSA_SelfCtrl: Self-Control. Non-significant Spearman correlations are blank.

SD_{age} = 1.08, age range: 14-19; 49.99% boys). Table 6 presents descriptive statistics for the study variables

Table 6
Sociodemographic Characteristics of Participants in Study 3

Variables	*n*
Age	286
14	29 (10%)
15	100 (35%)
16	108 (37%)
17	31 (11%)
18	10 (3.5%)
19	8 (2.8%)
Gender	286
Boy	143 (50.9%)
Girl	141 (49.8%)
Non-binary	1 (0.3%)
Prefer not to say	1 (0.3%)

Instruments

The DFSA (Rosič et al., 2022) adapted in Study 1 was administered.

Procedure

The same procedure as in Study 2 was followed.

Data Analysis

To evaluate temporal reliability of the DFSA subscales scores, intraclass correlation coefficients (ICC) for each dimension were computed to detect systematic measurement bias while verifying temporal stability of scores (Correa-Rojas, 2021). The ICC were calculated along with its 95% confidence interval using a two-way mixed-effects model, single measurement, and absolute agreement. Cutoff values of ICC values were: < .50 poor, .50 < .75 moderate, .75 < .90 good, and > .90 excellent reliability of the scores (Koo & Li, 2016).

To evaluate the longitudinal invariance of the DFSA measurement model between measurement time points (time 1 and 2, i.e. after 6 weeks), a series of progressively constrained CFAs was performed using MLR as the estimation method and full information maximum likelihood to handle missing values.

Results

Table 7 reports the ICCs and confidence intervals. Subscales for connectedness, authentic self-presentation, positive social comparison, civil participation, and self-control showed poor to moderate stability, indicating that scores are prone to fluctuate over time.

Table 8 presents fit indices for longitudinal invariance models of the DFSA. The configural, metric, and scalar models show adequate fit indices, with minimal changes in χ^2, CFI, and RMSEA. However, the strict model indicated a significant χ^2 difference. Although ΔRMSEA was within acceptable limits, the decrease in CFI exceeded the threshold. Hence, the DFSA demonstrated longitudinal invariance across factor loadings and item intercepts but not for unique item variances.

Table 7
Intraclass Correlation Coefficients and Confidence Intervals

Subscale	ICC	Lower CI	Upper CI	Classification
Connectedness	.467	.372	.553	Poor to Moderate
Authentic Self-Presentation	.504	.412	.585	Poor to Moderate
Positive Social Comparison	.464	.368	.550	Poor to Moderate
Civil Participation	.471	.375	.556	Poor to Moderate
Self-control	.599	.519	.668	Moderate

Note. ICC: Intraclass correlation coefficient. ICC was computed considering a single-measurement, absolute-agreement, two-way mixed effects model.

Discussion

This research had two aims: translating and adapting the DFSA and evaluating its psychometric properties in Spanish adolescents. Results showed that the Spanish DFSA is a promising tool for measuring digital flourishing, aligning with prior validations (Janicke-Bowles et al., 2023; Rosič et al., 2022; Schreurs & Vandenbosch, 2024; Yao et al., 2025).

Study 1 improved questionnaire comprehensibility by tailoring it to the Spanish context. While some items were easily understood, others posed difficulties, prompting further refinement. Based on cognitive interviews results, instructions were clarified, the language was simplified, and additional examples were provided to improve clarity. These adjustments laid the groundwork for the psychometric evaluation.

Table 8
Longitudinal Invariance Models

Model	χ^2	*df*	CFI	RMSEA	$\Delta\chi^2$	Δdf	*p*-value	ΔCFI	ΔRMSEA
Time 1	328.651	179	.916	.047	-	-	-	-	-
Time 2	320.362	179	.941	.047	-	-	-	-	-
Configural	649.01	358	.931	.047	-	-	-	-	-
Metric	673.04	374	.930	.046	17.401	16	.360	-.001	-.001
Scalar	693.33	390	.929	.046	20.502	16	.198	-.001	-.001
Strict invariance	815.31	411	.901	.052	74.380	21	<.001	-.027	.007

Note. $\Delta\chi^2$: Chi-square difference across the previous and the current model; Δdf: Degrees of Freedom Difference across the previous and the current model; *p*-value: Probability Value; ΔCFI: Change in Comparative Fit Index across the previous and the current model; ΔRMSEA: Change in Root Mean Square Error of Approximation across the previous and the current model.

In Study 2, the correlated five-factor model comprising connectedness, authentic self-presentation, positive social comparison, civil participation, and self-control, showed the best fit in the Spanish adolescent context and supports the conceptualization of digital flourishing as a set of interrelated but distinct dimensions. This finding aligns with prior validations of the scale in both adolescent and adult samples (Janicke-Bowles et al., 2023; Rosič et al., 2022), where the multidimensional structure consistently outperformed alternative models. In our sample, both the one-factor and the hierarchical models showed poorer fit indices compared to the five-factor solution, further supporting a multidimensional conceptualization of the construct over the use of a global DFSA score. Internal consistency was acceptable across subscales, except for connectedness, which was borderline-possibly due to its three-item length (Streiner, 2003).

The study also found strict measurement invariance across age groups, meaning the construct is measured equivalently in early, middle and late adolescents. As a result, observed differences between these age groups could probably be attributed to true differences in the underlying latent variable, rather than to variations in item interpretation (Meredith, 1993). Only metric measurement invariance was met across gender, indicating that the construct is conceptualized similarly by boys and girls. However, the lack of scalar invariance suggests discrepancies in item intercepts across gender, meaning that boys and girls may interpret items differently, potentially leading to biased comparisons of latent means (Blanco-Canitrot et al., 2018).

The DFSA's validity based on relationships to other variables was supported. The connectedness subscale correlated negatively with loneliness, a pattern consistent with prior research suggesting that digital communication can help foster a sense of belonging and reduce feelings of isolation (Trucharte et al., 2023; Vincent, 2016). Authentic self-presentation was positively associated with subjective authenticity, supporting that adolescents who feel able to act in accordance with their values and preferences online also perceive their digital self-presentation as more genuine (Ryan & Ryan, 2019; Schreurs & Vandenbosch, 2022). Positive social comparison online was positively associated with inspiration, consistent with studies showing that upward comparison in online contexts can evoke constructive and motivating emotional responses (Chang, 2022; Meier & Schäfer, 2018). Civil participation was inversely related to Internet aggression, indicating that adolescents who engage more frequently in polite and respectful digital communication report lower involvement in hostile online interactions (Lysenstøen et al., 2021; Werner et al., 2010). Finally, self-control correlated negatively with problematic social media use, echoing previous findings that highlight the role of self-regulatory difficulties in problematic patterns of social media engagement (Boer et al., 2020; Osatuyi & Turel, 2018). However, effect sizes were small (r = .05 to .20), finding not uncommon in media effects research (Meier & Reinecke, 2021). These low estimates may reflect moderate measurement error, especially in dimensions like positive social comparison, self-control, and civil participation, which showed lower reliability (DeVellis & Thorpe, 2021). This suggests a need to review and possibly expand these subscales.

It is worth noting the weak, albeit significant, relationship between positive social comparison and the need for competence. Conceptualization of the scale (Janicke-Bowles et al., 2023) proposes that enhancing competence in digital communication involves successfully organizing one's online social environment to reduce negative social comparisons and increase positive ones. However, in both the current study and the original validation, this subscale, while significant, shows the lowest correlation with the hypothesized basic psychological need (in this case with competence). This may be due to operationalization of the items. While items capture the received benefits from positive social comparisons, the scale does not address the presence of negative social comparisons, which may be equally important in assessing a sense of competence in digital interactions. Without considering both positive and negative social comparisons, the scale may fail to fully capture adolescents' ability to manage social dynamics in digital communication, which is central to the feeling of competence in this context. Similarly, all DFSA subscales were significantly associated with satisfaction with life, further supporting the scale's relevance in capturing key aspects of overall well-being (Janicke-Bowles et al., 2023; Kjell & Diener, 2021).

Study 3 showed poor to moderate temporal stability of the DFSA across six weeks. The ICC values suggest that scores fluctuate, potentially due to changes in school or family context, social dynamics, digital trends, or broader sociocultural factors (Magis-Weinberg et al., 2021). Given that the DFSA measures adolescents' digital communication experiences, such variability is not unexpected. Adolescence is a developmental period characterized by ongoing changes in self-concept, social habits, and digital engagement patterns, making adolescents more susceptible to variations in their responses (Berk, 2022). Moreover, recent research emphasizes that the time frame chosen for measurement plays an important role in how digital media uses and effects manifest. Media use and its effects can vary depending on the daily events, the distinction between weekdays and weekends, and even seasonal factors (Vandenbosch et al., 2025). It is therefore possible that a six-week interval is insufficient to capture meaningful temporal stability, and longer intervals should be considered in future research. For instance, study on digital flourishing fluctuations among adolescents found relatively stable patterns when assessments were spaced over one-year with four-month intervals (Rosič et al., 2024).

Longitudinal invariance testing showed scalar invariance over time, indicating that score changes reflect genuine shifts in the latent construct rather than interpretation differences (Mackinnon et al., 2022). However, residual invariance was not met, suggesting that item-level measurement error varied across time. Despite this, the DFSA appears suitable for longitudinal studies, although further research is needed.

The Spanish version of the DFSA offers educators and researchers a promising tool to assess the extent to which adolescents experience their digital communication as enriching and meaningful. While most available instruments emphasize problematic or excessive use, the DFSA offers a complementary, theory-based perspective by capturing five positive dimensions of digital communication. The results support its reliability, structural validity, and measurement invariance in the samples, allowing for use across diverse adolescent groups. In educational settings, the DFSA can help identify areas where students perceive greater or lesser fulfilment in their digital experiences, inform digital literacy programs, and support more balanced technology-related policies. Developed exclusively for research purposes, the scale is not intended for diagnostic or high-stakes decision-making. Instead, it promotes educational dialogue around adolescents' lived positive digital communication experiences, fostering a more holistic understanding of their

relationship with technology and supporting the development of healthier, more autonomous, and socially engaged digital habits.

This study has some limitations. First, the cognitive group interviews included fewer male than female participants. Additionally, the sample used to validate the DFSA was composed entirely of students from Valencia and Madrid, limiting the generalizability of the findings to Spanish adolescents as a whole. Moreover, lower internal consistency was found with the connectedness subscale. Future research may explore whether revisiting the original five-item subscale of social connectedness with Spanish adolescents would yield more reliable results than a three-item subscale (Janicke-Bowles et al., 2023). Moreover, DFSA is a self-report measure and captures reflections of adolescents' digital communication experiences rather than actual outcomes. This could lead to socially desirable responses (Janicke-Bowles et al., 2023). However, self-reported measures are frequently used in digital communication use research (Meier & Reinecke, 2021). Finally, while several of the scales used in Study 2 had validated Spanish versions, three instruments had not been formally validated in Spanish: the Satisfaction of Basic Psychological Needs (Girelli et al., 2019), the Virtual Self subscale (Bodroža & Jovanović, 2016), the Social Media-Induced Inspiration Scale (Meier & Schäfer, 2018), and the Internet Aggression Scale (Werner et al., 2010). These were included following the same approach used in the original adolescent validation of the DFSA (Rosič et al., 2022), but relying on non-validated translations is not considered best practice and may affect the accuracy and interpretability of the results. Future studies should further validate the DFSA with other validated measures in Spanish.

The DFSA focuses on positive digital experiences. Combining it with measures of digital drawbacks may clarify how benefits and harms coexist in media use (Vanden Abeele, 2021). This counterbalance is essential, as positive experiences alone do not capture the full scope of adolescent digital communication. Although the DFSA emphasizes need satisfaction via positive digital interactions, Basic Psychological Needs Theory (Ryan & Deci, 2017) suggests that experiences can also lead to need frustration. Future research should consider developing instruments to assess negative digital experiences linked to need frustration, offering a fuller picture of adolescents' digital lives within SDT. Additionally, cross-country comparisons could reveal how cultural differences shape digital flourishing. Understanding these variations would inform culturally tailored strategies to promote positive digital experiences among adolescents.

Author Contributions Statement

Alfredo Zarco-Alpuente: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Visualization, Writing - Original draft. **Víctor Ciudad-Fernández**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing - Original draft, Writing - Review & editing. **Jasmina Rosič:** Conceptualization, Methodology, Writing - Original draft, Writing - Review & editing. **Sophie H. Janicke-Bowles**: Conceptualization, Writing - Review & editing. **Tamara Escrivà-Martínez**: Conceptualization, Data curation, Writing - Review & editing**. Paula Samper-García**: Conceptualization, Methodology, Project administration, Supervision, Writing - Review & editing.

Funding

Alfredo Zarco-Alpuente and Víctor Ciudad-Fernández are supported by FPU grants FPU22/01588 and FPU21/00527, respectively, funded by the Ministry of Science, Innovation and Universities, Spain. The funding source had no involvement in the study design, data collection, analysis or interpretation, manuscript writing, or the decision to submit the article for publication.

Conflict of Interest

The authors declare that there are no conflicts of interest.

Data Availability Statement

This study was preregistered in November 2023 on the Open Science Framework (OSF) at https://osf.io/be4wh/?view_only=bc0e99ccd6334f66aaf463ccd7b0403b. Data, scripts, supplementary materials, and other resources are available on the same OSF page.

References

Berk, L. E. (2022). *Development through the lifespan*. Sage Publications.

Beutel, M. E., Klein, E. M., Brähler, E., Reiner, I., Jünger, C., Michal, M., Wiltink, J., Wild, P. S., Münzel, T., Lackner, K. J., & Tibubos, A. N. (2017). Loneliness in the general population: Prevalence, determinants and relations to mental health. *BMC Psychiatry*, *17*(1), 97. https://doi.org/10.1186/s12888-017-1262-x

Blanco-Canitrot, D., Alvarado, J. M., & Ondé, D. (2018). Consequences of disregarding metric invariance on diagnosis and prognosis using psychological tests. *Frontiers in Psychology, 9*, 167. https://doi.org/10.3389/fpsyg.2018.00167

Bodroža, B., & Jovanović, T. (2016). Validation of the new scale for measuring behaviours of Facebook users: Psycho-Social Aspects of Facebook Use (PSAFU). *Computers in Human Behavior*, *54*, 425-435. https://doi.org/10.1016/j.chb.2015.07.032

Boer, M., van den Eijnden, R. J., Boniel-Nissim, M., Wong, S. L., Inchley, J. C., Badura, P., ... & Stevens, G. W. (2020). Adolescents' intense and problematic social media use and their well-being in 29 countries. *Journal of Adolescent Health*, *66*(6), S89-S99. https://doi.org/10.1016/j.jadohealth.2020.02.014

Buchanan, E. M., & Scofield, J. E. (2018). Methods to detect low quality data and its implication for psychological research. *Behavior Research Methods*, *50*, 2586-2596. https://doi.org/10.3758/s13428-018-1035-6

Carpenter, S. (2018). Ten steps in scale development and reporting: A guide for researchers. *Communication Methods and Measures, 12*(1), 25-44. https://doi.org/10.1080/19312458.2017.1396583

Chang, C. (2022). Being inspired by media content: Psychological processes leading to inspiration. *Media Psychology*, *26*(1), 72-87. https://doi.org/10.1080/15213269.2022.2097927

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464-504. https://doi.org/10.1080/10705510701301834

Correa-Rojas, J. (2021). Intraclass correlation coefficient: Applications to estimate the temporal stability of a measuring instrument. *Ciencias Psicológicas, 15*(2), 1-16.

Datareportal. (2024). *Digital 2024: Global overview report*. https://datareportal.com/reports/digital-2024-global-overview-report

De la Fuente, R., Parra, A., & Sánchez-Queija, I. (2017). Psychometric properties of the Flourishing Scale and measurement invariance between two samples of Spanish university students. *Evaluation & the health professions*, *40*(4), 409-424. https://doi.org/10.1177/0163278717703446

Deci, E. L., & Ryan, R. M. (2000). The "what" and" why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*(4), 227-268. https://doi.org/10.1207/S15327965PLI1104_01

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale Development: Theory and applications* (5th ed.). Sage Publications.

Eisinga, R., Grotenhuis, M. T., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown?. *International Journal of Public Health*, *58*, 637-642. https://doi.org/10.1007/s00038-012-0416-3

Girelli, L., Cavicchiolo, E., Lucidi, F., Cozzolino, M., Alivernini, F., & Manganelli, S. (2019). Psychometric properties and validity of a brief scale measuring basic psychological needs satisfaction in adolescents. *Journal of Educational, Cultural and Psychological Studies, 20,* 215-229. https://doi.org/10.7358/ECPS-2019-020-GIRE

Gudka, M., Gardiner, K. L. K., & Lomas, T. (2023). Towards a framework for flourishing through social media: a systematic review of 118 research studies. *The Journal of Positive Psychology*, *18*(1), 86-105. https://doi.org/10.1080/17439760.2021.1991447

Hernández, A., Hidalgo-Montesinos, M. D., Hambleton, R. K., & Gómez-Benito, J. (2020). International Test Commission guidelines for test adaptation: A criterion checklist. *Psicothema, 32*(3), 390–398. https://hdl.handle.net/2445/185025

Hoareau, N., Bagès, C., & Guerrien, A. (2021). Cyberbullying, self-control, information, and electronic communication technologies: do adolescents know how to exercise self-control on the internet?. *International Journal of Bullying Prevention*, *5*(4), 306-316. https://doi.org/10.1007/s42380-021-00099-2

Holly, L., Wong, B. L. H., van Kessel, R., Awah, I., Agrawal, A., & Ndili, N. (2023). Optimising adolescent wellbeing in a digital age. *BMJ*, *380*, e068279. https://doi.org/10.1136/bmj-2021-068279

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

Huang, C. (2020). A meta-analysis of the problematic social media use and mental health. *International Journal of Social Psychiatry*, *68*(1), 12-33. https://doi.org/10.1177/0020764020978434

Janicke-Bowles, S. (2024). Digital flourishing in the US: Validation of the digital flourishing scale (DFS) and demographic exploration. *Communication studies, 75*(3), 322-341. https://doi.org/10.1080/10510974.2023.2289688

Janicke-Bowles, S. H., Buckley, T. M., Rey, R., Wozniak, T., Meier, A., & Lomanowska, A. (2023). Digital flourishing: Conceptualizing and assessing positive perceptions of mediated social interactions. *Journal of Happiness Studies*, *24*(3), 1013-1035. https://doi.org/10.1007/s10902-023-00619-5

Janicke-Bowles, S. H., Raney, A. A., Oliver, M. B., Dale, K. R., Zhao, D., Neumann, D., Clayton, R. B., & Hendry, A. A. (2022). Inspiration on social media: Applying an entertainment perspective to longitudinally explore mental health and well-being. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *16*(2), 1. https://doi.org/10.5817/CP2022-2-1

Jones, L. M., & Mitchell, K. J. (2015). Defining and measuring youth digital citizenship. *New media & society*, *18*(9), 2063-2079. https://doi.org/10.1177/1461444815577797

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2021). *semTools: Useful tools for structural equation modeling (Version 0.5-5) [R package]*. https://CRAN.R-project.org/package=semTools

Kassambara, A. (2019). *ggcorrplot: Visualization of a correlation matrix using 'ggplot2'* (Version 0.1.3) [R package]. https://CRAN.R-project.org/package=ggcorrplot

Kjell, O. N. E., & Diener, E. (2021). Abbreviated three-item versions of the satisfaction with life scale and the harmony in life scale yield as strong psychometric properties as the original scales. *Journal of Personality Assessment, 103*(2), 183-194. https://doi.org/10.1080/00223891.2020.1737093

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Kyriazos, T. A. (2018). Applied psychometrics: sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology*, *9*(08), 2207-2230.

Lysenstøen, C., Bøe, T., Hjetland, G. J., & Skogen, J. C. (2021). A Review of the Relationship Between Social Media Use and Online Prosocial Behavior Among Adolescents. *Frontiers in Psychology*, *12*, 579347. https://doi.org/10.3389/fpsyg.2021.579347

Mackinnon, S., Curtis, R., & O'Connor, R. (2022). A tutorial in longitudinal measurement invariance and cross-lagged panel models using lavaan. *Meta-Psychology*, *6*. https://doi.org/10.15626/MP.2020.2595

Magis-Weinberg, L., Ballonoff Suleiman, A., & Dahl, R. E. (2021). Context, development, and digital media: Implications for very young adolescents in LMICs. *Frontiers in Psychology, 12*, 632713. https://doi.org/10.3389/fpsyg.2021.632713

Manning, N., Penfold-Mounce, R., Loader, B. D., Vromen, A., & Xenos, M. (2017). Politicians, celebrities and social media: a case of informalisation? *Journal of Youth Studies*, *20*(2), 127-144. https://doi.org/10.1080/13676261.2016.1206867

McNeish, D. (2023). Psychometric properties of sum scores and factor scores differ even when their correlation is 0.98: A response to Widaman and Revelle. *Behavior Research Methods, 55*(8), 4269-4290. https://doi.org/10.3758/s13428-022-02016-x

Meier, A., & Reinecke, L. (2021). Computer-mediated communication, social media, and mental health: a conceptual and empirical meta-review. *Communication Research, 48*(8), 1182-1209. https://doi.org/10.1177/0093650220958224

Meier, A., & Schäfer, S. (2018). The positive side of social comparison on social network sites: How envy can drive inspiration on Instagram. *Cyberpsychology, Behavior, and Social Networking*, *21*(7), 411-417. https://doi.org/10.1089/cyber.2017.0708

Meier, A., Gilbert, A., Börner, S., & Possler, D. (2020). Instagram inspiration: How upward comparison on social network sites can contribute to well-being. *Journal of Communication, 70*(5), 721-743. https://doi.org/10.1093/joc/jqaa025

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525-543. https://doi.org/10.1007/BF02294825

O'Keeffe, G. S., & Clarke-Pearson, K. (2011). The impact of social media on children, adolescents, and families. *Pediatrics*, *127*(4), 800-804. https://doi.org/10.1542/peds.2011-0054

Ortuño-Sierra, J., Aritio-Solana, R., Chocarro de Luis, E., Nalda, F. N., & Fonseca-Pedrero, E. (2019). Subjective well-being in adolescence: New psychometric evidences on the satisfaction with life scale. *European Journal of Developmental Psychology, 16*(2), 236-244. https://doi.org/10.1080/17405629.2017.1360179

Osatuyi, B., & Turel, O. (2018). Tug of war between social self-regulation and habit: Explaining the experience of momentary social media addiction symptoms. *Computers in Human Behavior, 85*, 95-105. https://doi.org/10.1016/j.chb.2018.03.037

Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*(1), 136-144. https://doi.org/10.7334/PSICOTHEMA2013.259

Revelle, W. (2023). *psych: Procedures for personality and psychological research* (Version 2.3.9) [R package]. Northwestern University. https://CRAN.R-project.org/package=psych

Rosič, J., Janicke-Bowles, S. H., Carbone, L., Lobe, B., & Vandenbosch, L. (2022). Positive digital communication among youth: The development and validation of the digital flourishing scale for adolescents. *Frontiers in Digital Health*, *4*, 2. https://doi.org/10.3389/fdgth.2022.975557

Rosič, J., Schreurs, L., Janicke-Bowles, S. H., & Vandenbosch, L. (2024). Trajectories of digital flourishing in adolescence: The predictive roles of developmental changes and digital divide factors. *Child Development, 95*(5), 1586-1602. https://doi.org/10.1111/cdev.14101

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of statistical software*, *48*(1), 1-36.

Ryan, K., Gannon-Slater, N., & Culbertson, M. J. (2012). Improving survey methods with cognitive interviews in small-and medium-scale evaluations. *American Journal of Evaluation, 33*(3), 414-430. https://doi.org/10.1177/1098214012441499

Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness.* The Guilford Press. https://doi.org/10.1521/978.14625/28806

Ryan, W. S., & Ryan, R. M. (2019). Toward a social psychology of authenticity: Exploring within-person variation in autonomy, congruence, and genuineness using self-determination theory. *Review of General Psychology*, *23*(1), 99-112. https://doi.org/10.1037/gpr0000162

Savci, M., Ercengiz, M., & Aysan, F. (2018). Turkish adaptation of the social media disorder scale in adolescents. *Archives of Neuropsychiatry*, *55*(3), 248-255. https://doi.org/10.5152/npa.2017.19285

Schreurs, L., & Vandenbosch, L. (2022). The Development and Validation of Measurement Instruments to Address Interactions with Positive Social Media Content. *Media Psychology*, *25*(2), 262-289. https://doi.org/10.1080/15213269.2021.1925561

Schreurs, L., & Vandenbosch, L. (2024). *Enhancing adolescents' emotion regulation in the social media context: A cluster randomized controlled trial of the Vibe Check social media literacy intervention*. OSF. https://doi.org/10.31219/osf.io/kagyu

Smallenbroek, O., Zelenski, J. M., & Whelan, D. C. (2017). Authenticity as a eudaimonic construct: The relationships among authenticity, values, and valence. *The Journal of Positive Psychology*, *12*(2), 197-209. https://doi.org/10.1080/17439760.2016.1187198

Smith, D., Leonis, T., & Anandavalli, S. (2021). Belonging and loneliness in cyberspace: impacts of social media on adolescents' well-being. *Australian Journal of Psychology*, *73*(1), 12-23. https://doi.org/10.1080/00049530.2021.1898914

Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*(1), 99-103. https://doi.org/10.1207/S15327752JPA8001_18

Trucharte, A., Calderón, L., Cerezo, E., Contreras, A., Peinado, V., & Valiente, C. (2023). Three-item loneliness scale: psychometric properties and normative data of the Spanish version. *Current Psychology*, *42*(9), 7466-7474. https://doi.org/10.1007/S12144-021-02110-X/TABLES/7

Vanden Abeele, M. (2021). Digital wellbeing as a dynamic construct. *Communication Theory*, *31*(4), 932-955. https://doi.org/10.1093/CT/QTAA024

Vandenbosch, L., Beullens, K., Vanherle, R., & Schreurs, L. (2025). Digital media uses and effects: The contributing roles of time. *Journal of Children and Media, 19*(1), 71-76. https://doi.org/10.1080/17482798.2024.2438690

Vincent, E. A. (2016). Social media as an avenue to achieving sense of belonging among college students. *Vistas Online*, *36*, 1-14.

Werner, N. E., Bumpus, M. F., & Rock, D. (2010). Involvement in internet aggression during early adolescence. *Journal of Youth and Adolescence*, *39*, 607-619. https://doi.org/10.1007/s10964-009-9419-7

Yao, W., Hou, H., Yang, P., & Ni, S. (2025). The co-occurrence of adolescent smartphone addiction and academic burnout: The role of smartphone stress and digital flourishing. *Education and Information Technologies*, *30*(4), 4987-5007. https://doi.org/10.1007/s10639-024-13017-yhttps://doi.org/10.1007/s10639-024-13017-y

Ybarra, M. L., & Mitchell, K. J. (2004). Online aggressor/targets, aggressors, and targets: A comparison of associated youth characteristics. *Journal of Child Psychology and Psychiatry*, *45*(7), 1308-1316. https://doi.org/10.1111/j.1469-7610.2004.00328.x

GRUPO EDITORIAL DE PSICOFUNDACIÓN

Psicothema (2026) 38(1) 36-45

Psicothema

https://www.psicothema.com/es • ISSN 0214–9915 • eISSN 1886-144X

Colegio Oficial de Psicología del Principado de Asturias

Psicothema

Article

Assessing Impulsivity in Adolescents: Psychometric Properties of the Spanish Short S-UPPS-P

Esteve Montasell-Jordana[1,2], Eva Penelo[1], Laura Blanco-Hinojo[3,4], Beatriz Lanceta[2], Laura Gomàriz-Camacho[2], Mar Gràcia[2], Anna Soler[2], Jesús Pujol[3] and Joan Deus[1]

[1] Universidad Autònoma de Barcelona (Spain)
[2] ITA Salud Mental (Spain)
[3] Hospital del Mar (Spain)
[4] Barcelona Institute for Global Health (Spain)

ARTICLE INFO

Received: 30/04/2025
Accepted: 21/07/2025

Keywords:
Adolescents
Impulsivity
Normative data
S-UPPS-P
Validation

ABSTRACT

Background: The short S-UPPS-P is a 20-item self-report tool for assessing impulsivity in adolescents, differentiating five dimensions: Negative Urgency, Lack of Perseverance, Lack of Premeditation, Sensation Seeking, and Positive Urgency. This study aimed to evaluate the psychometric properties of the Spanish S-UPPS-P and to establish normative data for adolescents in Spain. **Method**: Participants were 8,944 adolescent students (ages 11–19) from 66 high schools and 789 adolescent psychotherapy patients from 7 centers. **Results**: The expected 5-factor model, evaluated with confirmatory factor analysis (CFA), showed insufficient fit (CFI and TLI ≤ .90, RMSEA = .076). However, an exploratory approach yielded satisfactory results (CFI and TLI ≥ .97, RMSEA ≤ .036), with full measurement invariance across age, gender and sample type. Internal consistency reliability ranged from moderate to excellent (ω = .67–.82). Convergent validity with the Barratt Impulsiveness Scale total score was satisfactory (*r* = .47–.59). No significant differences in scale scores were observed across gender, age, or sample type, providing the use of a single norm. **Conclusions**: These findings support the S-UPPS-P as a valid, reliable tool for assessing impulsivity in Spanish adolescents. The availability of standardized norms enhances its utility in clinical and educational contexts.

Evaluación de la Impulsividad en Adolescentes: Propiedades Psicométricas de la Versión Corta Española S-UPPS-P

Palabras clave:
Adolescencia
Impulsividad
Baremos
S-UPPS-P
Validación

RESUMEN

Antecedentes: El S-UPPS-P es un instrumento de 20 ítems para evaluar la impulsividad en adolescentes, diferenciando cinco dimensiones: Urgencia Negativa, Falta de Perseverancia, Falta de Premeditación, Búsqueda de Sensaciones y Urgencia Positiva. Este estudio evaluó sus propiedades psicométricas y estableció datos normativos en adolescentes españoles. **Método:** Participaron 8.944 estudiantes (11-19 años) de 66 institutos y 789 pacientes adolescentes de salud mental. **Resultados:** El modelo de cinco factores, evaluado mediante análisis factorial confirmatorio (AFC), mostró ajuste insuficiente (CFI y TLI ≤ .90, RMSEA = .076). Sin embargo, un enfoque exploratorio mostró resultados satisfactorios (CFI y TLI ≥ .97, RMSEA ≤ .036), con invariancia completa del modelo de medida en función de la edad, género y tipo de muestra. La consistencia interna fue moderada a excelente (ω = .67-.82), y la validez convergente con la Escala de Impulsividad de Barratt fue adecuada (*r* = .47-.59). No se hallaron diferencias significativas en las puntuaciones según género, edad o muestra, permitiendo el uso de un único baremo. **Conclusiones**: Estos resultados apoyan al S-UPPS-P como un instrumento válido y fiable para evaluar la impulsividad en adolescentes españoles. La disponibilidad de baremos aumenta su utilidad en contextos clínicos y educativos.

Cite as: Montasell-Jordana, E., Penelo, E., Blanco-Hinojo, L., Lanceta, B., Gomàriz-Camacho, L., Gràcia, M., Soler, A., Pujol, J., & Deus., J. (2026). Assessing impulsivity in adolescents: Psychometric properties of the Spanish short S-UPPS-P. *Psicothema, 38*(1), 36-45. https://doi.org/10.70478/psicothema.2026.38.04
Corresponding author: Esteve Montasell-Jordana, esteve.montasell@autonoma.cat
This article is published under Creative Commons License 4.0 CC-BY-NC-ND

Impulsivity is a multifaceted construct defined as "a predisposition toward rapid, unplanned reactions to internal or external stimuli [with diminished] regard to the negative consequences of these reactions to the impulsive individual or others" (Potenza, 2007, p. 5). It has been suggested that high impulsivity may be associated with cognitive impairments and various problem behaviors, as well as engaging in risky behaviors that could potentially contribute to the development of mental health problems (Potenza, 2007).

Adolescence is a developmental stage characterized by heightened emotional reactivity and poor inhibitory control, which makes adolescents more prone than older individuals to impulsive actions and experimentation with potentially risky behaviors, such as drug use, suicidal behaviour, early sexual activity, or delinquent and aggressive behaviors, (Caro-Cañizares et al., 2024; Duell & Steinberg, 2019). However, the availability of assessment tools specifically validated for this population remains limited (Kulendran et al., 2016). Whiteside and Lynam (2001) developed a conceptual framework for impulsivity within the context of the five-factor model of personality (Costa & McCrae, 1985). Based on the analysis of 17 impulsivity-related scales, they identified four distinct facets of impulsivity and created a multidimensional measure known as the UPPS Impulsive Behavior Scale, which includes Negative Urgency, Lack of Premeditation, Lack of Perseverance, and Sensation Seeking. This model was later expanded by Cyders and colleagues (2007) by incorporating Positive Urgency, resulting in the UPPS-P scale. The UPPS-P scale allows for assessment of multiple aspects of impulsive personality, capturing various expressions of impulsivity that are relevant to a range of clinical manifestations in youth, such as in mood disorders (Caro-Cañizares et al., 2024), fetal alcohol spectrum disorders (Kingdon et al., 2016; Mattson et al., 2019; Carrera et al., 2024), Attention-Deficit/Hyperactivity Disorder (Miller et al., 2010) or eating disorders (Mallorquí-Bagué et al., 2020). Notably, Urgency is a core component of impulsivity and a transdiagnostic risk factor for several mental disorders, particularly during developmental adolescence (Littlefield et al., 2016; Sonmez et al., 2024).

After the UPPS-P gained wide acceptance, shorter versions were developed (Billieux et al., 2012; Cyders et al., 2014), reducing the original 59-item scale to 20 items while maintaining the original 5-factor structure. These shorter versions (S-UPPS-P) are frequently used in clinical settings to support professional judgment and streamline multi-step assessments, thanks to their brevity and ease of administration (Rammstedt & Beierlein, 2014). Their reduced cognitive load and shorter completion time make them particularly suitable for adolescents in both clinical and educational contexts (Omrani et al., 2019). Adolescents, compared to adults, are more prone to boredom, cognitive fatigue, and inconsistent adherence to response scales (Fortgang & Cannon, 2022).

Previous research has shown that the 20-item and 5-factor model of the S-UPPS-P provides an acceptable fit in adolescent samples (Donati et al., 2021; Eray et al., 2023; Pechorro et al., 2021; Wang et al., 2020) when its internal structure is evaluated using confirmatory factor analysis (CFA), mostly considering indicators as continuous. Potential competing models (such as a single factor or three interrelated factors grouping Negative and Positive Urgency [as broad urgency] and combining Lack of Premeditation and of Perseverance [labelled as deficits in conscientiousness], while Sensation Seeking remaining separated) have shown to fit worse. Higher correlations have been observed between Negative and Positive Urgency, as well as between Lack of Premeditation and Lack of Perseverance. By contrast, Sensation Seeking is recognized as a distinct dimension of impulsivity, associated with motivational aspects such as novelty seeking, excitement, and arousal, and it operates quite independently of other traits (Billieux et al., 2012). Measurement invariance has been established across various demographic characteristics, including age and gender identities, in different countries and languages (Donati et al. 2021; Fournier et al., 2025; Pechorro et al., 2021; Wang et al., 2020). S-UPPS-P scores have demonstrated poor-acceptable to good internal consistency reliability across diverse languages, with coefficients ranging from .53 to .87 (Donati et al., 2021; Eray et al., 2023; Pechorro et al., 2021; Wang et al., 2020). Regarding convergent validity, low to moderate but statistically significant correlations have been reported between S-UPPS-P Negative and Positive Urgency and Lack of Premeditation and the Barratt Impulsiveness Scale (BIS) scores (Eray et al., 2023).

When comparing scale scores by gender, most studies involving adolescents have found no significant differences, although males tend to score slightly higher than females on the Sensation Seeking subscale (Wang et al., 2020). In terms of age, findings have been more heterogeneous in youth (Sonmez et al., 2024). For instance, in adolescents, Wang et al., (2020) identified differences across all subscale scores except Sensation Seeking. However, other authors have reported no significant differences based on age (Donati et al., 2021; Montasell-Jordana et al., 2025).

Although the shortened UPPS-P (S-UPPS-P) has been translated into many languages, adapted, and validated for use in adolescents (Donati et al., 2021; Eray et al., 2023; Pechorro et al., 2021; Wang et al., 2020), to our knowledge, it has been evaluated in adults (Candido et al., 2012), but no study has yet evaluated the psychometric properties of the S-UPPS-P for adolescents in Spanish. This study aimed to fill this gap by pursuing three specific objectives, both in a community and a clinical sample: a) to test the factor structure, measurement invariance across gender, age, and sample type, and internal consistency of the S-UPPS-P derived scale scores; b) to examine its convergent validity with an alternative self-reported measure of impulsivity (BIS-11-A); and c) to explore the relationship between S-UPPS-P scores and participant characteristics, specifically gender, age, and sample type, and accordingly, to provide normative data for the Spanish adolescent population. Based on previous findings of internal structure, we expect to obtain the best fit for the 5-factor model. We hypothesize a low correlation for Sensation Seeking and a medium correlation for the other S-UPPS-P scale scores with the total BIS-11-A score. We do not expect to find differences in S-UPPS-P scale scores based on age, gender or sample type due to the variety of results of the previous validation studies available.

Method

Participants

In this study, we utilized both a community and a clinical subsample to evaluate the psychometric properties of the Spanish S-UPPS-P scale for adolescents. Participants for the community subsample were recruited using a multi-stage cluster sampling from schools located throughout the territory of Catalonia, Spain. The database of the Department of Education of the Generalitat de Catalunya (Government of Catalonia, 2022a, 2022b) was used to

select schools of different types (private, public and subsidized), as well as different academic courses. Additionally, demographic information regarding population density and family income levels was obtained from the Institut d'Estadística de Catalunya (IDESCAT, 2022a, 2022b) to guide the clustering of the selected schools. A total of 66 secondary schools were randomly selected and considered for the study during the academic year 2021-2022. The inclusion criterion for participants enrolled in these schools was being aged between 11 and 19. Students were excluded if they were in special education or adapted courses, or if they had an insufficient level of reading comprehension in Spanish. For the clinical subsample, a convenience sampling method was used to enroll consecutively admitted inpatients receiving psychotherapy and individuals undergoing day hospital treatment from seven hospitals and day clinics within the (blind) network. These centers provide treatment for people with various mental disorders referred from the main public hospitals. The inclusion criterion for the clinical subsample was the same as those for the community one, with participants aged between 11 and 19. Patients were excluded if they had an IQ below 80 (as assessed by the Wechsler Intelligence Scale [WISC-V]) or the Wechsler Adult Intelligence Scale [WAIS-IV] following the internal protocol of the clinical centers) or if they had an inadequate level of reading comprehension in Spanish.

The initial sample comprised 9929 participants (9024 from community and 905 from clinical settings) who agreed to take part in the study. Data of participants who omitted information or left the administration blank during the data collection process (n = 64), those who fell outside the specified age range (n =108), and those who did not complete the tests (n =24) were excluded, resulting in a final sample of 9733 participants (8944 for the community subsample and 789 for the clinical subsample). Students self-reported socio-demographic information in an *ad hoc* survey, which also included questions on possible mental health disorders. Participants were asked to indicate any diagnoses provided by mental health professionals, referencing a detailed list of specific disorders, with an open-ended option for unlisted diagnoses. In fewer than 10% of schools, psychological disorders were identified by the school's psychological services following survey administration. For the clinical subsample, psychological disorders were diagnosed collaboratively by the Neuropsychology department and the Psychiatry department of Ita Salud Mental using the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition [DSM-V] (American Psychiatric Association, 2013) or the International Classification of Diseases (10th or 11th version) [ICD] (World Health Organization (WHO), 1992; 2024). Sample characteristics are displayed in Table 1.

Instruments

Sociodemographic data. This data was collected *ad-hoc* to characterize the sample, including different variables as place of birth, type of school, disorders self- reported, if they had siblings, or had repeated any course and socioeconomic status of the student population.

S-UPPS–P Impulsivity scale (Cyders et al., 2014; Verdejo-García et al., 2010). This self-report questionnaire consists of 20 items and aims to assess five distinct personality pathways to impulsive behavior: Negative Urgency (e.g., "When I feel rejected, I will often say things that I later regret"), Lack of Perseverance (e.g., "Once I get going on something I hate to stop"), Lack of Premeditation (e.g., "I like to stop and think things over before I do them"), Sensation Seeking (e.g., "I quite enjoy taking risks"), and Positive Urgency (e.g., "I tend to lose control when I am in a great mood"). Each item is rated on a 4-point Likert-type scale ranging from 1 (*strongly agree*) to 4 (*strongly disagree*). Subscale scores are calculated by summing the item responses (reversed when necessary) with higher scores indicating higher levels of each trait. Verdejo-García et al. (2010) used a college sample from Granada (Spain) exclusively to validate the long version of the UPPS-P in young adults.

BIS-11-A (Barratt Impulsiveness Scale, version 11 for Adolescents; Fossati et al., 2002). We used the Spanish version adapted for adolescents by Martínez-Loredo et al. (2015) to evaluate convergent validity, since this test is the most widely used psychometric instrument in the field of impulsivity. The BIS-11-A comprises 30 items measuring motor, unplanned, and attentional aspects of impulsivity. Each item in BIS-11-A presents a statement describing impulsivity-related thoughts or behaviors in different situations. The items are scored on a 4-point Likert-type scale, ranging from 1 (*rarely/never*) to 4 (*almost always/always*). The total score is obtained by summing the item responses, with items reversed when necessary, with higher scores indicating higher levels of impulsivity. In our sample, we found good internal consistency reliability, with an omega coefficient of .84. BIS-11-A views impulsivity as a more global, unidimensional construct involving motor, attentional, and planning-related aspects.

Procedure

The procedure received approval from the ethics committee of (CEEAH nº 6494) and also authorization from the Department of Education of the Government of Catalonia (Spain) for recruiting centers (Register: nº: 9067/490777/2021).

For the community subsample, initial contact was established with the school principals, who were provided with an overview of the research goals and a request for cooperation. Upon agreement to participate, each institution's administration reviewed and approved the detailed study protocol. An information sheet outlining the study was given to each participating institution, along with a video document explaining the study's characteristics, objectives, and guidelines for parental communication. A 2-week notice period was provided to parents, during which they could opt their minor children out of the study. The self-reported questionnaires and an *ad-hoc* survey for socio-demographics and mental health problems were administered collectively during a 1-hour academic session. A teacher assisted with the administration, and the first-author was present to oversee the process. The questionnaires were administered using an online platform to facilitate data collection. All students received an information sheet confirming that their data would be treated confidentially and used solely at the group level. In four centers, the *ad-hoc* socio-demographic survey did not include the section on mental health problems, and diagnoses of mental disorders were reported directly to the first-author by the school services, following their internal data protection protocols.

For the clinical subsample, an information sheet was provided to the centers with a document explaining the characteristics, objectives and procedures for subsequent data handling. Parental consent for minors (< 18) was obtained by email and also collected during

Table 1
Sociodemographic Characteristics of the Final Sample (N = 9733)

		Community (n = 8944)	Clinical (n = 789)
Age; *M* (*SD*)	(Years)	14.7 (1.5)	16.3 (1.7)
Gender; *n* (%)	Male	4376 (48.9%)	168 (21.3%)
	Female	4417 (49.4%)	610 (77.3%)
	Non-binary	151 (1.7%)	7 (0.9%)
	Not-reported	0 (0.0%)	4 (0.5%)
Place of birth; *n* (%)	Spain	8274 (92.5%)	668 (84.6%)
	Other European countries	163 (1.8%)	59 (7.5%)
	Outside Europe	507 (5.7%)	62 (7.9%)
Siblings; *n* (%)	Yes	7520 (84.1%)	663 (84.0%)
Socio-economic status [a]; *n* (%)	Low	1021 (11.5%)	[b]
	Medium-low	3392 (37.9%)	[b]
	Medium	1808 (20.2%)	[b]
	Medium-high	1471 (16.4%)	[b]
	High	1252 (14.0%)	[b]
Current education level; *n* (%)	Primary	0 (0.0%)	3 (0.3%)
	Mandatory secondary high school (ESO)	7529 (84.1%)	461 [c] (58.6%)
	Post obligatory High School pre university studies (ESPO)	1066 (12.0%)	197 (25.0%)
	Post obligatory basic professional education (PFI/FPB)	25 (0.3%)	5 (0.6%)
	Post obligatory formation for middle and superior grades (CFGM/CFGS)	324 (3.6%)	75 (9.5%)
	University	0 (0.0%)	47 (6.0%)
Type of school; *n* (%)	Public	3857 (43.1%)	[b]
	Subsidized	5004 (56%)	[b]
	Private	83 (0.9%)	[b]
Repetition course; *n* (%)	Yes	717 (8.0%)	179 (22.7%)
Disorder	Without disorder	7033 (78.6%)	0 (0.0%)
	Attention deficit hyperactivity disorder	491 (5.5%)	88 (11.1%)
	Language/learning impairment	468 (5.2%)	0 (0.0%)
	Anxiety	406 (4.6%)	47 (6.0%)
	Eating disorders	189 (2.1%)	240 (30.4%)
	Autism spectrum disorders	157 (1.8%)	97 (12.3%)
	Depression/mood disorder	151 (1.7%)	78 (9.9%)
	Borderline personality disorder	3 (0.0%)	77 (9.8%)
	Substance use disorder	0 (0.0%)	60 (7.6%)
	Posttraumatic stress disorder	2 (0.0%)	58 (7.3%)
	Fetal alcohol spectrum disorders	1 (0.0%)	42 (5.3%)
	Other	43 (0.5%)	2 (0.3%)
Treatment	Inpatients	NA	515 (65.3%)
	Day hospital	NA	274 (34.7%)

[a] based on IDESCAT database https://www.idescat.cat/pub/?id=rfdbc. [b] detail not available. [c] Each of the univariate descriptive analyses was performed using list-wise deletion. NA = Not Applicable. *Note.* Language/learning impairment include Developmental Oral Language disorder, Dyslexia, Dyscalculia, Dysorthographia; Eating Disorders include Anorexia I or II, Bulimia or Binge Disorder.

monthly parents' group meetings at each clinical center by the first-author. A 2-week notice period was given to parents, during which they could opt their minor children out of the study. The self-reported questionnaires and the *ad-hoc* survey for socio-demographics were administered collectively during a 1-hour group therapy session with the assistance of a psychologist or individually with any of the research authors who were clinicians. The questionnaires were administered in paper-and-pencil format. All patients received and signed an information sheet assuring them that their data would be treated confidentially and only be used at the group level.

Data Analysis

We conducted the analyses using SPSS 29 and MPlus 8.9 programs. Internal structure of S-UPPS-P items was analyzed with weighted least squares means and variance adjusted (WLSMV) estimator and, when applicable, theta parameterization. First, three models were analyzed with confirmatory factor analysis (CFA) to test whether a single-factor model (Model A1: all items loading on a single factor), a 5-factor model (Model A2: items loading on the expected five intercorrelated factors), or a 3-factor model (Model A3: three intercorrelated factors -broad urgency, lack of conscientiousness, and sensation-seeking-) showed the best fit to the data, following previous research on the S-UPPS-P items. Second, a cross-validation design was used to determine the dimensionality from a more non-restricted (or "exploratory") approach. This was done by splitting the sample randomly into two subsamples of approximately the same size. In the first subsample, exploratory factor analysis (EFA) with the extraction of 1 to 5 factors was conducted, with geomin rotation for multidimensional solutions

(Models B#). For determining the number of factors to retain, we relied on eigenvalues and Cattell's scree test, since parallel analysis is not available in Mplus for categorical indicators. Acceptable salient loadings were considered above .35. In the second subsample, exploratory structural equation modeling (ESEM; Asparouhov & Muthén, 2009) with target rotation was conducted to test if the best EFA solution could be replicated (Model C). ESEM is considered a more flexible approach than CFA because, with target rotation, ESEM estimates the factor loadings of all items on all factors while constraining non-target loadings to be as close to zero as possible. In contrast, CFA restricts each item to load solely onto its intended factor, with all cross-loadings on non-intended factors fixed at zero. For all the factor analyses aforementioned, the common fit indices were used to assess goodness-of-fit (Jackson et al., 2009): Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA). The following thresholds were applied (Brown, 2015): an excellent fit was defined as CFI and TLI > .95 and RMSEA < .05, while a moderate fit was considered for CFI and TLI > .90 and RMSEA < .08. And third, for the best-fitting model, measurement invariance of ESEM across gender, age, and sample type (community without disorder, community with any disorder, and clinical) was tested (Models D#, E#, and F#), following the standard sequence (e.g., Vandenberg & Lance, 2000). The process involved testing four models across each group of responses, the last three models each nested in the previous one: configural (resulting ESEM taken as baseline model, with all parameters free across groups except those for model identification), metric or weak invariance (fixing factor loading to be equal), scalar or strong invariance (fixing also item thresholds to be equal), and strict invariance (fixing also uniquenesses to be equal). The factor variance strategy was used for model identification (for detailed steps, see Ezpeleta & Penelo, 2015). Because group sizes were unequal, specific criteria were used to indicate a meaningful worsening of fit and, consequently, non-invariance when comparing nested models: decrease in CFI > .004 and increase in RMSEA > .02 (Chen, 2007). In other words, evidence for the more parsimonious model and, therefore, support for invariance at each step was considered if CFI and RMSEA were as good as or better than for the more complex model (i.e., less constrained): an increase in CFI or a decrease of up to .004 (change up to −.004), and a decrease in RMSEA or an increase of up to .02.

Omega coefficient (McDonald, 1999) was used for evaluating internal consistency reliability of the S-UPPS-P scale scores. The convergent validity with BIS-11-A impulsivity total score was assessed with Pearson's correlation coefficients.

Finally, differences across gender and age (2-factor mixed), and among sample type (1-way) were evaluated with analysis of variance (ANOVA) to establish the need for separate normative data by groups. To define the age stages, we based our categorization on WHO guidelines (2024), Specifically, early adolescence includes ages approximately 10 to 13, middle adolescence from 14 to 16, and late adolescence from 17 to 20, an age range that aligns with our sample distribution. Three criteria were combined to determine the relevant differences of these variables on raw scores. As the main criterion, η^2 effect-size was used applying Cohen's rules of thumb of 0.01 for small, 0.06 medium and 0.14 for large effect (Cohen, 1988). In addition, the following information was considered. The standard error of measurement (SE*m*) was obtained based on the reliability coefficient and standard deviation of the raw scores, and then the 95% CI or range of true scores around the SE*m* values was derived. Lastly, a difference greater than 5-6 points on raw subscale scores was considered as an indicator of practical importance. Normative data for each subscale score were then calculated on the relevant normative reference groups, using T-scores and percentile ranks.

Results

Missing responses for the 20 S-UPPS-P items were very low (Graham, 2009): 0.01%; only 10 participants (0.10%) exhibited missing values for one or more items. Item mean (and standard deviation) values ranged from 1.65 to 2.97 (0.73-1.19). Median (in absolute value) of skewness was 0.35 and kurtosis was 0.81. None of the items showed floor or ceiling effects.

Goodness-of-fit indexes for CFA were insufficient both for the 1-factor and 3-factor models (Table S1, Models A1 and A3: CFI and TLI ≤ .803; RMSEA ≥ .097), and better but not acceptable enough for the 5-factor model (Table S1, Model A2: CFI = .899; TLI = .880; RMSEA = .076, 90% CI [.075, .077]). Moving to an exploratory approach, and regarding EFA in the first subsample of the cross-validation design, the first four observed eingenvalues were above 1 (5.37-1.13), the fifth very little below (0.98), and from the sixth all were clearly below 1 (≤ 0.76). Cattell's scree test also suggested the extraction of three or five factors, the profile clearly flattening from the sixth factor onwards. The 5-factor solution with EFA showed the best fit (Table S1, Model B5: CFI = .985; TLI = .972; RMSEA ≤ 0.36, 90% CI [.034, .039]) and also showed the simplest and most interpretable loading structure (Table 2, left). Fit for this model (consisting of 20 items and five correlated factors) with ESEM in the second subsample was also satisfactory (Table S2 from supplementary material, Model C: CFI = .987, TLI = .975, RMSEA = .035, 90% CI [.032, .037]), and results for factor loadings and factor correlations were very similar (Table 2, right). Both with EFA/ESEM, the pattern of salient factor loadings of S-UPPS-P was as expected: all the items showed the highest factor loading on their intended factor, with values above .35 (all ≥ .41/.45); factor loadings on non-intented factors were all below .20, except for two/one items (.26-.27/.20, which could explain the poor fit, buy by very little, of the 5-factor model when analyzed with CFA). The expected pattern of factor correlations was also observed: .64/.69 between Urgency factors, .46/46 between Lack of Premeditation and Lack of Perseverance, and lower values involving Sensation Seeking (.10-.43/.12-.39 in absolute value).

Subsequently, the ESEM model was used as the baseline configural model for the tests of equivalence of factor loadings (weak or metric invariance), item thresholds (strong or scalar invariance), and item uniquenesses (strict invariance) across gender, age and sample type. Full weak, strong and strict measurement invariance was supported across all types of groups (CFI increased or at most decreased ≤ .004; RMSEA decreased or at most increased ≤ .002). These findings support the cross-group comparability of S-UPPP-P across gender, age and sample types (Table S1, Models D#, E#, and F#).

Table 2
Cross-validation Exploratory Factor Analysis (Standardized Parameters) for S-UPPS-P and Omega Coefficient

		EFA with geomin rotation (n = 4860)					ESEM with target rotation (n = 4873)				
Factor loadings [a]	Item (original numeration)	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
Negative urgency	6. *When I feel bad, I will…	**.56**	.17	−.04	.06	.03	**.59**	.18	−.11	.10	.02
	8. *Sometimes when I feel bad…	**.74**	.00	−.01	.03	.01	**.82**	−.07	−.12	.06	.03
	13. *When I am upset I often…	**.59**	.02	.19	−.08	.05	**.61**	−.03	.19	−.08	.05
	15. *When I feel rejected …	**.48**	.13	.09	−.08	−.03	**.51**	.10	.06	−.09	−.06
Positive urgency	3.* When I am in great mood …	.04	**.68**	.05	.05	.04	.07	**.64**	.04	.04	.03
	10.* I tend to lose control…	.03	**.80**	−.03	−.03	−.04	.05	**.76**	.00	−.04	−.03
	17. *Others are shocked…	−.04	**.69**	.00	.03	.02	−.03	**.71**	−.03	.03	.02
	20. *I tend to act without thinking…	.02	**.74**	.07	−.03	.02	.01	**.77**	.06	−.03	.03
Lack of premeditation	2. My thinking is usually careful…	.05	.04	**.41**	.26	.00	−.01	.01	**.45**	.20	.03
	5. I like to stop and think…	−.01	.01	**.71**	.06	.05	.01	.00	**.72**	.02	.07
	12. I tend to value and…	.03	.03	**.48**	.18	−.07	−.03	.07	**.58**	.10	−.10
	19. I usually think carefully…	−.01	−.01	**.83**	−.03	.01	−.01	.02	**.83**	−.07	.02
Lack of perseverance	1. I generally like to see things…	.00	.03	−.08	**.68**	.01	.04	.05	−.02	**.69**	.01
	4. Unfinished tasks...	−.12	−.01	.05	**.62**	.06	−.09	.05	.01	**.64**	.08
	7. Once I get going on something…	−.12	−.19	.02	**.46**	−.08	−.14	−.13	.00	**.48**	−.07
	11. I finish what I start.	.16	.00	.05	**.69**	−.01	.18	−.02	.16	**.59**	−.06
Sensation seeking	9. *I quite enjoy taking...	.09	.11	.03	.13	**.64**	.09	.10	.11	.01	**.61**
	14. *I welcome new and exciting…	.11	−.04	−.02	−.01	**.73**	.12	−.05	.03	−.06	**.68**
	16. *I would like to learn to fly...	−.27	.10	−.02	−.02	**.56**	−.18	.07	−.10	.02	**.58**
	18. *I would enjoy the sensation…	−.08	−.05	.04	−.04	**.70**	−.02	−.07	−.03	.02	**.77**
Factor correlations [b] and omega [c]											
	F1 (Negative urgency)	*.74*					*.74*				
	F2 (Positive urgency)	.64	*.82*				.69	*.82*			
	F3 (Lack of premeditation)	.33	.43	*.76*			.35	.44	*.75*		
	F4 (Lack of perseverance)	−.02	.13	.46	*.67*		−.05	.11	.46	*.68*	
	F5 (Sensation seeking)	.30	.43	.24	−.10	*.73*	.28	.39	.18	−.12	*.73*

* Inverse items reversed prior to analysis.
[a] In bold: Salient factor loading above ≥.35. Shaded cells indicate the factor in which the item was assigned, taken into account the content.
[b] For factor correlations: all p-values < .05
[c] In italics: internal consistency reliability (omega coefficient)

Internal consistency reliability ranged from moderate (.67-.68 for Lack of Perseverance) to excellent values (.82 for Positive Urgency) (Table 2, bottom). In terms of convergent validity with the BIS-11-A, the total score correlated highly-moderately with the theoretically most closely related S-UPPS-P subscale scores: .47 with Negative Urgency, .51 with Positive Urgency, and .59 with Lack of Premeditation. Lower correlations were observed for Lack of Perseverance (.27) and Sensation Seeking (.22).

Results from the 3 × 3 two-way ANOVA (gender [females, males, and non-binary] × age [11-13, 14-16, and 17-19 years] (Table S2 from supplementary material, top) and from the one-way ANOVA (sample type [community sample, clinical sample]) (Table S2 from supplementary material, bottom) for S-UPPS-P scores showed very small or null effects for all terms, including interaction for the former (all $\eta^2 \leq 0.033$). In addition, the 95% CI for range of "true" scores based on *SEm* was wider than the range between the highest and the lowest observed group mean (for cells with $n > 30$), which in turn did not exceed the threshold of 5-6 points considered as a cut point of practical importance for the raw scores. Taken as a whole, differences among gender, age and sample type were considered negligible. Therefore, we calculated norms based exclusively on the total sample for each derived scale score. T-scores and percentile ranks are provided in Table 3.

Table 3
Normative Data for the Spanish Adolescent S-UPPS-P (N = 9733)

	NeUr		PoUr		LPrm		LPrs		SeSe	
Score	T	Pc	T	Pc	T	Pc	T	Pc	T	Pc
4	30	2	37	8	31	2	32	3	28	1
5	33	5	40	20	35	8	37	10	32	4
6	36	10	43	30	39	15	41	20	35	9
7	40	17	47	41	44	26	46	35	38	14
8	43	26	50	53	48	42	50	52	42	22
9	47	38	53	65	52	59	55	69	45	31
10	50	50	57	74	56	73	59	82	48	41
11	53	62	60	82	60	84	64	91	51	53
12	57	73	64	89	64	91	68	96	55	65
13	60	83	67	94	68	96	73	98	58	76
14	64	90	70	97	72	98	77	99	61	85
15	67	95	74	99	77	99	82	99	64	92
16	70	98	77	99	81	99	86	99	68	98
M	10.02		7.98		8.57		7.96		10.59	
SD	2.95		2.95		2.43		2.22		3.06	
SEm	1.5		1.3		1.2		1.3		1.6	

Note. T: T-score; Pc: Percentile rank; NeUr: Negative urgency; PoUr: Positive urgency; LPrm: Lack of premeditation; LPrs: Lack of perseverance; SeSe: Sensation seeking. SEm: Standard Error of Measuremen

Discussion

The aim of the present study was to evaluate the psychometric properties of the S-UPPS-P scale and to provide normative data for the Spanish adolescent population. Overall, our results supported the expected internal structure, demonstrating equivalent across gender, age and sample type, along with acceptable internal consistency reliability. Regarding, convergent validity, the S-UPPS-P subscale scores showed moderate to high correlations with the total BIS total score, except for Lack of Perseverance. Furthermore, negligible or no differences were observed in raw scores across gender, age and sample type, allowing for the derivation of a single set of normative data for the entire sample.

The results obtained from the present adolescent sample supported the expected 5-factor internal structure of the S-UPPS–P items, consistent with the original UPPS-P model (Lynam et al., 2006; Whiteside & Lynam, 2001) and previous findings (Donati et al., 2021; Eray et al., 2023; Pechorro et al., 2021). Measurement invariance analyses provided key insights into the comparability of S-UPPS-P scale scores across gender, age, and sample type. Specifically, full measurement invariance was established across all groups, supporting the equivalence of factor loadings and thresholds, and allowing for meaningful group comparisons (Meredith, 1993). This finding partially aligns with previous research, which reported full measurement invariance across age and gender (Wang et al; 2020), only across gender (Donati et al., 2021; Fournier et al., 2024), or failed to achieve it (Pechorro et al., 2021). Our results suggest that the relationships between the items and their underlying latent constructs (e.g., impulsivity traits) are consistent across age, gender and sample types. To our knowledge, no previous study has examined the S-UPPS-P measurement invariance across clinical and community adolescent samples.

Regarding dimensionality, all items had a salient factor loading above .35 on their intended factor. Factor correlations ranged from moderate to strong, except for the value involving Sensation Seeking, which showed lower correlations, evidencing related but distinguishable factors, aligned with the theoretical model underlying the test, with varying magnitudes among different pairs of factors. In line with prior research, the strongest correlations were identified between dimensions more closely linked from a theoretical standpoint (Donati et al., 2021; Eray et al., 2023; Pechorro et al., 2021), such as Negative Urgency and Positive Urgency (Fisher-Fox et al., 2024), and Lack of Premeditation and Lack of Perseverance (Gomez & Watson, 2023). Predictably, Sensation Seeking showed low correlations with the other factors, supporting its distinct nature (Smith et al., 2007). The low correlation between Lack of Perseverance and both Urgency scale scores obtained as in previous research (Donati et al., 2021, Wang et al., 2020) may reflect two different processes as Lack of Perseverance and Lack of Premeditation (cognitive impulsivity) are linked to top-down processing, and Negative Urgency and Positive Urgency (emotional impulsivity) dimensions can be linked to bottom-up processing both linked as for example in Attention-Deficit/Hyperactivity Disorder [ADHD] (Gomez & Watson, 2023).

In relation to internal consistency, the subscale scores exhibited coefficient values ranging from moderate to excellent (ω between .67-.68 for Lack of Perseverance and .82 for Positive Urgency). Our findings align with those obtained in several validation studies of short versions in adolescent populations of the S-UPPS-P (Wang et al., 2020). Notably, Lack of Perseverance has presented smaller internal consistency coefficients than the other scale scores in other short studies in adolescents (Eray et al., 2023; Pechorro et al., 2021).

In assessing convergent validity, Lack of Premeditation showed a moderate to high correlation with the BIS-11-A global score, which was expected given that several items in the BIS-11-A specifically measure aspects of non-planning impulsiveness. The S-UPPS-P provides a more nuanced and clinically informative assessment between profiles or risk factors (e.g., emotional vs. cognitive impulsivity). However, contrary to our expectations Lack of Perseverance scores showed a low correlation with the BIS-11-A global score. Eray et al (2023) is the only study to report a low correlation between Lack of Perseverance and motor impulsivity subscale score of the BIS-11-A. A possible explanation for this low correlation may involve social desirability, as perseverance is seen as a valued trait in adolescence, potentially leading to biased responses (Carvalho et al., 2023; Holden & Passey, 2010; Schoenmakers et al., 2024; Wu, 2025). Additionally, the link between Perseverance and cognitive effort might partially account for this result. Given the ongoing development of executive functions during adolescence, adolescents may exhibit more variability in the capacity to sustain effort, which could attenuate its relationship with impulsivity measures such as the BIS-11-A (Fortgang & Cannon, 2022). Furthermore, the relatively lower reliability observed for the Lack of Perseverance subscale most likely has attenuated the convergent relation with the BIS scale score. Perseverance performance requires a certain level of sustained effort, but beyond a specific point increasing this effort does not lead to further improvement and instead remains constant. This nonlinear performance may partially explain the lower correlation value (Bandalos, 2018). This suggests that Perseverance may counterbalance the tendency to avoid cognitive effort, which often drives impulsivity, especially in adolescence (Patzelt et al., 2019). This is especially relevant in neurodevelopmental disorders like ADHD or fetal alcohol spectrum disorders (FASDs), where impairments in Perseverance contribute to risk-taking behaviors (Eray et al., 2023; Kingdon et al., 2016; Mattson et al., 2019; Miller et al., 2010; Wagner et al., 2024). Finally, Sensation Seeking exhibited a low correlation, as expected taking into account that the BIS-11-A does not specifically assess Sensation Seeking as one of its facets (Smith et al., 2007).

Since the observed differences by gender, age and sample type were considered negligible, normative data were ultimately calculated based on the total sample for each derived scale score. A single S-UPPS-P norm for adolescents aligns with previous studies reporting no literature of no gender differences (Montasell-Jordana et al., 2025; Fournier et al., 2025). This finding matches with the availability of undifferentiated norms by age and gender both in adults (Gialdi et al., 2021; Pinto et al., 2021) and in adolescents in Spain for the long version of the UPPS-P (Montasell-Jordana et al., 2025).

Several limitations must be recognized. We had no opportunity to corroborate self-reported diagnoses with professional clinical diagnoses for most of the community subsample. Resource constraints and limited access to comprehensive diagnostic data precluded the acquisition of accurate and valid diagnostic information from all participants through the psychological services of the schools. Moreover, additional resources for additional self-reported mental health or emotional well-being instruments and neuropsychological tasks to further explore

facets of impulsivity were also unavailable. The results from the non-binary group should be interpreted with caution due to the limited sample size, which may restrict the reliability of the estimates for this subgroup. Nonetheless, our results are based on a very large and diverse sample obtained through random stratification for the community sample, representing the adolescent population across family situations, school types, income levels and population densities. Furthermore, the prevalence of self-reported mental health problems in our samples closely aligns with findings from epidemiological studies of adolescents in Spain (Haro et al., 2006).

Despite these limitations, the present findings suggest that the scores for the Spanish version of the S-UPPS-P have acceptable validity and internal consistency in the adolescent population. Its reduced length may make it more suitable for clinical and survey administration among adolescents compared to longer versions (Omrani et al., 2019), which minimizes the time and effort required for respondents in this population (Fortgang & Cannon, 2022). Given that adolescence is a critical period for the emergence of impulsivity-related disorders, the S-UPPS-P may serve as a useful tool for early identification and risk assessment in clinical settings, and may complement screening efforts as well as therapeutic and clinical services (Fisher-Fox et al., 2024). Moreover, the normative data provided by this study offers a useful resource for future research. Given that adolescence is a critical period for the emergence of impulsivity-related disorders, the S-UPPS-P may serve as a useful tool for early identification and risk assessment in clinical settings (Um et al., 2018). Specifically, it can be incorporated into screening protocols in primary care or mental health services to detect adolescents exhibiting elevated levels of specific impulsivity traits (e.g. Negative Urgency or Lack of Premeditation) as for specific risk profiles for suicide behaviors (Lynam et al., 2011), which are associated with emotional dysregulation or externalizing behaviors.

Author Contributions

Esteve Montasell-Jordana: Conceptualization, Investigation, Validation, Supervision, Funding acquisition, Project administration, Resources, Writing – Original draft, Writing - Review & editing. **Eva Penelo:** Conceptualization, Methodology, Software, Data curation, Validation, Formal analysis, Supervision, Writing – Review & editing. **Laura Blanco-Hinojo**: Conceptualization, Data curation, Supervision, Visualization, Writing – Review & editing. **Beatriz Lanceta**: Investigation, Project administration, Resources. **Laura Gomàriz-Camacho**: Investigation, Project administration, Resources **Anna Soler**: Investigation, Project administration, Resources**. Jesús Pujol**: Conceptualization, Validation, Writing – Review & editing. **Joan Deus**: Conceptualization, Validation, Supervision, Project administration, Resources, Writing – Review & editing.

Acknowledgments

We thank all the participating high schools for their contribution to this study, as well as the families and patients from ITA Salud Mental.

Funding

This research was partly supported by grant (SGR 2021/00717) from the Agency of University and Research Funding Management (AGAUR) of the Catalonia Government. This funding source had no role in the design of this study, data collection, management, analysis, and interpretation of data, writing of the manuscript, and the decision to submit the manuscript for publication.

Declaration of Interests

The authors declare that there is no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon request. Supplementary material can be found here: https://figshare.com/s/29528f85a8c8ca5d4fff

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*(3), 397-438. https://doi.org/10.1080/10705510903008204

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences.* The Guilford Press

Billieux, J., Rochat, L., Ceschi, G., Carré, A., Offerlin-Meyer, I., Defeldre, A. C., Khazaal, Y., Besche-Richard, C., & Van der Linden, M. (2012). Validation of a short French version of the UPPS-P Impulsive Behavior Scale. *Comprehensive Psychiatry*, *53*(5), 609–615. https://doi.org/10.1016/j.comppsych.2011.09.001

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

Cándido, A., Orduna, E., Perales, J. C., Verdejo-García, A. J., & Billieux, J. (2012). Validation of a short Spanish version of the UPPS-P Impulsive Behaviour Scale. *Trastornos Adictivos*, *14*(3), 73-78. https://doi.org/10.1016/S1575-0973(12)70048-X

Caro-Cañizares, I., Brenlla, M. E., Carballo, J. J., Santos, M., Guija, J. A., & Giner, L. (2024). Suicidal behavior and stressful life events: The mediating role of the impulsivity-aggression-hostility triad through psychological autopsy. *Psicothema*, *36*(4), 380-388. https://doi.org/10.7334/psicothema2023.300

Carrera, P., Román, M., Cáceres, I., & Palacios, J. (2024). Internalizing problems in adopted Eastern European adolescents: The role of the informant, early adversity and post-adoption processes. *Psicothema*, *36*(2), 103-112. https://doi.org/10.7334/psicothema2023.152

Carvalho, C. B., Arroz, A. M., Martins, R., Costa, R., Cordeiro, F., & Cabral, J. M. (2023). "Help me control my impulses!": Adolescent impulsivity and its negative individual, family, peer, and community explanatory factors. *Journal of Youth and Adolescence*, *52*(12), 2545–2558. https://doi.org/10.1007/s10964-023-01837-z

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464–504. http://doi.org/10.1080/10705510701301834

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Elrbaum.

Costa, P. T., & McCrae, R. R. (1985). *The NEO Personality Inventory manual*. Psychological Assessment Resources.

Cyders, M. A., Smith, G. T., Spillane, N. S., Fischer, S., Annus, A. M., & Peterson, C. (2007). Integration of impulsivity and positive mood to predict risky behavior: Development and validation of a measure of positive urgency. *Psychological Assessment*, *19*(1), 107–118. https://doi.org/10.1037/1040-3590.19.1.107

Cyders, M. A., Littlefield, A. K., Coffey, S., & Karyadi, K. A. (2014). Examination of a short English version of the UPPS-P Impulsive Behavior Scale. *Addictive Behaviors*, *39*(9), 1372–1376. https://doi.org/10.1016/j.addbeh.2014.02.013

Donati, M. A., Beccari, C., Bacherini, A., Capitanucci, D., & Primi, C. (2021). Psychometric properties of the short UPPS-P scale in adolescents: Gender, age invariance, and validity among italian youth. *Addictive Behaviors*, *120*, 106987. https://doi.org/10.1016/j.addbeh.2021.106987

Duell, N., & Steinberg, L. (2019). Positive risk taking in adolescence. *Child Development Perspectives, 13*(1), 48–52. https://doi.org/10.1111/cdep.12310

Eray, Ş., Sigirli, D., Yavuz, B. E., Şahin, V., Liu, M., & Cyders, M. A. (2023). Turkish adaptation and validation of the Short-UPPS-P in adolescents and examination of different facets of impulsivity in adolescents with ADHD. *Child Neuropsychology*, *29*(3), 503–519. https://doi.org/10.1080/09297049.2022.2100338

Ezpeleta, L., & Penelo, E. (2015). Measurement invariance of oppositional defiant disorder dimensions in 3-year-old preschoolers. *European Journal of Psychological Assessment, 31*(1), 45–53. http://doi.org/10.1027/1015-5759/a000205

Fisher-Fox, L., Whitener, M., Wu, W., Cyders, M. A., & Zapolski, T. C. B. (2024). Urgency as a predictor of change in emotion dysregulation in adolescents. *Frontiers in Psychiatry*, *15*, 1451192. https://doi.org/10.3389/fpsyt.2024.1451192

Fortgang, R. G., & Cannon, T. D. (2022). Cognitive effort and impulsivity. *Personality and Individual Differences, 199,* 1–9. https://doi.org/10.1016/j.paid.2022.111830

Fossati, A., Barratt, E. S., Acquarini, E., & Di Ceglie, A. (2002). Psychometric properties of an adolescent version of the Barratt Impulsiveness Scale-11 for a sample of Italian high school students. *Perceptual and Motor Skills*, *95*(2), 621–635. https://doi.org/10.2466/pms.2002.95.2.621

Fournier, L., Bőthe, B., Demetrovics, Z., Koós, M., Kraus, S. W., Nagy, L., Potenza, M. N., Ballester-Arnal, R., Batthyány, D., Bergeron, S., Briken, P., Burkauskas, J., Cárdenas-López, G., Carvalho, J., Castro-Calvo, J., Chen, L., Ciocca, G., Corazza, O., Csako, R. I., Fernandez, D. P., … Billieux, J. (2025). Evaluating the factor structure and measurement invariance of the 20-item short version of the UPPS-P Impulsive Behavior Scale across multiple countries, languages, and gender identities. *Assessment*, *32*(5), 635–653. https://doi.org/10.1177/10731911241259560

Gialdi, G., Somma, A., Borroni, S., & Fossati, A. (2021). Factor structure, measurement invariance across gender sub-groups, and normative data for the Italian translation of the UPPS-P Impulsive Behavior Scale in Italian community-dwelling adults. *Mediterranean Journal of Clinical Psychology, 9*(2), e3004. https://doi.org/10.13129/2282-1619/MJCP-3004

Gomez, R., & Watson, S. (2023). Associations of UPPS-P negative urgency and positive urgency with ADHD dimensions: Moderation by lack of premeditation and lack of perseverance in men and women. *Personality and Individual Differences*, *206*, 112125. https://doi.org/10.1016/j.paid.2023.112125

Government of Catalonia, Departament Educació, Serveis Educatius. (2022a). *Centres educatius de Catalunya* [Educational Centers in Catalonia]. https://educacio.gencat.cat/ca/serveis-tramits/directoris-centres/

Government of Catalonia, Departament Educació, Estadistica de l'ensenyament. (2022b). *Estadistica d'Inici de curs 2022-2023* [Statistics for the start of the 2022-2023 academic year]. https://educacio.gencat.cat/web/.content/home/departament/estadistiques/estadistiques-ensenyament/estadistica-inici-curs/estadistica-inici-curs.xlsx

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549–576. http://doi.org/10.1146/annurev.psych.58.110405.085530

Haro, J. M., Palacín, C., Vilagut, G., Martínez, M., Bernal, M., Luque, I., Codony, M., Dolz, M., Alonso, J., & Grupo ESEMeD-España (2006). Prevalence of mental disorders and associated factors: Results from the ESEMeD-Spain study. *Medicina Clinica*, *126*(12), 445–451. https://doi.org/10.1157/13086324

Holden, R. R., & Passey, J. (2010). Socially desirable responding in personality assessment: Not necessarily faking and not necessarily substance. *Personality and Individual Differences, 49*(5), 446–450. https://doi.org/10.1016/j.paid.2010.04.015

IDESCAT (Institut d'Estadistica de Catalunya). (2022a). *Densitat de població. Comarques i Aran, àmbits i províncies* [Population density. Counties and Aran, areas and provinces]. https://www.idescat.cat/indicadors/?id=aec&n=15227

IDESCAT (Institut d'Estadistica de Catalunya). (2022b) *Renda familiar disponible bruta territorial* [Gross territorial disposable household income]. https://www.idescat.cat/pub/?id=rfdbc

Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods, 14*(1), 6–23. http://doi.org/10.1037/a0014694

Kingdon, D., Cardoso, C., & McGrath, J. J. (2016). Research Review: Executive function deficits in fetal alcohol spectrum disorders and attention-deficit/hyperactivity disorder - a meta-analysis. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *57*(2), 116–131. https://doi.org/10.1111/jcpp.12451

Kulendran, M., Patel, K., Darzi, A., & Vlaev, I. (2016). Diagnostic validity of behavioural and psychometric impulsivity measures: An assessment in adolescent and adult populations. *Personality and Individual Differences, 90,* 347–352. https://doi.org/10.1016/j.paid.2015.11.026

Littlefield, A. K., Stevens, A. K., Ellingson, J. M., King, K. M., & Jackson, K. M. (2016). Changes in negative urgency, positive urgency, and sensation seeking across adolescence. *Personality and Individual Differences*, *90*, 332–337. https://doi.org/10.1016/j.paid.2015.11.024

Lynam, D. R., Miller, J. D., Miller, D. J., Bornovalova, M. A., & Lejuez, C. W. (2011). Testing the relations between impulsivity-related traits, suicidality, and nonsuicidal self-injury: A test of the incremental validity of the UPPS model. *Personality Disorders: Theory, Research, and Treatment, 2*(2), 151–160. https://doi.org/10.1037/a0019978

Lynam, D. R., Smith, G. T., Whiteside, S. P., & Cyders, M. A. (2006). *The UPPS–P: Assessing five personality pathways to impulsive behavior* (Tech. Rep.). IN: Purdue University.

Mallorquí-Bagué, N., Testa, G., Lozano-Madrid, M., Vintró-Alcaraz, C., Sánchez, I., Riesco, N., Granero, R., Perales, J. C., Navas, J. F., Megías-Robles, A., Martínez-Zalacaín, I., Veciana de las Heras, M., Jiménez-Murcia, S., & Fernández-Aranda, F. (2020). Emotional and non-emotional facets of impulsivity in eating disorders: From anorexia nervosa to bulimic spectrum disorders. *European Eating Disorders Review, 28*(4), 410–422. https://doi.org/10.1002/erv.2734

Martínez-Loredo, V., Fernández-Hermida, J. R., Fernández-Artamendi, S., Carballo, J. L., & García-Rodríguez, O. (2015). Spanish adaptation and validation of the Barratt Impulsiveness Scale for early adolescents (BIS-11-A). *International Journal of Clinical and Health Psychology, 15(3),* 274–282. https://doi.org/10.1016/j.ijchp.2015.07.002

Mattson, S. N., Bernes, G. A., & Doyle, L. R. (2019). Fetal alcohol spectrum disorders: A review of the neurobehavioral deficits associated with prenatal alcohol exposure. *Alcoholism: Clinical and Experimental Research, 43(6),* 1046–1062. https://doi.org/10.1111/acer.14040

McDonald, R. P. (1999). *Test theory: A unified treatment*. Erlbaum.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58(4),* 525–543. https://doi.org/10.1007/BF02294825

Miller, D. J., Derefinko, K. J., Lynam, D. R., Milich, R., & Fillmore, M. T. (2010). Impulsivity and Attention Deficit-Hyperactivity Disorder: Subtype classification using the UPPS Impulsive Behavior Scale. *Journal of Psychopathology and Behavioral Assessment, 32(3),* 323–332. https://doi.org/10.1007/s10862-009-9155-z

Montasell-Jordana, E., Penelo, E., Blanco-Hinojo, L., Pujol, J., Billieux, J., & Deus, J. (2025). Psychometric properties and normative data of the Spanish UPPS-P Impulsive Behavior Scale in adolescents. *Journal of Personality Assessment, 107(2),* 234–243. https://doi.org/10.1080/00223891.2024.2399184

Omrani, A., Wakefield-Scurr, J., Smith, J., & et al., (2019). Survey development for adolescents aged 11–16 years: A developmental science based guide. *Adolescent Research Review, 4(4),* 329–340. https://doi.org/10.1007/s40894-018-0089-0

Patzelt, E. H., Kool, W., Millner, A. J., & Gershman, S. J. (2019). The transdiagnostic structure of mental effort avoidance. *Scientific Reports, 9*, 1689. https://doi.org/10.1038/s41598-018-37802-1

Pechorro, P., Revilla, R., Palma, V. H., Nunes, C., Martins, C., & Cyders, M. A. (2021). Examination of the SUPPS-P Impulsive Behavior Scale among male and female youth: Psychometrics and invariance. *Children, 8*(4), 283. https://doi.org/10.3390/children8040283

Pinto, A. L. de C. B., Garcia, M. S., Godoy, V. P., Loureiro, F. F., da Silva, A. G., & Malloy-Diniz, L. F. (2021). Psychometric properties and normative data of the Brazilian version of UPPS-P Impulsive Behavior Scale. *Current Research in Behavioral Sciences, 2*, 100052. https://doi.org/10.1016/j.crbeha.2021.100052

Potenza, M. N. (2007). To do or not to do? The complexities of addiction, motivation, self-control, and impulsivity. *The American Journal of Psychiatry, 164*(1), 4–6. https://doi.org/10.1176/ajp.2007.164.1.4

Rammstedt, B., & Beierlein, C. (2014). Can't we make it any shorter? The limits of personality assessment and ways to overcome them. *Journal of Individual Differences, 35*(4), 212–220. https://doi.org/10.1027/1614-0001/a000141

Schoenmakers, M., Tijmstra, J., Vermunt, J., & Bolsinova, M. (2024). Correcting for extreme response style: Model choice matters. *Educational and Psychological Measurement, 84*(1), 145–170. https://doi.org/10.1177/00131644231155838

Smith, G. T., Fischer, S., Cyders, M. A., Annus, A. M., Spillane, N. S., & McCarthy, D. M. (2007). On the validity and utility of discriminating among impulsivity-like traits. *Assessment, 14*(2), 155–170. https://doi.org/10.1177/1073191106295527

Sonmez, A. I., Garcia, J. Q., Thitiseranee, L., Blacker, C. J., & Lewis, C. P. (2024). Scoping review: Transdiagnostic measurement of impulsivity domains in youth using the UPPS Impulsive Behavior Scales. *Journal of the American Academy of Child and Adolescent Psychiatry, 63*(8), 789–812. https://doi.org/10.1016/j.jaac.2024.03.011

Um, M., Hershberger, A. R., Whitt, Z. T., & Cyders, M. A. (2018). Recommendations for applying a multi-dimensional model of impulsive personality to diagnosis and treatment. *Borderline Personality Disorder and Emotion Dysregulation, 5*, 6. https://doi.org/10.1186/s40479-018-0084-x

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

Verdejo-García, A., Lozano, O., Moya, M., Alcázar, M. A., & Pérez-García, M. (2010). Psychometric properties of a Spanish version of the UPPS-P Impulsive Behavior Scale: reliability, validity and association with trait and cognitive impulsivity. *Journal of Personality Assessment*, *92*(1), 70–77. https://doi.org/10.1080/00223890903382369

Wagner, D., Mason, S. G., & Eastwood, J. D. (2024). The experience of effort in ADHD: A scoping review. *Frontiers in Psychology*, *15*, 1349440. https://doi.org/10.3389/fpsyg.2024.1349440

Wang, Y., Long, J., Liu, Y., Liu, T., & Billieux, J. (2020). Psychometric properties of the Chinese SUPPS-P Impulsive Behavior Scale: Factor structure and measurement invariance across gender and age. *Frontiers in Psychiatry*, *11*, 529949. https://doi.org/10.3389/fpsyt.2020.529949

Whiteside, S. P., & Lynam, D. R. (2001). The Five Factor Model and impulsivity: Using a structural model of personality to understand impulsivity. *Personality and Individual Differences, 30*(4), 669–689. https://doi.org/10.1016/S0191-8869(00)00064-7

World Health Organization. (1992). *The ICD-10 classification of mental and behavioural disorders.* World Health Organization.

World Health Organization. (2024). *Clinical descriptions and diagnostic requirements for ICD-11 mental, behavioural and neurodevelopmental disorders.* World Health Organization.

Wu, P.-C. (2025). Stability of response styles in a personality measure: Evidence with clinical adolescents. *Counselling and Psychotherapy Research, 25*, 12859. https://doi.org/10.1002/capr.12859

GRUPO EDITORIAL DE PSICOFUNDACIÓN

Psicothema (2026) 38(1) 46-57

Psicothema

https://www.psicothema.com/es • ISSN 0214–9915 • eISSN 1886-144X

Colegio Oficial de Psicología del Principado de Asturias

Psicothema

Article

Psychometric Properties of the Teachers' Responses to Bullying Questionnaire (TRBQ) in Spanish Students

Laura Rodríguez-Pérez[1], Rosario Del Rey[1], Noemí García-Sanjuán[2] and Noelia Muñoz-Fernández[1]

[1] Universidad de Sevilla (Spain)
[2] Universidad Internacional de La Rioja (UNIR) (Spain)

ARTICLE INFO

Received: 05/05/2025
Accepted: 05/08/2025

Keywords:
Bullying
Teacher response
TRBQ
Measurement invariance

ABSTRACT

Background: Students' perceptions of teacher response play a critical role in addressing bullying, as they are closely linked to student involvement. However, no validated instruments currently exist in Spain to assess this construct adequately. This study aimed to validate the Teachers' Responses to Bullying Questionnaire (TRBQ) in Spain, examine its measurement invariance across educational levels, gender, and bullying roles, and to explore students' perceptions of teacher responses based on these variables. **Method:** A total of 1,241 students (48.8% girls; 48.3 % primary school; M_{age} = 12.00; *SD* = 1.79; range = 9–18 years) from southern Spain participated. **Results:** EFA revealed a three-factor structure—non-intervention, restorative psychoeducational strategies, and disciplinary methods—with good fit, confirmed through CFA. The instrument demonstrated satisfactory reliability and measurement invariance. Girls perceived teacher responses as more frequent. Restorative strategies were more common in primary school, while non-intervention was more prevalent in secondary school. No significant differences emerged for disciplinary methods. Non-involved students reported more restorative interventions, bullies-victims perceived more non-intervention; and aggressors reported greater use of disciplinary methods. **Conclusions:** The Spanish adaptation and validation of the TRBQ provides a valuable tool for assessing teacher responses to bullying and contributes to research and intervention in school contexts.

Propiedades Psicométricas del Teachers' Responses to Bullying Questionnaire (TRBQ) en Estudiantes Españoles

Palabras clave:
Acoso escolar
Respuesta del profesorado
TRBQ
Invarianza métrica

RESUMEN

Antecedentes: La percepción del alumnado sobre la respuesta del profesorado desempeña un papel fundamental en el acoso escolar, ya que se relaciona estrechamente con su implicación en el fenómeno. Sin embargo, en España no existen instrumentos validados que evalúen adecuadamente este constructo. Este estudio pretende validar el Teachers' Responses to Bullying Questionnaire (TRBQ) en España, examinar su invarianza métrica por nivel educativo, género y rol de implicación, y describir la respuesta del profesorado percibida en función de estas variables. **Método:** Participaron 1,241 estudiantes españoles (48.8% chicas; 48.3% de primaria; M_{edad} = 12.00; *DT* = 1.79; rango = 9-18 años). **Resultados**: El AFE reveló una estructura trifactorial—no intervención, estrategias psicoeducativas restaurativas y métodos disciplinarios—con un ajuste adecuado, confirmado por el AFC. El instrumento mostró una fiabilidad adecuada e invarianza métrica. Las chicas percibieron la intervención del profesorado como más frecuente. Las estrategias restaurativas fueron mayores en primaria, la no intervención en secundaria. El alumnado no implicado informó de más intervenciones restaurativas; los agresores-víctimas reportaron mayor no intervención; y los agresores mayor uso de métodos disciplinarios. **Conclusiones:** La adaptación española y validación del TRBQ representa una valiosa herramienta para evaluar la respuesta del profesorado al acoso escolar.

Cite as: Rodríguez-Pérez, L., Del Rey, R., García-Sanjuán, N., & Muñoz-Fernández, N.. (2026) Psychometric properties of the Teachers´ Responses to Bullying Questionnaire (TRBQ) in Spanish students. *Psicothema, 38*(1), 46-57. https://doi.org/10.70478/psicothema.2026.38.05
Corresponding author: Noelia Muñoz-Fernández, nmunoz2@us.es
This article is published under Creative Commons License 4.0 CC-BY-NC-ND

The school context is one of the main settings in which bullying occurs (Yoon et al., 2016), positioning teachers as key figures in detecting it and intervening. Teachers' responses to bullying have become an increasingly important area of study in recent years (Colpin et al., 2021; Demol et al., 2020, 2021). In many cases, teachers are the first adults that students turn to for help in a situation of victimization (Díaz-Aguado, 2023; Wachs et al., 2019). Thus, it is the responsibility of teachers, alongside other members of the educational community, to ensure that appropriate interventions are implemented. Despite its importance, a lack of consensus persists regarding how teacher responses to bullying should be conceptualized and measured (Colpin et al., 2021).

Teacher responses to bullying have been conceptualized in various ways. One of the most common distinctions is between active and passive responses: the former encompasses any strategy employed by the teacher to address the situation, while the latter refers to inaction or the lack of response (Demol et al., 2020; Song et al., 2018; Troop-Gordon & Ladd, 2015). Other scholars have distinguished between individual and group responses. Individual responses target the victim or aggressor directly—for instance, by offering support to the victim or applying disciplinary measures to the aggressor—whereas group responses engage the peer group or other adult figures through strategies such as group discussions or collaboration with external professionals (Troop-Gordon & Ladd, 2015; Wachs et al., 2019; Yoon et al., 2016). Finally, some scholars have explored the distinction between punitive and restorative approaches. Punitive responses include imposing sanctions on the aggressor or encouraging the victim to adopt a more assertive attitude, while restorative responses focus on repairing the harm by offering emotional support to the victim and encouraging the aggressor to acknowledge the impact of their behavior (Bauman et al., 2008; Burger et al., 2015; Kollerová et al., 2021; Rigby, 2014).

Research on how teachers respond to bullying also differs depending on the source of data. Early studies primarily relied on teacher self-reports, often assessing hypothetical responses or intention to intervene in bullying situations using vignettes or simulated scenarios (Burger et al., 2015; Chen, 2023; Collier et al., 2015; Dedousis-Wallace et al., 2013; Duong & Bradshaw, 2013). Actual responses were assessed to a much lesser extent (Troop-Gordon & Ladd, 2015). More recent research has examined teacher responses to bullying from the students' perspective (Colpin et al., 2021), providing insights into how such responses are perceived and interpreted by those directly affected (Demol et al., 2020). In contrast to teacher self-reports, research drawing on students' perceptions usually examines teachers' actual responses to bullying incidents (Denny et al., 2014; Berkowitz, 2013).

This latter approach is particularly valuable. Research has shown that teachers often overestimate the frequency of their response, either due to social desirability bias or because they may fail to recognize all instances of bullying (Campaert al., 2017; Yoon & Bauman, 2014). Moreover, the effectiveness of teacher responses may depend not only on the specific actions taken but also on how those responses are perceived by students (Devlesschouwer et al., 2025; Muñoz-Fernández et al., 2025a; Troop-Gordon et al., 2021a; Wachs et al., 2019).

In the international context, several instruments have been developed to assess teacher responses. One of the earliest is the Handling Bullying Questionnaire (HBQ; Bauman et al., 2008), which was designed to measure teachers' intended responses to hypothetical bullying scenarios. The original instrument includes 22 items and assesses five dimensions: working with the victim, working with the bully, ignoring the incident, enlisting other adults, and disciplining the bully (Bauman et al., 2008). The HBQ has been cross-culturally validated with factorial solutions identifying two (Grumm & Hein, 2012; Yoon et al., 2011), five (Burger et al., 2015), and six factors (Siddiqui et al., 2023), with moderate reliability reported across these studies. Despite its usefulness, HBQ lacks a student-report version and is limited to assessing hypothetical teacher responses.

To overcome the limitations associated with using hypothetical scenarios, Troop-Gordon and Ladd (2015) designed the Classroom Management Policies Questionnaire (CMPQ), a 56-item instrument that assesses strategies teachers use in real-life bullying situations. The CMPQ asks teachers to indicate which strategies they usually apply in their classroom practice with boys and girls, separately. The original CMPQ is organized into seven dimensions: contacting parents, separating students, punishing aggressors, suggesting avoidance, suggesting assertion, advising independent coping, and ignoring the incident. In the validation conducted by Troop-Gordon and Ladd (2015), the last two dimensions were merged, resulting in a six-factor solution with good reliability indices.

The Perceived Teacher Response Scale (PTRS; Troop-Gordon & Quenette, 2010)—a 24-item student version of the CMPQ that originally assesses six dimensions: contact parents, reprimand aggressors, advocate avoidance, advocate assertion, separate students, and advocate independent coping. Following a cross-validation process, the 'separate students' dimension was excluded from the final model. A recent analysis of the PTRS reintroduced the separate students' dimension but removed the punishment scale due to its low reliability (Troop-Gordon et al., 2021b). This suggests that the factorial structure of the PTRS may lack stability. Moreover, the instrument does not account for the use of victim support strategies, a response identified in the literature as one of the most effective and valued by students in bullying situations (Gregory et al., 2011; Van der Zanden et al., 2015; Zych et al., 2019).

While the CMPQ and PTRS are valid and reliable tools for assessing teacher responses to bullying, their use has been limited to specific cultural contexts and, to our knowledge, no cross-cultural adaptations have been reported to date. To address this gap, the Teachers' Response to Bullying Questionnaire (TRBQ) was developed and validated in various countries, including Italy, Belgium, the Philippines, and China, showing good psychometric properties (Campaert et al., 2017; Llego et al., 2024; Nappa et al., 2021; van Gils et al., 2022; Xiao & Hooi et al., 2024). These adaptations make the TRBQ particularly suitable for cross-cultural comparisons.

The TRBQ includes both a teacher self-report version (TRBQ-T; Muñoz-Fernández et al., 2025b) and a student-report version (TRBQ; Campaert et al., 2017; Nappa et al., 2021), allowing for meaningful comparisons across informants. In its original version, Campaert et al. (2017) assessed students' perceptions of teacher responses to bullying in primary school

settings, focusing on three domains: actions directed towards the bully, actions directed at the victim, and non-intervention. Strategies targeting the aggressor included group discussion, mediation, and disciplinary sanctions, while those aimed at the victim included victim support, mediation, and group discussion.

Subsequently, Nappa et al. (2021) developed a revised version of the TRBQ for use with secondary school students. This version streamlined the structure into three broader dimensions: non-intervention, disciplinary methods, and supportive/relational interventions—the latter encompassing group discussion, mediation, and victim support. Building on this work, van Gils et al. (2022) extended the empirical validation of the revised TRBQ with a sample of primary school students in Italy and Belgium. Through comparisons of different factor structures, their findings supported a five-factor structure—non-intervention, disciplinary methods, group discussion, mediation, and victim support—as the best-fitting solution.

These discrepancies in factorial solutions—such as the three-factor model proposed by Nappa et al. (2021) for secondary students and the five-factor model supported by van Gils et al. (2022) for primary students—may reflect both cultural differences (e.g., between Italy and Belgium) and developmental differences between student age groups. Moreover, the TRBQ has not yet been adapted or validated in Spain. Therefore, it is necessary to examine its structure in Spain, including both primary and secondary students, to advance the empirical evidence on the TRBQ.

Beyond examining the psychometric properties of instruments such as the TRBQ, it is also crucial to consider the student-level variables that may influence how teacher responses are perceived. While these instruments aim to capture general trends in students' views, individual characteristics and contextual factors could significantly shape these perceptions. Factors such as gender, educational level, and bullying involvement role may impact how students interpret teacher actions. Although empirical evidence on these moderate effects remains limited, prior studies suggest these variables are closely associated with students' involvement in bullying and may therefore also influence how they perceive adult responses.

Regarding educational level, some studies indicate a higher prevalence of bullying involvement among younger students, particularly in the final years of primary school (van Aalst et al., 2022; van der Zanden et al., 2015; van Gils et al., 2023). However, teacher responses are often perceived as more effective in primary school settings (Kärna et al., 2011). In contrast, older students appear more likely to report victimization to teachers (ten Bokkel et al., 2021). These findings highlight the need for instruments capable of capturing differences across educational stages.

Regarding gender, boys are more often involved as aggressors or bully-victims (Ordóñez-Ordóñez & Narváez, 2020), while girls are often involved as victims (Chocarro & Garaigordobil, 2019; Li et al., 2020). However, findings regarding gender differences in perceived success of teacher responses remain inconclusive (Rigby, 2020; Wachs et al., 2019), underscoring the importance of adopting a gender-sensitive perspective and developing tools that facilitate gender-based comparisons.

As for bullying roles, literature typically distinguishes between aggressors, victims, and bystanders (Harbin et al., 2018; Salmivalli, 2010). However, roles are dynamic and can shift over time (Mendoza-González et al., 2020). Recent research has noted a rise in the bully-victim profile—students who are both victims and aggressors (Burger et al., 2015; Quintana-Orts et al., 2023; Romera et al., 2011)—surpassing the prevalence of the pure roles (Andrade et al., 2021; Sung et al., 2018). This emerging profile has sparked growing research interest due to its complexity. Furthermore, bullying roles may influence how students perceive the success of teacher responses, although findings are still scarce and inconsistent (Berkowitz, 2013; Johander et al., 2024; Wachs et al., 2019). These insights emphasize the need for measurement tools that demonstrate invariance across key variables such as educational level, gender, and bullying role. Measurement invariance ensures that the instrument assesses the same constructs in equivalent ways across different groups, allowing for meaningful comparisons of students' perceptions of teacher responses.

To address existing gaps in the literature and advance the field, the general aim of the present study is to contribute to the understanding of teacher response from the student perspective by adapting and validating the Teachers' Response to Bullying Questionnaire (TRBQ) with a sample of primary and secondary school students in southern Spain. The specific aims of this study are: 1) to explore the most appropriate factorial solution of the TRBQ in our context; 2) to test the measurement invariance of TRBQ across educational level, students' gender, and bullying involvement role; and 3) to describe the students' perceptions of teacher responses according to these variables. Given the lack of consistent evidence regarding the factorial structure of the TRBQ and the absence of prior validation studies in the Spanish context, an exploratory approach was adopted in this study. Previous research has reported varying structures—three factors in primary and secondary school settings (Campaert et al., 2017; Nappa et al., 2021) and five factors in more recent work on primary education (van Gils et al., 2022)—which may reflect cultural or developmental differences. Similarly, no prior studies have examined measurement invariance by gender, educational level, or bullying role using the TRBQ or related instruments. Therefore, this study does not test specific hypotheses but is grounded in existing classifications of teacher responses to bullying that inform the theoretical framework of the TRBQ.

Method

Participants

This study employed a cross-sectional design with a cluster sampling method. The sample consisted of 1,241 students (48.8% girls; $n = 605$), aged between 9 and 18 years ($M = 12.00$; $SD = 1.79$), from 72 classes across 11 schools in Andalusia, Spain. Regarding educational level, 48.3% of the students were in Primary Education ($n = 600$), and 51.7% were in Compulsory Secondary Education ($n = 641$). More specifically, the participants were distributed across the following grade levels: 5th grade ($n = 325$) and 6th grade ($n = 258$) in Primary Education; and 1st to 4th grades of Secondary Education—1st ESO ($n = 179$), 2nd ESO ($n = 191$), 3rd ESO ($n = 143$), and 4th ESO ($n = 128$).

Instruments

Teacher Responses to Bullying

The *Teachers' Responses to Bullying Questionnaire* (TRBQ; Nappa et al., 2021; van Gils et al., 2022) was adapted to Spanish. This instrument assesses students' perceptions of teacher responses in bullying situations using a 5-point Likert scale (*1 = Never, 2 = Almost never, 3 = Sometimes, 4 = Often, 5 = Always*). The instrument begins with the following prompt: "*What did your main teacher do, or what do you think they would do, in response to a bullying case in your class or school?*". In the Spanish educational context, the main teacher refers to the teacher responsible for overseeing the class group, often serving as the primary point of contact for both students and families. This wording was designed to capture both direct experiences (i.e., when students had witnessed teacher responses to actual bullying episodes) and general perceptions or expectations (i.e., in cases where they had not personally observed such situations). This approach enables the assessment of students' perceptions of teacher responses regardless of their direct exposure to bullying. The original TRBQ consists of 15 items. However, in the Spanish adaptation, the original item 14 ("My teacher reports the bullying episode to the principal or the parents") was split into two separate items, one referring to reporting the incident to the principal (item 14, see Table 1) and the other to the parents (item 15, see Table 1). As a result, the TRBQ in the Spanish version of the TRBQ comprises 16 items. The psychometric properties of the TRBQ are reported in the Results section.

Bullying

The *European Bullying Intervention Project Questionnaire* (EBIP-Q; Ortega-Ruiz et al., 2016) was used to assess bullying involvement. This instrument consists of 14 items that assess the frequency of students' engagement in victimization and aggression, using a 5-point Likert scale (*0 = Never, 1 = Almost never, 2 = Sometimes, 3 = Often, 4 = Always*). Based on the responses, students were classified into four involvement roles: victim, aggressor, bully-victim, and non-involved. The classification was performed using cut-off points based on previous studies (Ortega-Ruiz et al., 2016). Students were categorized as victims, who reported having suffered some behavior once or twice a month or more often in the last two months, or as aggressors if they reported engaging in aggression with the same frequency. Those who met both criteria were classified as bully-victims. Students who did not meet either threshold were classified as not involved. The psychometric properties of the EBIP-Q were tested in the original study (Ortega-Ruiz et al., 2016) and subsequent research (Rodriguez-Hidalgo et al., 2019), identifying two factors: victimization and aggression. In this study, internal consistency was adequate (α = .84 for victimization; α = .85 for aggression).

Procedure

Data was collected using paper-and-pencil questionnaires administered during a 30-minute session held during regular school hours. Participation in the study was voluntary and required informed consent from the students' families, assent from students under the age of 14, and informed consent from those aged 14 and older, along with the necessary permissions from the participating schools. Anonymity was assured for all participants. The research was approved by the Research Ethics Committee of the Universidad de Sevilla (0562-N23).

Data collection took place between October and December 2023. In all schools, the administration was carried out by a trained research team following standardized instructions to ensure consistency.

Data Analysis

Analyses were conducted using SPSS Statistics 29 and Mplus 8.4. Descriptive statistics and item normality (skewness ±2; kurtosis ±7; George & Mallery, 2010) were first examined. A cross-validation approach was applied: an Exploratory Factor Analysis (EFA) was performed on a randomly selected subsample (n = 593), followed by a Confirmatory Factor Analysis (CFA) on the remaining subsample (n = 625) to test the TRBQ's structure in Spain. EFA used Robust Maximum Likelihood (MLR) estimation and GEOMIN oblique rotation. Factor retention was based on parallel analysis, requiring a minimum of three items per factor with loadings ≥ .30; cross-loading items (difference ≤ .10) were removed (Floyd & Widaman, 1995). Model fit was evaluated using

Table 1
Distribution of Responses for Each TRBQ Item

Items My main teacher…	Never *n* (%)	Almost never *n* (%)	Sometimes *n* (%)	Often *n* (%)	Always *n* (%)
Ignores bullying	896 (74.0%)	144 (11.9%)	89 (7.3%)	31 (2.6%)	51 (4.2%)
Does not notice when bullying occurs	641 (53.6%)	231 19.3%)	163 (13.6%)	62 (5.2%)	100 (8.4%)
Let the students solve it on their own.	508 (42.2%)	239 (19.9%)	292 (24.3%)	70 (5.8%)	95 (7.9%)
Helps the students involved to resolve the bullying.	194 (16.3%)	63 (5.3%)	145 (12.2%)	209 (17.5%)	581 (48.7%)
Talks about bullying with the whole class	206 (21.9%)	145 (12.2%)	217 (18.3%)	192 (16.2%)	372 (31.4%)
Discuss with the class how much the victim can suffer because of bullying	211 (17.7%)	85 (7.1%)	236 (19.8%)	213 (17.2%)	444 (37.3%)
Encourages the students to make peace	130 (10.9%)	63 (5.3%)	151 (12.7%)	251 (21.1%)	597 (50.1%)
Helps the (involved) students find a solution to the bullying episode	106 (8.9%)	51 (4.3%)	132 (11.1%)	214 (17.9%)	690 (57.8%)
Encourages other students in the class to comfort and support the victim	194 (16.3%)	122 (10.2%)	192 (16.1%)	202 (17.0%)	481 (40.4%)
Tries to help the victim	109 (9.1%)	39 (3.3%)	138 (11.5%)	159 (13.3%)	750 (62.8%)
Comforts the victim.	130 (11.1%)	57 (4.9%)	150 (12.8%)	169 (14.4%)	669 (56.9%)
Tells the bully/bullies that their behavior is unacceptable.	191 (16.2%)	94 (8.0%)	175 (14.8%)	157 (13.3%)	563 (47.7%)
Takes disciplinary actions against the bully/bullies.	116 (9.7%)	70 (5.9%)	141 (11.8%)	177 (14.8%)	688 (57.7%)
Reports the bullying episode to the principal.	135 (11.5%)	80 (6.8%)	198 (16.9%)	159 (13.6%)	597 (51.1%)
Reports the bullying episode to the families.	126 (10.7%)	65 (5.5%)	190 (16.1%)	174 (14.6%)	624 (53.0%)
Explains what bullying is and discusses it with the class.	182 (15.2%)	83 (7.0%)	189 (15.8%)	184 (15.4%)	556 (46.6%)

established thresholds: CFI > .90, RMSEA and SRMR < .08, and χ^2/df < 5 (Hu & Bentler, 1999; Wheaton et al., 1977).

Internal consistency was assessed using Composite Reliability (CR), with .60 as the minimum for exploratory research (Bagozzi & Yi, 1988; Hair et al., 2014). Convergent validity was examined through CFA loadings (≥ .40), and Average Variance Extracted (AVE) was calculated (Weiss, 2011); AVE ≥ .50 was preferred, though .40 was acceptable if CR exceeded .60 (Fornell & Larcker, 1981; Huang et al., 2013).

To compare teacher responses across educational level (Primary: n = 600; Secondary: n = 641), gender (boys: n = 624; girls: n = 605), and bullying roles (victims: n = 382; aggressors: n = 57; non-involved students, n = 635; and bully-victims: n = 144), measurement invariance was tested. Measurement invariance testing included three steps: 1) configural invariance, assessing whether the model structure is the same across groups; 2) metric invariance, assessing whether groups interpret the items in the same way; and 3) scalar invariance, assessing whether factor means can be validly compared across groups. Configural, metric, and scalar invariance were evaluated using Chen's (2007) criteria: ΔCFI < .010 and ΔRMSEA < .015 indicated full invariance. Partial invariance was tested by freeing non-invariant parameters. The MLR estimator was used due to non-normal data distributions.

To examine group differences in students' perceptions of teacher responses, one-way and factorial ANOVAs were conducted. Effect sizes were interpreted using η^2: small (< .01), moderate (.01–.06), and large (> .14) (Cohen, 1988). A significance level of p < .05 was applied. Post hoc comparisons used the Bonferroni correction. Interaction effects between educational level, gender, and bullying role were explored via factorial ANOVA.

Missing data ranged from 2.4% to 5.8% per item. All models were estimated using Full Information Maximum Likelihood (FIML) to handle missing data without imputation (Enders & Bandalos, 2001).

Results

Descriptive Analysis

The detailed frequencies for each response category of TRBQ are provided in Table 1. Table 2 summarizes the descriptive statistics, skewness, and kurtosis for each item. Most items displayed acceptable levels of univariate normality. However, item 1 showed considerable deviations from normality, with high values of skewness and kurtosis across both samples. This suggests that students rarely perceive their teacher as ignoring bullying, leading to a strong concentration of responses at the lower end of the scale.

Exploratory Factor Analysis (EFA)

One- to four-factorial solutions were examined to evaluate the progressive model fit (see Table 3). The one- and two-factor models presented a poor fit. The three-factor model showed a clear improvement, with further enhancement observed in the four-factor model. However, the four-factor solution was not retained, as it did not meet the criterion of having at least three items per factor. The three-factor model was selected based on the results of the parallel analysis, which supported a three-factor structure. Upon further inspection, items 12 and 16 presented cross-loadings, with similar factor loadings on multiple factors. Consequently, both items were progressively removed. After their exclusion, the final model demonstrated good fit (see Table 3).

The first factor included items 1 to 3, reflecting a lack of teacher action in bullying situations, and was labeled as Non-Intervention (NI). The second factor, comprising items 4 to 11, encompassed strategies such as group discussion, victim support, and mediation, and was labeled Restorative Psychoeducational (RP). The third factor,

Table 2
Descriptive Statistics, Skewness, and Kurtosis of TRBQ Items in the EFA and AFC Subsamples

Item	EFA (n = 593)			CFA (n = 625)		
	M (SD)	Skewness (SE)	Kurtosis (SE)	*M (SD)*	Skewness (SE)	Kurtosis (SE)
1	0.49 (1.01)	2.17 (0.10)	3.90 (0.20)	0.49 (1.00)	2.19 (0.09)	4.01 (0.19)
2	0.96 (1.25)	1.16 (0.10)	0.23 (0.20)	0.95 (1.26)	1.18 (0.09)	0.25 (0.19)
3	1.17 (1.24)	0.79 (0.10)	-0.28 (0.20)	1.15 (1.23)	0.82 (0.09)	-0.24 (0.19)
4	2.77 (1.50)	-0.86 (0.10)	-0.75 (0.20)	2.78 (1.50)	-0.87 (0.09)	-0.75 (0.19)
5	2.22 (1.53)	-0.23 (0.10)	-1.41 (0.20)	2.23 (1.52)	-0.23 (0.10)	-1.40 (0.19)
6	2.51 (1.48)	-0.54 (0.10)	-1.09 (0.20)	2.51 (1.49)	-0.54 (0.10)	-1.11 (0.19)
7	2.93 (1.36)	-1.33 (0.10)	0.51 (0.20)	2.95 (1.35)	-1.10 (0.09)	-0.05 (0.19)
8	3.10 (1.30)	-1.33 (0.10)	0.51 (0.20)	3.10 (1.29)	-1.33 (0.10)	0.53 (0.19)
9	2.50 (1.49)	-0.51 (0.10)	-1.19 (0.20)	2.51 (1.49)	-0.52 (0.09)	-1.17 (0.19)
10	3.16 (1.31)	-1.43 (0.10)	0.72 (0.20)	3.17 (1.30)	-1.45 (0.09)	-1.17 (0.19)
11	3.00 (1.37)	-1.17 (0.10)	0,02 (0.20)	3.03 (1.36)	-1.21 (0.10)	0.15 (0.20)
12	2.66 (1.50)	-0.66 (0.10)	-1.04 (0.20)	2.67 (1.50)	-0.67 (0.09)	-1.03 (0.19)
13	3.01 (1.35)	-1.14 (0.10)	-0.03 (0.20)	3.03 (1.34)	-1.16 (0.10)	0.01 (0.19)
14	2.81 (1.42)	-0.83 (0.10)	-0.67 (0.20)	2.83 (1.41)	-0.85 (0.10)	-0.63 (0.20)
15	2.88 (1.39)	-0.95 (0.10)	-0.44 (0.20)	2.88 (1.40)	-0.95 (0.10)	-0.46 (0.20)
16	2.68 (1.48)	-0.72 (0.10)	-0.92 (0.20)	2.67 (1.49)	-0.70 (0.10)	-0.96 (0.19)

Table 3
EFA, CFA, and the Measurement Invariance of TRBQ

Models	S–B χ^2	df	χ^2/df	CFI	ΔCFI	RMSEA [90% CI]	ΔRMSEA	SRMR	BIC	AIC	Decision
EFA (1 factor)	702.45	104	6.75	0.819		0.099 [0.092-0.105]		0.077	28645.22	28434.73	
EFA (2 factors)	453.98	89	5.10	0.889		0.083 [0.076-0.091]		0.044	28420.60	28144.33	
EFA (3 factors)	284.31	75	3.79	0.937		0.069 [0.060-0.077]		0.035	28259.32	27921.66	
EFA (4 factors)	170.67	62	2.75	0.967		0.054 [0.045-0.064]		0.023	28175.42	27889.70	
EFA (without items 12 and 16)	204.49	52	3.93	0.946		0.070 [0.060-0.081]		0.031	24458.41	24164.60	
CFA (3 factors, without items 12 and 16)	301.07	74	4.06	0.914		0.070 [0.062-0.078]		0.049	26367.98	26168.28	
CFA (3 factors, Nappa et al., 2021)	395.59	101	3.91	0.903		0.068 [0.061-0.075]		0.050	30378.54	30152.21	
CFA (5 factors, van Gils et al., 2022)	278.36	94	2.96	0.939		0.056 [0.048-0.064]		0.045	30276.18	30018.79	
						Educational level					
Configural invariance	615.50	148	4.15	0.914		0.072 [0.066-0.078]		0.053	50539.72	50080.28	Accepted
Metric invariance	631.25	159	3.97	0.913	0.001	0.070 [0.064-0.076]	0.002	0.055	50477.46	50074.17	Accepted
Scalar invariance	694.16	170	4.08	0.904	0.009	0.071 [0.066-0.077]	0.001	0.058	50466.22	50119.08	Accepted
						Gender					
Configural invariance	636.14	148	4.29	0.909		0.074 [0.068-0.080]		0.054	50369.58	49910.95	Accepted
Metric invariance	645.17	159	4.05	0.909	0.000	0.071 [0.066-0.077]	-0.003	0.055	50295.02	49892.44	Accepted
Scalar invariance	659.43	170	3.87	0.909	0.000	0.069 [0.064-0.075]	-0.002	0.055	50225.66	49879.14	Accepted
						Bullying roles					
Configural invariance	825.92	296	2.79	0.908		0.077 [0.071-0.083]		0.056	50629.32	49712.51	Accepted
Metric invariance	860.06	329	2.61	0.907	0.001	0.073 [0.067-0.079]	0.004	0.060	50432.20	49683.47	Accepted
Scalar invariance	901.36	362	2.38	0.906	0.001	0.070 [0.065-0.076]	0.003	0.061	50234.17	49653.52	Accepted

Note. S-B χ^2 = Satorra-Bentler chi-square; χ^2/df = Satorra-Bentler chi-square/degrees of freedom; CFI = comparative fit index; ΔCFI = difference in CFI between the two models examined; RMSEA = root mean information criteria; 90% CI = confidence interval RMSEA; ΔRMSEA = difference in RMSEA between the two models compared; SRMR = standardized root mean square residual; BIC = Bayesian information criterion; AIC = Akaike Information Criterion.

consisting of items 13 to 15, reflected punitive responses and was labeled Disciplinary Methods (DM; see Table 4). The total variance explained was 53.72%: NI (9.33%), RP (28.91%), and DM (15.49%).

Table 4
Factor Loadings and Communalities From the EFA

Item	Factor 1 = NI	Factor 2 = RP	Factor 3 = DM	Communality
TR1	**0.734**	-0.265	0.008	0.55
TR2	**0.688**	-0.004	-0.122	0.48
TR3	**0.492**	0.060	-0.118	0.26
TR4	0.085	**0.599**	-0.066	0.33
TR5	0.202	**0.378**	0.006	0.21
TR6	0.053	**0.524**	0.073	0.34
TR7	-0.002	**0.761**	0.010	0.59
TR8	-0.086	**0.882**	-0.004	0.76
TR9	0.020	**0.599**	0.079	0.43
TR10	-0.012	**0.719**	0.200	0.74
TR11	0.004	**0.611**	0.294	0.70
TR13	-0.004	0.373	**0.515**	0.66
TR14	0.083	0.001	**0.901**	0.82
TR15	-0.005	0.155	**0.693**	0.65

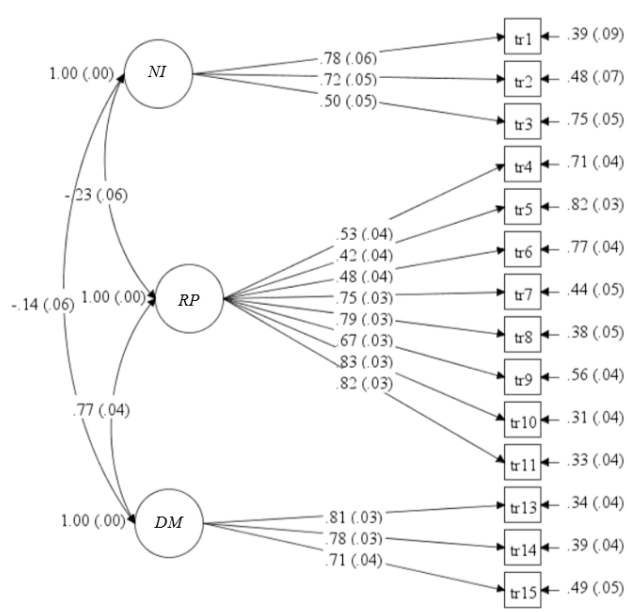
Note. NI = Non-intervention; RP = Restorative Psychoeducational; DM = Disciplinary methods.

Confirmatory Factor Analysis (CFA)

A CFA was performed to validate the three-factor structure (NI, RP, and DM) identified in the EFA. The model showed acceptable fit indices (see Table 3). Additionally, its fit was compared to two alternative models previously reported in the literature: the three-factor structure proposed by Nappa et al., (2021) and the five-factor structure by van Gils et al. (2022). The results indicated that the model derived from the EFA showed the best fit, as evidenced by the lowest AIC and BIC values (see Table 3).

All factor correlations were below .80, indicating adequate discriminant validity. Standardized factor loadings were statistically significant, ranging from .42 to .83. Both CR and AVE values met acceptable thresholds, further supporting the internal consistency and convergent validity of the factors (see Figure 1).

Figure 1
Three-factor Model of the TRBQ



Note. NI = Non-intervention; RP = Restorative psychoeducational; DM = Disciplinary methods; all values shown in the diagram are standardized; CR [NI = 0.71; RP = 0.86; DM = 0.81]; AVE [NI = 0.45; RP = 0.46; DM = 0.58].

Measurement Invariance

Differences in the Perception of Teacher Responses

Significant differences in students' perceptions of teacher responses were observed across educational level, gender, and bullying role (see Table 5). Regarding educational level, non-intervention was perceived as more frequent among secondary school students. Primary school students perceived higher levels of restorative psychoeducational responses. The effect sizes were small in both cases. No significant differences were found between educational levels in perceptions of disciplinary methods.

Concerning gender, girls perceived all forms of teacher response as more frequent than boys, although the effect sizes were small across comparisons.

Regarding bullying roles, non-intervention was perceived as more frequent by students identified as bully-victims, with a small effect size. Non-involved students perceived restorative psychoeducational responses as more frequent, whereas bully-victims perceived them as less frequent. Finally, aggressor students perceived greater use of disciplinary methods. The effect size was small across comparisons (see Table 5).

Additionally, interaction effects of educational level, gender, and bullying role on students' perceptions of teacher responses were examined. For non-intervention, a three-way interaction effect between educational level, gender, and bullying role was significant ($F(3, 1177) = 4.81$, $p = .002$, $\eta^2 = .012$). In primary school, female bully-victims perceived higher levels of teacher non-intervention, whereas in secondary school, male bully-victims and aggressors perceived the highest levels of non-intervention. Regarding restorative psychoeducational responses, a significant interaction effect was found between educational level and gender ($F(1, 1174) = 4.26$, $p = .039$, $\eta^2 = .004$). Male students in primary school perceived more restorative psychoeducational responses than male students in secondary school. All the effect sizes were small. No significant interaction effects were found for perceptions of disciplinary methods.

Discussion

Although an increasing number of studies confirm that teacher response is crucial to prevent and stop the development of bullying cases, there are no validated instruments in Spain to analyze it validly and reliably. Therefore, the main objective of this study was to contribute to the field of research on teacher response to bullying by adapting and analyzing the psychometric properties of the Teachers' Responses to Bullying Questionnaire (TRBQ).

The first objective was to explore the structure of TRBQ in Spain to further explore the underlying dimensions. The EFA identified a three-factor solution: non-intervention, restorative psychoeducational strategies (including group discussion, victim support, and mediation), and disciplinary methods. Items 12 and 16 were removed due to cross-loadings on multiple factors, likely because both referred to actions that could plausibly fit into more than one response category. The three-factor structure was subsequently confirmed through the CFA, whose fit indices were adequate. These results support the validity of the TRBQ as an appropriate instrument for assessing students' perceptions of teacher responses to bullying in the Spanish educational context.

Similarly, the factorial solution identified aligns with previous studies, such as Nappa et al. (2021) in Italy, where a three-factor structure was also found in a sample of secondary school students, encompassing non-intervention, relational or supportive responses, and disciplinary methods. These findings suggest a cross-cultural convergence in students' perception, as similar structuring of teacher responses to bullying emerges in both Spain and Italy. Notably, restorative psychoeducational strategies are rarely applied in isolation; instead, they are typically combined—integrating victim support, mediation, and group discussion. This tendency to employ multiple responses aligns with recent studies indicating that teachers often employ a combination of responses rather than relying on a single response, as concluded from studies based on teacher reports (Burger et al., 2015) and student reports (Muñoz-Fernández et al., 2025a; van Gils et al., 2024).

The second objective of the study was to analyze the measurement invariance of the TRBQ across educational level, gender, and bullying role. The results indicated full measurement invariance across all comparisons, supporting TRBQ's validity for assessing students' perceptions of teacher responses regardless of whether the students are boys or girls, in primary or secondary education, or involved in bullying as aggressors, victims, bully-victims, or not involved. To date, few studies have examined measurement invariance based on bullying roles, marking this work an innovative and relevant contribution in this field. Moreover, these findings not only reinforce the instrument's psychometric robustness but also highlight its utility as a versatile tool, suitable for use across diverse educational contexts and student profiles—enhancing its applicability and potential for cross-group comparisons.

Regarding the third objective, the study analyzed differences in students' perceptions of teacher responses. The results indicated that

Table 5
Perceived Teacher Responses Across Educational Level, Gender, and Bullying Roles

	Educational level					Gender					Bullying roles						
	Primary *M (SD)*	Secondary *M (SD)*	*F(df)*	*p*	η^2	Boys *M (SD)*	Girls *M (SD)*	F(df)	p	η^2	Not involved *M (SD)*	Aggressors *M (SD)*	Victims *M (SD)*	Bully-victims *M (SD)*	*F(df)*	*p*	η^2
NI	0.82 (0.82)	0.93 (1.02)	F(1, 1215) = 3.91	.034	.004	0.80 (0.92)	0.92 (0.92)	F(1, 1210) = 5.28	.001	.013	0.72 (0.81)	1.05 (0.99)	0.92 (0.92)	1.37 (1.16)	F(3, 1199) = 21.51	.001	.051
RP	2.95 (0.88)	2.61 (1.12)	F(1, 1212) = 33.46	<.001	.027	2.72 (1.10)	2.86 (0.92)	F(1, 1206) = 2.89	.034	.007	2.89 (0.97)	2.82 (0.85)	2.71 (1.05)	2.42 (1.15)	F(3, 1196) = 8.84	.001	.022
DM	2.89 (1.15)	3.00 (1.24)	F(1, 1203) = 2.37	.124	.002	2.87 (1.27)	3.04 (1.10)	F(1, 1210) = 3.39	.017	.008	3.08 (1.16)	3.31 (0.88)	2.79 (1.23)	2.56 (1.29)	F(3, 1189) = 11.30	.001	.028

Note. NI = Non-intervention; RP = Restorative psychoeducational; DM = Disciplinary methods; Sample sizes: Primary ($n = 600$), Secondary ($n = 641$), Boys ($n = 624$), Girls ($n = 605$), Not involved ($n = 635$), Aggressors ($n = 57$), Victims ($n = 382$), Bully-victims ($n = 144$).

restorative psychoeducational strategies were perceived as more frequent among primary school students, whereas non-intervention was more commonly perceived among secondary school students. Additionally, the analysis of the interaction between educational level and gender in restorative psychoeducational responses revealed that primary school boys perceived greater use of this strategy compared to secondary school boys. This difference may be explained by the greater sensitivity and proactive attitudes of primary school teachers in addressing bullying situations (Sokol et al., 2016; van Aalst et al., 2024), potentially related to differences in teacher training. In Spain, primary school teachers complete a four-year university degree that includes coursework in child development, pedagogy, psychology, and classroom management. In contrast, secondary school teachers typically hold a subject-specific degree followed by a one-year postgraduate program in education (Real Decreto 1834/2008).

Concerning the use of disciplinary methods, the lack of significant differences in perceptions between primary and secondary school students aligns with previous research indicating that teachers at both educational levels tend to resort to disciplinary strategies when aiming to restore order and enforce clear consequences (Bauman et al., 2008; Yoon & Bauman, 2014). Moreover, this tendency could be explained by the fact that disciplinary strategies represent a more traditional and immediately applicable response, whereas the proper implementation of restorative psychoeducational strategies might require specific skills and training.

In terms of gender, the results of this study showed that girls perceived teachers' responses as more frequent than boys. One possible explanation, consistent with previous research, is that girls tend to consider bullying as a more serious problem, which may make them more attentive to, and more likely to report, teachers' responses (Sokol et al., 2016). Additionally, girls are more often involved in bullying as victims (Chocarro & Garaigordobil, 2019; Li et al., 2020), have greater academic engagement, and have more positive perceptions of their teachers in both academic and relational aspects (King, 2016). These factors may contribute to their greater sensitivity to teacher responses.

While the overall pattern showed girls perceiving greater teacher responses, the interaction effects revealed important nuances. Specifically, female bully-victims in primary school perceived the highest levels of non-intervention. This may be explained by their closer relationships with teachers (Furrer & Skinner, 2003), which could foster higher expectations of support. When these expectations are unmet, perceptions of teacher inaction may be particularly salient.

In contrast, boys perceived all teacher responses as less frequent than girls. This perception could be partly explained by a lower tendency among boys to seek help (Bjereld et al., 2024), as well as by the association between being a boy and a higher probability of experiencing a failed response, both in the role of aggressor and victim (Johander et al., 2024). This interpretation is supported by the interaction effects observed: in secondary school, male students involved in bullying—both as aggressors and bully-victims—reported the highest perceptions of teacher inaction. These patterns suggest that students' roles in bullying, combined with their gender, shape how they interpret teachers' responses. These findings underscore the importance of ensuring that teachers' responses are equally visible and effective for all students, and they point to the need for further research into the reasons why boys, especially those involved in bullying, tend to report lower awareness of teacher intervention.

Regarding bullying roles, non-involved students perceived more restorative psychoeducational strategies, while those involved as bully-victims tended to perceive a greater lack of response. This may suggest that students not involved in bullying dynamics are more receptive to teacher responses. In contrast, bully-victim students may perceive a systematic absence of response, reinforcing feelings of ambivalence and neglect. This highlights the urgent need to address this complex profile, which often poses challenges for teachers in terms of identification and appropriate response.

Meanwhile, students identified as aggressors reported a higher perception of disciplinary methods, consistent with previous studies that highlight the predominance of punitive approaches when addressing this group (Byers et al., 2011; Campaert et al., 2017; Rigby, 2014; Yoon et al., 2016). However, these results may suggest the need to work with aggressive students from a psychoeducational perspective, enabling them to recognize the harm they have caused and to take responsibility to change the situation.

Despite its contribution, some limitations should be acknowledged. First, the reliance on student self-reports may be subject to halo effects (Spooren et al., 2013), as perceptions of teacher responses could be influenced by personal relationships or past experiences, potentially compromising objectivity. Future research should adopt multi-informant designs to cross-validate student reports with data from teachers or families. Second, although a three-factor structure was validated, the restorative strategies dimension comprises more items and subtypes than the disciplinary and non-intervention dimensions. Future work should aim to balance the scale by expanding items in the latter dimensions. Third, the TRBQ lacks specific items targeting restorative responses toward aggressors, despite growing evidence supporting psychoeducational approaches for this group (Johander et al., 2021). Developing a dedicated subscale would enhance the instrument's comprehensiveness. Additionally, the proportion of students identified as victims or bully-victims (43%) exceeds national averages. This discrepancy likely stems from methodological differences: our study included students reporting victimization once or twice a month (Solberg & Olweus, 2003), while national data (MEFP, 2022) used a stricter 'once a week' criterion and considered only pure victims. Harmonizing frequency and classification criteria across studies would improve comparability and prevalence estimates. Finally, the study's cross-sectional design and regional sample (southern Andalusia) limit generalizability. Future longitudinal research with broader and more diverse samples is needed to confirm the instrument's stability and applicability across educational and cultural contexts.

This study represents one of the first contributions in Spain to validate an instrument for assessing teacher responses to bullying from the students' perspective. The adaptation of the TRBQ enables meaningful cross-cultural comparisons with other countries where it has been validated. The TRBQ demonstrates sensitivity to key variables such as gender, educational stage, and the bullying roles, making it a versatile tool for exploring different student profiles. Furthermore, it can serve as a valuable resource for evaluating the effectiveness of programs in general, and teacher training programs in particular, by measuring changes in teacher responses following targeted interventions (Van Verseveld et al., 2019). The findings can serve for designing school policies and prevention strategies tailored to the unique characteristics of the student population and the specific educational stages.

Author Contributions

Laura Rodríguez-Pérez: Conceptualization, Formal analysis, Methodology, Writing - Original draft, Writing - Review and editing. **Rosario Del Rey**: Conceptualization, Methodology, Writing – Review and editing. **Noemí García-Sanjuán**: Conceptualization, Writing - Original draft. **Noelia Muñoz-Fernández**: Funding Acquisition, Project administration, Formal Analysis, Methodology, Supervision, writing - Review and editing.

Funding

This work has received funding from the MICIU/AEI/10.13039/501100011033 and the European Union NextGenerationEU/PRTR, which financed the grant of Noelia Muñoz-Fernández [RYC2021-034977-I]. This funding source had no role in the design of this study, data collection, management, analysis, and interpretation of data, writing of the manuscript, and the decision to submit the manuscript for publication.

This research was funded by the VII Plan Propio de Investigación y Transferencia of Universidad de Sevilla [VII PPIT-US]. The Universidad de Sevilla financed the grant of Laura Rodríguez-Pérez [VII-PPITUS].

Declaration of Interests

The authors declare that there is no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon request.

References

Andrade, B., Guadix, I., Rial, A., & Suárez, F. (2021). *Impacto de la tecnología en la adolescencia: Relaciones, riesgos y oportunidades. Un estudio comprensivo e inclusivo hacia el uso saludable de las TRIC* [Impact of technolog y on adolescence: Relationships, risks and opportunities. A comprehensive and inclusive study towards the healthy use of ICTs]. UNICEF España, Universidad de Santiago de Compostela y Consejo General de Colegios Profesionales de Ingeniería en Informática. https://www.unicef.es/publicacion/impactode-la-tecnologia-en-la-adolescencia

Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, *16*(1), 74–94. https://doi.org/10.1007/BF02723327

Bauman, S., Rigby, K., & Hoppa, K. (2008). US teachers' and school counsellors' strategies for handling school bullying incidents. *Educational Psychology, 28*(7), 837–856. https://doi.org/10.1080/01443410802379085

Berkowitz, R. (2013). Student and teacher responses to violence in school: The divergent views of bullies, victims, and bully-victims. *School Psychology International, 34*(5), 547-560. https://doi.org/10.1177/0143034313511012

Bjereld, Y., Thornberg, R., & Hong, J. S. (2024). Why don't all victims tell teachers about being bullied? A mixed methods study on how direct and indirect bullying and student-teacher relationship quality are linked with bullying disclosure. *Teaching and Teacher Education, 148*, 104664. https://doi.org/10.1016/j.tate.2024.104664

Burger, C., Strohmeier, D., Spröber, N., Bauman, S., & Rigby, K. (2015). How teachers respond to school bullying: An examination of self-reported intervention strategy use, moderator effects, and concurrent use of multiple strategies. *Teaching and Teacher Education, 51*, 191-202. https://doi.org/10.1016/j.tate.2015.07.004

Byers, D. L., Caltabiano, N. J., & Caltabiano, M. L. (2011). Teachers' attitudes towards overt and covert bullying, and perceived efficacy to intervene. *Australian Journal of Teacher Education, 36*(11), 105-119. https://doi.org/10.14221/ajte.2011v36n11.1

Campaert, K., Nocentini, A., & Menesini, E. (2017). The efficacy of teachers' responses to incidents of bullying and victimization: The mediational role of moral disengagement for bullying. *Aggressive Behavior, 43*(6), 483–492. https://doi.org/10.1002/ab.21706

Chocarro, E., & Garaigordobil, M. (2019). Bullying y cyberbullying: diferencias de sexo en víctimas, agresores y observadores [Bullying and cyberbullying: gender differences in victims, aggressors and observers]. *Pensamiento Psicológico, 17*(2), 57-71. https://doi.org/10.11144/Javerianacali.PPSI17-2.bcds

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*, 464-504. https://doi.org/10.1080/10705510701301834

Chen, L. M. (2023). Exploring the impact of multiple predictors on teachers' willingness to intervene in relational bullying: The moderating role of victim-blaming tendency. *Journal of School Violence*, *23*(3), 363–375. https://doi.org/10.1080/15388220.2023.2299984

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2.ª ed.). Lawrence Erlbaum Associates.

Colpin, H., Bauman, S., & Menesini, E. (2021). Teachers' responses to bullying: Unravelling their consequences and antecedents. Introduction to the special issue. *European Journal of Developmental Psychology, 18*(6), 781-797. https://doi.org/10.1080/17405629.2021.1954903

Collier, K. L., Bos, H. M. W., & Sandfort, T. G. M. (2015). Understanding teachers' responses to enactments of sexual and gender stigma at school. *Teaching and Teacher Education, 48*, 34-43. https://doi.org/10.1016/j.tate.2015.02.002

Dedousis-Wallace, A., Shute, R., Varlow, M., Murrihy, R., & Kidman, T. (2013). Predictors of teacher intervention in indirect bullying at school and outcome of a professional development presentation for teachers. *Educational Psychology, 34*(7), 862–875. https://doi.org/10.1080/01443410.2013.785385

Demol, K., Verschueren, K., Salmivalli, C., & Colpin, H. (2020). Perceived teacher responses to bullying influence students' social cognitions. *Frontiers in Psychology, 11*, 592582. https://doi.org/10.3389/fpsyg.2020.592582

Demol, K., Verschueren, K., Jame, M., Lazard, C., & Colpin, H. (2021). Student attitudes and perceptions of teacher responses to bullying: An experimental vignette study. *European Journal of Developmental Psychology*, *18*(6), 814-830. https://doi.org/10.1080/17405629.2021.1896492

Denny, S., Peterson, E. R., Stuart, J., Utter, J., Bullen, P., Fleming, T., Ameratunga, S., Clark, T., & Milfont, T. (2014). Bystander intervention, bullying, and victimization: A multilevel analysis of New Zealand high

schools. *Journal of School Violence, 13*(1), 1-26. https://doi.org/10.1080/15388220.2014.910470

Devleeschouwer, C., Tolmatcheff, C., & Galand, B. (2025). Classmates and teachers matter: Effects of class norms and teachers' responses on bullying behaviors. *International Journal of Bullying Prevention*. https://doi.org/10.1007/s42380-025-00288-3

Díaz-Aguado, M. J. (2023). *Indicadores para evaluar y mejorar la convivencia escolar* [Indicators for evaluating and improving school coexistence]. Ministerio de Educación, Formación Profesional y Deportes. https://www.educacionfpydeportes.gob.es/dam/jcr:0cfbe305-d319-45ff-97fd-c1d1851674aa/indicadores-para-evaluar-y-mejorar-la-convivencia-escolar.pdf

Duong, J., & Bradshaw, C. P. (2013). Using the extended parallel process model to examine teachers' likelihood of intervening in bullying. *Journal of School Health, 83*(6), 422-429. https://doi.org/10.1111/josh.12046

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8*(3), 430–457. https://doi.org/10.1207/S15328007SEM0803_5

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*(3), 286–299. https://doi.org/10.1037/1040-3590.7.3.286

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of marketing research, 18*(1), 39-50. https://journals.sagepub.com/doi/abs/10.1177/002224378101800104

Furrer, C., & Skinner, E. (2003). Sense of relatedness as a factor in children's academic engagement and performance. *Journal of Educational Psychology, 95*(1), 148–162. https://doi.org/10.1037/0022-0663.95.1.148

George, D., & Mallery, P. (2010). *SPSS for windows step by step: A simple guide and reference* (10th ed.). Pearson.

Gregory, A., Cornell, D., Fan, X., Sheras, P. Shih, T., & Huang, F. (2011). Authoritative school discipline: High school practices associated with lower bullying and victimization. *Journal of Educational Psychology, 102*(2), 483-496. https://doi.org/10.1037/a0018562

Grumm, M., & Hein, S. (2012). Correlates of teachers' ways of handling bullying. *School Psychology International*, *34*(3), 299-312. https://doi.org/10.1177/0143034312461467

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2014). *A primer on partial least squares structural equation modeling (PLS-SEM)*. Sage Publications.

Harbin, S. M., Kelley, M. L., Piscitello, J., & Walker, S. J. (2018). Multidimensional bullying victimization scale: development and validation. *Journal of School Violence, 18*(1), 146-161. https://doi.org/10.1080/15388220.2017.1423491

Huang, C.C., Wang, Y.-M., Wu, T.-W., & Wang, P.-A. (2013). An empirical analysis of the antecedents and performance consequences of using the Moodle platform. *International Journal of Information and Education Technology, 3*(2), 217-221. https://www.ijiet.org/papers/267-IT0040.pdf

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

Johander, E., Turunen, T., Garandeau, C. F., & Salmivalli, C. (2021). Different approaches to address bullying in KiVa schools: Adherence to guidelines, strategies implemented, and outcomes obtained. *Prevention Science*, *22*(3), 299-310. https://doi.org/10.1007/s11121-020-01178-4

Johander, E., Turunen, T., Garandeau, C. F., & Salmivalli, C. (2024). Interventions that failed: Factors associated with the continuation of bullying after a targeted intervention. *International Journal of Bullying Prevention*, *6*(4), 421-433. https://doi.org/10.1007/s42380-023-00169-7

Kärnä, A., Voeten, M., Little, T. D., Poskiparta, E., Kaljonen, A., & Salmivalli, C. (2011). A large-scale evaluation of the KiVa antibullying program: grades 4-6. *Child development*, *82*(1), 311–330. https://doi.org/10.1111/j.1467-8624.2010.01557.x

King, R. B. (2016). Gender differences in motivation, engagement, and achievement are related to students' perceptions of peer—but not of parent or teacher—attitudes toward school. *Learning and Individual Differences, 52*, 60–71. https://doi.org/10.1016/j.lindif.2016.10.006

Kollerová, L., Soukup, P., Strohmeier, D., & Caravita, S. C. S. (2021). Teachers' active responses to bullying: Does the school collegial climate make a difference? *European Journal of Developmental Psychology*, *18*(6), 912–927. https://doi.org/10.1080/17405629.2020.1865145

Li, R., Lian, Q., Su, Q., Li, L., Xie, M. & Hu, J. (2020). Trends and sex disparities in school bullying victimization among U.S. youth, 2011-2019. *BMC Public Health, 20*(1), 15-83. https://doi.org/10.1186/s12889-020-09677-3

Llego, J., Samson, M. J., Gabriel, E., Corpus, J., Bustillo, K. G., & Villar, J. (2024). Nursing faculty members' response to bullying in the eyes of their students: a pilot study in Pangasinan. *Nurse Education Today, 138*, 106195. https://doi.org/10.1016/j.nedt.2024.106195

Mendoza-González, B., Delgado, I. & Atenas, M. (2020). Student profile not involved in bullying: description based on gender stereotypes, parenting practices, cognitive-social strategies and food over-intake. *Anales de Psicología, 36*(3), 483-491. https://doi.org/10.6018/analesps.337011

Ministerio de Educación y Formación Profesional. (2022). *Estudio estatal sobre la convivencia escolar en educación primaria* [State Study of School Coexistence in Primary Education]. Secretaría General Técnica, Subdirección General de Documentación y Publicaciones. https://sede.educacion.gob.es/publiventa/descarga.action?f_codigo_agc=23029

Muñoz-Fernández, N., Wachs, S., Marcenaro, N., & Del Rey, R. (2025a). An exploratory study on the perceived effectiveness of teacher interventions in bullying: Insights from the students' perspective. *The British Journal of Educational Psychology.* https://doi.org/10.1111/bjep.12778

Muñoz-Fernández, N., Nappa, M., Stefanelli, F., & Palladino, B.E. (2025b). A validation study of Teachers' Responses to Bullying Questionnaire (TRBQ–T) in two large samples of teachers. *International Journal of Bullying Prevention*. https://doi.org/10.1007/s42380-025-00315-3

Nappa, M. R., Palladino, B. E., Nocentini, A., & Menesini, E. (2021). Do the face-to-face actions of adults have an online impact? The effects of parent and teacher responses on cyberbullying among students. *European Journal of Developmental Psychology*, *18*(6), 798–813. https://doi.org/10.1080/17405629.2020.1860746

Ordóñez-Ordóñez, M. C., & Narváez, M. R. (2020). Self-esteem in adolescents involved in bullying situations. *MASKANA, 11*(2), 27-33. https://doi.org/10.18537/mskn.11.02.03

Ortega-Ruiz, R., Del Rey, R., & Casas, J. A. (2016). Assessing bullying and cyberbullying: Spanish validation of EBIPQ and ECIPQ. *Psicología Educativa*, *22*(1), 71–79. https://doi.org/10.1016/j.pse.2016.01.004

Quintana-Orts, C., Mora-Merchán, J.A., Muñoz-Fernández, N., & Del Rey, R. (2023). Bully-victims in bullying and cyberbullying: An analysis of school-level risk factors. *Social Psychology of Education, 27,* 587–609. https://doi.org/10.1007/s11218-023-09846-3

Real Decreto 1834/2008, de 8 de noviembre, por el que se definen las condiciones de formación para ejercer la docencia en educación secundaria obligatoria, bachillerato, formación profesional y enseñanzas

de régimen especial [Royal Decree 1834/2008, of November 8, defining the training requirements for teaching in compulsory secondary education, upper secondary education, vocational training, and special education]. *Boletín Oficial del Estado*, n. 287, November 28, 2008, pages 47587 to 47593. https://www.boe.es/eli/es/rd/2008/11/08/1834

Rigby, K. (2014). How teachers address cases of bullying in schools: a comparison of five reactive approaches. *Educational Psychology in Practice*, *30*(4), 409–419. https://doi.org/10.1080/02667363.2014.949629

Rigby, K. (2020). How teachers deal with cases of bullying at school: What victims say. *International Journal of Environmental Research and Public Health, 17*(7), 2338. https://doi.org/10.3390/ijerph17072338

RodríguezHidalgo, A. J., Alcívar, A., & HerreraLópez, M. (2019). Traditional bullying and discriminatory bullying around special educational needs: Psychometric properties of two instruments to measure it. *International Journal of Environmental Research and Public Health, 16*(1), 142. https://doi.org/10.3390/ijerph16010142

Romera, E., Del Rey, R. & Ortega, R. (2011). Factors associated with involvement in bullying: A study in Nicaragua. *Psychosocial Intervention, 20*(2), 161-170. https://doi.org/10.5093/in2011v20n2a4

Salmivalli, C. (2010). Bullying and the peer group: A review. *Aggression and Violent Behavior, 15*(2), 112-120. https://doi.org/10.1016/j.avb.2009.08.007

Siddiqui, S., Schultze-Krumbholz, A., & Hinduja, P. (2023). Practices for dealing with bullying by educators in Pakistan: Results from a study using the handling bullying questionnaire. *Cogent Education*, *10*(2), 2236442. https://doi.org/10.1080/2331186X.2023.2236442

Solberg, M. E., & Olweus, D. (2003). Prevalence estimation of school bullying with the Olweus Bully/Victim Questionnaire. *Aggressive Behavior, 29*(3), 239–268. https://doi.org/10.1002/ab.10047

Sokol, N., Bussey, K., & Rapee, R. M. (2016). The impact of victims' responses on teacher reactions to bullying. *Teaching and Teacher Education, 55*, 78–87. https://doi.org/10.1016/j.tate.2015.11.002

Song, K.-H., Lee, S.-Y., & Park, S. (2018). How individual and environmental factors influence teachers' bullying intervention. *Psychology in the Schools, 55(9),* 1086–1097. https://doi.org/10.1002/pits.22151

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, *83*(4), 598-642. https://doi.org/10.3102/0034654313496870

Sung, Y. H., Chen, L. M., Yen, C. F., & Valcke, M. (2018). Double trouble: The developmental process of school bully-victims. *Children and Youth Services Review*, *91*, 279–288. https://doi.org/10.1016/j.childyouth.2018.06.025

ten Bokkel, I. M., Stoltz, S. E. M. J., van den Berg, Y. H. M., Orobio de Castro, B., & Colpin, H. (2021). Speak up or stay silent: Can teacher responses towards bullying predict victimized students' disclosure of victimization? *European Journal of Developmental Psychology, 18*(6), 831–847. https://doi.org/10.1080/17405629.2020.1863211

Troop-Gordon, W., & Quenette, A. (2010). Children's perceptions of their teacher's responses to students' peer harassment: Moderators of victimization-adjustment linkages. *Merrill-Palmer Quarterly*, *56*(3), 333–360. https://doi.org/10.1353/mpq.0.0056

Troop-Gordon, W. & Ladd, G. W. (2015). Teachers' victimization-related beliefs and strategies: Associations with students' aggressive behavior and peer victimization. *Journal of Abnormal Child Psychology, 43*(1), 45–60. https://doi.org/10.1007/s10802-013-9840-y

Troop-Gordon, W., Chambless, K., & Brandt, T. (2021a). Peer victimization and well-being as a function of same-ethnicity classmates and classroom social norms: Revisiting person × environment mismatch theory. *Developmental Psychology, 57*(12), 2050–2066. https://doi.org/10.1037/dev0001161

Troop-Gordon, W., Kaeppler, A. K., & Corbitt-Hall, D. J. (2021b). Youth's expectations for their teacher's handling of peer victimization and their socioemotional development. *The Journal of Early Adolescence*, *41*(1), 13-42. https://doi.org/10.1177/0272431620931192

van Aalst, D. A. E., Huitsing, G., & Veenstra, R. (2022). A systematic review on primary school teachers' characteristics and behaviors in identifying, preventing, and reducing bullying. *International Journal of Bullying Prevention, 4*(3), 124–137. https://doi.org//10/s42380-022-00145-7

Van Aalst, D. A. E., Huitsing, G., & Veenstra, R. (2024). Understanding teachers' likelihood of intervention in bullying situations: Testing the theory of planned behavior. *International Journal of Bullying Prevention.* https://doi.org/10.1007/s42380-024-00209-w

Van der Zanden, P. J., Denessen, E. J., & Scholte, R. H. (2015). The effects of general interpersonal and bullying-specific teacher behaviors on pupils' bullying behaviors at school. *School Psychology International*, *36*(5), 467–481. https://doi.org/10.1177/0143034315592754

van Gils, F. E., Colpin, H., Verschueren, K., Demol, K., ten Bokkel, I. M., Menesini, E., & Palladino, B. E. (2022). Teachers' responses to bullying questionnaire: A validation study in two educational contexts. *Frontiers in Psychology, 13*, 830850. https://doi.org/10.3389/fpsyg.2022.830850

van Gils, F., Verschueren, K., Demol, K., Bokkel, I. & Colpin, H. (2023). Teachers´ bullying-related cognitions as predictors of their responses to bullying among students. *British Journal Educational Psychology, 93*(5), 513-530. https://doi.org/10.1111/bjep.12574

van Gils, F. E., Demol, K., Verschueren, K., ten Bokkel, I. M., & Colpin, H. (2024). Teachers' responses to bullying: A person-centered approach. *Teaching and Teacher Education*, *148*, 104660. https://doi.org/10.1016/j.tate.2024.104660

Van Verseveld, M. D. A., Fukkink, R. G., Fekkes, M., & Oostdam, R. J. (2019). Effects of antibullying programs on teachers' interventions in bullying situations: A meta-analysis. *Psychology in the Schools*, *56*(10), 1522–1539. https://doi.org/10.1002/pits.22283

Wachs, S., Bilz, L., Niproschke, S., & Schubarth, W. (2019). Bullying intervention in schools: A multilevel analysis of teachers' success in handling bullying from the students' perspective. *Journal of Early Adolescence*, *39*, 642–668. https://doi.org/10.1177/0272431618780423

Wheaton, B., Muthen, B., Alwin, D. F., & Summers, G. F. (1977). Assessing reliability and stability in panel models. *Sociological methodology*, *8*, 84-136.

Weiss, B. A. (2011). *Reliability and validity calculator for latent variables* [Computer software]. George Washington University. https://blogs.gwu.edu/weissba/teaching/calculators/reliability-validity-for-latent-variables-calculator/

Xiao, M., & Hooi, L. B. (2024). Teacher interventions on Bullying: Mediating Role of Classroom Climate. *Advances in Education, Humanities and Social Science Research, 12*(1) 277-289. https://doi.org/10.56028/aehssr.12.1.277.2024

Yoon, J., Bauman, S., Choi, T., & Hutchinson, A. S. (2011). How South Korean teachers handle an incident of school bullying. *School Psychology International*, *32*(3), 312-329. https://doi.org/10.1177/0143034311402311

Yoon, J. S., & Bauman, S. (2014). Teachers: A critical but overlooked component of bullying prevention and intervention. *Theory Into Practice, 53*(4), 308–314. https://doi.org/10.1080/00405841.2014.947226

Yoon, J., Sulkowski, M. & Bauman, S. (2016): Teachers´ responses to bullying incidents: Effects of teacher characteristics and contexts. *Journal of School Violence, 15*(1), 91-113. https://doi.org/10.1080/15388220.2014.963592

Zych, I., Farrington, D. P., & Ttofi, M. M. (2019). Protective factors against bullying and cyberbullying: A systematic review of meta-analyses. *Aggression and violent behavior, 45*, 4-19. https://doi.org/10.1016/j.avb.2018.06.008

Psicothema

Volumen 38, no. 1

PUBLISHED BY

MEMBER OF

SPONSORED BY