

# **Psicothema**



**Volume 37, no. 3**

**ISSN: 0214-9915 • eISSN 1886-144X**

**Colegio Oficial de Psicología del Principado de Asturias**



**Editor-in-Chief:** Laura E. Gómez. Univ. de Oviedo, Spain  
**Deputy Editor:** Rebeca Cerezo. Univ. de Oviedo, Spain  
**Associate Editors:** Susana Al-Halabi. Univ. de Oviedo, Spain  
 Jorge L. Arias, 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  
**Honorary Editor:** José Muñiz, Univ. Nebrija, Spain

**Managing Editor:** Gloria García-Fernández, Univ. de Oviedo, Spain  
**Editorial Office:** Leticia García, COPPA, Spain  
 M.ª Ángeles Gómez, COPPA, Spain

## ADVISORY BOARD

Olivia Afonso. Oxford Brookes Univ., UK  
 Leandro Almeida. Univ. do Minho, Portugal  
 David Álvarez. Univ. de Oviedo, Spain  
 Marta Álvarez Cañizo. Univ. de Valladolid, Spain  
 Rui Alves. Univ. of Porto, Portugal  
 Antonio M. Amor González. Univ. de Salamanca, Spain  
 Constantino Arce. Univ. de Santiago, 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  
 Elisardo Becoña. Univ. de Santiago, Spain  
 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  
 Juan L. Castejón. Univ. de Alicante, Spain  
 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  
 Concepción Fernández. 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  
 Antonio León García Izquierdo. Univ. de Oviedo, 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  
 Eve Griffin. National Suicide Research Fdn., Ireland  
 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  
 Valentina Ladera. Univ. de Salamanca, Spain  
 Christa Labouliere. Columbia University, USA  
 Juan Alfonso Lara Torralbo, Univ. de Córdoba, Spain  
 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  
 Urbano Lorenzo Seva. Univ. Rovira i Virgili, Spain  
 Verónica Marina Guillén. Univ. de Cantabria, Spain  
 Eduardo Martín Cabrera. Univ. de La Laguna, Spain  
 Víctor Martínez Loreda, Univ. de Sevilla, Spain  
 Ana Miranda. Univ. de Valencia, Spain  
 Santiago Monleón Verdú. Univ. de Valencia, Spain  
 Fabia Morales Vives. Univ. Rovira i Virgili, Spain  
 M.ª 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 Paino. Univ. de Oviedo, Spain  
 Alicia Pérez de Albéniz. Univ. de La Rioja, Spain  
 M.ª Carmen Pérez Fuentes. Univ. de Almería, Spain  
 Gina Quirarte. UNAM, Mexico  
 Celestino Rodríguez. Univ. de Oviedo, Spain  
 Susana Rodríguez, Univ. da Coruña, Spain  
 Javier Rodríguez Ferreiro. Univ. de Barcelona, Spain  
 M.ª Fe Rodríguez Muñoz. UNED, Spain  
 Sonia Romero. Univ. Aut. de Madrid, Spain  
 Pedro Rosário. Univ. do Minho, Portugal  
 M.ª Consuelo Sáiz Manzanares. Univ. de Burgos, Spain  
 Louis A. Sass. Univ. Rutgers, USA  
 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  
 Linda C. Sobell. Univ. Nova Southeastern, 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  
 Andreu Vigil. Univ. Rovira i Virgili, Spain  
 Jianzhong Xu. Mississippi State Univ., USA  
 Jin H. Yoon. Univ. of Texas, USA  
 Izabela Zych. Univ. de Córdoba, Spain

**Psicothema** is indexed by Social Sciences Citation Index (WOS), Scopus, Google Scholar, SciELO, Dialnet, EBSCO Essentials Academic Search Premier, PubMed, PsycINFO, IBECs, Redinet, Psycodoc, 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  
 Ildelfonso 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>



## 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](http://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. Please 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](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](mailto: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 than 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 ResearcherID, 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 sent by email to [psicothema@cop.es](mailto: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 change 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, postprint, and editorial version (immediately after publication). Exploitation rights and Psicothema's self-archiving permissions may be found at <https://dulginea.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, Psycodoc, 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

Navigating multi-language assessments: Best practices for test development, linking, and evaluation Louise Badham, María Elena Oliveri and Stephan G. Sireci.....	1-11
When to use bootstrap- <i>F</i> in one-way repeated measures ANOVA: Type I error and power María J. Blanca, Roser Bono, Jaume Arnau, F. Javier García-Castro, Rafael Alarcón and Guillermo Vallejo .....	12-22
ChatGPT simulated patient: Use in clinical training in Psychology Ana Sanz, José Luis Tapia, Eva García-Carpintero, J. Francisco Rocabado and Lorena M. Pedrajas.....	23-32
Emotionally tough, sexting rough: Relationship between callous unemotional traits and aggravated sexting in 11 countries Mara Morelli, Fau Rosati, Elena Cattelino, Flavio Urbini, Roberto Baiocco, Dora Bianchi, Fiorenzo Laghi, Maurizio Gasseau, Piotr Sorokowski, Michal Misiak, Martyna Dziekan, Heather Hudson, Alexandra Marshall, Thanh Truc Nguyen, Lauren Mark, Kamil Kopecky, René Szotkowski, Ezgi Toplu Demirtaş, Joris Van Ouytsel, Koen Ponnet, Michel Walrave, Tingshao Zhu, Ya Chen, Nan Zhao, Xiaoqian Liu, Alexander Voiskounsky, Nataliya Bogacheva, Maria Ioannou, John Synnott, Kalliopi Tzani-Pepelasis, Vimala Balakrishnan, Moses Okumu, Eusebius Small, Silviya Pavlova Nikolova, Michelle Drouin, Alessandra Ragona and Antonio Chirumbolo.....	33-44
Assessing positive organizational culture: Psychometric properties of the POCS Javier Barría-González, Jaime García-Fernández, Ricardo Pérez-Luco and Álvaro Postigo.....	45-53

Article

# Navigating Multi-Language Assessments: Best Practices for Test Development, Linking, and Evaluation

Louise Badham<sup>1</sup> , María Elena Oliveri<sup>2</sup>  and Stephan G. Sireci<sup>3</sup> 

<sup>1</sup>International Baccalaureate (United Kingdom)

<sup>2</sup>Purdue University (USA)

<sup>3</sup>University of Massachusetts Amherst (USA)

## ARTICLE INFO

Received: 15/10/2024

Accepted: 20/03/2025

### Keywords:

Comparability  
Cross-lingual assessment  
Linking tests  
Test translation  
Validity

## ABSTRACT

**Background:** Developing assessments in multiple languages is hugely complex, impacting every stage from test development to scoring, and evaluating scores. Different approaches are needed to examine comparability and enhance validity in cross-lingual assessments. **Method:** A review of literature and practices relating to different methods used in cross-lingual assessment is presented. **Results:** There has been a shift from source-to-target language translation to developing items in multiple languages simultaneously. Quantitative and qualitative methods are used to link and evaluate assessments across languages and provide validity evidence. **Conclusions:** This article provides practitioners with an overview and research-based recommendations relating to test development, linking, and validation of assessments produced in multiple languages.

## Evaluaciones Multilingües: Mejores Prácticas para su Desarrollo, Vinculación y Evaluación

## RESUMEN

**Antecedentes:** El desarrollo de evaluaciones en varios idiomas es enormemente complejo y afecta a todas las etapas, desde el desarrollo de las pruebas hasta la puntuación y la evaluación de las puntuaciones. Se necesitan diferentes enfoques para examinar la comparabilidad y mejorar la validez de las evaluaciones interlingües. **Método:** Se presenta una revisión de la literatura y las prácticas relacionadas con los métodos utilizados en diferentes áreas de la evaluación interlingüística. **Resultados:** Se ha pasado de la traducción del idioma de origen al idioma de destino al desarrollo simultáneo de ítems en varios idiomas. Se utilizan métodos cuantitativos y cualitativos para vincular y evaluar las evaluaciones en varios idiomas y proporcionar pruebas de validez. **Conclusiones:** Este artículo proporciona a los profesionales una visión general y recomendaciones de la literatura relacionada con el desarrollo de pruebas, la vinculación y la validación de evaluaciones producidas en varios idiomas.

### Palabras clave:

Comparabilidad  
Evaluación interlingüística  
Pruebas de vinculación  
Traducción de exámenes  
Validez

## Introduction

Developing assessments in multiple languages is complex, impacting test development, scoring, and evaluating results. Ensuring fairness and validity across languages requires considering linguistic structures, sociolinguistic factors, and educational policies that shape assessment outcomes in diverse global settings. For instance, as the second most spoken language in the United States, Spanish versions of large-scale assessments are designed to accommodate emergent bilingual students. However, linguistic differences between English and Spanish, including verb conjugation complexity and sentence structure require careful adaptation to ensure construct equivalence. Additionally, regional dialectal differences among Spanish speakers from Spain, Mexico, the Caribbean, and Central and South America must be addressed to avoid cultural bias.

Multilingual assessment practices in other global contexts further highlight the need for tailored approaches. In Canada, where English and French are official languages, assessments must ensure validity across linguistic groups while accounting for language-specific conceptual distinctions. In sub-Saharan Africa, where indigenous languages coexist with colonial languages (English, French, and Portuguese), educational policies influence assessments in local languages versus global languages. Test developers must navigate complex decisions about language prioritization, given variations in literacy levels and educational access. Considerations are particularly complex in international large-scale assessments (ILSAs) spanning diverse linguistic and cultural contexts.

This article reviews cross-lingual assessment methods and practices, providing guidance in identifying and mitigating biases against linguistic or cultural groups. Bias in assessments can disadvantage certain groups, making it crucial to consider global and local linguistic variations in translations (van de Vijver & Poortinga, 2005). Such bias can have far-reaching social consequences, such as limiting educational outcomes or career progression. Therefore, assessment development and evaluations must be robust and rigorous to minimize bias and allow valid score-based inferences (Ercikan & Lyons-Thomas, 2013).

Our review covers: (a) multi-language test development; (b) “linking” different language versions of assessments; and (c) evaluating results of cross-lingual assessments. Our review extends previous work by Sireci et al. (2016)—which reviews some of the same sources—by connecting linking and evaluation methods with test development approaches and exploring new techniques from emerging technologies. Whilst the scope of the study was limited to cognitive skills in multi-language educational assessments, the issues addressed apply across multiple settings, such as personnel selection in multinational corporations.

## Adapting Tests Across Languages

As highlighted in the *Standards for Educational and Psychological Testing*, “simply translating a test from one language to another does not ensure that the translation produces a version of the test that is comparable in content and difficulty” (AERA et al., 2014, p.60). Therefore, whilst *translation* is often used to describe the process of adjusting tests into other languages, the term *adaptation* is preferable (Hambleton, 2005; ITC, 2017). Adaptation reflects that the process accounts for cultural relevance, aiming to maximize validity in target language assessments (Ercikan & Por, 2020). *Transadaptation* is also

used, but is redundant, having the same definition as adaptation. Thus, “adaptation” is used here, although “translation” is sometimes used interchangeably, or to describe part of the adaptation process.

## Approaches to Developing Multi-Language Tests

Approaches include: (a) *adapting* tests from one (*source*) language to another (*target*) language(s); (b) *simultaneous development*, where multilingual teams create and adapt items together; and (c) *parallel development*, where each language version is developed separately.

(*Successive*) *adaptation* involves developing a monolingual, source-language test version, which is translated into one or more target languages (Rogers et al., 2003). The process may include “back-translation” (Brislin, 1970), where tests are translated from source to target language and back again, then source versions are compared to verify whether the original meaning has been retained. *Simultaneous development* is a form of adaptation, but rather than developing a source language first, multilingual committees develop and immediately adapt items across languages (Tanzer, 2005). In *parallel development*, content is developed independently in each language according to common specifications. Rather than using common items, the approach to defining and representing constructs on multi-language assessments is designed to be comparable. Some items may also be adapted to maximize construct comparability (Ercikan & Lyons-Thomas, 2013).

## Linking and Comparing Tests Across Languages

Cross-lingual assessment literature discusses different levels of “equivalence” or score comparability. Examples include “structural,” “metric,” and “scalar” equivalence, which sit within the broader areas of *linking* or *equating* test scores (Sireci et al., 2016). In equating, scores from different test forms are adjusted and placed onto a common scale and can theoretically be considered interchangeable (Lord, 1980). Equating requires measurement of the same construct, and that tests are developed from common content specifications. Adapted tests typically involve the same construct and content specifications. However, translated content cannot be considered “common”, so strict equating of translated tests is impossible (Dorans & Middleton, 2012; Sireci, 1997). “Weaker” forms of equating known as “linking” have fewer assumptions and are usually sought for multi-language assessments (Sireci, 1997; Sireci et al., 2016).

Sireci (1997) identified three cross-lingual linking designs: (a) *separate monolingual groups* where each group takes the language version it was developed for; (b) *matched monolingual groups* where examinees from different languages are matched on external criteria (e.g., socioeconomic status) rather than using anchor items; or (c) *bilingual groups* where bilingual examinees either take both language versions, or are randomly assigned one language. Related to linking, are methods that *evaluate comparability* across multi-language assessments (Sireci et al., 2016). Rather than seeking to establish a relationship between assessments, these approaches examine whether tests measure the same construct in the same way across language groups. The most common method is differential item functioning (DIF) (Zumbo, 2015), which examines how different groups with similar abilities respond to the same items.

DIF can, for example, be used to examine cross-language item comparability, or cross-country variations arising from cultural differences (Ercikan, 2002).

## International Guidelines

The International Test Commission (ITC) provides internationally recognized guidelines on assessment practices, such as *Guidelines for Translating and Adapting Tests* (ITC, 2017). For emerging technology-based approaches, *Guidelines for Technology-Based Assessments* also aim “to ensure fair and valid assessment in a digital environment” (ITC & Association of Test Publishers, 2022, p.1). Informed by these guidelines, our review explored how cross-lingual assessment theory and methods have been applied in practice.

### Method

#### Search Parameters

To identify relevant literature, we used the keywords: “translation,” “adaptation,” “transadaptation,” “cross-lingual,” and “dual-language,” combined with “test,” and “assessment.” Literature citing “ITC” was also included. Citation histories for key articles were reviewed to identify influential research and emerging themes. Grey literature from international assessment organizations (e.g., adaptation guidelines) was included to illustrate multi-language assessment practices. Key publications are provided in the References.

#### Search Process

Our multi-step search strategy combined database searches with manual reviews of high-impact journals. We searched four major academic databases: Educational Resource Information Center (ERIC), PsycInfo, Web of Science, and EBSCO to ensure broad coverage of peer-reviewed literature on test adaptation, translation, and cross-lingual validity. ERIC was particularly relevant for educational research, while PsycInfo captured studies on psychological and linguistic aspects of assessment. Web of Science and EBSCO broadened disciplinary scope, ensuring that emerging research trends in related fields were considered.

We compiled a vetted bibliography of studies that met our inclusion criteria, then conducted a secondary manual review of top-ranked journals in educational measurement, assessment, and cross-cultural psychology (e.g., *International Journal of Testing*). Journals were selected based on impact factor, citation influence, and reputations for publishing high-quality research in test adaptation and multilingual assessment. Combining systematic database searches with a targeted review of high-impact journals allowed us to capture broad trends and in-depth discussions from specialized sources. By integrating diverse methodological perspectives, our review reflected both empirical research and theoretical insights valuable for practitioners and researchers in multilingual assessments.

#### Screening and Filtering

An initial 618,761 records were filtered by publication type (journal), field of study (education and related disciplines), and language (English), to 51,744. Selections were further streamlined by reviewing citation counts to prioritize highly cited and influential

studies, while recognizing that recent publications (2022–2024) may have lower citation counts. Publications that did not contribute new information were excluded due to saturation. Selected publications prioritized studies that contributed new theoretical, methodological, or empirical knowledge relevant to multilingual assessment adaptation. These were categorized according to key cross-lingual assessment areas (Table 1).

We read, analyzed, and synthesized selected publications to extract key themes, methodologies, and best practices related to multi-language assessments.

**Table 1**  
*Sources Reviewed*

Methods used to...	# Publications
(a) create exams for use in multiple languages	16
(b) adapt exams from one language into other languages	17
(c) link different language versions of exams	23
(d) evaluate comparability of scores in multi-language exams	24

## Results

Selected publications were stratified by their focus with respect to: (a) “Test Development,” involving methods for creating or adapting multi-language assessments; (b) “Test Score Linking,” comprising cross-lingual linking methods; and (c) “Evaluating Comparability” including measurement invariance studies at test and item-levels, and computational linguistic techniques.

### Test Development

The test adaptation literature spans over 50 years and includes discussions of the pros and cons of different models for developing multi-language tests (e.g., Hambleton, 1994; van der Vijver & Tanzer, 1997). “Test development” encompasses *adapting* and *creating* multi-language tests, as both refer to processes of constructing assessments for use in different languages. Different test development models are presented here, with examples of ILSAs that illustrate them in practice.

#### Adaptation

Adaptation is the most common approach to developing multi-language assessments (Ercikan & Por, 2020), and is used for virtually all ILSAs (Ebbs & Koršňáková, 2016). Multi-language ILSAs using adaptation include the Programme for International Student Assessment (PISA) (OECD, 2016, 2024), the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS) (Ebbs et al., 2021; Ebbs & Koršňáková, 2016; Martin et al., 2020).

TIMSS and PIRLS are delivered in over 50 languages around the world (Ebbs et al., 2021, Martin et al., 2020), using a decentralized translation approach. National research or study centers adapt assessments into national language(s) following agreed procedures. Translator(s) and reviewer(s) must have experience of the cultural context and working with students in the target demographic, which helps mitigate risks of (dis)advantaging respondents from using direct translations. Koršňáková et al. (2020) illustrated this approach in an Arabic version of TIMSS for Middle East and North African countries, which accommodated cultural and regional

variations in language. Translators from different Arabic-speaking countries each produced initial translations, and a reviewer cross-checked translations to select the best version. Finally, an expert panel reviewed and refined the translation for the target audience. This approach helped ensure cross-linguistic, cross-national, and cross-cultural equivalence, without which one cannot achieve a quantitative cross-cultural comparison (Dept et al., 2017; ITC, 2017; Koršňáková et al., 2020).

PISA is also delivered globally in over 50 languages (OECD, 2024). Assessments are first developed in English and French source versions (Grisay et al., 2009; OECD, 2016, 2024), which are both translated into the target language, before being reconciled into a final, target-language version. This method helps identify linguistic discrepancies during adaptation. Some assessments are cross-checked against other verified language versions to increase cultural relevance (e.g., Catalan, Galician and Basque compared to Spanish versions (OECD, 2024)). Back-translation is also used (ibid), which is a useful quality control check (ITC, 2017). However, idiosyncratic features of source languages translated into target languages can go unnoticed (El Masri et al., 2016), or high-quality back-translations may mask issues from poorer quality initial translations (Koršňáková et al., 2020). Grisay (2003) demonstrated double-translation's advantages over back-translation in a PISA reading passage, where irony was lost in literal translations—an issue identified in double-translation reconciliation, but would likely have been missed in back-translation.

### Challenges in Adaptation

Zhao et al. (2018) created a typology of language translation errors in PISA items to examine characteristics of specific source-target language combinations. In reviewing error types from English-to-Spanish translations, they found one required modification, 14 were eliminated, and 11 new error types were identified in English-to-Chinese translations. Different error types also occurred when translating science versus mathematics items, indicating different content areas create different translation challenges. Thus, different translation approaches may be needed for different content areas or language combinations.

When investigating cross-lingual comparability in PISA science items, El Masri et al. (2016) observed different word frequencies can make “common” words more challenging in some languages than others. They exemplified this with “crescent moon.” “Crescent” is a high-frequency word in French (due to the famous pastry) and Arabic (being a symbol of Islam, the dominant religion in Arabic-speaking countries), but low-frequency—so more cognitively challenging—in English. Additionally, they observed biases can arise from inherent linguistic complexities and “untranslatable language idiosyncrasies” (p. 440), rather than any fault in adaptation processes. For example, “abbreviation incongruence” (p. 444) can occur where universal Latin script abbreviations (e.g., chemical elements) are used in non-Latin script assessments (e.g., Arabic), placing an additional cognitive burden on students in those languages. Similarly, Lu and Sireci (2007) identified “differential speededness” in translated assessments, where some languages require more words than others to express the same meaning. Consequently, examinees may require more time to take assessments in some languages than others.

### Simultaneous Development

To mitigate issues arising from linguistic and cultural differences, checklists (e.g., Hambleton & Zenisky, 2011) can support quality assurance in adaptation. Additionally, linguistic and cultural considerations can be integrated directly into adaptation. *Simultaneous development* involves developers from different languages and cultures throughout test development, thereby ensuring “maximum linguistic and cultural decentering” in the process (Tanzer, 2005, p.238). Linguistic and cultural nuances, such as dialectical differences, can be identified during adaptation. For example, “aula” (classroom) is used in U.S. Spanish and Spain, but “salón” is preferred in Mexico. Such subtle differences could impact comprehension and familiarity, especially for students in specific educational settings. In addition to improving cultural relevance, Rogers et al. (2010) suggested integrating such information from different linguistic and cultural groups can make test development more efficient, by minimizing review time later in the process.

### Parallel Development

Parallel development is a relatively rare approach. One example is International Baccalaureate (IB) exams, which are offered in up to 75 languages (IBO, 2024). Whilst many IB assessments (e.g., Sciences) are adapted across languages, others including Literature are created in parallel. Cross-lingual comparability is supported through common test specifications providing guidance on content, cognitive areas to be measured, quantity and format of items (IBO, 2018). Comparability is further enhanced through translation templates and cross-language “assessment editing meetings”, where authors from different languages review and discuss exam drafts together to align standards (Sireci & Oliveri, 2023). Parallel development naturally removes many challenges inherent in adaptation—including translation errors and untranslatable language idiosyncrasies—since each language is developed independently, rather than representing a source version.

The *ITC Guidelines for Translating and Adapting Tests* (2017) specify the legitimacy of assessing constructs across cultural/linguistic groups must be established in multi-language assessments. In their checklist for operationalizing ITC guidelines, Hernández et al. (2020) suggested different test versions are preferable where constructs are not generalizable across populations. Parallel approaches may be more suitable in these situations. Yet, with different content in each language, fewer statistical methods are available to investigate comparability (Badham & Furlong, 2023), thus parallel development represents “a compromise between comparability and cultural authenticity” (Ercikan & Lyons-Thomas, 2013, p.552).

### GenAI and Multi-Language Test Development

Recently, dramatic surges in generative AI (GenAI) have presented innovative opportunities for developing multi-language assessments. Duolingo, with over 40 languages offered through its language learning app (Blanco, 2024) offers an example. Goodwin et al. (2023) demonstrated how GenAI could support multi-language item development simultaneously and at scale. Expert judgement and open-source corpora were used to train multilingual large language models (LLMs) on extensive word-sets to create prototype Duolingo listening and reading items in French



and Spanish. GenAI has the potential to transform multilingual test development practices, by improving efficiencies for large-scale assessments, decreasing costs, and reducing workload (Hao et al., 2024). AI-supported automated item generation can “mitigate security risks and avoid overexposure of test content” (ITC, 2022, p.135), but also faces challenges including copyright and intellectual ownership (Hao et al., 2024). Additionally, there are validity concerns, as LLMs can reflect bias inherent in internet-scraped data, and models may be “biased for or against particular groups and/or produce poor outputs in under-represented languages” (ibid, p.26). Despite current limitations, GenAI offers enormous potential for supporting multi-language test development.

### Test Score Linking

In many cases, score scales from multilingual assessments are linked in some fashion to facilitate interpretations and comparisons. There are several appropriate cross-lingual linking and data collection designs (Table 2) that are helpful for specific assessment contexts (ITC, 2017). We briefly describe these designs next.

In *separate monolingual groups* designs, links are formed using anchor items assumed to be comparable across languages. This assumption is generally verified via statistical analyses of DIF across languages (i.e., items not flagged for DIF are used as anchor items). However, “this justification is somewhat circular, because DIF analyses assume the variable on which examinees are matched is free of construct and method bias” (Sireci et al., 2016, p.189). Thus, it does not rule out the possibility of unidirectional bias (e.g., all items are more difficult in one language). Nevertheless, it is a helpful validation check, and has been used to link scores, or evaluate cross-lingual comparability using item response theory (IRT) (e.g., OECD, 2024).

Separate monolingual groups are used to link Psychometric Entrance Tests (PET), high-stakes college admissions exams in Israel. PET verbal and quantitative reasoning tests are adapted from Hebrew into five other languages, using adaptation and parallel development. Quantitative reasoning items are translated and DIF procedures conducted for each language using Hebrew as the reference group. However, vocabulary and analogy items are constructed uniquely in each language (Allalouf et al., 2009), since they are too different across languages to be translated (Allalouf et al., 1999). Cross-lingual DIF analyses are conducted to select items to link scales across languages. Items comprising linking anchors must demonstrate a correlation of  $>0.80$  with respect to their item difficulty parameters across languages. This is less stringent than most equating studies (e.g., Kolen & Brennan, 2004), illustrating the lower level of linking being conducted (Sireci et al., 2016). The PET example demonstrates

tests using parallel development may be statistically linked using adapted items as linking anchors.

A qualitative variation on separate monolingual groups is *social moderation*. Social moderation is the lowest level of linking, where expert judgement is used to form a link or common standard of achievement across languages. IB “cross-language standardization meetings” are an example, where examiners from parallel language versions discuss and align marking standards together (Sireci & Oliveri, 2023). Similarly, Davis et al. (2008) convened separate language panels of experts to set pass/fail standards on English and French high school reading and writing tests. The standards set on each exam resulted in about 1–6% differences in pass rates, illustrating that parallel standard setting processes using social moderation may be used to set credible standards on parallel language versions. Whilst parallel cross-lingual assessments can be linked using social moderation, this offers a “weaker” level of linking (Sireci et al., 2016) compared to methods such as IRT.

Matched monolingual groups designs have been used rarely (e.g., Milman et al., 2018), due to challenges identifying valid external criteria that can be considered equivalent across language groups, and exhibit sufficient overlap for matching purposes. Bilingual groups have been used in small-scale contexts, but have limited practical applicability in large-scale assessments due to limited availability of bilingual examinees. Ong and Sireci (2008) used a bilingual design where examinees took English and Malay 9<sup>th</sup> grade math tests in counterbalanced order. Overall, 7 of 40 items were flagged for DIF. Next, they performed linking using several methods including linear, equipercentile, and IRT, both with and without using DIF items as part of the equating anchor. The equating resulted in a 2-point adjustment across languages with DIF items included, and a 1-point difference without them. Such results underscore the need to screen items for DIF prior to linking (ibid; Sireci et al., 2016).

### Evaluating Comparability

Considerable literature focuses on evaluating comparability (measurement invariance) across adapted tests, both at item-(using DIF) and test-score levels (using dimensionality procedures). Many focus on ILSAs such as PISA, TIMSS and PIRLS. Comparability studies have different purposes, including establishing measurement equivalence to justify linking procedures, or evaluating adaptation processes by identifying translation issues.

### Evaluating Invariance Across Languages

Rapp and Allalouf (2003) used a *double-linking plan*—where a test form is equated to two other forms—to evaluate whether

**Table 2**  
Cross-Lingual Linking Designs (adapted from Sireci, 1997)

Design	Assumptions	Examples
Separate monolingual groups	No systematic method bias exists across all items, which justifies DIF analyses	Allalouf et al. (2009); Angoff & Cook (1988); Hulin et al. (1982); Hulin & Mayer (1986); OECD (2024); Woodcock & Muñoz-Sandoval (1993)
Matched monolingual groups	Valid matching criteria sufficiently account for group differences. Overlap of distributions on these criteria are sufficient for matching	Milman et al. (2018)
Bilingual groups	Bilingual examinees are sufficiently representative of monolingual groups, with roughly equal proficiency across languages	Boldt (1969); Cascallar & Dorans (2005); CTB-McGraw Hill (1988); Ong & Sireci (2008); Sireci & Berberoglu (2000); Sukin et al. (2015)

the PET linking process introduced equating error. When double-linking studies are conducted in a single language, typically the two separate equating results are averaged. However, Rapp and Allaouf used within-language equating to establish a baseline equating error, before comparing it to the cross-lingual equating error. The verbal test contained a pair of parallel sections, which could be equated within each target language using a common person design (same-language equating), and to their Hebrew (source language) counterparts using linking items. Rapp and Allaouf assumed differences between the within-language and across-language equating results would reflect the instability associated with their cross-lingual linking. The average equating difference across test forms in the first target language was about ten times that observed for within-language equating forms. They concluded the within- and across-language double-linking design was useful for evaluating cross-lingual linking stability, hypothesizing various reasons for instability, including translation differences, cultural familiarity, item position effects, and different anchor test lengths.

As Sireci et al. (2016) highlight, the PET research illustrates how cross-lingual linking has been evaluated on high-stakes tests. Lower-stakes tests, including TIMSS and PISA use similar approaches (i.e., translated items, DIF-screening, and common-item linking). The approach has limitations, “as the viability of the linking anchor cannot be unequivocally established. The linking anchor may have items that differ across languages but escape DIF detection, or it may underrepresent the construct the test is designed to measure” (ibid, p.191). Allalouf et al. (2009) questioned whether “a superior, no-DIF link with an inferior representation of content” was preferable, “or an inferior link (that includes some DIF items) with a superior representation of content” (p.105).

### *Assessing Invariance of Dimensionality*

Multi-group confirmatory factor analysis (MGCFA) has been widely used to evaluate construct bias and score comparability in cross-lingual assessment (Davidov, 2011; van de Vijver et al., 2019), partly because it can handle multiple groups in a single analysis. Multidimensional Scaling (MDS) has also frequently been used to explore comparability from a dimensionality perspective in cross-lingual research (e.g., Robin et al., 2003; Wolff et al., 2011). MDS is useful as data from multiple groups can be analyzed concurrently to determine the structural similarities (and differences) across groups by using an individual differences MDS analysis and evaluating the group weights to modify the common structure for each group (Sireci, 2005; Sireci & Wells, 2010; Sireci et al., 2016). Asparouhov and Muthén’s (2014) alignment procedure can also accommodate multiple groups (e.g. countries and languages), and is more flexible than MGCFA, as it does not require parameters to be exactly equal across groups (van de Vijver et al., 2019). It accommodates partial invariance, by seeking patterns of parameter estimates that allow small variations between parameters, but only minimal large differences. The estimation stops when the overall amount of non-invariant parameters is minimised, providing “the best possible comparability that can be achieved with the given data” (ibid, p.16).

As discussed in Sireci et al. (2016), an advantage of MDS over MGCFA is its exploratory nature, so dimensionality of the assessment

does not need to be specified in advance. This is helpful when the dimensionality is “unknown, or the hypothesized dimensionality is not widely supported” (ibid, p.193). The disadvantage is that MDS is solely descriptive—it provides no statistical test to evaluate structural differences across groups (Fischer & Fontaine, 2011)—necessitating reliance primarily on visual interpretations and descriptive indices (Sireci et al., 2016). Dimensionality structure across cultures is not sufficient to ensure score comparability. Therefore, MDS is useful for exploratory purposes, but on its own, is insufficient for establishing measurement equivalence to justify linking procedures. The strictness of MGCFA—requiring exact equality of parameters across groups—is also a drawback, since this requirement is rarely met in practice (van de Vijver et al., 2019). The flexibility of the alignment procedure makes it more practical than MGCFA for real-life data analysis. Whilst a useful investigative technique, as it allows partial invariance, it is insufficient for establishing measurement equivalence. Therefore, such methods are typically combined with procedures like DIF to justify linking.

### *Evaluating Invariance of Adapted Items*

DIF procedures are commonly used to evaluate cross-lingual comparability at item-level, often combined with structural equivalence or qualitative analyses to help interpret cross-lingual differences. Grisay et al. (2009) evaluated deviations of item difficulty parameters for countries participating in PISA and PIRLS reading assessments, from “international” item parameters using the global population. Despite a large commonality across the global and country-specific item difficulty parameters and only a modest level of DIF on average, higher magnitudes of DIF were noted for non-Indo-European languages (e.g., Arabic and Chinese). Gökçe et al. (2021) also investigated whether DIF on TIMSS math was associated with differences between language families and cultures. They compared DIF across three language-country combinations: (a) same language, but different countries, (b) same countries, but different languages, and (c) different languages and countries. With more distant cultures and language families, the presence of DIF increased. The magnitude of DIF was greatest when both language and country differed, and smallest when languages were same, but countries were different.

Ercikan and Koh (2005) investigated English and French TIMSS math and science assessments across countries using DIF and structural equivalence using MGCFA. There was a lack of equivalence at both structural and item-levels, with substantial DIF found in some comparisons (e.g., 79% of science items flagged for DIF across France and the U.S.). The global fit indices associated with the MGCFA illustrated relatively worse fit of the models to the data where the greatest amount of DIF was observed. Ercikan and Koh cautioned against making cross-lingual comparisons when substantial DIF and inconsistencies in test structure are observed across translated assessments. Similarly, Oliveri et al. (2012) evaluated item- and test-level comparability of English and French PISA mathematics problem-solving subtests. Although 3 of 10 items functioned differentially across languages, when aggregating these results to evaluate differential test functioning, they found comparable test characteristic curves, suggesting comparability overall. Their study illustrates the importance of considering invariance at both test and item-levels, as

item-level differences may balance out, causing no apparent effect at test-level (Wainer et al., 1991, Sireci et al., 2016).

Allalouf et al. (1999) followed DIF analyses with qualitative investigations. Hebrew-Russian bilingual content specialists and translators investigated items flagged for DIF in Russian translations of PET verbal reasoning items. They identified four potential causes of DIF: word familiarity and frequency across languages; content changes due to translation; item format; and cultural relevance. Similarly, Gierl and Khaliq (2001) examined English and French, 6th and 9<sup>th</sup>-grade math and social tests, where bilingual content specialists hypothesized potential sources of DIF on flagged items. Translators then categorized items flagged for DIF on a subsequent assessment into the hypothesized categories, illustrating how previously identified sources of DIF could be used to explain subsequently flagged items. The identified sources of DIF also aligned with Allalouf et al. (1999) study, although different languages were involved (Sireci et al., 2016).

### Computational Linguistics

Computational linguistics is increasingly used to investigate cross-lingual differences. El Masri et al. (2016) used computational linguistics to identify linguistic intricacies across languages in PISA science items. They noted idiosyncrasies may be overlooked in expert review-based quality assurance processes, and recommended computational linguistics tools (e.g., Educational Testing Service's *Text-Evaluator Tool*) for evaluating text complexity and identifying differences across translated assessments. Similarly, McGrane et al. (2022) used computational linguistics to examine linguistic complexity across languages in IB science exams. Natural Language Processing (NLP) techniques using a multilingual text processing framework were used to analyze large DIF items across

languages. Differences in linguistic complexity explained up to 11% of DIF results. They recommended that text analysis tools be used during item development to examine item complexity across languages. AI-based NLP techniques can be particularly useful in test development contexts where piloting may be infeasible (e.g. due to reduced timelines) (ITC, 2022).

### Discussion

Our review illustrated different approaches to develop, link, and evaluate cross-language comparability. Adaptation is most common, with iterative, team-based approaches preferred over back-translation. Simultaneous item development helps prevent language prioritization, and identifies cross-lingual and cross-cultural issues during adaptation processes. Parallel development, though rare, is useful when adaptation cannot adequately capture constructs. Emerging GenAI tools show promise but raise concerns over intellectual ownership and potential biases in LLMs.

Test development balances comparability and cultural authenticity (Ercikan & Lyons-Thomas, 2013). Adapted tests enhance comparability through anchor items, but face challenges in translation and ensuring cultural relevance. Parallel development largely removes challenges with translation and language differences, thereby maximizing cultural authenticity. However, with fewer statistical techniques available, comparability and linking are inherently weaker. Hybrid approaches—such as adapting items in parallel tests—offer a compromise between comparability and cultural authenticity, as stronger linking can be established with adapted items as anchors across languages. (e.g. Allalouf, 2009).

Empirical studies have evaluated comparability of dimensionality, items, and achievement level standards from cross-lingual tests (Table 3).

**Table 3**  
*Selected Summary of Comparability Studies*

Citation	Context	Validity Evidence	Statistical Analyses	Findings
Alatli (2020, 2022)	PISA science & reading	Internal structure	DIF, MGCFA	Only structural invariance held. Approx. 35% of science items exhibited DIF due to translation issues; 5 of 7 reading items displayed DIF across China and Turkey.
Allalouf et al. (1999)	PET verbal tests	Internal structure, Test content	DIF	DIF explained by differential difficulty caused by translation, item format, or cultural relevance.
Cascallar & Dorans (2005)	SAT, PAA, & ESLAT	Relations to other variables	Multiple regression	Bilinguals used to compute predicted scores on SAT from PAA and ESLAT.
Davis et al. (2008)	High school reading & writing	Test content	n/a	Setting standards on each test simultaneously using bilingual translators and facilitators to ensure consistent processes across languages.
Ercikan & Koh (2005)	TIMSS math and science	Internal structure	DIF, MGCFA	Structure of assessments was inconsistent across languages in some countries and substantial DIF was found.
Gierl & Khaliq (2001)	Math and social studies tests	Internal structure, Test content	DIF	Bilingual translators and content specialists identified causes of DIF, confirmed by content and statistical analyses on a similar test.
Grisay et al. (2009)	PISA & PIRLS reading	Internal structure	DIF	Greater DIF for non-Indo-European languages.
Gökçe et al. (2021)	TIMSS math	Internal structure	DIF	As differences between language families and cultures increased, observed DIF increased.
McGrane et al. (2022)	IB sciences	Internal structure, Test content	DIF, NLP	Linguistic complexity accounted for up to 11% of variance of DIF.
Oliveri et al. (2012)	TIMSS math	Internal structure	DIF, MGCFA	Whilst 3 of 10 items functioned differentially across languages, DIF did not manifest at test score level.
Rapp & Allalouf (2003)	PET verbal test	Internal structure	Equating analyses	Equating error across language versions was 10x larger than within-language equating error.



Studies focusing on internal structure as sources of validity evidence were most common, using DIF procedures to evaluate item invariance and MGCFA to evaluate structural (dimensional) equivalence. Computational linguistics techniques including text analysis tools offer opportunities for evaluating cross-lingual comparability post-hoc and during test development. Most cross-lingual assessment research indicates many items are differentially difficult across languages, but also that differences are not in one systematic direction, and sufficient comparability likely exists. Some degree of non-invariance must be expected in cross-lingual assessment, as it is unrealistic to assume all items will function equally across all subpopulations (Oliveri & von Davier, 2011, 2014, 2017). Having most, but not all, items from different languages on the same scale is more realistic, and likely sufficient for most comparability needs (ibid). No studies focusing on validity evidence based on testing consequences were found, which is an area recommended for future research.

Adaptation/development approaches have different benefits and drawbacks, including different analyses being available for linking and evaluating comparability (Table 4).

Selection of appropriate multi-language assessment methods depends on the specific context of assessments (e.g. content area, language combinations, or large-scale versus small-scale). The importance of score comparability will always depend on the test purpose, and the decisions and actions taken based on scores. The advantages and challenges for different multi-language development approaches presented here may guide practitioners to choose the most appropriate approach for their contexts. We hope this review, and the many studies referenced, help test developers and evaluators build more valid cross-lingual assessments.

**Table 4**  
*Benefits and Challenges of Multi-Language Development Approaches*

Development approach	Benefits	Challenges
(Successive) adaptation	Stronger link across languages, Established statistical methods to investigate equivalence (e.g. DIF).	Cultural relevance & authenticity, Translation errors, Language differences (e.g. language idiosyncrasies, word frequencies, differential speededness).
Simultaneous development	Stronger link across languages, Established statistical methods to investigate equivalence (e.g. DIF), Linguistic and cultural decentering, Reduced review time.	Language differences (e.g. language idiosyncrasies, word frequencies, differential speededness).
Parallel development	Cultural relevance & authenticity, Removes risk of translation errors, Reduces impact of language differences.	Weaker link across languages, Labour intensive, Harder to investigate comparability statistically.
GenAI	Time efficient, Cost effective, Reduced labour, Reduced security risk, Lowers exposure of test content.	Copyright/intellectual ownership, Risk of bias, Not sufficiently developed in all languages.

## Author Contributions

**Louise Badham:** Conceptualization, Methodology, Project administration, Formal analysis, Writing\_Review and Editing. **Maria Elena Oliveri:** Methodology, Investigation, Formal analysis, Writing\_Review and Editing. **Stephen G. Sireci:** Funding acquisition, Methodology, Investigation, Formal analysis, Writing – original draft.

## Funding

This research was funded by the International Baccalaureate (IB). The first author is an employee of the IB, and participated throughout the study.

## Declaration of Interests

The authors declare there are no conflicts of interest. The views expressed are those of the authors and not to be taken as views of the IB.

## Data Availability Statement

The data supporting this review are available within the cited references.

## References

- Alatli, B. (2020). Cross-cultural measurement invariance of the items in the Science Literacy Test in the Programme for International Student Assessment (PISA-2015). *International Journal of Education and Literacy Studies*, 8(2), 16–27.
- Alatli, B. (2022). An investigation of cross-cultural measurement invariance and item bias of PISA 2018 reading skills items. *International Online Journal of Education and Teaching*, 9(3), 1047–1073.
- Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement*, 36(3), 185–198. <https://doi.org/10.1111/j.1745-3984.1999.tb00553.x>
- Allalouf, A., Rapp, J., & Stoller, R. (2009). Which item types are better suited to the linking of verbal adapted tests? *International Journal of Testing*, 9(2), 92–107. <https://doi.org/10.1080/15305050902880686>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Angoff, W. H., & Cook, L. L. (1988). Equating the scores of the Prueba de Aptitud Académica and the Scholastic Aptitude Test (Report No. 88-2). *ETS Research Report Series*. <https://doi.org/10.1002/j.2330-8516.1988.tb00259.x>
- Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(4), 495–508. <https://doi.org/10.1080/10705511.2014.919210>
- Badham, L., & Furlong, A. (2023). Summative assessments in a multilingual context: What comparative judgment reveals about comparability across different languages in Literature. *International Journal of Testing*, 23(2), 111–134. <https://doi.org/10.1080/15305058.2022.2149536>
- Blanco, C. (2024). *2024 duolingo language report*. Duolingo. <https://blog.duolingo.com/2024-duolingo-language-report/>
- Boldt, R. F. (1969). *Concurrent validity of the PAA and SAT for bilingual Dade School County high school volunteers (College Entrance*







- Examination Board Research and Development Report 68-69, No. 3). Educational Testing Service.
- Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-cultural Psychology*, 1(3), 185–216. <https://doi.org/10.1177/135910457000100301>
- Cascallar, A. S., & Dorans, N. J. (2005). Linking scores from tests of similar content given in different languages: An illustration of methodological alternatives. *International Journal of Testing*, 5(4), 337–356. [https://doi.org/10.1207/s15327574ijt0504\\_1](https://doi.org/10.1207/s15327574ijt0504_1)
- CTB/McGraw-Hill (1988). *Spanish assessment of basic education: Technical report*. McGraw Hill.
- Davidov, E. (2011). Nationalism and constructive patriotism: A longitudinal test of comparability in 22 countries with the ISSP. *International Journal of Public Opinion Research*, 23(1), 88–103. <https://doi.org/10.1093/ijpor/edq031>
- Davis, S. L., Buckendahl, C. W., & Plake, B. S. (2008). When adaptation is not an option: An application of multilingual standard setting. *Journal of Educational Measurement*, 45(3), 287–304. <https://doi.org/10.1111/j.1745-3984.2008.00065.x>
- Dept, S., Ferrari, A., & Halleux, B. (2017). Translation and cultural appropriateness of survey material in large-scale assessments. In P. Lietz, J. Cresswell, K. Rust and R. Adams (Eds.), *Implementation of large-scale education assessments* (pp. 168–191). Wiley. <https://doi.org/10.1002/9781118762462.ch6>
- Dorans, N. J., & Middleton, K. (2012). Addressing the extreme assumptions of presumed linkings. *Journal of Educational Measurement*, 49(1), 1–18. <https://doi.org/10.1111/j.1745-3984.2011.00157.x>
- EBBS, D., & Koršňáková, P. (2016). Translation and translation verification for TIMSS 2015. In Martin, M. O., Mullis I. V. & Martin H. (Eds.), *Methods and procedures in TIMSS 2015* (pp. 7.1–7.16). TIMSS & PIRLS International Study Center, Boston College.
- EBBS, D., Flicop, S., Hidalgo, M. M., & Netten, A. (2021). Systems and instrument verification in PIRLS 2021. In *Methods and procedures: PIRLS 2021 technical report* (pp. 5.1–5.24). TIMSS & PIRLS International Study Center, Boston College. <https://doi.org/10.6017/lse.tpisc.tr2103.kb2485>
- El Masri, Y. H., Baird, J.-A., & Graesser, A. (2016) Language effects in international testing: The case of PISA 2006 science items. *Assessment in Education: Principles, Policy & Practice*, 23(4), 427–455. <https://doi.org/10.1080/0969594X.2016.1218323>
- Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing*, 2(3–4), 199–215. <https://doi.org/10.1080/15305058.2002.9669493>
- Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, 5(1), 23–35. [https://doi.org/10.1207/s15327574ijt0501\\_3](https://doi.org/10.1207/s15327574ijt0501_3)
- Ercikan, K., & Lyons-Thomas, J. (2013). Adapting tests for use in other languages and cultures. In K. Geisinger, B. Bracken, J. Carlson, J.-I. Hansen, N. Kuncel, S. Reise, & M. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology*, Vol. 3. *Testing and assessment in school psychology and education* (pp. 545–569). American Psychological Association. <https://doi.org/10.1037/14049-026>
- Ercikan, K., & Por, H. (2020). Comparability in multilingual and multicultural assessment contexts. In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability of large-scale educational assessments: Issues and recommendations* (pp. 205–225). National Academy of Education Press. <https://naeducation.org/wp-content/uploads/2020/06/Comparability-of-Large-Scale-Educational-Assessments.pdf>
- Fischer, R., & Fontaine, J. R. J. (2011). Methods for investigating structural equivalence. In D. Matsumoto and F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 179–215). Cambridge University Press. <https://doi.org/10.1017/CBO9780511779381.010>
- Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement*, 38(2), 164–187. <https://doi.org/10.1111/j.1745-3984.2001.tb01121.x>
- Gökçe, S., Berberoglu, G., Wells, C. S., & Sireci, S. G. (2021). Linguistic distance and translation differential item functioning on Trends in International Mathematics and Science Study mathematics assessment items. *Journal of Psychoeducational Assessment*, 39(6), 728–745. <https://doi.org/10.1177/07342829211010537>
- Goodwin, S., Bilsky, L., Mulcaire, P., & Settles, B. (2023, 26–28 April). *Machine learning applications to develop tests in multiple languages simultaneously and at scale* [Conference presentation]. Association of Language Testers in Europe 8<sup>th</sup> International Conference, Madrid, Spain.
- Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing*, 20(2), 225–240. <https://doi.org/10.1191/0265532203lt2540a>
- Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, 2, 63–83.
- Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, 10(3), 229–244.
- Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Lawrence Erlbaum Publishers.
- Hambleton, R. K., & Zenisky, A. L. (2011). Translating and adapting tests for cross-cultural assessments. In D. Matsumoto, & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 46–74). Cambridge University Press.
- 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>
- Hernández, A., Hidalgo, 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://doi.org/10.7334/psicothema2019.306>
- Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology*, 67(6), 818–825. <https://doi.org/10.1037/0021-9010.67.6.818>
- Hulin, C. L., & Mayer, L. J. (1986). Psychometric equivalence of a translation of the Job Descriptive Index into Hebrew. *Journal of Applied Psychology*, 71(1), 83–94. <https://doi.org/10.1037/0021-9010.71.1.83>
- International Baccalaureate Organization (2018). *Assessment principles and practices—Quality assessments in a digital age*. International Baccalaureate Organization. [https://ibo.org/globalassets/new-structure/about-the-ib/pdfs/dp-final-statistical-bulletin-may-2024\\_en.pdf](https://ibo.org/globalassets/new-structure/about-the-ib/pdfs/dp-final-statistical-bulletin-may-2024_en.pdf)
- International Baccalaureate Organization (2024). *The IB Diploma Programme and Career-Related Programme: May 2024 assessment*

- session final statistical bulletin. International Baccalaureate Organization. [https://ibo.org/globalassets/new-structure/about-the-ib/pdfs/the-ib-dp-and-cp-statistical-bulletin\\_en.pdf](https://ibo.org/globalassets/new-structure/about-the-ib/pdfs/the-ib-dp-and-cp-statistical-bulletin_en.pdf)
- International Test Commission. (2017). *ITC Guidelines for translating and adapting tests (2nd edition)*. International Test Commission. [https://www.intestcom.org/files/guideline\\_test\\_adaptation\\_2ed.pdf](https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf)
- International Test Commission and Association of Test Publishers (2022). *Guidelines for technology-based assessments*. International Test Commission and Association of Test Publishers. <https://www.intestcom.org/upload/media-library/tba-guidelines-final-2-23-2023-v4-167785144642TgY.pdf>
- Kolen, M., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices (2nd edition)*. Springer-Verlag.
- Koršňáková, P., Dept, S., & Ebbs, D. (2020). Translation: The preparation of national language versions of assessment instruments. In H. Wagemaker (Ed.), *Reliability and validity of international large-scale assessment: Understanding IEA's comparative studies of student achievement, volume 10* (pp. 85–111). IEA Research for Education, Springer, Cham. [https://doi.org/10.1007/978-3-030-53081-5\\_6](https://doi.org/10.1007/978-3-030-53081-5_6)
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Publishers.
- Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice*, 26(4), 29–37. <https://doi.org/10.1111/j.1745-3992.2007.00106.x>
- Martin, M. O., von Davier, M., & Mullis, I. V. (2020). Methods and procedures: TIMSS 2019 technical report. *International Association for the Evaluation of Educational Achievement*. <https://timssandpirls.bc.edu/timss2019/methods/>
- McGrane, J., Kayton, H., Double, K., Woore, R., & El Masri, Y. (2022). *Is science lost in translation? Language effects in the International Baccalaureate Diploma Programme science assessments*. Oxford University Centre for Educational Assessment. <https://ibo.org/globalassets/new-structure/research/pdfs/ib-dp-science-translation-final-report.pdf>
- Milman, L. H., Faruqi-Shah, Y., Corcoran, C. D., & Damele, D. M. (2018). Interpreting mini-mental state examination performance in highly proficient bilingual Spanish–English and Asian Indian–English speakers: Demographic adjustments, item analyses, and supplemental measures. *Journal of Speech, Language, and Hearing Research*, 61(4), 847–856.
- OECD. (2016). PISA 2018 translation and adaptation guidelines. OECD Publishing. <https://www.oecd.org/content/dam/oecd/en/about/programmes/edu/pisa/pisa-database/survey-implementation-tools/pisa-2018/PISA-2018-TRANSLATION-AND-ADAPTATION-GUIDELINES.pdf>
- OECD. (2024). *PISA 2022 Technical Report*. OECD Publishing. [https://www.oecd.org/en/publications/pisa-2022-technical-report\\_01820d6d-en.html](https://www.oecd.org/en/publications/pisa-2022-technical-report_01820d6d-en.html)
- Oliveri, M. E., Olson, B., Ercikan, K., & Zumbo, B. D. (2012). Methodologies for investigating item- and test-level measurement equivalence in international large-scale assessments. *International Journal of Testing*, 12(3), 203–223. <https://doi.org/10.1080/15305058.2011.617475>
- Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, 53(3), 315–333.
- Oliveri, M. E., & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing*, 14(1), 1–21. <https://doi.org/10.1080/15305058.2013.825265>
- Oliveri, M. E., & von Davier, M. (2017). Analyzing invariance of item parameters used to estimate trends in international large-scale assessments. In H. Jiao & R.W. Lissitz (Eds.), *Test fairness in the new generation of large-scale assessment* (pp.121–146). Information Age Publishing.
- Ong, S. L., & Sireci, S. G. (2008). Using bilingual students to link and evaluate different language versions of an exam. *US-China Education Review*, 5(11), 37–46.
- Rapp, J., & Allalouf, A. (2003). Evaluating cross-lingual equating. *International Journal of Testing*, 3(2), 101–117. [https://doi.org/10.1207/S15327574IJT0302\\_1](https://doi.org/10.1207/S15327574IJT0302_1)
- Robin, F., Sireci, S. G., & Hambleton, R. K. (2003). Evaluating the equivalence of different language versions of a credentialing exam. *International Journal of Testing*, 3(1), 1–20. [https://doi.org/10.1207/S15327574IJT0301\\_1](https://doi.org/10.1207/S15327574IJT0301_1)
- Rogers, W. T., Gierl, M. J., Tardif, C., Lin, J., & Rinaldi, C. M. (2003). Differential validity and utility of successive and simultaneous approaches to the development of equivalent achievement tests in French and English. *Alberta Journal of Educational Research*, 49(3), 290–304. <https://doi.org/10.11575/ajer.v49i3.54986>
- Rogers, W. T., Gierl, M. J., Tardif, C., Lin, J., & Rinaldi, C. M. (2010). Validity of the simultaneous approach to the development of equivalent achievement tests in English and French. *Applied Measurement in Education*, 24(1), 39–70. <https://doi.org/10.1080/08957347.2011.532416>
- Sireci, S. G. (1997). Problems and issues in linking tests across languages. *Educational Measurement: Issues and Practice*, 16(1), 12–19. <https://doi.org/10.1111/j.1745-3992.1997.tb00581.x>
- Sireci, S. G. (2005). Using bilinguals to evaluate the comparability of different language versions of a test. In R.K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 117–138). Lawrence Erlbaum Publishers.
- Sireci, S. G., & Berberoglu, G. (2000). Using bilingual respondents to evaluate translated- adapted items. *Applied Measurement in Education*, 13(3), 229–248. [https://doi.org/10.1207/S15324818AME1303\\_1](https://doi.org/10.1207/S15324818AME1303_1)
- Sireci, S. G., & Oliveri, M. E. (2023). *A Critical Review of the International Baccalaureate Organization's Multilingual Assessment Processes and Best Practices' Recommendations [Report for the IB]*. International Baccalaureate Organization.
- Sireci, S. G., Rios, J. A., & Powers, S. (2016). Comparing test scores from tests administered in different languages. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 181–202). Routledge.
- Sireci, S. G., & Wells, C. S. (2010). Evaluating the comparability of English and Spanish video accommodations for English language learners. In P. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations* (pp. 33–68). Council of Chief State School Officers.
- Sukin, T., Sireci, S. G., & Ong, S. L. (2015). Using bilingual examinees to evaluate the comparability of test structure across different language versions of a mathematics exam. *Actualidades en Psicología*, 29(119), 131–139. <http://doi.org/10.15517/ap.v29i119.19244>
- Tanzer, N. K. (2005). Developing tests for use in multiple languages and cultures: A plea for simultaneous development. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 235–263). Lawrence Erlbaum Publishers.
- van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, 47(4), 263–279. <https://doi.org/10.1016/j.erap.2003.12.004>

- van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39–64). Lawrence Erlbaum Publishers.
- van de Vijver, F., Avvisati, F., Davidov, E., Eid, M., Fox, J. P., Le Donné, N., Lek, K., Meuleman, B., Paccagnella, M., & Van de Schoot, R. (2019). *Invariance analyses in large-scale studies*. OECD Publishing.
- Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement*, 28(3), 197–219. <https://doi.org/10.1111/j.1745-3984.1991.tb00354.x>
- Wolff, H. G., Schneider-Rahm, C. I., & Forret, M. L. (2011). Adaptation of a German multidimensional networking scale into English. *European Journal of Psychological Assessment*, 27(4), 244–250. <https://doi.org/10.1027/1015-5759/a000070>
- Woodcock, R. W., & Muñoz-Sandoval, A. F. (1993). An IRT approach to cross-language test equating and interpretation. *European Journal of Psychological Assessment*, 9(3), 233–241.
- Zhao, X., Solano-Flores, G., & Qian, M. (2018). International test comparisons: Reviewing translation error in different source language-target language combinations. *International Multilingual Research Journal*, 12(1), 17–27. <https://doi.org/10.1080/19313152.2017.1349527>
- Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Astivia, O. L. O., & Ark, T. K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly*, 12(1), 136–151. <https://doi.org/10.1080/15434303.2014.972559>

Article

# When to use Bootstrap- $F$ in One-Way Repeated Measures ANOVA: Type I Error and Power

María J. Blanca<sup>1</sup> , Roser Bono<sup>2,3</sup> , Jaume Arnau<sup>3</sup> , F. Javier García-Castro<sup>4</sup> ,  
Rafael Alarcón<sup>1</sup> , and Guillermo Vallejo<sup>5</sup> 

<sup>1</sup>University of Malaga (Spain)

<sup>2</sup>Institute of Neurosciences, University of Barcelona (Spain)

<sup>3</sup>University of Barcelona (Spain)

<sup>4</sup>Universidad Loyola Andalucía (Spain)

<sup>5</sup>University of Oviedo (Spain)

## ARTICLE INFO

Received: 18/12/2024  
Accepted: 24/02/2025

### Keywords:

Bootstrap- $F$   
Within-subject design  
Greenhouse-Geisser adjustment  
Huynh-Feldt adjustment  
Robustness

## ABSTRACT

**Background:** With repeated measures, the traditional ANOVA  $F$ -statistic requires fulfillment of normality and sphericity. Bootstrap- $F$  ( $B-F$ ) has been proposed as a procedure for dealing with violation of these assumptions when conducting a one-way repeated measures ANOVA. However, evidence regarding its robustness and power is limited. Our aim is to extend knowledge about the behavior of  $B-F$  with a wider range of conditions. **Method:** A simulation study was performed, manipulating the number of repeated measures, sample sizes, epsilon values, and distribution shape. **Results:**  $B-F$  may become conservative with higher values of epsilon, and liberal under extreme violation of both normality and sphericity and small sample sizes. In these cases,  $B-F$  may be used with a more stringent alpha level (.025). The results also show that power is affected by sphericity: the lower the epsilon value, the larger the sample size required to ensure adequate power. **Conclusions:**  $B-F$  is robust under non-normality and non-sphericity with sample sizes larger than 20-25.

## Cuándo Usar $F$ -Bootstrap en ANOVA Unifactorial de Medidas Repetidas: Error de Tipo I y Potencia

## RESUMEN

**Antecedentes:** El estadístico  $F$  del ANOVA de medidas repetidas requiere el cumplimiento de los supuestos de normalidad y esfericidad. El procedimiento  $F$ -bootstrap ( $F-B$ ) se ha propuesto como alternativa al ANOVA cuando se violan estos supuestos. Sin embargo, la evidencia empírica sobre su robustez y potencia es limitada. El objetivo es analizar el comportamiento de  $F-B$  en un mayor número de condiciones. **Método:** Se realizó un estudio de simulación, manipulando el número de medidas repetidas, tamaño muestral, valores de  $\epsilon$  y forma de la distribución. **Resultados:** El procedimiento  $F-B$  resulta conservador con valores altos de  $\epsilon$ , y puede llegar a ser liberal bajo una violación extrema de la normalidad y esfericidad con tamaño muestral pequeño. En estos casos,  $F-B$  puede utilizarse con un nivel de alfa más restrictivo (.025). Los resultados también muestran que la potencia se ve afectada por la esfericidad: cuanto menor es el valor de  $\epsilon$ , mayor es el tamaño muestral necesario para garantizar una potencia adecuada. **Conclusiones:** El procedimiento  $F-B$  es robusto en condiciones de no normalidad y no esfericidad con tamaños de muestra superiores a 20-25.

### Palabras clave:

Remuestreo  
Diseño intrasujeto  
Ajuste Greenhouse-Geisser  
Ajuste Huynh-Feldt  
Robustez



## Introduction

Bootstrapping is a computing-intensive method introduced by Efron (1979) and colleagues (e.g., Efron & Gong, 1983; Efron & Tibshirani, 1993) that basically involves drawing random samples from the original dataset with replacement, and then computing the sample distribution for a given statistic for each bootstrap sample. This resampling process enables the estimation of confidence intervals, standard errors, and hypothesis tests, providing a robust alternative to traditional parametric methods. The method has a wide range of applications, including comparison of means tests, correlation and regression, multilevel analysis, mediation and moderation, graph analysis, time series analysis, and survival analysis (Chernick & LaBudde, 2011; Christensen & Golino, 2021; Hayes, 2017; Vallejo et al., 2013; Wilcox, 2022). The increasing popularity of bootstrapping for statistical inference has seen it gradually incorporated into the most common statistical software, such as R, SAS and IBM SPSS.

Bootstrap can be used in conjunction with different statistical procedures, including those derived from the general linear model such as regression analysis and analysis of variance (ANOVA), to make inferences about a population. In the case of ANOVA, this involves generating the empirical sampling distribution for the  $F$ -statistic by repeatedly resampling with replacement from the dataset, rather than using the theoretical distribution of the statistic. Because bootstrap does not rely on the parametric assumptions of normality and homogeneity of variance, it is particularly useful when these assumptions are violated (Chernick, 2008).

Simulation studies are valuable tools that involve running numerous random data sets to assess how a statistic performs under various conditions. Robustness in terms of Type I error is typically interpreted using Bradley's liberal criterion (1978), which considers a statistic to be robust if its Type I error rate is between 2.5% and 7.5% for an alpha of 5%.

When repeated measures are involved, traditional ANOVA (RM-ANOVA) requires normality and sphericity. Simulation studies have shown that the  $F$ -statistic of RM-ANOVA is generally robust to non-normality when the sphericity assumption is met (Berkovits et al., 2000; Blanca et al., 2023a; Keselman et al., 1996; Kherad-Pajouh & Renaud, 2015). Blanca et al. (2023a) found that the test was robust in 99.95% of the 1786 conditions studied, and also that the Type I error rate was only greater than .075 (specifically, .078) in the case of a design with four repeated measures, extreme departure from normality (skewness  $\gamma_1 = 2.31$ , kurtosis  $\gamma_2 = 8$ ), and  $N = 10$ . However, RM-ANOVA is very sensitive to sphericity violation, rendering it a liberal test (Berkovits et al., 2000; Blanca et al., 2023b; Haverkamp & Beauducel, 2017, 2019; Voelkle & McKnight, 2012).

To control Type I error when sphericity is violated, the use of adjusted  $F$ -tests, such as the Greenhouse-Geisser ( $F$ -GG) and Huynh-Feldt ( $F$ -HF) adjustments, has been proposed. These two procedures modify the degrees of freedom of the  $F$ -statistic by a multiplicative factor, known as epsilon ( $\varepsilon$ ), making it a more demanding test. The value of  $\varepsilon$  is considered an indicator of the amount by which the data depart from sphericity, and it ranges between  $1/k-1$  and 1, where  $k$  is the number of repeated measures. When the data meet the sphericity assumption,  $\varepsilon = 1$ , and the greater the departure from this value the greater the violation

of sphericity.  $F$ -GG and  $F$ -HF differ in how  $\varepsilon$  is computed, and the decision over which procedure to use is controversial. Indeed, there is evidence for the superiority of both  $F$ -GG (Voelkle & MacKnight, 2012) and  $F$ -HF (Haverkamp & Beauducel, 2017, 2019; Oberfeld & Franke, 2013), while some studies have found that both offer reasonable control of Type I error (Berkovits et al., 2000; Muller et al., 2007). A value-based strategy has also been proposed based on the expected value of  $\varepsilon$ . For example, Huynh and Feldt (1976) recommend using  $F$ -GG if  $\varepsilon$  is less than .75, and  $F$ -HF for  $\varepsilon$  greater than .75. More recently, Blanca et al. (2023b) established another cut-off point based on the results of a simulation study with normal data and a larger number of manipulated conditions than were considered in the aforementioned studies, taking the Greenhouse-Geisser  $\varepsilon$  estimation ( $\hat{\varepsilon}$ ) as reference. They suggested, as a general rule, using  $F$ -GG because it is more conservative, although in the event of discrepant results from the two procedures, they recommend using  $F$ -GG for  $\hat{\varepsilon}$  values below .60, and  $F$ -HF for  $\hat{\varepsilon}$  values of .60 or higher.

When normality and sphericity are simultaneously violated, the behavior of adjusted  $F$ -tests depends on several factors, namely sample size and the degree of violation of both sphericity and normality. Blanca et al. (2024) found that although the aforementioned rule generally holds under non-normality and non-sphericity, there are two exceptions in which neither  $F$ -GG nor  $F$ -HF is robust: a) With  $N \leq 10$ ,  $\hat{\varepsilon} \leq .60$ , and severe deviation from normality ( $\gamma_1 = 1.41$ ,  $\gamma_2 = 3$ ) and, b) with  $N \leq 30$ ,  $\hat{\varepsilon} \leq .60$ , and extreme deviation from normality ( $\gamma_1 = 2$ ,  $\gamma_2 = 6$  and 8). These authors discuss several available analytic alternatives, none of which is free of criticism, highlighting that bootstrapping may be the most promising alternative according to results obtained in other studies (e.g., Berkovits et al., 2000).

Berkovits et al. (2000) proposed a bootstrap method for one-way repeated measures ANOVA, referred to as bootstrap- $F$  ( $B$ - $F$ ), which generates the bootstrap sample from centered data. They conducted a simulation study to analyze the behavior of this procedure in terms of Type I error with a four repeated measures design, introducing different values of sample size (10, 15, 30, and 60) and epsilon (.48, .57, .75, and 1). Distribution shape was also manipulated so as to include both normal data and distributions labeled as showing slight ( $\gamma_1 = 1$ ,  $\gamma_2 = 0.75$ ), moderate ( $\gamma_1 = 1.75$ ,  $\gamma_2 = 3.75$ ), and severe ( $\gamma_1 = 3$ ,  $\gamma_2 = 21$ ) deviation from normality. The results showed that  $B$ - $F$  was a robust alternative under violations of sphericity and normality, even in small samples and with severe non-normality, with Type I error rates below 7.5% in all conditions manipulated. However, the test became conservative in some cases with  $\varepsilon = 1$ .

To our knowledge, the behavior of the  $B$ - $F$  test proposed by Berkovits et al. (2000) has scarcely been investigated with one-way designs, although it has been studied with split-plot designs. Vallejo et al. (2006) performed a simulation study in which they tested this procedure with a 3x4 design with  $N = 30$ , 45, and 60,  $\varepsilon = .50$ , .75, and 1, and the same non-normal distributions as Berkovits et al. (2000). The findings were consistent with those of Berkovits et al. (2000), insofar as the test was robust under non-sphericity and non-normality but tended to be conservative with high values of  $\varepsilon$ . These results were subsequently confirmed by Vallejo et al. (2010) using a 3x4 design with  $N = 30$  and 45, and  $\varepsilon = .50$ , in which they found that  $B$ - $F$  controlled Type I error with different non-normal distributions.

Overall, the empirical evidence suggests that *B-F* is a robust procedure for dealing with violations of normality and sphericity. However, this evidence is limited as published simulation studies include a small number of manipulated conditions in terms of repeated measures, sample sizes, sphericity, and distribution shapes. The aim of the present study is therefore to extend knowledge about the robustness and power of *B-F* by considering a wider range of conditions. To this end, we included designs with 3, 4, and 6 repeated measures, sample sizes from 10 to 180,  $\hat{\epsilon}$  values from the corresponding lower bound to 1, and six distributions representing slight to extreme deviations from normality.

### Bootstrap-*F*

The goal in using this procedure is to estimate an appropriate critical value when the null hypothesis is true. This is done by centering the data in each repeated measure condition, randomly generating *B* bootstrap samples with replacement from the centered data in each condition, computing the statistics for each bootstrap sample generated, and obtaining an estimate of the distribution of the statistic (Wilcox, 2003, p. 379). Berkovits et al. (2000) consider that the *B-F* procedure comprises the following steps:

1. Organize data in a matrix of *N* participants x *K* measurement occasions. To test the null hypothesis of equality of means among repeated measures, compute the *F*-statistic based on original data, labeled as observed  $F_o$ .

$$\begin{bmatrix} X_{11} & \cdots & X_{1k} \\ \vdots & \ddots & \vdots \\ X_{N1} & \cdots & X_{Nk} \end{bmatrix}$$

The data with 3 repeated measures shown in Table 1 provide an illustration of the procedure. The observed  $F_o$  is 92.19.

2. Center the data with the aim of estimating an appropriate critical value of the *F*-statistics, subtracting the respective mean of the *k*th level of the repeated measure from each observation:  $C_{ij} = X_{ik} - \bar{X}_{.k}$ . This matrix will have the same distributional properties and the same covariance

matrix as the original data (Berkovits et al., 2000). The data matrix is now:

$$\begin{bmatrix} C_{11} & \cdots & C_{1k} \\ \vdots & \ddots & \vdots \\ C_{N1} & \cdots & C_{Nk} \end{bmatrix}$$

Table 2 displays the data matrix with centered data (for the example shown in Table 1).

3. With the centered data, generate *B* bootstrap samples with replacement by randomly sampling *N* rows of data.

$$\begin{bmatrix} C_{11}^* & \cdots & C_{1k}^* \\ \vdots & \ddots & \vdots \\ C_{N1}^* & \cdots & C_{Nk}^* \end{bmatrix}$$

In the example, we would generate 599 bootstrap samples, although for illustrative purposes, only 2 are displayed in Table 3.

**Table 2**  
Data Matrix With Centered Data (for the Example Shown in Table 1)

ID	Centered 1	Centered 2	Centered 3
1	2.08	-0.25	-0.33
2	-1.92	0.75	0.67
3	-0.92	-2.25	-1.33
4	-0.92	0.75	-0.33
5	0.08	-0.25	-2.33
6	-0.92	-0.25	0.67
7	-2.92	-0.25	0.67
8	0.08	1.75	0.67
9	1.08	0.75	1.67
10	3.08	-0.25	-0.33
11	0.08	-1.25	-0.33
12	1.08	0.75	0.67

**Table 1**  
Data Matrix for Illustrative Purposes

ID	Measure 1	Measure 2	Measure 3
1	13	6	4
2	9	7	5
3	10	4	3
4	10	7	4
5	11	6	2
6	10	6	5
7	8	6	5
8	11	8	5
9	12	7	6
10	14	6	4
11	11	5	4
12	12	7	5
M	10.92	6.25	4.33
SD	1.67	1.05	1.07

**Table 3**  
Bootstrap Samples 1 and 2 With Centered Data (C)

Bootstrap sample 1				Bootstrap sample 2			
ID	C 1	C 2	C 3	ID	C 1	C 2	C 3
1	2.08	-0.25	-0.33	2	-1.92	0.75	0.67
6	-0.92	-0.25	0.67	7	-2.92	-0.25	0.67
10	3.08	-0.25	-0.33	4	-0.92	0.75	-0.33
5	0.08	-0.25	-2.33	9	1.08	0.75	1.67
3	-0.92	-2.25	-1.33	8	0.08	1.75	0.67
11	0.08	-1.25	-0.33	7	-2.92	-0.25	0.67
12	1.08	0.75	0.67	5	0.08	-0.25	-2.33
8	0.08	1.75	0.67	6	-0.92	-0.25	0.67
3	-0.92	-2.25	-1.33	5	0.08	-0.25	-2.33
11	0.08	-1.25	-0.33	2	-1.92	0.75	0.67
10	3.08	-0.25	-0.33	12	1.08	0.75	0.67
8	0.08	1.75	0.67	3	-0.92	-2.25	-1.33

$F_1^* = 3.12$

$F_2^* = 2.66$

4. Compute  $F$ -statistics with data from each bootstrap sample, labeled as  $F_1^*$ , ...,  $F_B^*$ , thus creating the empirical sampling distribution of the  $F$ -statistic. The  $F$ -statistics for bootstrap samples 1 and 2 are equal to  $F_1^* = 3.12$  and  $F_2^* = 2.66$ , respectively. Sort the  $F^*$  values in ascending order. Suppose that we obtain a set of  $F^*$  values after performing 599 bootstrap samples. We then sort these values in ascending order, resulting in the following ranking: (1)  $F_3^* = 0.95$ , (2)  $F_2^* = 2.66$ , (3)  $F_1^* = 3.12$ , ..., (569)  $F_{328}^* = 5.03$ , ..., (599)  $F_{430}^* = 52.34$ .
5. Estimate the critical value  $F_c^*$ , where  $c = (1 - \alpha)B$ . The  $F_c^*$  of step 1 is compared with this critical value, and hence the null hypothesis is rejected if  $F_o \geq F_c^*$ . For instance, with  $\alpha = .05$  and  $B = 599$  bootstrap samples,  $c = .95 * 599 = 569.05$ . The  $F^*$  in position 569 will thus be the critical value  $F_c^*$ . The proportion of  $F^*$  values that are larger than the observed  $F_o$  represents the bootstrap  $p$ -value (Berkovits et al., 2000; Vallejo et al., 2010). The null hypothesis of equality of means is rejected if this  $p$ -value is less than or equal to .05. In the example, the  $F^*$ -statistic in position 569 is the critical value:  $F_c^* = 5.03$ . As  $F_o = 92.19$  is larger than 5.03, the null hypothesis of equality of means among repeated measures is rejected. There is no  $F^*$  value larger than  $F_o$ , yielding a  $p < .001$ .

The procedure can be performed using the WRS2 library of R (Mair & Wilcox, 2020), with the rmanovab function and without using trimmed means.

## Method

### Instrument

A simulation study was carried out using the interactive matrix language (IML) module of SAS 9.4. A series of macros was designed to generate data. Non-normal data were generated using the procedure proposed by Fleishman (1978), which applies a polynomial transformation that simulates data with specific values of skewness and kurtosis. To simulate data with different degrees of sphericity violation, we generated a series of unstructured covariance matrices with different values of  $\hat{\epsilon}$  for each repeated measure condition. The unstructured matrix was used because it is the most general covariance structure (Kowalchuk et al., 2004) and is typically found in longitudinal behavioral data (Arnau et al., 2014; Bono et al., 2010). The probability of the values associated with  $B$ - $F$  was calculated using PROC GLM of SAS (more details about the simulation procedure with SAS are available upon request from the corresponding author). Five thousand replications were performed for each condition manipulated with  $B = 599$  bootstraps, as used elsewhere (Vallejo et al., 2006, 2010). This number was selected based on the recommendation that  $\alpha$  should be a multiple of  $(B + 1)^{-1}$  (Wilcox, 2022). In addition, simulation studies suggest that in terms of probability coverage, there is little or no advantage to using  $B > 599$  when  $\alpha = .05$  (Wilcox, 2022).

### Procedure

Type I error rates were recorded, reflecting the percentage of false rejections of the null hypothesis at the 5% significance level. Robustness of  $B$ - $F$  was assessed based on Bradley's (1978) liberal criterion, which considers a procedure to be robust if the Type I error rate is between 2.5% and 7.5% for a nominal alpha of 5%.

The procedure is considered conservative if the Type I error rate is below the lower bound, and liberal if it is above the upper bound. This criterion was chosen because it is widely used in simulation studies and in research focused on repeated measures (e.g., Arnau et al., 2012; Berkovits et al., 2000; Keselman et al., 1999; Kowalchuk et al., 2004; Livacic-Rojas et al., 2010; Oberfeld & Franke, 2013; Vallejo et al., 2006, 2010, 2011), thus facilitating the comparison of results across similar studies.

The variables manipulated for a one-way design were:

1. Number of repeated measures ( $K$ ): The repeated measures were 3, 4, and 6.
2. Total sample size ( $N$ ): Sample sizes were 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, 100, 120, 150, and 180.
3. Epsilon ( $\hat{\epsilon}$ ): The Greenhouse-Geisser estimation of epsilon was used (Box, 1954; Geisser & Greenhouse, 1958; Greenhouse & Geisser, 1959). Depending on the number of repeated measures,  $\hat{\epsilon}$  values ranged from the lower limit to 1. For  $K = 3$ ,  $\hat{\epsilon}$  values were .50, .60, .70, .80, .90, and 1; for  $K = 4$ , they were .33, .40, .50, .60, .70, .80, .90, and 1; and for  $K = 6$ , they were .20, .30, .40, .50, .60, .70, .80, .90, and 1.
4. Distribution shape: Six distributions were used, representing slight to extreme deviations from normality, chosen from among those used by Blanca et al. (2024). Skewness and kurtosis values are shown in Table 4.

Empirical power was also calculated as the percentage rejection of the null hypothesis at a significance level of 5%. It was analyzed by selecting mean values with a linear pattern in which the means increase linearly and proportionally to each other (e.g., 0, 0.5, 1), with medium effect size,  $f \approx 0.25$ . The number of repeated measures and  $N$  were the same as those for Type I error. Epsilon values ( $\hat{\epsilon}$ ) ranged from the lower limit to .90. To simplify the study, distributions 2, 3, and 6 were selected so as to represent the variability of performance of  $B$ - $F$  with respect to Type I error (i.e.,  $B$ - $F$  performed similarly in distributions 3 and 4, and also in distributions 5 and 6). These distributions correspond to moderate ( $\gamma_1 = 1, \gamma_2 = 1.50$ ), severe ( $\gamma_1 = 1.41, \gamma_2 = 3$ ), and extreme deviation from normality ( $\gamma_1 = 2.31, \gamma_2 = 8$ ).

**Table 4**  
Skewness ( $\gamma_1$ ) and Kurtosis ( $\gamma_2$ ) Coefficients for Each Simulated Distribution

Distribution	Type	$\gamma_1$	$\gamma_2$
1	-	0.4	0.8
2	Gamma ( $\alpha = 4$ )	1	1.50
3	Gamma ( $\alpha = 2$ )	1.41	3
4	Gamma ( $\alpha = 1.5$ )	1.63	4
5	Exponential	2	6
6	Gamma ( $\alpha = 0.75$ )	2.31	8

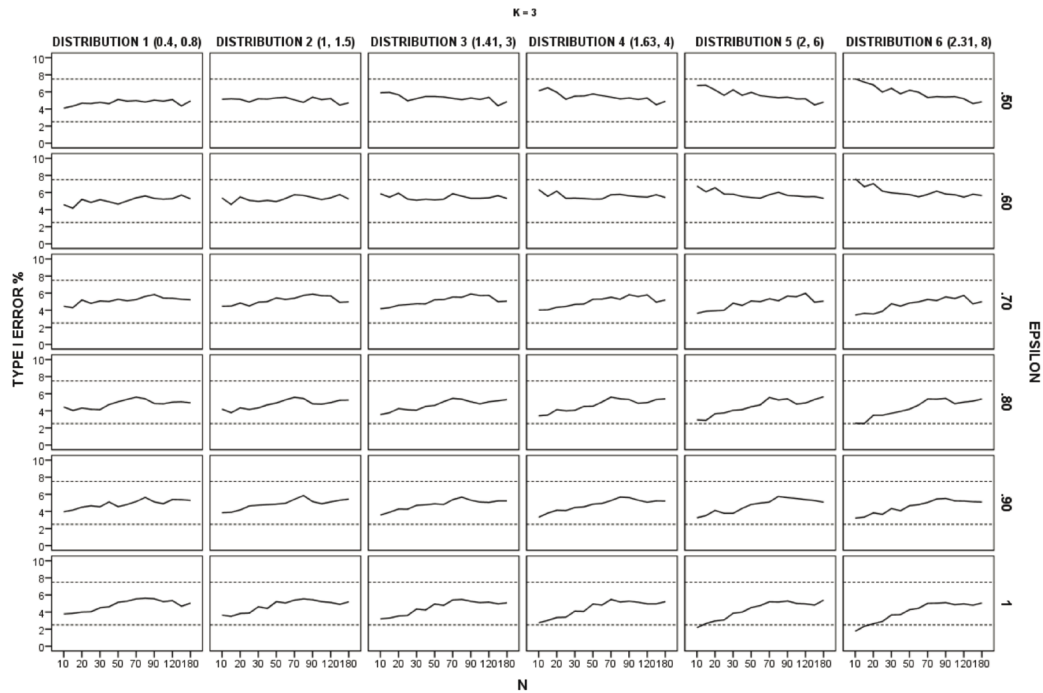
## Results

### Type I Error Rate

Type I error rates for each  $K$ , distribution,  $N$ , and  $\hat{\epsilon}$  value are displayed in Figures 1-3 (detailed tables can be found at <https://dx.doi.org/10.24310/riuma.37706>). The results are summarized in Table 5. For  $K = 3$  and 4,  $B$ - $F$  is robust with distributions 1-4, with maximum values of  $\gamma_1$  and  $\gamma_2$  equal to 1.63 and 4, respectively. For the distribution with  $\gamma_1 = 2$  and  $\gamma_2 = 6$ , the procedure is conservative

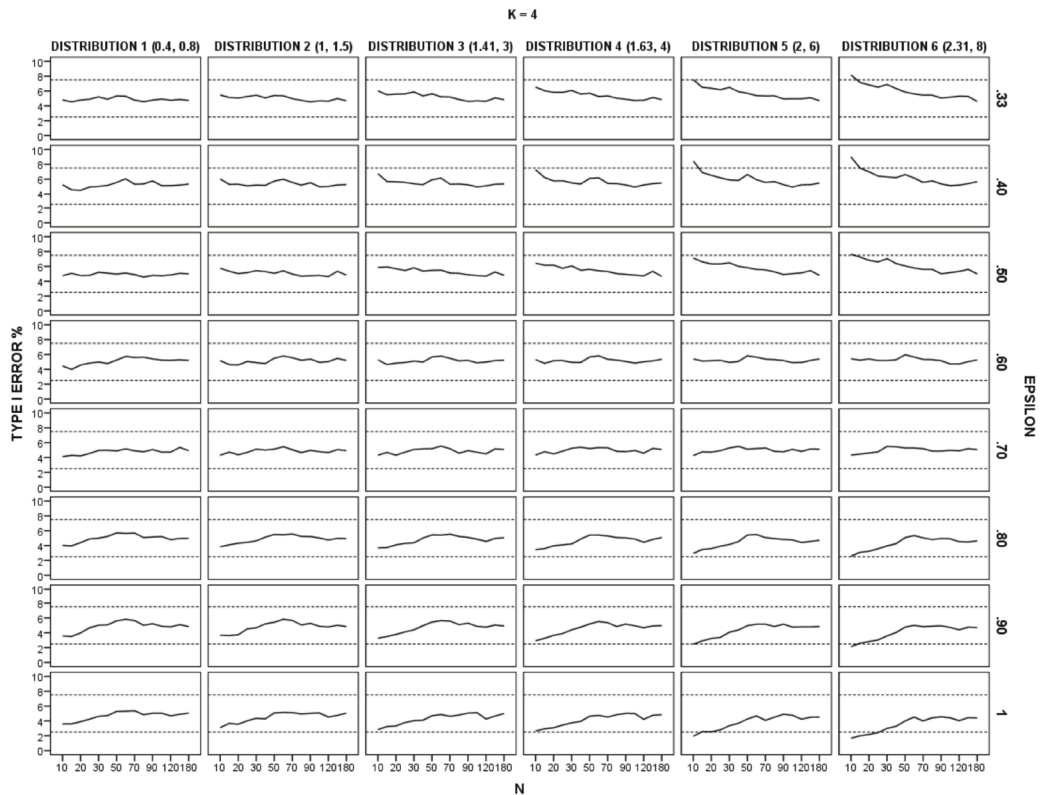


**Figure 1**  
Type I Error Rate (Percentage) for  $K = 3$  as a Function of Distribution Shape,  $N$ , and  $\hat{\epsilon}$

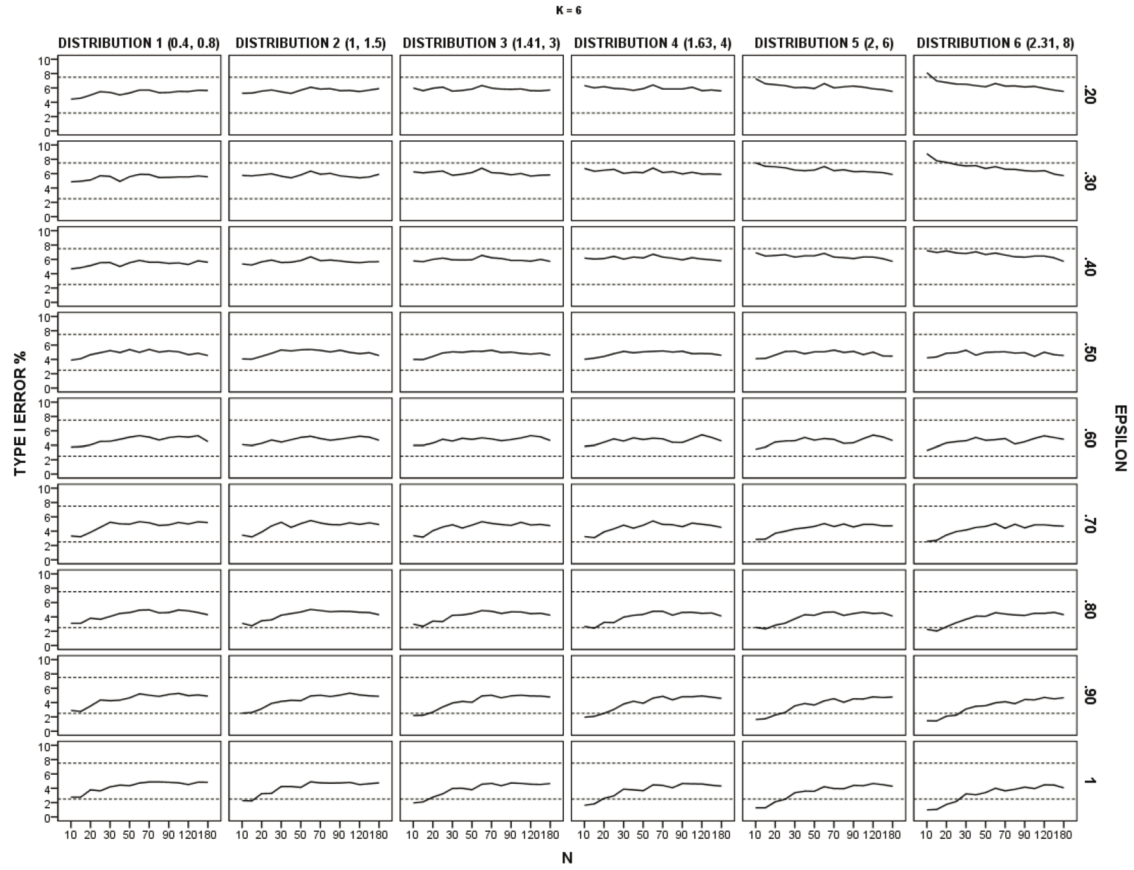


Note. In parentheses: skewness and kurtosis coefficients.

**Figure 2**  
Type I Error Rate (Percentage) for  $K = 4$  as a Function of Distribution Shape,  $N$ , and  $\hat{\epsilon}$



Note. In parentheses: skewness and kurtosis coefficients.

**Figure 3**Type I Error Rate (Percentage) for  $K = 6$  as a Function of Distribution Shape,  $N$ , and  $\hat{\epsilon}$ 

Note. In parentheses: skewness and kurtosis coefficients.

**Table 5**

Summary of the Results Obtained for Type I Error

D	$\gamma_1$	$\gamma_2$	$K = 3$	$K = 4$	$K = 6$
1	0.4	0.8	Robust	Robust	Robust
2	1	1.5	Robust	Robust	C: $\hat{\epsilon} = 1, N = 10-15$ Otherwise robust
3	1.41	3	Robust	Robust	C: $\hat{\epsilon} = .90, N = 10-15$ $\hat{\epsilon} = 1, N = 10-15$ Otherwise robust
4	1.63	4	Robust	Robust	C: $\hat{\epsilon} = .80, N = 15$ $\hat{\epsilon} = .90, N = 10-15$ $\hat{\epsilon} = 1, N = 10-15$ Otherwise robust
5	2	6		L: $\hat{\epsilon} = .33, N = 10$ $\hat{\epsilon} = .40, N = 10$ C: $\hat{\epsilon} = .90, N = 10$ $\hat{\epsilon} = 1, N = 10$ Otherwise robust	L: $\hat{\epsilon} = .30, N = 10$ C: $\hat{\epsilon} = .80, N = 15$ $\hat{\epsilon} = .90, N = 10-20$ $\hat{\epsilon} = 1, N = 10-25$ Otherwise robust
6	2.31	8	L: $\hat{\epsilon} = .50, N = 10$ $\hat{\epsilon} = .60, N = 10$ C: $\hat{\epsilon} = 1, N = 10-15$ Otherwise robust	L: $\hat{\epsilon} = .33, N = 10$ $\hat{\epsilon} = .40, N = 10$ $\hat{\epsilon} = .50, N = 10$ C: $\hat{\epsilon} = .90, N = 10$ $\hat{\epsilon} = 1, N = 10-25$ Otherwise robust	L: $\hat{\epsilon} = .20, N = 10$ $\hat{\epsilon} = .30, N = 10-20$ C: $\hat{\epsilon} = .80, N = 10-15$ $\hat{\epsilon} = .90, N = 10-25$ $\hat{\epsilon} = 1, N = 10-25$ Otherwise robust

Note. D: Distribution; C: Conservative; L: Liberal;  $\gamma_1$ : Skewness;  $\gamma_2$ : Kurtosis.

with high values of  $\hat{\epsilon}$  ( $\hat{\epsilon} = 1$  for  $K = 3$ , and  $\hat{\epsilon} \geq .90$  for  $K = 4$ ) and small sample size ( $N = 10$ ). This tendency to be conservative is also found for both  $K = 3$  and  $K = 4$  for the distribution with  $\gamma_1 = 2.31$  and  $\gamma_2 = 8$  for high values of  $\hat{\epsilon}$  and small sample size. However, with this distribution  $B-F$  tends to be liberal with  $N = 10$  and lower values of epsilon ( $\hat{\epsilon} \leq .60$  for  $K = 3$  and  $\hat{\epsilon} \leq .50$  for  $K = 4$ ).

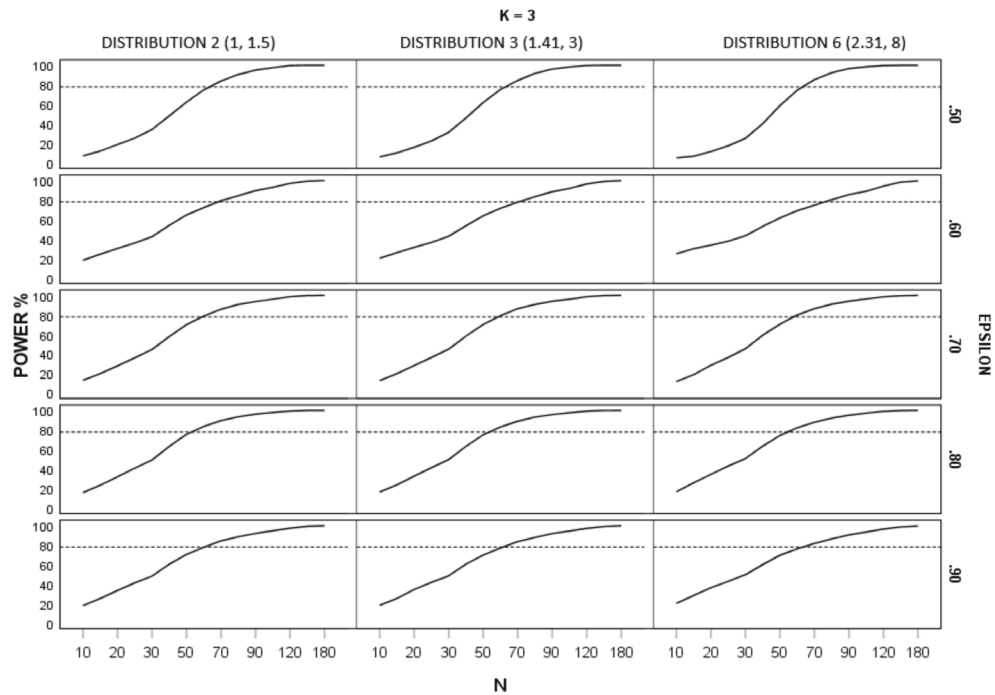
For  $K = 6$ ,  $B-F$  is only robust under all conditions for the distribution with slight deviation from normality ( $\gamma_1 = 0.4, \gamma_2 = 0.8$ ). With the remaining distributions, the test tends to be liberal with extreme deviation from normality, lower values of epsilon, and small sample size. For the distribution with  $\gamma_1 = 2$  and  $\gamma_2 = 6$ ,  $B-F$  is liberal with  $\hat{\epsilon} = .30$  and  $N = 10$ , whereas in the case of the distribution with  $\gamma_1 = 2.31$  and  $\gamma_2 = 8$ ,  $B-F$  is liberal with  $\hat{\epsilon} = .20$  and  $N = 10$  and with  $\hat{\epsilon} = .30$  and  $N = 10-20$ . In addition, and as with  $K = 3$  and 4, it tends to be conservative with high values of  $\hat{\epsilon}$  and small sample size. The worst scenario is with extreme deviation from normality ( $\gamma_1 = 2.31, \gamma_2 = 8$ ), in which  $B-F$  is conservative with  $\hat{\epsilon} = .80$  and  $N = 10-15$ , and with  $\hat{\epsilon} \geq .90$  and  $N = 10-25$ .

### Statistical Power

Empirical power for each  $K$ , distribution,  $N$ , and  $\hat{\epsilon}$  value are displayed in Figures 4-6 (detailed tables can be found at <https://dx.doi.org/10.24310/riuma.37706>). Table 6 shows the sample size at which a power of 80% is reached. As expected, the results show

**Figure 4**

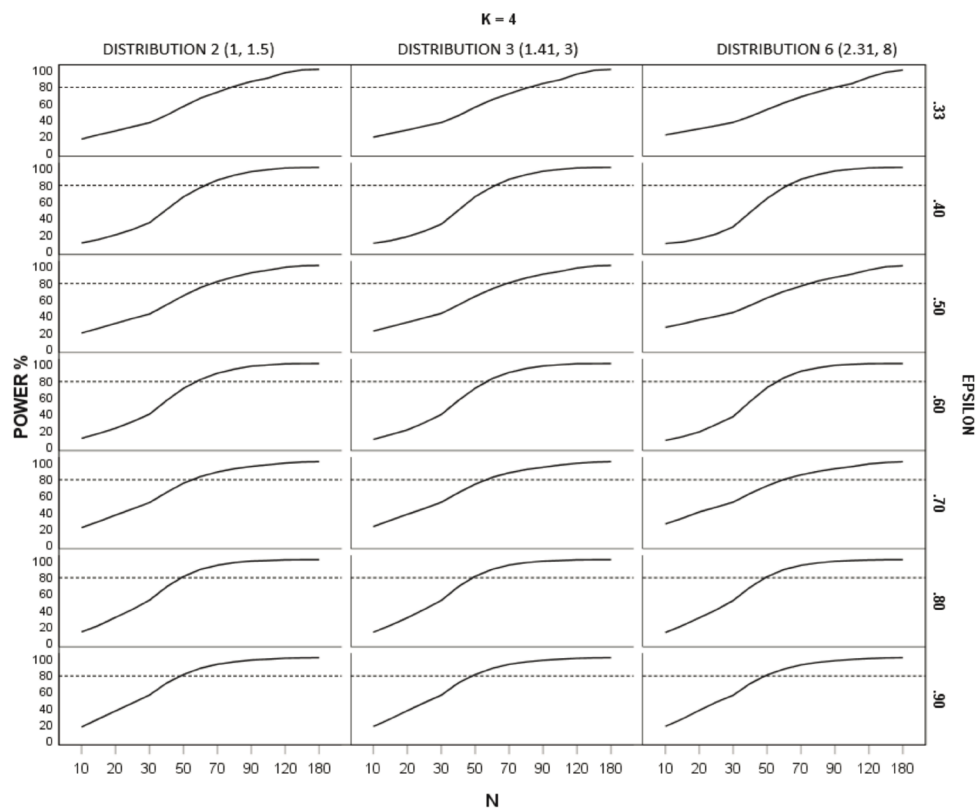
Power (Percentage) for  $K = 3$  as a Function of Distribution Shape,  $N$ , and  $\hat{\epsilon}$



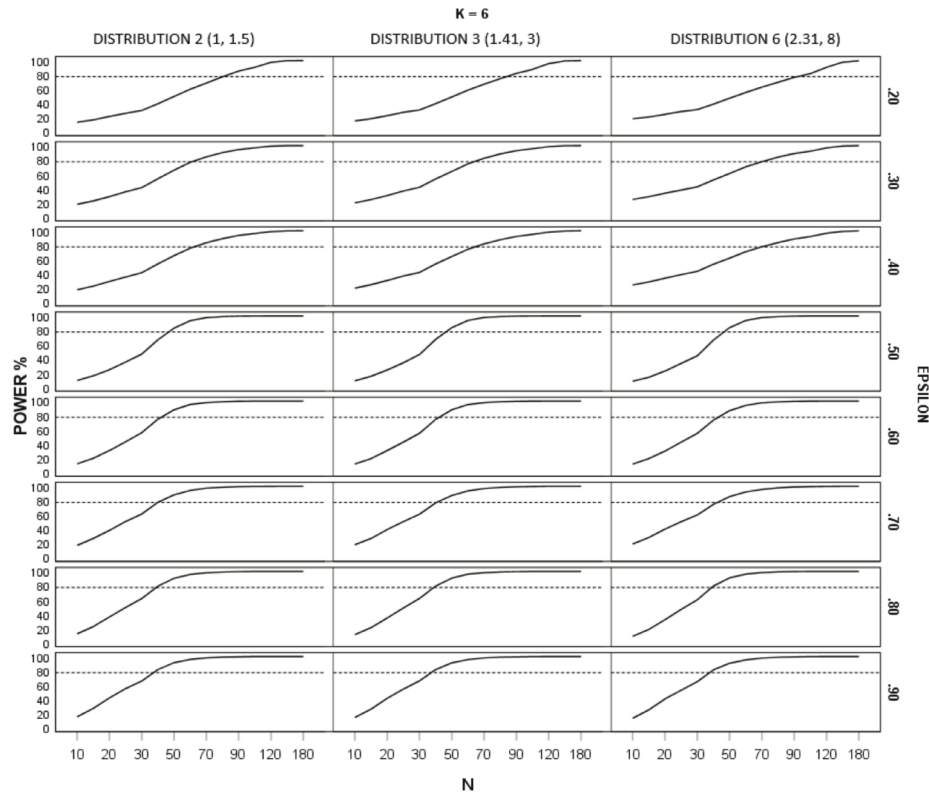
Note. In parentheses: skewness and kurtosis coefficients.

**Figure 5**

Power (Percentage) for  $K = 4$  as a Function of Distribution Shape,  $N$ , and  $\hat{\epsilon}$



Note. In parentheses: skewness and kurtosis coefficients.

**Figure 6**Power (Percentage) for  $K = 6$  as a Function of Distribution Shape,  $N$ , and  $\hat{\epsilon}$ 

Note. In parentheses: skewness and kurtosis coefficients.

**Table 6**

Sample Size at Which a Power of 80% is Reached as a Function of Distribution Shape,  $\hat{\epsilon}$ , and Number of Repeated Measures ( $K$ )

$K$	$\hat{\epsilon}$	Distribution 2 ( $\gamma_1 = 1$ ; $\gamma_2 = 1.5$ )	Distribution 3 ( $\gamma_1 = 1.41$ ; $\gamma_2 = 3$ )	Distribution 6 ( $\gamma_1 = 2.31$ ; $\gamma_2 = 8$ )
3	.50	70	70	70
	.60	80	80	80
	.70	70	70	70
	.80	60	60	60
	.90	70	70	70
4	.33	90	90	100
	.40	70	70	70
	.50	70	80	80
	.60	60	60	60
	.70	60	60	70
	.80	60	60	60
	.90	60	60	60
6	.20	90	90	100
	.30	70	70	80
	.40	70	70	80
	.50	50	50	50
	.60	50	50	50
	.70	50	50	50
	.80	50	50	50
	.90	40	40	40

Note.  $\gamma_1$ : skewness;  $\gamma_2$ : kurtosis.

that power increases with sample size and also that it is affected by  $\hat{\epsilon}$  values, such that a large sample is required to reach adequate power with  $\hat{\epsilon}$  values close to the lower bound. Overall, for  $K = 3$ , this power is achieved with 60-80 participants across all distributions and  $\hat{\epsilon}$  values. For  $K = 4$ , this power is reached in a range of 70-100 participants when  $\hat{\epsilon} \leq .50$ , and 60-70 when  $\hat{\epsilon} \geq .60$ . For  $K = 6$ , 70-100 participants are required when  $\hat{\epsilon} \leq .40$ , and 40-50 for  $\hat{\epsilon} \geq .50$ .

### Discussion

The aim of this study was to extend knowledge about the robustness and power of  $B-F$  by considering a wider range of conditions than has been the case previously. To this end, we simulated designs with 3, 4, and 6 repeated measures, sample sizes from 10 to 180,  $\hat{\epsilon}$  values from the corresponding lower bound to 1, and six distributions representing slight to extreme deviations from normality.

Regarding robustness, the results show that Type I error rates vary as a function of the number of repeated measures, distribution shape, epsilon value, and sample size. For  $K = 3$ ,  $B-F$  is robust for distributions with  $\gamma_1 \leq 1.63$  and  $\gamma_2 \leq 4$  in all conditions manipulated. However, with higher values of  $\gamma_1$  and  $\gamma_2$  the procedure becomes conservative with  $\hat{\epsilon} = 1$  and  $N = 10$ . With the most extreme deviation from normality ( $\gamma_1 = 2.31$  and  $\gamma_2 = 8$ ), the procedure is liberal with smaller values of  $\hat{\epsilon}$  and  $N = 10$ . These results indicate that for  $K = 3$ ,  $B-F$  remains robust with violation of both sphericity and normality for distributions with  $\gamma_1 \leq 1.63$  and  $\gamma_2 \leq 4$ , but with more severe deviations from normality a sample size larger than 10 is required for low values of  $\hat{\epsilon}$ .

For  $K = 4$ ,  $B-F$  is again robust for distributions with  $\gamma_1 \leq 1.63$  and  $\gamma_2 \leq 4$  in all conditions manipulated. With higher values of  $\gamma_1$  and  $\gamma_2$  and for  $N = 10$  the procedure becomes conservative with  $\hat{\epsilon} = 1$ , and liberal with low values of  $\hat{\epsilon}$ . Overall,  $B-F$  is suitable for use with extreme deviation from both normality and sphericity when sample size is larger than 10.

For  $K = 6$ ,  $B-F$  is robust with very slight deviation from normality ( $\gamma_1 = 0.4$  and  $\gamma_2 = 0.8$ ) in all conditions studied. With the other distributions considered, it is conservative with high values of  $\hat{\epsilon}$  and small sample size. A tendency towards liberality appears with severe deviation from normality,  $\gamma_1 \geq 2$  and  $\gamma_2 \geq 6$ , with small values of  $\hat{\epsilon}$  ( $\hat{\epsilon} \leq .30$ ) and small sample sizes ( $N = 10$  and  $20$ ). These results indicate that  $B-F$  may be used under extreme deviation from both normality and sphericity when sample size is larger than 20.

The results regarding liberality of  $B-F$  appear to contradict those of Berkovits et al. (2000), who found that the procedure was robust under all manipulated conditions. However, their study was conducted under more limited conditions (specifically,  $K = 4$  and  $\epsilon > .48$ ) than was the case here. Consistent with Berkovits et al. (2000), our results for  $K = 4$  and  $\hat{\epsilon} = .50$  likewise show that  $B-F$  is robust under all non-normality conditions. Our findings are also in line with those reported by Vallejo et al. (2006, 2010) when using a 3x4 split-plot design and  $\hat{\epsilon} \geq .50$ . The tendency we observed for  $B-F$  to be conservative with high  $\hat{\epsilon}$  values was also documented in both these previous studies.

As a general rule, the first point to consider is that  $B-F$  may become conservative with higher values of  $\hat{\epsilon}$  (e.g.,  $\hat{\epsilon} \geq .80$  for  $K = 6$ ), in which case adjusted  $F$ -tests, such as Greenhouse-Geisser and Huynh-Feldt adjustments, may be a better option (Blanca et al., 2023b, 2024). Second,  $B-F$  is suitable for use under violation of both sphericity and normality for distributions with  $\gamma_1 \leq 1.63$  and  $\gamma_2 \leq 4$ . With non-normal distributions of these characteristics,  $B-F$  is superior to adjusted  $F$ -tests insofar as the latter have shown a tendency to be liberal with  $N = 10$  and low values of  $\hat{\epsilon}$  (Blanca et al., 2024). Third, with more extreme deviation from normality,  $B-F$  yields reliable results if  $N > 20$ . More specifically,  $B-F$  requires  $N > 10$  for  $K = 3$  if  $\hat{\epsilon} \leq .60$  and for  $K = 4$  if  $\hat{\epsilon} \leq .50$ , whereas  $N > 20$  is required for  $K = 6$  if  $\hat{\epsilon} \leq .30$ . In these scenarios,  $B-F$  is slightly superior to adjusted  $F$ -tests as the latter require  $N > 30$  (Blanca et al., 2024).

A possible option in those scenarios where  $B-F$  is liberal (e.g., under extreme violation of both normality and sphericity and small sample size) is to use a more stringent alpha level. This solution has been proposed previously with other statistical tests (Blanca et al., 2018; Keppel & Wickens, 2004; Tabachnick & Fidell, 2007). Here we conducted simulations of these cases (see Table 5), considering nominal alpha levels of 2.5% and 1%, and computing  $B-F$  (results are shown in Table 7). In general, a nominal alpha level of 2.5% is sufficient to keep the Type I error rate for  $B-F$  within [2.5%, 7.5%] in all conditions. It is important to clarify here that using Bradley's liberal criterion implies that the researcher assumes that the actual significance level is between 2.5% and 7.5% for the corresponding nominal value (5%, or 2.5% when a more stringent alpha level is used).

As for empirical power, our results show that power increases with sample size, reflecting the well-known relationship between the two. We also found that deviation from normality did not affect the power of  $B-F$ . However, it is more sensitive to non-

**Table 7**

Type I Error Rates for Bootstrap- $F$  (in Percentages) for a Nominal Alpha of 2.5% (1% in Parentheses) in the Conditions Under Which it is not Robust at the 5% Nominal Alpha Level ( $\gamma_1$ : Skewness;  $\gamma_2$ : Kurtosis)

K	$\hat{\epsilon}$	N	$\gamma_1 = 2, \gamma_2 = 6$	$\gamma_1 = 2.31, \gamma_2 = 8$
3	.50	10		4.62 (2.72)
	.60	10		4.60 (2.48)
4	.33	10	5.00 (3.16)	5.68 (3.82)
	.40	10	5.70 (3.52)	6.42 (4.30)
	.50	10		4.76 (2.82)
6	.20	10		5.46 (3.64)
	.30	10	5.26 (3.40)	6.06 (4.00)
	.30	15		5.26 (3.34)
	.30	20		5.14 (3.06)

sphericity: the greater the violation of sphericity, with  $\hat{\epsilon}$  values close to the lower bound, the larger the sample size required to ensure adequate power. For example, and assuming 80% power to be adequate (Cooper & Garson, 2016; Kirk, 2013), a sample size of 90-100 is required for  $K = 6$  and  $\hat{\epsilon} = .20$ , whereas for  $\hat{\epsilon} = .60$ , 50 participants are sufficient to reach 80% power for a medium effect size. If we compare these results with those reported by Blanca et al. (2024) for the two adjusted  $F$ -tests, then  $B-F$  seems to have greater power in some cases as it is less affected by non-normality.

In conclusion, the  $B-F$  procedure offers an alternative for the analysis of repeated measures data with a nominal alpha of 5% under certain conditions specified in Table 5, which researchers may consult to decide if it is a correct option given the characteristics of their data. As a rule of thumb, and to ensure that  $B-F$  remains robust under non-normality and non-sphericity, a  $N > 20$  is required to maintain Type I error rates  $\leq 7.5\%$ . In the event of extreme violations of both normality and sphericity and  $10 \leq N \leq 20$ ,  $B-F$  may be used if a more stringent alpha level (e.g., 2.5%) is considered. It should also be noted that with high  $\hat{\epsilon}$  values the procedure may become conservative and require a  $N > 25$ . Researchers may consult Table 6 to determine the sample size at which 80% power is reached as a function of the number of repeated measures and other data characteristics.

Researchers may also wish to consider other alternatives to  $B-F$ , including the adjusted  $F$ -tests mentioned above, as well as classical non-parametric tests such as the Friedman test, multivariate analysis, and the linear mixed model (LMM). However, simulation studies have shown that these procedures also have limitations and can become liberal with violations of sphericity and small sample sizes (Berkovits et al., 2000; Blanca et al., 2023b, 2024; Harwell & Serlin, 1994; Haverkamp & Beauducel, 2017, 2019; Hayoz, 2007). A further limitation of the LMM relates to problems identifying the true structure of the covariance matrix (Brown & Prescott, 2006). An interesting line of future research would therefore be to compare these procedures and to analyze how they perform when used in conjunction with the bootstrap method.

This study has a number of limitations that need to be acknowledged. First, the results are applicable only to the conditions studied here, that is, to designs containing 3, 4, and 6 repeated measures, and to non-normal distributions with values of skewness and kurtosis coefficients up to 2.31 and 8, respectively. Although



these conditions reflect a wide range of real-life scenarios, future research might focus on exploring the performance of  $B$ - $F$  in designs with a larger number of repeated measures, in more complex experimental designs that incorporate both within- and between-subject factors, and with distributions showing greater deviation from normality. Investigation of these scenarios will provide a deeper understanding of the applicability of the procedure in various research contexts. Second, we have considered the unstructured covariance matrix as being the most general structure. Further research might include other types of structures that contemplate serial correlation, such as autoregressive, heterogeneous autoregressive, Toeplitz, etc. This would help to extend knowledge about the robustness of  $B$ - $F$  under different dependency structures. Third, the data simulated here include complete cases without accounting for the presence of missing values. The importance of detecting patterns of missing data and mechanisms of loss, as well as selecting an appropriate imputation method, is widely acknowledged (Berglund & Heeringa, 2014; Vallejo et al., 2011). A possible avenue for further research would therefore be to analyze both Type I error and power of  $B$ - $F$  with different patterns of missing data and different imputation methods. Finally, the present study focuses on the comparison of untrimmed means, so it would be interesting to explore the performance of  $B$ - $F$  with trimmed means. Outliers often pose difficulties in data analysis, and the use of trimmed means is a procedure that can deal with this problem (Wilcox, 2022).

#### Author Contributions

**María J. Blanca:** Conceptualization, Methodology, Writing – Original draft, Formal Analysis. **Roser Bono:** Methodology, Software, Writing – Review and Editing. **Jaume Arnau:** Software, Writing – Review and Editing. **F. Javier García-Castro:** Methodology, Writing – Review and Editing. **Rafael Alarcón:** Methodology, Formal Analysis, Writing – Review and Editing. **Guillermo Vallejo:** Software, Writing – Review and Editing.

#### Acknowledgements

The authors would like to thank Macarena Torrado for her collaboration in this study.

#### Funding

This research was supported by the Ministry of Science and Innovation (grant PID2020-113191GB-I00 from the MCIN/AEI/ 10.13039/501100011033 and by funding from the Regional Government of Andalusia to Consolidated Research Group CTS110). This funding source had no role in the design of this study, data collection, management, analysis, and interpretation of data, writing of the manuscript, or the decision to submit the manuscript for publication.

#### Declaration of Interests

The authors declare that there are no conflicts of interest.

#### Data Availability Statement

Data are available at <https://dx.doi.org/10.24310/riuma.37706>






#### References

- Arnau, J., Bendayan, R., Blanca, M. J., & Bono, R. (2014). Should we rely on the Kenward–Roger approximation when using linear mixed models if the groups have different distributions? *British Journal of Mathematical and Statistical Psychology*, 67(3), 408–429. <https://doi.org/10.1111/bmsp.12026>
- Arnau, J., Bono, R., Blanca, M. J., & Bendayan, R. (2012). Using the linear mixed model to analyze nonnormal data distributions in longitudinal designs. *Behavior Research Methods*, 44(4), 1224–1238. <https://doi.org/10.3758/s13428-012-0196-y>
- Berglund, P., & Heeringa, S. (2014). *Multiple imputation of missing data using SAS*. SAS Institute Inc.
- Berkovits, I., Hancock, G., & Nevitt, J. (2000). Bootstrap resampling approaches for repeated measure designs: Relative robustness to sphericity and normality violations. *Educational and Psychological Measurement*, 60(6), 877–892. <https://doi.org/10.1177/00131640021970961>
- Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2018). Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit? *Behavior Research Methods*, 50(3), 937–962. <https://doi.org/10.3758/s13428-017-0918-2>
- Blanca, M. J., Arnau, J., García-Castro, F. J., Alarcón, R., & Bono, R. (2023a). Non-normal data in repeated measures: Impact on Type I error and power. *Psicothema*, 35(1), 21–29. <https://doi.org/10.7334/psicothema2022.292>
- Blanca, M. J., Arnau, J., García-Castro, F. J., Alarcón, R., & Bono, R. (2023b). Repeated measures ANOVA and adjusted  $F$ -tests when sphericity is violated: Which procedure is best? *Frontiers in Psychology*, 14, Article 1192453. <https://doi.org/10.3389/fpsyg.2023.1192453>
- Blanca, M. J., Alarcón, R., Arnau, J., García-Castro, F. J., & Bono, R. (2024). How to proceed when both normality and sphericity are violated in repeated measures ANOVA. *Anales de Psicología / Annals of Psychology*, 40(3), 466–480. <https://doi.org/10.6018/analesps.594291>
- Bono, R., Arnau, J., & Vallejo, G. (2010). Modelización de diseños split-plot y estructuras de covarianza no estacionarias: un estudio de simulación [Modeling split-plot data and nonstationary covariance structures: A simulation study]. *Escritos de Psicología / Psychological Writings*, 3(3), 1–7. <https://doi.org/10.5231/Psy.Writ.2010.2903>
- Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems II. Effect of inequality of variance and of correlation of error in the two-way classification. *Annals of Mathematical Statistics*, 25(3), 484–498. <https://doi.org/10.1214/aoms/1177728717>
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2), 144–152. <https://doi.org/10.1111/j.2044-8317.1978.tb00581.x>
- Brown, H., & Prescott, R. (2006). *Applied mixed models in medicine* (2nd edition). Wiley.
- Chernick, M. R. (2008). *Bootstrap methods: A guide for practitioners and researchers* (2nd ed.). John Wiley & Sons, Inc.
- Chernick, M. R., & LaBudde, R. A. (2011). *An introduction to bootstrap methods with applications to R*. John Wiley & Sons, Inc.
- Christensen, A. P., & Golino, H. (2021). Estimating the stability of psychological dimensions via bootstrap exploratory graph analysis: A Monte Carlo simulation and tutorial. *Psych*, 3(3), 479–500. <https://doi.org/10.3390/psych3030032>
- Cooper, J. A., & Garson, G. D. (2016). *Power analysis*. Statistical Associates Blue Book Series.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1), 1–26. <http://www.jstor.org/stable/2958830>

- Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician*, 37(1), 36-48. <https://doi.org/10.2307/2685844>
- Efron, B., & Tibshirani, R. J., (1993). *An introduction to the bootstrap*. Chapman & Hall.
- Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43(4), 521-532. <https://doi.org/10.1007/BF02293811>
- Geisser, S., & Greenhouse, S. W. (1958). An extension of Box's results on the use of the *F* distribution in multivariate analysis. *The Annals of Mathematical Statistics*, 29(3) 885-891. <https://doi.org/10.1214/aoms/1177706545>
- Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika* 24(2), 95-112. <https://doi.org/10.1007/BF02289823>
- Haverkamp, N., & Beauducel, A. (2017). Violation of the sphericity assumption and its effect on Type-I error rates in repeated measures ANOVA and multi-level linear models (MLM). *Frontiers in Psychology*, 8, Article 1841. <https://doi.org/10.3389/fpsyg.2017.01841>
- Haverkamp, N., & Beauducel, A. (2019). Differences of Type I error rates for ANOVA and multilevel-linear-models using SAS and SPSS for repeated measures designs. *Meta-Psychology*, 3, Article MP.2018.898. <https://doi.org/10.15626/mp.2018.898>
- Harwell, M. R., & Serlin, R. C. (1994). A Monte Carlo study of the Friedman test and some competitors in the single factor, repeated measures design with unequal covariances. *Computational Statistics & Data Analysis*, 17(1), 35-49. [https://doi.org/10.1016/0167-9473\(92\)00060-5](https://doi.org/10.1016/0167-9473(92)00060-5)
- Hayoz, S. (2007). Behavior of nonparametric tests in longitudinal design. *15th European young statisticians meeting*. [http://matematicas.unex.es/~idelpuerto/WEB\\_EYSM/Articles/ch\\_stefanie\\_hayoz\\_art.pdf](http://matematicas.unex.es/~idelpuerto/WEB_EYSM/Articles/ch_stefanie_hayoz_art.pdf)
- Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Publications.
- Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*, 1(1), 69-82. <https://doi.org/10.2307/1164736>
- Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Prentice Hall.
- Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1999). A comparison of recent approaches to the analysis of repeated measurements. *British Journal of Mathematical and Statistical Psychology*, 52(1), 63-78. <https://doi.org/10.1348/000711099158964>
- Keselman, J. C., Lix, L. M., & Keselman, H. J. (1996). The analysis of repeated measurements: A quantitative research synthesis. *British Journal of Mathematical and Statistical Psychology*, 49(2), 275-298. <https://doi.org/10.1111/j.2044-8317.1996.tb01089.x>
- Kherad-Pajouh, S., & Renaud, O. (2015). A general permutation approach for analyzing repeated measures ANOVA and mixed-model designs. *Statistical Papers*, 56(4), 947-967. <https://doi.org/10.1007/s00362-014-0617-3>
- Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Sage Publications.
- Kowalchuk, R. K., Keselman, H. J., Algina, J., & Wolfinger, R. D. (2004). The analysis of repeated measurements with mixed-model adjusted *F* tests. *Educational and Psychological Measurement*, 64(2), 224-242. <https://doi.org/10.1177/0013164403260196>
- Livacic-Rojas, P., Vallejo, G., & Fernández, P. (2010). Analysis of Type I error rates of univariate and multivariate procedures in repeated measures designs. *Communications in Statistics – Simulation and Computation*, 39(3), 624-664. <https://doi.org/10.1080/03610910903548952>
- Mair, P., & Wilcox, R. (2020). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods*, 52, 464-488. <https://doi.org/10.3758/s13428-019-01246-w>
- Muller, K., Edwards, L., Simpson, S., & Taylor, D. (2007). Statistical tests with accurate size and power for balanced linear mixed models. *Statistics in Medicine*, 26(19), 3639-3660. <https://doi.org/10.1002/sim.2827>
- Oberfeld, D., & Franke, T. (2013). Evaluating the robustness of repeated measures analyses: The case of small sample sizes and nonnormal data. *Behavior Research Methods*, 45(3), 792-812. <https://doi.org/10.3758/s13428-012-0281-2>
- Tabachnick, B. G., & Fidell, L. S. (2007). *Experimental design using ANOVA*. Thomson Brooks/Cole.
- Vallejo, G., Ato, M., Fernández, P., & Livacic-Rojas, P. (2013). Multilevel bootstrap analysis with assumptions violated. *Psicothema*, 25(4), 520-528. <https://doi.org/10.7334/psicothema2013.58>
- Vallejo, G., Cuesta, M., Fernández, M., & Herrero, F. (2006). A comparison of the bootstrap-*F*, improved general approximation, and Brown-Forsythe multivariate approaches in a mixed repeated measures design. *Educational and Psychological Measurement*, 66(1), 35-62. <https://doi.org/10.1177/0013164404273943>
- Vallejo, G., Fernández, M. P., Livacic-Rojas, P. E., & Tuero-Herrero, E. (2011). Comparison of modern methods for analyzing repeated measures data with missing values. *Multivariate Behavioral Research*, 46(6), 900-937. <https://doi.org/10.1080/00273171.2011.625320>
- Vallejo, G., Fernández, M. P., Tuero, E., & Livacic-Rojas, P. E. (2010). Análisis de medidas repetidas usando métodos de remuestreo [Analyzing repeated measures using resampling methods]. *Anales de Psicología / Annals of Psychology*, 26(2), 400-409.
- Voelkle, M. C., & McKnight, P. E. (2012). One size fits all? A Monte-Carlo simulation on the relationship between repeated measures (M) ANOVA and latent curve modeling. *Methodology*, 8(1), 23-38. <https://doi.org/10.1027/1614-2241/a000044>
- Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. Gulf Professional Publishing.
- Wilcox, R. R. (2022). *Introduction to robust estimation and hypothesis testing*. Academic Press.

Article

## ChatGPT Simulated Patient: Use in Clinical Training in Psychology

Ana Sanz<sup>1</sup> , José Luis Tapia<sup>2,3</sup> , Eva García-Carpintero<sup>1</sup> , J. Francisco Rocabado<sup>2</sup>   
and Lorena M. Pedrajas<sup>4</sup> 

<sup>1</sup>UNIE Universidad (Spain)

<sup>2</sup>Universidad Nebrija (Spain)

<sup>3</sup>Universitat de València (Spain)

<sup>4</sup>Universidad Alfonso X El Sabio (Spain)

### ARTICLE INFO

Received: 15/11/2024

Accepted: 17/02/2025

#### Keywords:

Artificial intelligence

Standardized patients

Psychology education

Clinical simulation

Student confidence

### ABSTRACT

**Background:** Incorporating artificial intelligence (AI) as standardized patients (SPs) in psychology education may enhance experiential learning and student confidence. The aim of the study was to analyze the effectiveness of using AI-based simulations to develop communication skills and influence psychology students' affective state. **Method:** A mixed-methods intervention study was conducted with 31 third-year psychology students. Participants engaged in clinical simulations using ChatGPT as an SP. Quantitative data on affective state, communication attitudes, and perceptions of knowledge and skills were collected pre- and post-intervention via questionnaires. Qualitative data were obtained through open-ended questions and a focus group. Data were analyzed using repeated measures ANOVA and thematic analysis. **Results:** Significant reductions in negative affect and increases in perceived knowledge and skills were observed post-intervention. No significant changes were found in communication attitudes. Qualitative findings supported the quantitative results, indicating improved confidence and reduced anxiety during simulated patient interactions. **Conclusions:** Utilizing AI as SPs is an effective pedagogical tool that enhances experiential learning, increases student confidence in professional skills, and positively influences the affective state. This innovative approach offers a valuable supplement to traditional teaching methods in psychology education.

## Paciente Simulado con ChatGPT: Utilización en la Formación Clínica en Psicología

### RESUMEN

**Antecedentes:** La integración de la inteligencia artificial (IA) como pacientes estandarizados (PE) en la educación en psicología puede mejorar el aprendizaje experiencial y la confianza de los estudiantes. Este estudio analizó la efectividad de simulaciones basadas en IA para desarrollar habilidades de comunicación e influir en el estado afectivo de estudiantes de psicología. **Método:** Estudio de intervención de métodos mixtos con 31 estudiantes de tercer año, utilizando ChatGPT como PE. Se recopilaban datos cuantitativos sobre estado afectivo, actitudes hacia la comunicación y percepciones de conocimiento y habilidades antes y después de la simulación. También se obtuvieron datos cualitativos mediante preguntas abiertas y un grupo focal. Los datos se analizaron mediante ANOVA de medidas repetidas y análisis temático. **Resultados:** Los resultados mostraron una disminución significativa en el afecto negativo y un aumento en la percepción de conocimiento y habilidades tras la simulación. Los hallazgos cualitativos respaldaron estos resultados, indicando mayor confianza y menor ansiedad en las interacciones con el paciente simulado. **Conclusiones:** El uso de IA como PE es una herramienta pedagógica eficaz que complementa los métodos tradicionales, mejora el aprendizaje experiencial, refuerza la confianza en habilidades profesionales y tiene un impacto positivo en el estado afectivo de los estudiantes.

#### Palabras clave:

Inteligencia artificial

Pacientes estandarizados

Educación en psicología

Simulación clínica

Confianza del estudiante



## Introduction

Simulation-based education in the health sciences has become a well-established pedagogical approach, offering a dynamic, practice-focused learning experience (Sezgin & Bektas, 2023). This method provides a secure and supervised setting wherein students can develop, practice and hone essential clinical abilities without compromising patient safety or experiencing the stressors associated with direct interaction with actual patients. Additionally, it enables students to cultivate assurance in their capacity to perform effectively in intricate clinical scenarios, equipping them with the skills to navigate the complexities of genuine healthcare settings (Ton et al., 2024).

One of the most promising strategies within this approach is simulation with standardized patients (SPs) (Flanagan & Cummings, 2023). SPs, individuals trained to accurately represent various clinical conditions, have been demonstrated to be effective tools in both formative and summative teaching (Hillier et al., 2023). These SPs are capable of realistically simulating a variety of symptoms, behaviors, and emotions, providing direct and invaluable feedback on the students' performance during simulation sessions (Gerzina & Stovsky, 2023). The use of SPs has been demonstrated to be an efficacious method for the enhancement of clinical skills, decision-making and communication in students of health sciences (Johnson et al., 2020). Recent research indicates that the use of SP-based simulation has the potential to enhance the educational experience and facilitate active learning (Burrell et al., 2023; Dawood et al., 2024; Monahan et al., 2024).

Anxiety and confidence in patient care situations are factors that impact students' clinical performance. Students entering practice frequently encounter deficiencies in their knowledge, clinical skills and competencies in patient communication, which can give rise to feelings of insecurity and anxiety in the clinical setting. Simulation in a controlled environment can facilitate the development of greater confidence and stability, which in turn enhances performance in patient care. Research has identified several factors that influence students' perceptions of confidence and safety after simulation experiences (Basnet et al., 2024; Hawkins & Tredgett, 2016; Yu et al., 2021). In fact, Clinard (2022), all students indicated that these simulations significantly enhanced their confidence in treating patients, particularly in complex scenarios. This practical experience enables them to reinforce their ability and confidence in their own clinical skills. Similarly, Moss (2023) found that students exhibited a notable enhancement in their confidence ratings following the simulation, along with expressing high levels of satisfaction with the educational experience. Such exercises assist students in managing their fears and anxieties prior to encountering authentic care environments. In a focus group session, teachers also discussed how they perceived an increase in students' practical skills and overall satisfaction, which lends further support to the idea that simulation is an effective method to prepare students for clinical care (Carrero-Planells et al., 2021).

Notwithstanding its advantages, clinical simulation practices remain less prevalent within the Bachelor's Degree in Psychology compared to other health sciences fields, such as Nursing (García-Carpintero et al., 2024). A review of the literature revealed that studies on simulation in psychology are relatively scarce in Spain, with only a few investigations by Ruiz-Rodríguez et al. (2016) and Rodríguez et al. (2021). Moreover, recently published findings by López & López-Chicheri (2024) highlight that incorporating this

pedagogical approach enhances psychology students' self-efficacy in their competencies and increases their satisfaction with experiential learning. The authors further emphasize the need to extend the duration of simulations with SP within the psychology program. This would provide students with more comprehensive clinical experience, ensuring they are well-prepared before entering formal placements and engaging in direct patient contact. It is interesting to mention the recent publication by Baile (2024), which aims to validate a patient profile in psychology generated with artificial intelligence.

Advancements in artificial intelligence (AI), particularly the emergence of Large Language Models (LLMs), have generated new opportunities across various industries, including clinical simulation. LLMs, such as ChatGPT developed by OpenAI®, are sophisticated AI systems trained on vast amounts of text, allowing them to understand and generate natural language in a way that mimics human communication. By leveraging deep learning techniques, these models engage in conversational and adaptive interactions, responding fluidly to a wide range of inputs. In clinical simulation, LLMs can act as virtual patients, offering realistic, personalized interactions that enrich the educational experience and support the development of essential clinical competencies in psychology students (Isaza-Restrepo et al., 2018; Scherr et al., 2023). Human and AI-based simulated patients each offer dynamic, context-driven interactions but differ in how they adapt, display emotions, and provide feedback. Whereas human patients exhibit genuine emotional responses shaped by cultural and social contexts, AI simulations rely on programmed algorithms. Early chatbots like ELIZA (Weizenbaum, 1966) and AIML-based systems established foundational conversational structures but were limited by rule-based designs and minimal contextual awareness (Gutiérrez-Maldonado et al., 2008; Peñaloza-Salazar et al., 2011; Rizzo et al., 2011; see Gutiérrez-Maldonado et al., 2017, for a recent demonstration of an AIML-based system in VR). The emergence of large language models (LLMs) significantly expanded chatbot capabilities, allowing for more natural, flexible dialogue and enhanced contextual depth. This innovation makes simulated patients particularly valuable for training: they offer immediate feedback, reduce costs, and enable large-scale practice environments (Liu et al., 2023). Building on studies such as Scherr et al. (2023), which used ChatGPT for general clinical training, this approach tailors LLM-powered simulated patients to the field of psychology. While Scherr's study focuses on general medical scenarios, this application specifically trains students in psychology, enabling them to practice diagnosis and intervention for psychological disorders through interactive simulations. This approach harnesses AI's text-processing adaptability to provide students with lifelike chat encounters, allowing for more personalized learning and expanding opportunities for objective assessment, ultimately enhancing students' cognitive growth and self-efficacy in simulated scenarios (Morcela, 2022).

The main objective of the study was to analyse the feasibility and effectiveness of integrating artificial intelligence as a common pedagogical tool in the teaching of psychology. This general objective is in turn divided into two specific objectives:

1. To determine the impact of clinical simulation, through the use of AI-driven as SPs, on the improvement of self-perception of knowledge, students' affective state and communication skills.

- To explore the students' perspective on the appropriateness of AI-driven as SPs and the learning opportunity derived from its use.

## Method

A mixed-methods intervention study design was employed, incorporating a qualitative component (Fetters et al., 2013). Following an integrated concurrent design (Curry & Nunez-Smith, 2015), quantitative methodology was used in the first phase of data analysis, with qualitative methodology employed in the second phase. Qualitative data were collected post-intervention to elucidate potential mechanisms and explain the quantitative results. Quantitative and qualitative data were methodologically integrated by embedding one within the other, and jointly interpreted and reported through narrative and combined presentation approaches (Johnson, 2019). The study utilized ChatGPT-3.5 (OpenAI, 2023) as a large language model to simulate patient interactions during the intervention.

The study was conducted in accordance with the principles of the Declaration of Helsinki of the World Medical Association (WMA), and the protocol was approved by the Research Ethics Committee of University Nebrija (approval number UNNE-2024-0020). All participants were thoroughly informed, given the opportunity to ask questions, and provided their consent through signed forms for the focus group recording and for their inclusion in the quantitative and qualitative studies.

## Participants

To determine the required sample size, a statistical power analysis was conducted using G\*Power software (v3.1.9.7; Faul et al., 2007), employing the 'ANOVA: Repeated measures, within factors' statistical test appropriate for repeated measures. This analysis was performed under the assumption of a medium effect size ( $f = 0.3$ ), in the absence of specific prior data. A significance level of 0.05 and a planned power of 0.85 were set, establishing that a total of 27 participants would be necessary to reliably assess the pre and post-intervention changes.

The study sample consisted of 31 third-year psychology students (74% female) from a private university, recruited through convenience sampling. One participant identified as non-binary. Mean age was 21.03 years ( $SD = 1.43$ ).

For the qualitative phase, a sampling method based on the information power criteria was used. This approach suggests that the more relevant the information provided by the sample is to the study, the fewer participants are needed (Moser & Korstjens, 2018). Therefore, the same participants recruited for the intervention in the quantitative phase who agreed to participate were included in the focus group ( $n = 12$ ). None of them withdrew from the study.

## Instruments

The study employed a mixed-methods design, combining quantitative and qualitative approaches to thoroughly evaluate the intervention. Four quantitative instruments were used (see [https://osf.io/se7dq/?view\\_only=891d4fb6d1304f6597496bf69b29319](https://osf.io/se7dq/?view_only=891d4fb6d1304f6597496bf69b29319)), two of which were specifically created for this study. These instruments provided precise data on participants'

knowledge, professional competencies, social impact, attitudes toward communication, and emotional well-being.

**PANAS: Positive and Negative Affect Schedule** (Watson et al., 1988) (adaptation to Spanish, López-Gómez et al., 2015). A 20-item questionnaire that measures individuals' positive (PA) and negative (NA) affects. The items are divided into two subscales: one for positive affects (such as joy and enthusiasm) and another for negative affects (such as sadness and irritability). Each item is rated on a scale from 1 to 5, where 1 indicates that the *affect has not been experienced* at all and 5 indicates a *very intense experience*. The PANAS is widely used in both academic research and clinical applications to assess emotional well-being. The direct score ranges from 20 to 100. In a general sample from Spain (López-Gómez et al., 2015) the Pearson's bivariate correlation between the PA and NA scales was  $-0.19$  ( $p < 0.001$ ) and Cronbach's alpha was 0.92 for Positive Affect Scale and 0.88 for Negative Affect Scale. The item-total correlations of the PA factor ranged from 0.67 to 0.74, while those of the NA factor ranged between 0.52 and 0.69.

**HCAS: Healthcare Communication Attitudes Scale** (Escribano et al., 2021). An 11-item scale designed to assess healthcare professionals' attitudes towards communication in clinical settings. Each item is rated on a scale from 1 to 5, with higher scores indicating a more positive attitude towards effective communication. The HCAS helps identify professionals' perceptions and predispositions regarding the importance of communication in patient care, facilitating the implementation of training and development programs that enhance these critical skills in clinical practice. The direct score ranges from 11 to 55. In a sample of 255 nursing students with an average age was 22.66 years ( $SD = 4.75$ ) and 82% were female, the internal consistency of the scale was adequate (0.75), and the data fit well with the model (CFI = 0.99; TLI = 0.99; RMSEA = .01 95% CI [.00-.05]). The overall instrument score poorly correlated with the self-efficacy in communication skills variable.

**PIES: Perception and Impact Evaluation Scale.** An ad hoc tool consisting of three items designed to measure students' self-perception of their knowledge, professional competencies, and the social impact of their field of study. Each item is rated on a scale from 1 to 5, where 1 indicates a *low level of perception* or impact and 5 indicates a *high level*. This instrument aims to evaluate the development of key competencies and social awareness among students, providing valuable data to improve educational programs and pedagogical interventions. The direct score ranges from 3 to 15.

**SPI-MET: Simulated Patient Interaction Measurement & Evaluation Tool.** An 11-item questionnaire designed ad hoc to evaluate healthcare professionals' performance in interactions with simulated patients. The items focus on aspects such as linguistic adequacy and emotional expression, with each item rated on a scale from 0 to 5, where 0 indicates *inadequate performance* and 5 indicates *excellent performance*. In addition, the instrument includes an extra item to assess the overall adequacy of the tool as a simulated patient, rated on a scale from 0 to 10, where 0 indicates poor performance and 10 indicates excellent performance.

The collection of qualitative data was carried out using two main methods: participants' responses to an open-ended question included in the SPI-MET ("What would you add or improve about the activity?") and a focus group consisting of 12 participants conducted after the intervention, led by an observer and a moderator. This focus group encouraged participant interaction, fostering the emergence of

diverse opinions and perceptions. Information was gathered using a question guide developed from a prior literature review, focusing on specific topics of interest (Table 1). This facilitated an in-depth exploration of students' perceptions and experiences related to the categories proposed by the aforementioned instruments. The focus group was audio recorded with prior consent from the participants, lasting 57 minutes. Additionally, researchers' field notes were used as a secondary source of information to provide more detailed insights and support the data obtained in the focus group. The qualitative methodology offered a rich and contextualized perspective on their interaction with artificial intelligence in a clinical simulation setting.

**Table 1**  
*Categories and Focus Group Questions*

Categories	Focus group questions
Keen	How did you feel during the activity interacting with ChatGPT? How do you think this activity has influenced your safety? How do you think simulation has influenced your anxiety?
Patient Communication	How would you evaluate the way ChatGPT communicated as a patient? Did you find ChatGPT's behavior as a patient realistic and appropriate for the activity? Do you think ChatGPT's answers to your questions were appropriate and consistent with the situation?
Utility of the tool	How would you rate ChatGPT's performance as a simulated patient? How would you describe the degree of difficulty you experienced using ChatGPT during the activity? Were there any technical or interface aspects that made the tool difficult to use? Do you think this activity is useful to improve your skills in the subject?
Overall satisfaction	What level of overall satisfaction did you experience with the activity as a whole? What aspects of the activity would you highlight as positive or negative?

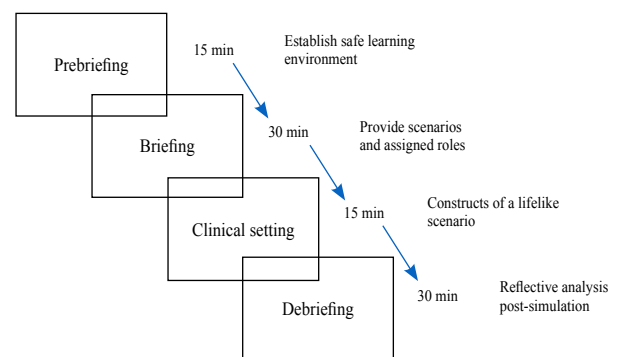
## Procedure

Before the experimental session, a clinical case was developed based on a pathology previously studied by students in the Mood Disorders module. Three clinical psychologists reviewed the case to ensure its suitability, after which the AI training process began. This involved two primary tasks: first, equipping the AI with sufficient information to address students' questions accurately, and second, fine-tuning its responses to maintain a consistent patient role. To create a realistic persona, the AI was given a detailed life history and personal profile, enhancing the coherence and authenticity of its responses (Pedrajas et al., 2024). Careful selection of verbs, instructions, and specific prompts shaped the AI's responses to embody the character and communication style needed for the exercise. The AI's identity was further defined by essential sociodemographic traits and communication aspects, supported by a dedicated chat and clear clinical context to prevent inconsistent responses. After establishing a coherent character profile, a pilot test was conducted with 10 subjects outside the experimental group to identify any unusual responses from the large language model (LLM). This preliminary test ensured the model's stability and reliability in delivering consistent, relevant answers. The clinical

case is accessible at [https://osf.io/se7dq/?view\\_only=891d4fb6d1304f6597496bfe69b29319](https://osf.io/se7dq/?view_only=891d4fb6d1304f6597496bfe69b29319).

With all materials prepared, the experiment was conducted in a single 2-hour session as part of the students' curriculum, facilitated by their regular professor and supported by two additional instructors. Each student had access to a computer and interacted with the same case study, in a psychological assessment first interview simulation scenario; however, each interaction was unique due to the LLM's adaptive responses. The clinical simulation process followed structured stages: *Prebriefing*, *Briefing*, *Clinical Setting*, and *Debriefing* (Duff et al., 2024; Kolbe et al., 2015) (Figure 1). Evaluation instruments, including PIES, PANAS, and HCAS, were administered individually both before the Prebriefing and after the Debriefing to assess changes across the simulation. The SPI-MET and focus group assessments, designed specifically for post-simulation feedback, were conducted only at the end of the session.

**Figure 1**  
*Stages of the Clinical Simulation Process*



## Data Analysis

The quantitative analysis assessed internal consistency using Cronbach's alpha and McDonald's omega for instrument reliability. Sensitivity analysis evaluated sample size adequacy (Lakens, 2022; Perugini et al., 2018). Repeated measures ANOVA compared pre- and post-scores to examine intervention effectiveness. Due to small sample size and limited gender diversity, gender variables were excluded. Bayesian hypothesis testing strengthened evidence for each instrument (Rouder et al., 2009, 2012).

Thematic analysis identified excerpts relevant to the research question (Nowell et al., 2017). Open-ended responses were descriptively coded, triangulated with focus group transcripts, and categorized manually, ensuring study reliability (Moser & Korstjens, 2018). Categories reflected variables measured by instruments to verify consistency. Adhering to COREQ guidelines (Tong et al., 2007), the mixed-methods approach enhanced quantitative reliability and deepened understanding of the phenomenon.

## Results

The internal consistency of the positive and negative affect subscales of the PANAS, as well as the HCAS scale, was evaluated using Cronbach's alpha and McDonald's omega coefficients. The results indicated excellent internal consistency for both PANAS subscales in both pre and post assessments (Cronbach's  $\alpha > 0.85$ ).

and McDonald's  $\omega > 0.86$ ). In contrast, the internal consistency of the HCAS scale was moderate to low, with Cronbach's alpha values of 0.570 (pre) and 0.500 (post), and McDonald's omega values of 0.695 (pre) and 0.701 (post).

### Changes in Affective States

Sensitivity analysis on the PANAS scores was conducted using G\*Power software (Faul et al., 2007) to determine the minimum detectable effect size for this study design. The analysis used four measures F tests, with an alpha level ( $\alpha$ ) of 0.05, an expected power of 0.85, and a total sample size of 31 participants. The result indicated that the minimum detectable effect corresponded to  $F = 0.23$ , with the critical F-value set at  $F_{critical}(3, 90) = 2.71$ . Therefore, F values equal to or larger than this threshold indicate a statistically meaningful result, corresponding to a probability of less than 5% under the null hypothesis.

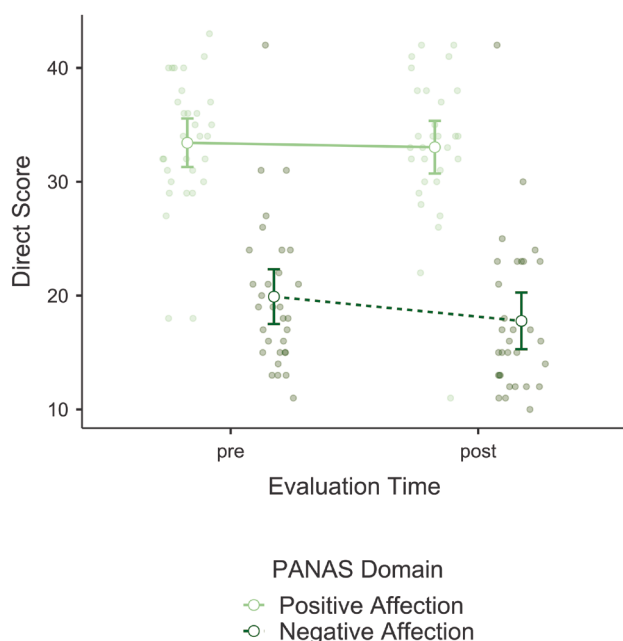
Prior to conducting the analysis an inspection of distributional assumptions indicated no significant violations. Consequently, we proceeded with the planned analysis. The repeated measures ANOVA revealed significant main effects for both Affect (positive and negative) and evaluation time (pre- and post-simulation) on PANAS scores (Figure 2). A significant main effect of Affect was found ( $F(1,30) = 55.17, p < .001, \eta^2p = 0.648$ ) indicating that positive affect scores were significantly higher than negative affect scores ( $M_{Diff} = 14.4$ ), with a large effect size demonstrating substantial impact. Additionally, a significant main effect of Evaluation Time was observed ( $F(1,30) = 15.67, p < .001, \eta^2p = 0.343$ ), showing lower score after the intervention ( $M_{Diff} = 1.26$ ). Furthermore, the interaction between affect and evaluation time was also significant ( $F(1,30) = 5.06, p < .032, \eta^2p = 0.144$ ). Post hoc analysis showed that negative affect scores decreased significantly from pre- to post-

simulation ( $M_{Diff} = 2.129, SE = 0.447, p_{Bonf} < .001$ ), while positive affect scores did not show significant differences from pre- to post-simulation ( $M_{Diff} = 0.387, SE = 0.550, p_{Bonf} = 1.000$ ).

The qualitative analysis reinforces the quantitative results, showing a clear prevalence of positive affective responses (19) over negative ones (15). Some participants highlighted difficulties in fully engaging due to the virtual nature of the interaction, noting that *"it's a more superficial situation than having the patient face-to-face"* (GF:30), suggesting that the lack of direct contact may influence the perception of authenticity in the experience. However, both in the focus group and open-ended responses, the positive impact of this practice on the development of professional skills was confirmed. For example, *"it helps us improve our therapeutic skills and become familiar with some cases, to get some practical preparation before internships, especially for those of us who want to specialize in clinical practice. It really helps us lose that 'fear' of facing a patient, even if it's just a simulation"* (P31), and in stress management, *"although it also helped ease my nerves not seeing the patient's face"* (GF:7). A trend towards reduced anxiety among students was also observed, compared to face-to-face interactions with real patients: *"being able to ask directly is really helpful and makes you feel less shy than if they were in front of you"* (GF:28). These findings suggest that AI can provide a less intimidating and more accessible learning environment, reducing anxiety and facilitating more effective development of clinical skills.

To further explore the effects on PANAS scores, we conducted Bayesian hypothesis testing comparisons to compute the Bayes Factors for the comparison of positive and negative affect scores across pre- and post-evaluation conditions. For negative affect scores, the Bayes Factor for the alternative hypothesis ( $BF_{10}$ ) was 532.675, indicating extreme evidence in favor of the presence of a significant difference between pre- and post-simulation scores. Conversely, for positive affect scores, the Bayes Factor for the alternative hypothesis ( $BF_{10}$ ) was 0.241, providing anecdotal evidence against a significant difference. The Bayes Factor for the null hypothesis ( $BF_{01}$ ) was 4.153 for positive affect scores, supporting the absence of a meaningful difference across conditions. These results suggest a strong effect of evaluation time on negative affect but no substantial changes in positive affect scores.

**Figure 2**  
Mean Scores for PANAS Positive and Negative Domains with Error Bars Representing the 95% Confidence Intervals



### Attitudes Toward Communication

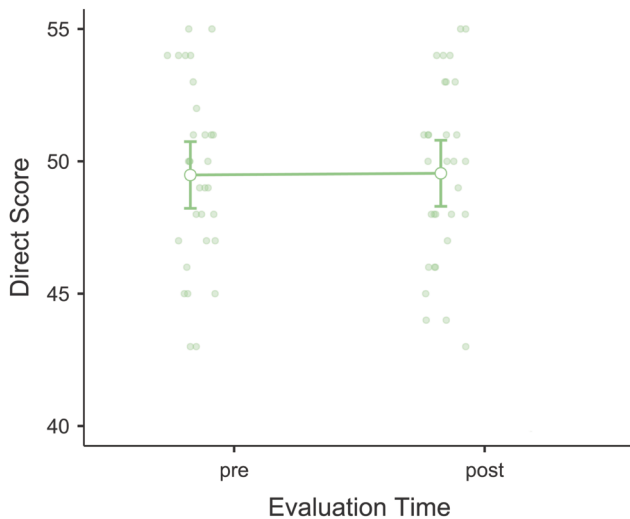
For the HCAS scores, the sensitivity analysis was conducted using two measures with an alpha level of 0.05, an expected power of 0.85, and a sample size of 31 participants. The analysis determined that the minimum detectable effect size corresponded to  $F = 0.28$ , with a critical F-value of  $F_{critical}(1, 30) = 4.17$ , indicating the threshold for statistical significance at the given parameters.

A previous analysis of the normality distributional assumptions showed that normality was compromised. Thus, a complementary non-parametric analysis is presented to support our findings. The repeated measures ANOVA conducted on the HCAS scale revealed no significant main effect of evaluation time ( $F(1, 30) = 0.017, p = .896, \eta^2p = 0.001$ ). Similarly, the non-parametric analysis showed no significant differences on evaluation time ( $W(31) = 127, p = 1.000$ ). Both tests indicate that there were no significant changes in attitudes toward healthcare communication from pre- to post-intervention, suggesting that the intervention did not influence these attitudes measurably (Figure 3).



**Figure 3**

Mean Scores for HCAS with Error Bars Representing the 95% Confidence Intervals



The Bayes Factor for the alternative hypothesis  $BF_{10} = 0.193$ , indicating substantial evidence against a significant difference between pre- and post-evaluation scores. For the null hypothesis  $BF_{01} = 5.178$ , providing strong support for the absence of differences in HCAS scores across the two time points. These results suggest that there is no meaningful change in HCAS scores from pre- to post-simulation, providing robust evidence in favor of the null hypothesis. In this case, the qualitative data suggest that students perceive this practice as an opportunity to refine their already acquired skills. Some participants highlighted the ability to steer the direction of the conversation during the simulation, noting that they “*were able to practice changing the direction of the conversation based on the patient’s responses*” (GF:56). Additionally, the usefulness of these practices for applying theoretical knowledge and gaining confidence was emphasized: “*These types of practices help us put into practice all the theoretical knowledge we acquire and help us gain confidence in ourselves*” (P13). However, there is no reference to the acquisition of new skills.

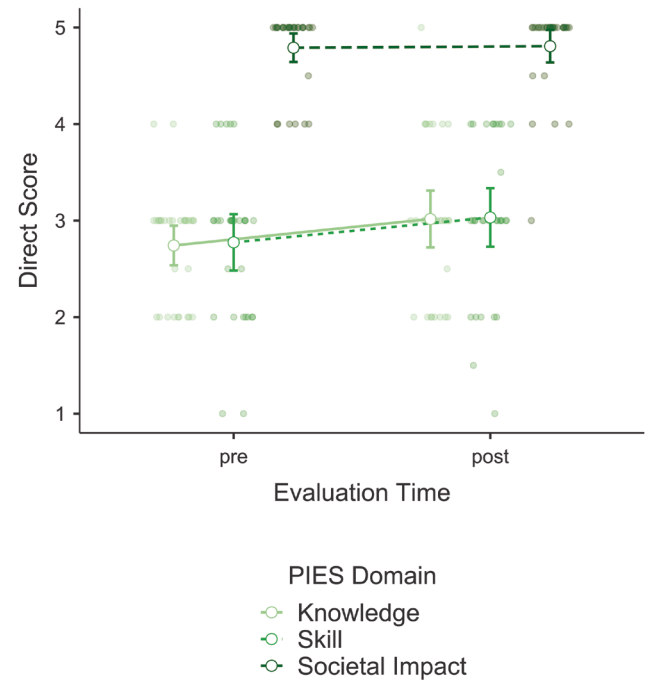
### Perceptions of Knowledge, Skills, and Social Value

Similarly, analyzing the PIES scores with six measures and the same parameters showed that the minimal detectable effect size corresponds to  $F = 0.20$  ( $F_{critical}(1, 30) = 2.28$ ).

Prior to conducting the analysis, an inspection of distributional assumptions and Mauchly’s test of sphericity indicated no significant violations, allowing us to proceed with the planned analysis. The repeated measures ANOVA for the PIES revealed significant main effects for both domain and evaluation time (Figure 4). A substantial effect of domain on PIES scores ( $F(2, 60) = 150.58$ ,  $p < .001$ ,  $\eta^2p = .834$ ), indicated significant variability across the domains of knowledge, skills, and social value, highlighting a large effect size. Additionally, a significant main effect of evaluation time ( $F(1, 30) = 7.52$ ,  $p = .010$ ,  $\eta^2p = 0.200$ ), showed notable changes in scores from pre- to post-evaluation, suggesting a medium effect size. However, the interaction between domain and evaluation time was not significant ( $F(2, 60) = 2.84$ ,  $p = .067$ ,  $\eta^2p = 0.086$ ),

**Figure 4**

Mean Scores for PIES Domains with Error Bars Representing the 95% Confidence Intervals



indicating a small effect size and suggesting that changes over time did not differ significantly across the domains. Post hoc comparisons revealed that both knowledge ( $M = 2.74$ ,  $SD = 0.561$ ) and skills scores ( $M = 2.77$ ,  $SD = 0.794$ ) were significantly higher than social value scores ( $M = 4.79$ ,  $SD = 0.404$ ) (all  $p < .001$ ), with large effect sizes evident in these differences. Paired samples t-tests further revealed that knowledge perception scores increased significantly from pre- ( $M = 2.74$ ,  $SD = 0.561$ ) to post-intervention ( $M = 3.02$ ,  $SD = 0.801$ ), ( $t(30) = 2.655$ ,  $p = .013$ ,  $d = -0.477$ ), indicating a medium effect size. Similarly, skills perception scores increased significantly from pre- ( $M = 2.77$ ,  $SD = 0.794$ ) to post-intervention ( $M = 3.03$ ,  $SD = 0.826$ ), ( $t(30) = 2.108$ ,  $p = .043$ ,  $d = 0.379$ ), also reflecting a medium effect size. In contrast, social value impact scores showed no significant difference from pre- ( $M = 4.79$ ,  $SD = 0.404$ ) to post-intervention ( $M = 4.81$ ,  $SD = 0.460$ ), ( $t(30) = -0.329$ ,  $p = .745$ ,  $d = 0.059$ ), indicating a very small effect size. The qualitative analysis suggest that students perceive the simulation as an exercise comparable to a clinical interview with a real patient, which has allowed them to establish smooth communication and guide the interview towards the most relevant topics in the context of the case: “*It gave me the chance to practice not going blank and managing the process of organizing my thoughts*” (GF:55). Additionally, they reported having applied the basic therapeutic skill of empathy during the exercise, despite it being an AI-based experience. This helped them identify the main clinical manifestations of the case, explore the problem’s history, and suggest a potential psychopathological diagnosis “*We share a common fear, and these practices help you understand your tools*” (GF:62), reflecting how the activity boosted their confidence in managing their clinical skills. Several students emphasized that the activity provided a valuable opportunity to apply

the theoretical knowledge they had acquired in a simulated practical setting: *"I think it's a great activity and, overall, a fantastic initiative that, in my view, should be done more often"* (P1).

Finally, the Bayesian paired samples t-tests for the alternative hypothesis ( $BF_{10} = 4.389$ ) indicate moderate evidence in favor of a significant difference between pre- and post-evaluation PIES scores. Conversely, the Bayes Factor for the null hypothesis ( $BF_{01} = 0.228$ ) provides weak evidence against the absence of differences. These results suggest that there is a notable change in PIES scores across the evaluation periods, with evidence supporting the presence of a significant effect.

### Evaluation of the AI as a Simulated Patient

The post-simulation evaluation using SPI-MET assessed the students' perceptions of the performance of the large language model (LLM) acting as a simulated patient. Out of the total participants, two did not complete the instrument, resulting in 29 valid responses. Descriptive analysis revealed a mean SPI-MET score of 3.99 ( $SD = 0.597$ ), indicating a generally favorable assessment of the LLM's performance as a simulated patient. Similarly, the mean score given by the students for the additional question, "How would you rate the tool's ability as a simulated patient on a scale of 0 to 10?" was 8.28 ( $SD = 1.13$ ). The frequency distribution showed that 3.4% of the students rated the LLM's performance as 6, 24.1% rated it as 7, 31.0% rated it as 8, 24.1% rated it as 9, and 17.2% rated it as 10. These results indicate that the majority of students rated the LLM's performance highly, with most ratings falling between 7 and 9, suggesting a generally positive perception of the LLM's effectiveness in simulating patient interactions. These data are corroborated in the focus group comments, which highlight how realistic the practice felt: *"It didn't seem impersonal, the responses were long because we weren't face-to-face, but the language used was conversational"* (GF:17). Another participant added: *"It seemed very realistic to me"* (GF:61). The ease of maintaining a meaningful conversation was emphasized, and it was noted that the system's ability to provide detailed responses was likely due to the lack of direct visual communication. Several students pointed out that the language used was notably straightforward: *"The language didn't seem unrealistic to me, but it was a bit formal and direct"* (GF:13). However, some expressed that the IA responses could feel cold and repetitive, with a certain robotic quality, though this did not significantly impact the empathetic nature of the interaction *"The patient repeated the same thing several times, even when asked to elaborate"* (P23).

### Discussion

In psychology education, theoretical knowledge of mental disorders and interventions must be paired with practical experiences to prepare students for real-world challenges. Clinical simulations bridge theory and practice, fostering competencies in a safe environment. However, research on their effectiveness in psychology is limited (Ruiz-Rodríguez et al., 2016; Rodríguez et al., 2021). This study assessed the feasibility and educational value of using AI-based simulations in a first-session patient interview scenario.

This study revealed several notable findings. Quantitatively, significant reductions in negative affect (PANAS) were observed post-intervention, emphasizing the ability of AI-based simulations to mitigate anxiety and stress in a controlled, low-risk environment.

Qualitative data reinforced the observed reductions in students' negative affect, indicating that this intervention helped to mitigate stress and anxiety—emotions that can negatively impact clinical performance. These outcomes align with previous research demonstrating the effectiveness of simulation-based training in reducing anxiety and building confidence compared to traditional methods (Abbott et al., 2021; Oliveira et al., 2022). However, no significant changes were detected in positive affect, potentially reflecting a ceiling effect or suggesting that the intervention primarily targeted stress reduction rather than enhancing positive emotional states. This fact is reflected in stress and coping theories (Lazarus & Folkman, 1984), which propose that interventions targeting perceived stress—such as simulated practice—can effectively lower negative affect without necessarily increasing positive emotions.

Similarly, the analysis of PIES scores demonstrated a significant improvement in students' perceptions of their knowledge and clinical skills, further supporting the pedagogical value of AI-simulated patients. Such an approach aligns with existing theoretical models, such as experiential learning theory (Kolb, 2014), which emphasizes the importance of hands-on, reflective practice in skill acquisition. Qualitative findings provided additional depth to these results. Students reported feeling more confident and better prepared to handle clinical scenarios after the intervention, likely due to the controlled environment that simulations offer (Elendu et al., 2024). They appreciated the opportunity to apply theoretical knowledge in a simulated practical setting, particularly valuing the structured feedback and safe environment that allowed them to refine their communication and diagnostic skills. These observations align with previous studies on virtual patients, which have highlighted their value in developing essential health science skills such as clinical interviewing and reasoning (Sezer et al., 2023; Cho & Kim, 2024; Jeon et al., 2024). Conversely, scores from the HCAS indicated no significant changes in communication attitudes, suggesting that this aspect might require more prolonged or varied interventions for measurable improvements.

As per the evaluation of the AI as a simulated patient, students described their interactions with ChatGPT as strikingly similar to real-life conversations. They valued the natural language and conversational flow, although some noted that the lack of visual interaction allowed for more detailed verbal responses. A few students did report that certain responses felt somewhat repetitive or lacked emotional depth, indicating that while ChatGPT performs effectively as a simulated patient, improvements in emotional expressiveness and naturalism are still possible.

The results suggest that incorporating AI-simulated patients can foster an active learning environment where students practice basic clinical and communication skills in controlled, simulated scenarios (Alrashidi et al., 2023; Liu et al., 2023; Farina et al., 2024). These initial findings indicate that such simulations may contribute to a more dynamic and participatory learning process, providing opportunities for students to apply theoretical knowledge in a practical setting (Higgins et al., 2021). While preliminary, the present findings point to the potential of AI-based simulation, specifically using ChatGPT, as a complementary pedagogical tool in psychology education. This approach offers a promising addition to traditional methods, providing opportunities for experiential learning that are scalable and adaptable. The value of integrating new technologies

into psychological and educational training has been highlighted by Elosua et al. (2023), suggesting that such tools can support student preparedness and confidence in professional skills.

This study highlights limitations, notably the rapid evolution of AI technologies, which complicates their long-term applicability in clinical education. The research evaluated AI as a simulated patient for basic therapist competencies, such as communication, emotional engagement, and interaction, but did not address its potential in enhancing diagnostic accuracy or advanced therapeutic skills. It also lacked comparisons between AI modalities like audio systems and chatbots, which could enhance realism. Ensuring AI aligns with psychometric standards and mental health frameworks is essential (Elosua et al., 2023).

Future research should examine AI-based simulations' transferability to clinical settings, their impact on diagnostic and decision-making skills, and integration into broader pedagogical frameworks. Advances in NLP and machine learning, such as AIML and LLMs, offer increased flexibility and interaction complexity but face challenges in ensuring safety, explainability, and domain-specific accuracy. Combining rule-based and LLM approaches and integrating conversational models into real-world environments with human-like agents presents both opportunities and challenges (Talbot & Rizzo, 2019).

#### Author Contributions

**Ana Sanz:** Conceptualization, Investigation, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing. **Jose Luis Tapia:** Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Eva Garcia-Carpintero:** Conceptualization, Formal analysis, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. **Francisco Rocabado:** Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Lorena Pedrajas:** Conceptualization, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing.

#### Funding

This work was supported by the State Plan for Scientific and Technical Research and Innovation of the Government of Spain (FPU19/02239); and by the Ministry of Science, Innovation and Universities, the State Research Agency, and the European Social Fund Plus (JDC2023-052644-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.

#### Declaration of Interests

The authors declare that there is no conflict of interest.

#### Data Availability Statement

The data supporting the findings of this research are accessible at the following link: [https://osf.io/se7dq/?view\\_only=891d4fb6d1304f6597496bfe69b29319](https://osf.io/se7dq/?view_only=891d4fb6d1304f6597496bfe69b29319)

#### References

- Abbott, E. F., Laack, T. A., Licatino, L. K., Wood-Wentz, C. M., Warner, P. A., Torsher, L. C., Newman, J. S., & Rieck, K. M. (2021). Comparison of dyad versus individual simulation-based training on stress, anxiety, cognitive load, and performance: a randomized controlled trial. *BMC Medical Education*, 21(1), Article 367. <https://doi.org/10.1186/s12909-021-02786-6>
- Alrashidi, N., Pasay An, E., Alrashedi, M. S., Alqarni, A. S., Gonzales, F., Bassuni, E. M., Pangket, P., Estadilla, L., Benjamin, L. S., & Ahmed, K. E. (2023). Effects of simulation in improving the self-confidence of student nurses in clinical practice: a systematic review. *BMC Medical Education*, 23(1), Article 815. <https://doi.org/10.1186/s12909-023-04793-1>
- Baile, J. I. (2024). Patient with depression created by freely accessible artificial intelligence for the teaching of Psychology. Preliminary study of its validity. *Tecnología, Ciencia y Educación*, 27, 7-42. <https://doi.org/10.51302/tce.2024.19069>
- Basnet, S., Shrestha, S. P., Shrestha, R., Shrestha, A. P., Shrestha, A., Sahu, S., Mhatre, B., & Silwal, P. (2024). Effect of simulation-based emergency airway management education on the knowledge, skills and perceived confidence of medical interns. *Annals of Medicine and Surgery*, 86(9), 5191-5198. <https://doi.org/10.1097/MS9.0000000000002376>
- Burrell, S. A., Ross, J. G., D'Annunzio, C., & Heverly, M. (2023). Standardized patient simulation in an oncology symptom management seminar-style course: prelicensure nursing student outcomes. *Journal of Cancer Education: the official journal of the American Association for Cancer Education*, 38(1), 185-192. <https://doi.org/10.1007/s13187-021-02096-x>
- Carrero-Planells, A., Pol-Castañeda, S., Alamillos-Guardiola, M. C., Prieto-Alomar, A., Tomás-Sánchez, M., & Moreno-Mulet, C. (2021). Students and teachers' satisfaction and perspectives on high-fidelity simulation for learning fundamental nursing procedures: A mixed-method study. *Nurse Education Today*, 104, Article 104981. <https://doi.org/10.1016/j.nedt.2021.104981>
- Clinard, E. S. (2022). Increasing student confidence with medically complex infants through simulation: a mixed methods investigation. *American Journal of Speech-language Pathology*, 31(2), 942-958. [https://doi.org/10.1044/2021\\_AJSLP-21-00234](https://doi.org/10.1044/2021_AJSLP-21-00234)
- Cho, M. K., & Kim, M. Y. (2024). The effect of virtual reality simulation on nursing students' communication skills: a systematic review and meta-analysis. *Frontiers in Psychiatry*, 15, Article 1351123. <https://doi.org/10.3389/fpsy.2024.1351123>
- Curry, L., & Nunez-Smith, M. (2015). *Mixed methods in health sciences research. A practical primer*. SAGE Publications, Inc.
- Dawood, E., Alshutwi, S. S., Alshareif, S., & Shereda, H. A. (2024). Evaluation of the effectiveness of standardized patient simulation as a teaching method in psychiatric and mental health nursing. *Nursing Reports*, 14(2), 1424-1438. <https://doi.org/10.3390/nursrep14020107>
- Duff, J. P., Morse, K. J., Seelandt, J., Gross, I. T., Lydston, M., Sargeant, J., Dieckmann, P., Allen, J. A., Rudolph, J. W., & Kolbe, M. (2024). Debriefing methods for simulation in healthcare: a systematic review. *Simulation in Healthcare: Journal of the Society for Simulation in Healthcare*, 19(1S), S112-S121. <https://doi.org/10.1097/SIH.0000000000000765>
- Escribano, S., Juliá-Sanchis, R., García-Sanjuán, S., Congost-Maestre, N., & Cabañero-Martínez, M. J. (2021). Psychometric properties of the attitudes towards medical communication scale in nursing students. *PeerJ*, 9, Article e11034. <https://doi.org/10.7717/peerj.11034>
- Elendu, C., Amaechi, D. C., Okatta, A. U., Amaechi, E. C., Elendu, T. C., Ezech, C. P., & Elendu, I. D. (2024). The impact of simulation-based training in

- medical education: A review. *Medicine*, 103(27), Article e38813. <https://doi.org/10.1097/MD.00000000000038813>
- Elosua, P., Aguado, D., Fonseca-Pedrero, E., Abad, F. J., & Santamaría, P. (2023). New trends in digital technology-based psychological and educational assessment. *Psicothema*, 35(1), 50-57. <https://doi.org/10.7334/psicothema2022.241>
- Farina, C. L., Moreno, J., & Schneidereith, T. (2024). Using simulation to improve communication skills. *The Nursing clinics of North America*, 59(3), 437-448. <https://doi.org/10.1016/j.cnur.2024.02.007>
- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G\* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175-191. <https://doi.org/10.3758/BF03193146>
- Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs-principles and practices. *Health Services Research*, 48(6 Pt 2), 2134-2156. <https://doi.org/10.1111/1475-6773.12117>
- Flanagan, O. L., & Cummings, K. M. (2023). Standardized patients in medical education: a review of the literature. *Cureus*, 15(7), Article e42027. <https://doi.org/10.7759/cureus.42027>
- García-Carpintero Blas, E., Vélez-Vélez, E., Gómez-Moreno, C., Martínez-Arce, A., Tovar-Reinoso, A., Rodríguez-Gómez, P., Vaquero Velerdas, L., & López-Martín, I. (2024). Simulation with a standardised patient to reduce stigma towards people with schizophrenia spectrum disorder among nursing students: A quasi-experimental study. *Archives of Psychiatric Nursing*, 52, 24-30. <https://doi.org/10.1016/j.apnu.2024.07.015>
- Gerzina, H. A., & Stovsky, E. (2023). Standardized patient assessment of learners in medical simulation. In *StatPearls [Internet]*. StatPearls Publishing.
- Gutiérrez-Maldonado, J., Alsina-Jurnet, I., Rangel-Gómez, M. V., Aguilar-Alonso, A., Jarne-Esparcia, A. J., Andrés-Pueyo, A., & Talam-Caparrós, A. (2008). Virtual intelligent agents to train abilities of diagnosis in Psychology and Psychiatry. In G. A. Tsihrintzis, M. Virvou, R. J. Howlett, & L. C. Jain (Eds.), *New directions in intelligent interactive multimedia* (pp. 497-505). Springer. [https://doi.org/10.1007/978-3-540-68127-4\\_51](https://doi.org/10.1007/978-3-540-68127-4_51)
- Gutierrez-Maldonado, J., Andres-Pueyo, A., Jarne, A., Talam, A., Ferrer, M., & Achotegui, J. (2017). Virtual reality for training diagnostic skills in Anorexia Nervosa: A usability assessment. In S. Lackey, & J. Chen (Eds.), *Virtual, augmented and mixed reality* (pp. 239-247). Springer International Publishing. [https://doi.org/10.1007/978-3-319-57987-0\\_19](https://doi.org/10.1007/978-3-319-57987-0_19)
- Hawkins, A., & Tredgett, K. (2016). Use of high-fidelity simulation to improve communication skills regarding death and dying: a qualitative study. *BMJ Supportive & Palliative Care*, 6(4), 474-478. <https://doi.org/10.1136/bmjspcare-2015-001081>
- Higgins, M., Madan, C., & Patel, R. (2021). Development and decay of procedural skills in surgery: A systematic review of the effectiveness of simulation-based medical education interventions. *The Surgeon: Journal of the Royal Colleges of Surgeons of Edinburgh and Ireland*, 19(4), e67-e77. <https://doi.org/10.1016/j.surge.2020.07.013>
- Hillier, M., Williams, T. L., & Chidume, T. (2023). Standardization of standardized patient training in medical simulation. In *StatPearls [Internet]*. StatPearls Publishing.
- Isaza-Restrepo, A., Gómez, M. T., Cifuentes, G., & Argüello, A. (2018). The virtual patient as a learning tool: a mixed quantitative qualitative study. *BMC Medical Education*, 18(1), Article 297. <https://doi.org/10.1186/s12909-018-1395-8>
- Jeon, Y., Choi, H., Lee, U., & Kim, H. (2024). Technology-based interactive communication simulation addressing challenging communication situations for nursing students. *Journal of Professional Nursing: Official Journal of the American Association of Colleges of Nursing*, 53, 71-79. <https://doi.org/10.1016/j.profnurs.2024.05.002>
- Johnson S. L. (2019). Impact, growth, capacity-building of mixed methods research in the Health Sciences. *American Journal of Pharmaceutical Education*, 83(2), Article 7403. <https://doi.org/10.5688/ajpe7403>
- Johnson, K. V., Scott, A. L., & Franks, L. (2020). Impact of standardized patients on first semester nursing students self-confidence, satisfaction, and communication in a simulated clinical case. *SAGE Open Nursing*, 6, Article 2377960820930153. <https://doi.org/10.1177/2377960820930153>
- Kolb, D. A. (2014). *Experiential learning: Experience as the source of learning and development*. FT press.
- Kolbe, M., Grande, B., & Spahn, D. R. (2015). Briefing and debriefing during simulation-based training and beyond: Content, structure, attitude and setting. *Best Practice & Research. Clinical Anaesthesiology*, 29(1), 87-96. <https://doi.org/10.1016/j.bpa.2015.01.002>
- Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, 8(1), Article 33267. <https://doi.org/10.1525/collabra.33267>
- Lazarus, R. S. & Folkman, S. (1984). *Stress, appraisal, and coping* (Vol. 464). Springer.
- Liu, X., Wu, C., Lai, R., Lin, H., Xu, Y., Lin, Y., & Zhang, W. (2023). ChatGPT: when the artificial intelligence meets standardized patients in clinical training. *Journal of Translational Medicine*, 21, Article 447. <https://doi.org/10.1186/s12967-023-04314-0>
- López, A.I., & López-Chicheri, I. (July 2024). *Entrenamiento en habilidades de comunicación orientadas a la intervención mediante simulación clínica en psicología (SCP) [Intervention-oriented communication skills training through clinical simulation in Psychology (SCP)]*. I Congreso Interprofesional de Simulación-OneHealth. <https://simulacion-onehealth.org/ponencia/entrenamiento-en-habilidades-de-comunicacion-orientadas-a-la-intervencion-mediante-simulacion-clinica-en-psicologia-scp/>
- López-Gómez, I., Hervás, G., & Vázquez, C. (2015). Adaptación de las “Escalas de afecto positivo y negativo” (PANAS) en una muestra general española. [Adaptation of the Positive and Negative Affect Scales (PANAS) in a general Spanish sample]. *Psicología Conductual*, 23(3), 529-548.
- Monahan, L., Eaves, C. L., Watson, J. C., Friese, J., McKenna, L., & Estrada-Ibarra, E. (2024). Improving adolescent psychosocial assessment through standardized patient simulation: an interdisciplinary quality improvement initiative. *International Journal of Environmental Research and Public Health*, 21(3), Article 283. <https://doi.org/10.3390/ijerph21030283>
- Morcela, O. A. (2022). ChatGPT: la IA está aquí y nos desafía [ChatGPT: AI is here and challenging us]. *AACINI-Revista Internacional de Ingeniería Industrial*, 6. <https://riii.fi.mdp.edu.ar/index.php/AACINI-RIII/article/view/67>
- Moser, A., & Korstjens, I. (2018). Series: Practical guidance to qualitative research. Part 3: Sampling, data collection and analysis. *The European Journal of General Practice*, 24(1), 9-18. <https://doi.org/10.1080/13814788.2017.1375091>
- Moss, C. R. (2023). Neonatal fragile skin: novel use of simulation to improve knowledge and confidence for neonatal nurse practitioner students. *Nurse Educator*, 48(4), E122-E125. <https://doi.org/10.1097/NNE.0000000000001354>
- Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, 16(1). <https://doi.org/10.1177/1609406917733847>
- Oliveira, G., Oliveira, F. S. E., Coelho, A. S. G., Cavalcante, A. M. R. Z., Vieira, F. V. M., Fonseca, L. M. M., Campbell, S. H., & Aredes, N. D. A. (2022). Effect of simulation on stress, anxiety, and self-confidence in nursing students: Systematic review with meta-analysis and meta-



- regression. *International Journal of Nursing Studies*, 133, Article 104282. <https://doi.org/10.1016/j.ijnurstu.2022.104282>
- OpenAI. (2023). *ChatGPT (March 14 version) [Large language model]*. <https://chat.openai.com>
- Pedrajas, M. L., Sanz, A., García-Carpintero, E., Martínez, E., Uceda, S. (2024). Training GPT as a standardized patient. In M. D. Díaz-Noguera, C. Hervás-Gómez, & F. Sánchez-Vera (Coords.), *Artificial Intelligence and Education* (pp.189-204). Octaedro. <https://doi.org/10.36006/09643-1-12>
- Peñaloza-Salazar, C., Gutierrez-Maldonado, J., Ferrer-Garcia, M., Garcia-Palacios, A., Andres-Pueyo, A., & Aguilar-Alonso, A. (2011). Simulated interviews 3.0: virtual humans to train abilities of diagnosis –usability assessment. *Studies in Health Technology and Informatics*, 167, 165-169. <https://doi.org/10.3233/978-1-60750-766-6-165>
- Perugini, M., Gallucci, M., & Costantini, G. (2018). A practical primer to power analysis for simple experimental designs. *Revue Internationale de Psychologie Sociale*, 31(1), 1-23. <https://dx.doi.org/10.5334/IRSP.181>
- Rizzo, A., Lange, B., Buckwalter, J. G., Forbell, E., Kim, J., Sagae, K., Williams, J., Difede, J., Rothbaum, B. O., Reger, G., Parsons, T., & Kenny, P. (2011). SimCoach: An intelligent virtual human system for providing healthcare information and support. *International Journal on Disability and Human Development*, 10(4), 277-281. <https://doi.org/10.1515/IJDHD.2011.046>
- Rodríguez, S., Condés, E., & Arriaga, A. (2021). Irrupción de la simulación clínica online en tiempos de COVID-19. Una experiencia ilustrativa de asignatura en el Grado de Psicología. *Revista de la Fundación de Educación Médica*, 24(2), 101-104. <https://dx.doi.org/10.33588/fem.242.1118>
- Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, 56(5), 356-374. <https://doi.org/10.1016/j.jmp.2012.08.001>
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16, 225-237. <https://doi.org/10.3758/PBR.16.2.225>
- Ruiz-Rodríguez, J., Bados López, A., Fusté Escolano, A., García-Grau, E., Saldaña García, C., & Lluch Canut, T. (2016). Aprendizaje experiencial de habilidades terapéuticas y análisis de su utilidad en función de la personalidad. *Behavioral Psychology*, 24, 405-422. <http://hdl.handle.net/2445/155366>
- Scherr, R., Halaseh, F. F., Spina, A., Andarib, S., & Rivera, R. (2023). ChatGPT interactive medical simulations for early clinical education: case study. *JMIR Medical Education*, 9, Article e49877. <https://doi.org/10.2196/49877>
- Sezer, B., Sezer, T. A., Teker, G. T., & Elcin, M. (2023). Developing a virtual patient: design, usability, and learning effect in communication skills training. *BMC Medical Education*, 23(1), Article 891. <https://doi.org/10.1186/s12909-023-04860-7>
- Sezgin, M. G., & Bektas, H. (2023). Effectiveness of interprofessional simulation-based education programs to improve teamwork and communication for students in the healthcare profession: A systematic review and meta-analysis of randomized controlled trials. *Nurse Education Today*, 120, Article 105619. <https://doi.org/10.1016/j.nedt.2022.105619>
- Talbot, T., & Rizzo, A. "Skip". (2019). Virtual human standardized patients for clinical training. In A. "Skip" Rizzo & S. Bouchard (Eds.), *Virtual reality for psychological and neurocognitive interventions* (pp. 387-405). Springer New York. [https://doi.org/10.1007/978-1-4939-9482-3\\_17](https://doi.org/10.1007/978-1-4939-9482-3_17)
- Ton, D. N. M., Duong, T. T. K., Tran, H. T., Nguyen, T. T. T., Mai, H. B., Nguyen, P. T. A., Ho, B. D., & Ho, T. T. T. (2024). Effects of standardized patient simulation and mobile applications on nursing students' clinical competence, self-efficacy, and cultural competence: A quasi-experimental study. *International Journal of Environmental Research and Public Health*, 21(4), Article 515. <https://doi.org/10.3390/ijerph21040515>
- Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care: Journal of the International Society for Quality in Health Care*, 19(6), 349-357. <https://doi.org/10.1093/intqhc/mzm042>
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology*, 54(6), 1063-1070. <https://doi.org/10.1037/0022-3514.54.6.1063>
- Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1), 36-45. <https://doi.org/10.1145/365153.365168>
- Yu, J. H., Chang, H. J., Kim, S. S., Park, J. E., Chung, W. Y., Lee, S. K., Kim, M., Lee, J. H., & Jung, Y. J. (2021). Effects of high-fidelity simulation education on medical students' anxiety and confidence. *PloS one*, 16(5), Article e0251078. <https://doi.org/10.1371/journal.pone.0251078>

Article

## Emotionally Tough, Sexting Rough: Relationship Between Callous Unemotional Traits and Aggravated Sexting in 11 Countries

Mara Morelli<sup>1</sup>, Fau Rosati<sup>1</sup>, Elena Cattelino<sup>2</sup>, Flavio Urbini<sup>3</sup>, Roberto Baiocco<sup>1</sup>,  
Dora Bianchi<sup>1</sup>, Fiorenzo Laghi<sup>1</sup>, Maurizio Gasseau<sup>2</sup>, Piotr Sorokowski<sup>4</sup>, Michal Misiak<sup>4,5</sup>,  
Martyna Dziekan<sup>6</sup>, Heather Hudson<sup>7</sup>, Alexandra Marshall<sup>8</sup>, Thanh Truc Nguyen<sup>9</sup>, Lauren Mark<sup>9</sup>,  
Kamil Kopecky<sup>10</sup>, René Szotkowski<sup>10</sup>, Ezgi Toplu Demirtaş<sup>11</sup>, Joris Van Ouytsel<sup>12</sup>, Koen Ponnet<sup>13</sup>,  
Michel Walrave<sup>14</sup>, Tingshao Zhu<sup>15</sup>, Ya Chen<sup>15</sup>, Nan Zhao<sup>15</sup>, Xiaoqian Liu<sup>15</sup>,  
Alexander Voiskounsky<sup>16</sup>, Nataliya Bogacheva<sup>17</sup>, Maria Ioannou<sup>18</sup>, John Synnott<sup>18</sup>,  
Kalliopi Tzani-Pepelasis<sup>18</sup>, Vimala Balakrishnan<sup>19</sup>, Moses Okumu<sup>20</sup>, Eusebius Small<sup>21</sup>,  
Silviya Pavlova Nikolova<sup>22</sup>, Michelle Drouin<sup>23</sup>, Alessandra Ragona<sup>1</sup>, and Antonio Chirumbolo<sup>1</sup>

<sup>1</sup>Sapienza University of Rome (Italy), <sup>2</sup>University of Aosta Valley (Italy), <sup>3</sup>LUMSA University (Italy), <sup>4</sup>University of Wrocław (Poland),

<sup>5</sup>University of Oxford (United Kingdom), <sup>6</sup>Adam Mickiewicz University (Poland), <sup>7</sup>University of Central Arkansas (USA),

<sup>8</sup>University of Arkansas for Medical Sciences (USA), <sup>9</sup>University of Hawaii at Manoa (Honolulu),

<sup>10</sup>Palacký University Olomouc (Czech Republic), <sup>11</sup>MEF University (Turkey), <sup>12</sup>Arizona State University (USA),

<sup>13</sup>Ghent University (Belgium), <sup>14</sup>University of Antwerp (Belgium), <sup>15</sup>Chinese Academy of Sciences (China),

<sup>16</sup>Moscow State University (Russia), <sup>17</sup>Sechenov University (Russia), <sup>18</sup>University of Huddersfield (Queensgate, United Kingdom)

<sup>19</sup>University of Malaya (Malaysia), <sup>20</sup>University of Illinois Urbana-Champaign (USA), <sup>21</sup>University of Texas (USA),

<sup>22</sup>Medical University-Varna (Bulgary), <sup>23</sup>Indiana University – Purdue University Fort Wayne (USA)

### ARTICLE INFO

Received: 05/09/2024

Accepted: 03/02/2025

#### Keywords:

Sexting

Non-consensual

Callousness

Unemotional

Young adults

### ABSTRACT

**Background:** Sexting is now widely acknowledged as a common sexual behavior among adolescents and young adults. However, the occurrence of abusive interactions, such as non-consensual sexting, warrants attention. Prevalence rates of non-consensual sexting vary between countries, influenced by gender and age. The present study examined the relationship between three facets of callous-unemotional (CU) traits (i.e., callousness, uncaring, and unemotional) and the sharing of non-consensual sexts across different relationship contexts (i.e., acquaintances, strangers, or partners). **Method:** Data were drawn from a cross-countries project encompassing 11 countries: Belgium, China, Czech Republic, Ireland, Italy, Malaysia, Poland, Russia, Turkey, Uganda, and the USA. The sample comprised 6093 young adults (3682 girls; 2401 boys), aged 13 to 30 ( $M = 20.35$ ;  $SD = 3.63$ ). **Results:** Results from a logistic mixed-model indicate that CU traits predict non-consensual sexting, with high callousness and uncaring, and low unemotional traits associated with non-consensual sexting involving partners and strangers. Younger individuals and women were more likely to engage in all forms of non-consensual sexting compared to older individuals and men. **Conclusions:** It is important to promote sexual education programs to increase emotional self-awareness and challenge gender stereotypes in order to reduce adverse outcomes associated with sexting.

Cite as: Morelli, M., Rosati, F., Cattelino, E., Urbini, F., Baiocco, R., Bianchi, D., Laghi, F., Gasseau, M., Sorokowski, P., Misiak, M., Dziekan, M., Hudson, H., Marshall, A., Nguyen, T. T., Mark, L., Kopecky, K., Szotkowski, R., Demirtaş, E. T., Van Ouytsel, J., Ponnet, K., Walrave, M., ... Chirumbolo, A. (2025). Emotionally tough, sexting rough: Relationship between callous unemotional traits and aggravated sexting in 11 countries. *Psicothema*, 37(3), 33-44. <https://doi.org/10.70478/psicothema.2025.37.22>

Corresponding author: Mara Morelli, [mara.morelli@uniroma1.it](mailto:mara.morelli@uniroma1.it)

This article is published under Creative Commons License 4.0 CC-BY-NC-ND

## Dureza Emocional y Sexting Rudo: Relación Entre los Rasgos Insensibles y no Emocionales y el Sexting Agravado en 11 Países

### RESUMEN

#### Palabras clave:

Sexting  
No consentido  
Insensibilidad  
Sin emociones  
Adultos jóvenes

**Antecedentes:** El sexting es un comportamiento sexual común entre adolescentes y adultos jóvenes, pero el sexting no consensuado merece atención debido a sus implicaciones abusivas. La prevalencia de este fenómeno varía según país, género y edad. Este estudio analizó cómo las tres facetas de los rasgos de insensibilidad emocional (insensibilidad, despreocupación y falta de emotividad) se relacionan con el envío de sexting no consensuado en diferentes contextos (conocidos, desconocidos o parejas). **Método:** Participaron 6093 adultos jóvenes (3682 mujeres, 2401 hombres) de 13 a 30 años ( $M = 20.35$ ;  $SD = 3.63$ ) en un estudio multinacional realizado en 11 países: Bélgica, China, República Checa, Irlanda, Italia, Malasia, Polonia, Rusia, Turquía, Uganda y Estados Unidos. **Resultados:** Los rasgos de insensibilidad emocional predicen el sexting no consensuado, especialmente altos niveles de insensibilidad y despreocupación, y bajos niveles de falta de emotividad en interacciones con parejas y desconocidos. Las mujeres y las personas jóvenes mostraron mayor probabilidad de participar en sexting no consensuado en comparación con hombres y personas mayores. **Conclusiones:** Es crucial implementar programas de educación sexual que fomenten la conciencia emocional y cuestionen los estereotipos de género, contribuyendo a reducir las consecuencias negativas del sexting no consensuado.

### Introduction

#### Sexting Behaviors

Sexting, defined as the sharing of sexually suggestive or provocative content via new technologies (Chalfen, 2009), has garnered increasing research attention, particularly concerning adolescents and young adults. This body of research has illuminated both the positive and negative impacts of sexting on sexual development and mental health (Mori et al., 2019; Temple & Lu, 2018).

Sexting is examined through two main perspectives: “experimental” and “aggravated.” Experimental sexting is seen as normative and consensual, occurring within romantic relationships and associated with sexual exploration, primarily observed during adolescence and young adulthood (Bianchi et al., 2019; Drouin & Landgraff, 2012). Aggravated sexting involves harmful motives, such as unauthorized sharing of sexts (Morelli et al., 2023a; Walker & Sleath, 2017), and is associated with aggressive behaviors like cyberbullying and revenge, as well as risky sexual behavior and online victimization (e.g., Gámez-Guadix & de Santisteban, 2018).

There is a gap in the literature regarding a cross-cultural perspective on the associations related to sexting. Most studies on sexting, including consensual and non-consensual forms, have been limited to single countries with few cross-country investigations. Efforts have been made to address this gap (Morelli et al., 2020, 2021), and recent research has revealed varying prevalence rates of sexting across different countries, likely influenced by cultural values within specific societies (Morelli et al., 2021). These cultural values can shape online behaviors, including sexting.

Cultural differences can significantly influence both the frequency and the forms of sexting behaviors. According to some interpretations, sexting is more prevalent in cultural contexts where sexual experiences occur at an earlier age and where a sexist culture with rigid binary gender roles is predominant (Gil-Llario et al., 2021). Research suggests that in more traditional societies, where gender differences are heightened, boys are more likely to engage in sexting

compared to girls (Baumgartner et al., 2014). Nevertheless, some research, while highlighting variations in sexting practices across different countries, found that women’s vulnerability to sexting remains unchanged (Gassó et al., 2021). Additional research emphasizes other characteristics that may impact sexting behaviors, such as gender, age, and personality traits.

With regard to age and gender differences, boys and young adolescents are more frequently implicated in aggravated behaviors, such as non-consensual sexting (i.e., the sharing of sexting images without consent), compared to girls and older individuals (Morelli et al., 2021; Mori et al., 2020). Research highlighted similar age and gender differences in consensual sexting behaviors (Livingstone & Görzig, 2014). More specifically, older adolescents exhibit a higher likelihood of sexting compared to younger counterparts (Gewirtz-Meydan et al., 2018; Madigan et al., 2018a; Mori et al., 2022), and while early research studies found that boys were more likely to sext than girls (Baumgartner et al., 2010), more recent research shows the opposite trend (Gewirtz-Meydan et al., 2018; Mori et al., 2022).

In the early part of the last decade, sexting was on the rise among youth (Madigan et al., 2018a), but recent reviews indicate that sexting rates have stabilized (Mori et al., 2022). Age is an important variable to consider, as younger individuals (e.g., adolescents) may exhibit greater disinhibition and a higher tendency toward risky behaviors, potentially transforming exploratory sexting into problematic behavior. However, increased attention from researchers focused on the associations between sexting and mental health, relationship issues, and negative consequences like worry, regret, and shame (Drouin et al., 2017; Mori et al., 2019). Research continues to explore the adverse effects of sexting on youth and young adults’ well-being, with a predominant focus on the victim’s perspective. Only a few studies have examined the correlates of aggravated sexting perpetration (Morelli et al., 2021, 2023b).

Recent meta-analyses indicated that young people engaged in non-consensual sexting were about 15% (Mori et al., 2020), and 18% (Madigan et al., 2018b). Morelli et al. (2021) cross-cultural study revealed that over 20% of adolescents and young adults

engaged in non-consensual sexting in the Czech Republic, Ireland, Malaysia, Russia, and Uganda. Lower percentages were observed in China, the USA, Italy, Poland, Belgium, and Turkey.

However, no previous studies have delineated the distinct targets of non-consensual sexting, which differ based on the depicted victim in the forwarded or shared content: acquaintances, stranger, or partner. Thus, it is unknown how personality traits relate to sharing non-consensual sexts across different relationship contexts.

## Callous-Unemotional Traits and Sexting

Callous-unemotional (CU) traits consist of personality characteristics reflecting affective deficits, including shallow affect, lack of empathy and remorse, low responsiveness to others' emotional cues, and minimal concern about one's behavior (Frick et al., 2014). These traits manifest through three key components: callousness (i.e., lack of empathy, guilt, and remorse, particularly evident in disregard for others during violent or illegal actions); uncaring (i.e., indifference towards one's actions and others' feelings, and disregard for rules and emotional states of others); unemotional (i.e., shallow or deficient affect, and lack of emotional expression) (Kimonis et al., 2008).

The Inventory of Callous-Unemotional Traits (ICU; Frick, 2004) is commonly used to assess callous-unemotional (CU) traits. Research utilizing this inventory has shown varying levels of CU traits, with the unemotional aspect consistently demonstrating weaker associations with antisocial behavior, delinquency, aggression, and psychopathy compared to levels of uncaring or callous features (Waller et al., 2014). These traits play a crucial role in defining the affective core components of psychopathy during adulthood (Hare & Neumann, 2008).

The stability of CU traits throughout life, from childhood to adulthood, is highlighted (Fontaine et al., 2010). These traits are linked to reduced capacity for prosocial emotional responsiveness among youth with CU traits (Waller et al., 2020). Individuals with high CU traits are more likely to engage in antisocial behavior, including aggression and sexual violence (Frick & White, 2008), and to have risky sexual relationships (Carlson et al., 2015). Elevated CU traits in youth lead to reduced emotional responses to distress cues and muted fear responses to risky situations (Pardini et al., 2003), compromising their ability to assess consequences and impairing decision-making abilities (Fanti et al., 2013; Pardini et al., 2003). CU traits are also strong predictors of physical aggression, relational aggression, and bullying (Helfritz & Stanford, 2006; Centifanti et al., 2015; Fanti et al., 2013).

Non-consensual sexting has been associated with both behavioral and emotional issues (Gámez-Guadix & de Santisteban, 2018), as well as low trait emotional intelligence (Morelli et al., 2023b, 2023c). Studies have investigated the relationship between sexting behaviors and personality traits, including using models such as HEXACO and the Dark Triad (Morelli et al., 2020, 2021). Research suggests that low levels of Honesty/Humility and conscientiousness may contribute to aggravated sexting (Morelli et al., 2020). Additionally, involvement in non-consensual sexting has been linked to traits like Machiavellianism, narcissism, and psychopathy (Morelli et al., 2021). As shown in these studies, personality traits are sometimes fundamental in understanding risky behaviors, especially in relational contexts. Empirically investigating their correlations can be crucial for prevention efforts.

Only one cross-sectional study has explored the link between CU traits and non-consensual sexting among preadolescents and adolescents, indicating a significant association with callousness and uncaring traits (Barroso et al., 2021). However, due to scale's reliability issues, data on the unemotional dimension were excluded from the analyses. Additionally, the study relied solely on a single-item measure to assess non-consensual sexting. No previous studies have examined CU traits in relation to various forms of non-consensual sexting considering the victim's identity or involved participants from multiple countries.

Aggravated sexting has been analyzed from a theoretical perspective (Dodaj & Sesar, 2020), through the collection of data from law enforcement agencies to outline different profiles of aggravated sexting (Wolak & Finkelhor, 2011) and to highlight its controversial aspects (Salter et al., 2013). These theoretical works have been followed by empirical studies conducted at the national level (Barroso et al., 2021; Bianchi et al., 2019; Van Ouytsel et al., 2021), but there remains a lack of cross-cultural research that integrates samples from diverse cultural contexts. Hence, this study addresses these research gaps by incorporating data from countries with significantly different cultural backgrounds, aiming to investigate aggravated sexting and enhance the generalizability of the findings.

## The Present Study

The study aims to investigate the correlation between CU traits (callousness, uncaring, unemotional) and various forms of non-consensual sexting (sharing or posting sexts of one's partner, acquaintances, or strangers without their consent) across 11 countries among adolescents and young adults. Building upon previous studies (Barroso et al. 2021; Fanti et al., 2009; Kokkinos & Voulgaridou, 2017; Wright et al., 2019), it is hypothesized that callousness and uncaring traits will positively correlate with non-consensual sexting, while unemotional traits will not. Specifically, we hypothesize that callousness and uncaring traits will predict non-consensual sexting (Barroso et al., 2021) in all its forms, including interactions with acquaintances, strangers, and partners, whereas unemotional traits will be unrelated (Fanti et al., 2009; Kokkinos & Voulgaridou, 2017; Wright et al., 2019). We further hypothesize an age effect, with older individuals engaging in sexting more frequently than younger individuals (Gewirtz-Meydan et al., 2018; Madigan et al., 2018a; Mori et al., 2022). Finally, we do not have a clear hypothesis regarding gender, as some studies suggest that males engage in sexting more frequently than females (Baumgartner et al., 2010; Morelli et al., 2021; Mori et al., 2020), while others report the opposite (Gewirtz-Meydan et al., 2018; Mori et al., 2022).

## Method

### Participants

The data utilized in the present study were derived from a larger cross-countries project focused on sexting. Data collection encompassed 11 countries: Belgium, China, Czech Republic, Ireland, Italy, Malaysia, Poland, Russia, Turkey, Uganda, and the USA. The study comprised a total of 6093 participants, with 3682 girls and 2401 boys (ten participants did not indicate their gender), averaging 20.35 years old ( $SD = 3.63$ ; range = 13 to 30 years



old). Regarding relationship status, approximately 81.8% ( $n = 4983$ ) reported currently having or having had a dating partner, while 17.5% ( $n = 1069$ ) reported never having had a dating partner. Descriptive statistics for participants from each country are presented in Table 1. The participants from each country constituted independent samples, with no repetition in measurements.

The G\*Power software conducted an a priori power analysis to determine the necessary sample size for each country, aiming for adequate statistical power and minimizing Type II Error. For bivariate level, assuming a small to medium effect size ( $r = .20$ ), an alpha level of .05, and a power of .80, a minimum of 194 participants per country was required. Therefore, each country aimed to collect at least 200 participants. For multiple regression analysis with 11 predictors, requiring a sample size of 1267 for an alpha level of .05, a power of 80%, and a small expected effect size of  $f^2 = 0.02$  (i.e., a conservative worst-case scenario), the global sample size of 6093 in this study exceeded this requirement, ensuring sufficient statistical power.

**Table 1**  
*Sample Characteristics by Country*

Countries	Sample size	Range	Age	Gender	
			<i>M(SD)</i>	girls	boys
Belgium	505	14-30	19.17 (3.42)	344	161
China	361	17-30	21.27 (2.64)	220	141
Czech Republic	733	13-30	19.51 (3.16)	469	264
Ireland	271	13-17	15.05 (0.69)	0	271
Italy	805	13-30	20.85 (4.25)	474	330
Malaysia	305	14-30	22.09 (2.16)	229	76
Poland	1075	13-30	20.8 (4.18)	543	532
Russia	278	15-30	19.79 (3.31)	208	70
Turkey	601	18-30	22.65 (2.95)	419	176
Uganda	226	14-20	17.29 (1.31)	137	86
USA	933	18-30	20.74 (2.36)	639	294

*Note.* Few participants failed to report their gender.

## Instruments

### Socio-Demographic Information

Participants provided information regarding their age, gender (girls were coded as 0, boys as 1), and dating relationship status (participants who had never had a partner were coded as 0, while those who currently have or have had a partner were coded as 1).

### Sexing Behaviors

Sexing is defined as sharing sexually suggestive or provocative messages/photos/videos via mobile phones, or internet social networking sites. The frequency of various aggravated sexting behaviors in which participants engaged over the past year was assessed using 12 items selected from the Sexting Behaviors Questionnaire (SBQ; Morelli et al., 2016). Each item was rated on a 5-point Likert scale ranging from 1 (*never*) to 5 (*always or almost daily*). Three dimensions of aggravated sexting were examined: a) Sending or posting sexts of acquaintances without

their consent (4 items, Cronbach's alpha = .86; reliability for each country ranging from .50 to .96). b) Sending or posting sexts of strangers without their consent (4 items, Cronbach's alpha = .83; reliability for each country ranging from .62 to .93). c) Sending or posting sexts of one's partner without their consent (4 items, Cronbach's alpha = .88; reliability for each country ranging from .50 to .93). As the items pertained to the frequency of behaviors, the variables did not exhibit a normal distribution. Consequently, each dimension was dichotomized thereafter, with 0 indicating that participants had never engaged in sexting, and 1 indicating that participants had engaged in sexting at least once.

### Callous-Unemotional Traits

The Callous Unemotional (CU) traits were assessed using the Inventory of Callous Unemotional Traits (ICU; Kimonis et al., 2008), a 24-item self-report questionnaire. Participants responded to items on a Likert scale ranging from 0 (*not at all true*) to 3 (*totally true*). CU traits represent the affective dimension of psychopathy (Frick et al., 2003; Hare & Neumann, 2008) and include a lack of empathy, guilt, and emotional expression. The scale includes three sub-scales: callousness, that is the absence of empathy and remorse (9 items; Cronbach's alpha of .69; reliability for each country ranging from .57 to .80), unemotional that is the lack of emotional expressiveness (5 items Cronbach's alpha of .76; reliability for each country ranging from .50 to .79), and uncaring that measures insensitivity toward others' emotions and performance (8 items; Cronbach's alpha of .60; reliability for each country ranging from .63 to .85).

### Procedure

Researchers from various countries were contacted by the Italian group coordinating the project and asked to sign a scientific agreement outlining sample size, characteristics, and procedures. An English questionnaire was distributed, with non-English speaking countries translating and back-translating the survey. The study followed Declaration of Helsinki guidelines and gained approval from the Ethics Committee of the Sapienza University of Rome, Italy (protocol code 405, 11/23 and 07.22.2015).

Participants completed an online survey, with underage individuals recruited from public schools after obtaining parental consent. Young adults were recruited from universities and through snowball sampling. Participants provided consent at the beginning of the survey by clicking on "Yes, I give my consent to participate in the study and to the use of my data for research purposes", ensuring anonymity and privacy due to the sensitive nature of the data. Only fully completed questionnaires were considered valid. Response rates varied by country, ranging from 85% to 100%. The use of online test administration can significantly contribute to addressing the three critical aspects mentioned by the reviewer: controlling the administration of tests, standardizing the administration, and minimizing errors. Firstly, online platforms allow for enhanced control of test administration through automation and structured protocols with uniform instructions and environment control. Secondly, online platforms inherently promote standardization as every participant receives the same version of the test, ensuring uniformity. Moreover, in tests with fixed-response formats, automated scoring eliminates the possibility of scoring bias or human error. Last but not least, online

platforms incorporate features to reduce human and procedural errors, enhancing the reliability of the data and reducing the errors in the administration.

## Data Analysis

Initially, descriptive statistics, frequencies, and correlations among variables were computed. Subsequently, we investigated how the three CU traits (i.e., Callousness, Uncaring, and Unemotional) predicted different forms of aggravated sexting behaviors (i.e., sending or posting sexts without consent of acquaintances, strangers, and relationship partners, while controlling for gender and age. As participants were nested in various countries, and the dependent variables were dichotomous, we conducted a generalized logistic mixed model for each of the three dependent variables, with Country serving as the grouping variable. In our model, the fixed effects predictors included the two demographic variables (age in years and gender, coded as 0 = *female*, 1 = *male*), the three CU traits (i.e., Callousness, Uncaring, and Unemotional), a fixed intercept, and one random intercept for each country.

The logistic mixed-effects model was adopted to appropriately handle the dichotomous nature of the dependent variables, address the nested data structure, and ensure robust and generalizable findings across the 11 countries included in the study. Firstly, we considered the nature of the Dependent Variables. The dependent variables in this study are dichotomous (e.g., sexting behaviors categorized as “present” or “absent”). Logistic regression is the appropriate statistical technique for analyzing relationships involving binary outcomes, as it models the probability of an event occurring. Secondly, we had a Multilevel Structure of the Data: The dataset includes participants from 11 different countries, introducing a multilevel structure where individuals (Level 1) are nested within countries (Level 2). This creates potential contextual effects and between-country variability that must be accounted for to avoid violating independence assumptions. A mixed-effects model is well-suited for this purpose as it allows us to control for clustering effects by including random intercepts for countries. Overall, this model increased statistical power and precision. In fact, by explicitly modeling the nested data structure, the mixed-effects model provides more accurate parameter estimates and standard errors. Ignoring the multilevel structure could result in underestimated standard errors and inflated Type I error rates. Moreover, the inclusion of random effects allows us to quantify

and account for the variability attributable to countries, improving the generalizability of the results across different cultural or national contexts.

Additionally, we considered possible interactions between the demographic variables and the CU traits by including interaction terms as six additional fixed effects: age\*Callousness, age\*Uncaring, age\*Unemotional, sex\*Callousness, sex\*Uncaring, and sex\*Unemotional. Following suggestions from various authors (e.g., Aiken & West, 1991; Cohen et al., 2013), variables were mean-centered. To interpret the findings of potential interactions between variables, a simple slope analysis was also conducted. As non-consensual sexting of one’s partner without their consent included items about sexting behaviors with a dating partner, the analysis for this variable was conducted only on the subsample of participants who reported having or having had a dating partner ( $n = 4974$ ). The exact number of observations for each analysis will be provided in each table. Analyses were performed through Jamovi version 2.4.11 (the Jamovi project, 2023) and the Jamovi module GAMLj3 (Gallucci, 2019).

## Results

### Descriptive Statistics and Correlations

With regard to prevalence of aggravated sexting across relationship contexts, individuals reported sending or posting sexts without consent at least once of acquaintances (12.9%,  $n = 786$ ), strangers, (21.5%,  $n = 1310$ ), and relationship partners (9.3%,  $n = 462$ ).

Correlations, means, and standard deviations of the investigated variables are summarized in Table 2. Both Callousness and Uncaring traits showed significant and positive correlations with all measured aggravated sexting behaviors, whereas the Unemotional dimension did not exhibit any significant correlation.

### CU Traits and Sending or Posting Sexts of Acquaintances Without Their Consent

As previously mentioned, three generalized logistic mixed models were conducted to examine how CU traits (i.e., Callousness, Uncaring, and Unemotional) predicted three different forms of aggravated sexting behaviors: sending or posting sexts of one’s partner without their consent, of acquaintances without their consent, and of strangers without their consent, while controlling

**Table 2**  
Correlations Among Variables

	1	2	3	4	5	6	7	8
1. Gender	1							
2. Age	-.02	1						
3. Callousness	.13**	.01	1					
4. Uncaring	.14**	-.14**	.21**	1				
5. Unemotional	.15**	-.09**	.19**	.25**	1			
6. Sharing sext of acquaintances without their consent	.11**	-.08**	.15**	.12**	-.01	1		
7. Sharing sext of strangers without their consent	.14**	-.06**	.11**	.12**	.02	.54**	1	
8. Sharing sext of one’s partner without their consent <sup>a</sup>	.11**	-.04**	.19**	.11**	.01**	.52**	.35**	1

Note. \*  $p < .05$ ; \*\*  $p < .01$ . Gender was coded as 0 = girls and 1 = boys. <sup>a</sup> Correlations for Sext of one’s partner were run on a subsample of  $n = 4974$  participants who currently have or have had a partner in the past.



for gender, and age. Additionally, interaction terms between the demographic variables and the CU traits were included in the model.

The first logistic mixed model examined sending or posting sexts of acquaintances without their consent, which explained about 11% of the variance ( $R$ -square marginal = 0.11;  $R$ -square conditional = 0.14). Results of the analysis are presented in Tables 3 and 4.

Both gender and age emerged as significant predictors, with males and younger participants tending to send/post more sexts of acquaintances without their consent. Callousness and Uncaring traits were significant predictors (Table 3): participants who scored higher on Callousness and Uncaring were 2.6 and 1.5 times, respectively, more likely to send/post more sexts of acquaintances without their consent. The Unemotional trait emerged as a negative significant predictor: participants who scored higher on this trait had a 31% lower probability of sending/posting more sexts of acquaintances without their consent.

Notably, a significant interaction was observed between age and Callousness (see Table 3). To elucidate this interaction effect, a simple slope analysis was conducted. When the level of age

was higher (Mean+1 $\cdot$ SD), the effect of Callousness on sending/posting more sexts of acquaintances without their consent was more pronounced ( $O.R.$  = 3.15,  $p$  < .001) compared to the effect observed when the level of age was lower ( $O.R.$  = 2.15,  $p$  < .001). It appeared that higher scores of Callousness were associated with sending/posting more sexts of acquaintances without their consent, particularly among older participants (refer to Figure 1).

### CU Traits and Sending or Posting Sexts of Strangers Without Their Consent

The second logistic mixed model examined the impact of CU traits on sending or posting sexts of strangers without their consent, accounting for approximately 6% of the variance ( $R$ -square marginal = 0.06;  $R$ -square conditional = 0.12). Results of the analysis are depicted in Tables 4 and 5. Gender emerged as a significant predictor, with male participants more inclined to send/post more sexts of strangers without their consent, while age did not exhibit statistical significance. Callousness and Uncaring traits remained significant positive predictors (Table 5): participants

**Table 3**

*Sending or Posting Sexts of Acquaintances Without Their Consent: Fixed Effects Parameter Estimates*

	<i>B</i>	<i>SE</i>	<i>exp(B)</i>	<i>95% Exp(B) CI</i>		<i>z</i>	<i>p</i>
Gender	0.53	0.09	1.69	1.42	2.01	5.93	<.001
Age	-0.06	0.01	0.94	0.92	0.97	-4.59	<.001
Callousness	0.96	0.09	2.60	2.18	3.11	10.52	<.001
Uncaring	0.43	0.08	1.55	1.33	1.80	5.59	<.001
Unemotional	-0.38	0.07	0.69	0.59	0.79	-5.14	<.001
Gender * Callousness	0.27	0.17	1.31	0.94	1.84	1.59	.11
Gender * Uncaring	0.02	0.14	1.02	0.77	1.35	0.14	.89
Gender * Unemotional	-0.23	0.15	0.80	0.60	1.06	-1.58	.11
Age * Callousness	0.05	0.03	1.05	1.00	1.11	2.09	.036
Age * Uncaring	-0.01	0.02	0.99	0.95	1.03	-0.51	.61
Age * Unemotional	-0.01	0.02	0.99	0.95	1.03	-0.34	.74

Note.  $N$  = 6083; Gender was coded as 0 = girls and 1 = boys.

**Table 4**

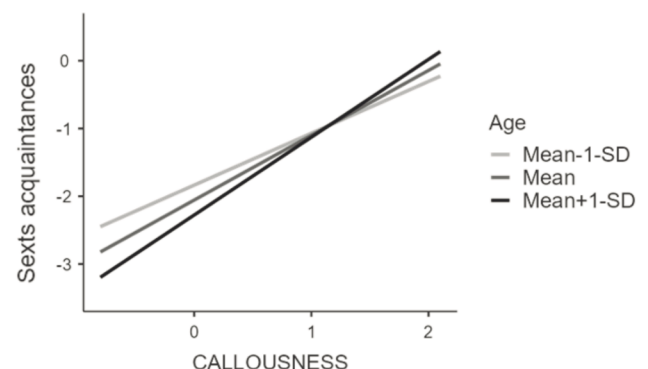
*Sending or Posting Sexts of Acquaintances/Strangers/Partners Without Their Consent: Estimates of Random Components*

		<i>SD</i>	<i>Variance</i>	<i>ICC</i>
Acquaintances	Intercept	0.31	0.097	0.029
Country	Residuals	1.00	1.00	.
Strangers	Intercept	0.46	0.21	0.06
Country	Residuals	1.00	1.00	.
Partners	Intercept	0.42	0.17	0.05
Country	Residuals	1.00	1.00	.

Note.  $N$  = 6083 (acquaintances and strangers);  $N$  = 4974 (partners); groups: COUNTRY; ICC = Intra Class Correlation.

**Figure 1**

*The Effect of Callousness on Sending or Posting Sexts of Acquaintances Without Their Consent in Function of Age*



**Table 5***Sending or Posting Sexts of Strangers Without Their Consent: Fixed Effects Parameter Estimates*

	<i>B</i>	<i>SE</i>	<i>exp(B)</i>	<i>95% Exp(B) CI</i>		<i>z</i>	<i>p</i>
Gender	0.58	0.07	1.79	1.56	2.06	8.27	<.001
Age	-0.02	0.01	0.98	0.96	1.00	-1.89	0.06
Callousness	0.63	0.08	1.87	1.60	2.18	7.99	<.001
Uncaring	0.27	0.06	1.31	1.16	1.49	4.31	<.001
Unemotional	-0.18	0.06	0.84	0.75	0.94	-3.08	.002
Gender * Callousness	0.16	0.15	1.18	0.88	1.58	1.11	.27
Gender * Uncaring	0.15	0.12	1.16	0.92	1.47	1.26	.21
Gender * Unemotional	-0.14	0.12	0.87	0.69	1.09	-1.21	.23
Age * Callousness	0.01	0.02	1.01	0.97	1.05	0.52	.60
Age * Uncaring	-0.03	0.02	0.97	0.94	1.00	-1.77	.08
Age * Unemotional	0.00	0.02	1.00	0.97	1.03	-0.04	.96

Note. *N* = 6083; Gender was coded as 0 = girls and 1 = boys.

scoring higher on Callousness and Uncaring were 1.87 and 1.31 times, respectively, more likely to send/post more sexts of strangers without their consent. The Unemotional trait emerged as a significant negative predictor: participants with higher scores on this trait had a 16.5% lower probability of sending/posting more sexts of strangers without their consent. No interaction effects were observed.

#### CU Traits and Sending or Posting Sexts of one's Partner Without Their Consent

The last logistic mixed model investigated the effect of CU traits on sending or posting sexts of one's partner without their consent and explained about 13% of the variance (*R*-square marginal = 0.13; *R*-square conditional = 0.17). Results of the analysis are displayed in [Tables 4](#) and [6](#). Both gender and age emerged as significant predictors, with males and younger participants tending to send/post more sexts of one's partner without their consent. Callousness and Uncaring traits were significant positive predictors ([Table 6](#)):

participants who scored higher on Callousness and Uncaring were 3.05 and 1.68 times, respectively, more likely to send/post more sexts of one's partner without their consent. Consistent with previous findings, the Unemotional trait emerged as a negative significant predictor: participants who scored higher on this trait had about 31% lower probability of sending/posting more sexts of one's partner without their consent.

Remarkably, a significant interaction was observed between gender and Unemotional trait (see [Table 6](#)). To interpret this interaction effect, a simple slope analysis was conducted. In males, the impact of the Unemotional trait on sending/posting more sexts of one's partner without their consent was more pronounced (*O.R.* = 0.56, *p* < .001) compared to the non-significant effect observed among females (*O.R.* = 0.86, *p* = .28). In other words, higher scores on the Unemotional trait were negatively associated with sending/posting sexts of one's partner without their consent among males, whereas this association was absent among females (refer to [Figure 2](#)).

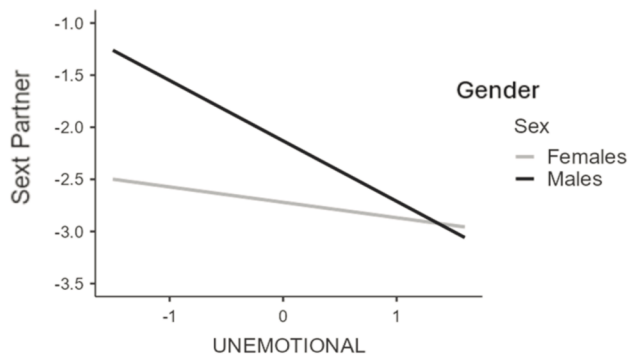
**Table 6***Sending or Posting Sexts of one's Partner Without Their Consent: Fixed Effects Parameter Estimates*

	<i>B</i>	<i>SE</i>	<i>exp(B)</i>	<i>95% Exp(B) CI</i>		<i>z</i>	<i>p</i>
Gender	0.59	0.12	1.80	1.43	2.27	5.02	<.001
Age	-0.04	0.02	0.96	0.93	1.00	-2.12	.034
Callousness	1.12	0.11	3.05	2.45	3.80	9.97	<.001
Uncaring	0.52	0.10	1.68	1.38	2.04	5.22	<.001
Unemotional	-0.36	0.10	0.70	0.57	0.84	-3.74	<.001
Gender * Callousness	0.25	0.21	1.29	0.85	1.96	1.19	.24
Gender * Uncaring	0.12	0.19	1.13	0.78	1.63	0.65	.52
Gender * Unemotional	-0.43	0.19	0.65	0.45	0.95	-2.25	.02
Age * Callousness	0.02	0.03	1.02	0.96	1.08	0.67	.50
Age * Uncaring	0.01	0.03	1.01	0.95	1.06	0.26	.80
Age * Unemotional	-0.04	0.03	0.96	0.91	1.01	-1.62	.11

Note. *N* = 4974; Gender was coded as 0 = girls and 1 = boys.

**Figure 2**

*The Effect of Unemotional on Sending or Posting Sexts of one's Partner Without Their Consent in Function of Gender*



## Discussion

Sexting is now widely acknowledged as a common sexual behavior among adolescents and young adults (Bianchi et al., 2017; Wachs et al., 2017). However, similar to other forms of sexual exploration, concerns may arise in specific circumstances, such as when explicit messages or images exchanged with an individual are shared without their knowledge or consent (Ringrose et al., 2013). In such cases, it is crucial to recognize the occurrence of abusive interactions, commonly referred to as non-consensual sexting (Barrense-Dias et al., 2020). Previous research has identified certain traits and experiences associated with non-consensual sexting, including behavioral and emotional difficulties, callousness, and histories of neglect and abuse during childhood (Barroso et al., 2021; Marinho et al., 2023).

There is a scarcity of research concerning non-consensual sexting, particularly regarding the investigation of the depicted person's identity in the shared sext. The present study contributes to the existing knowledge on non-consensual sexting by examining its association with CU traits, which represent the affective core components of psychopathy in adulthood. In doing so, this study enhances understanding of this phenomenon in a relatively underexplored area. Specifically, this research examines three potential victims involved in non-consensual sexting: romantic partners, acquaintances, and strangers. Additionally, the study focuses on three distinct CU traits: callousness, uncaring, and unemotional. Since CU traits are acknowledged as a risk factor for persistent antisocial behaviors among some youth (Viding & Kimonis, 2018), investigating the connections between these traits and non-consensual sexting can offer vital insights for identifying at-risk youth and implementing timely, targeted prevention interventions.

Previous studies examining the relationship between CU traits and non-consensual sexting have been limited in their scope, as they used a single-item measure to assess non-consensual sexting (instead of delineating different relationship contexts) and were limited to a single sample (Barroso et al., 2021; Marinho et al., 2023), thereby lacking in reliability. In contrast, our study employed a multi-item measure specifically administered for the present investigation, which has demonstrated good reliability in recent research (Morelli et al., 2023a, 2023b, 2023c) and collected primary data

from a total of 6093 adolescents and young adults (aged 13-30) across 11 countries worldwide.

Regarding the prevalence of aggravated sexting across different relational contexts, a higher propensity to engage in non-consensual sexting with strangers, rather than with acquaintances or partners, has emerged. This may confirm a tendency to experiment with risky forms of sexting outside of one's significant relationship (Dev et al., 2022).

The findings of the present study show that CU traits are key predictors of non-consensual sexting. As hypothesized, callousness and uncaring predicted an increased likelihood of engaging in all types of non-consensual sexting investigated (involving acquaintances, strangers, and partners). Conversely, lower unemotional traits predicted greater involvement in non-consensual sexting only concerning partners and strangers. These results align partially with previous research on CU traits and other forms of harassment (i.e., bullying, cyberbullying): callousness and uncaring were positively linked to bullying behaviors, whereas unemotional traits were found to be unrelated (Fanti et al., 2009; Kokkinos & Voulgaridou, 2017; Wright et al., 2019).

From a personality perspective, these findings can be interpreted through the lens of established theoretical models, such as the Dark Triad, HEXACO, and the Five Factor Model. Dark Triad traits—Narcissism, Machiavellianism, and Psychopathy—share similarities with CU traits, especially callousness and lack of empathy, which may increase the likelihood of harmful online behaviors like non-consensual sexting (March et al., 2017). Following the HEXACO model, high callousness and uncaring traits may correspond to low honesty-humility, indicating a propensity for exploiting others and engaging in unethical behavior (Morelli et al., 2020). Our expectation that callousness and uncaring traits would have differing associations compared to the unemotional trait concerning non-consensual sexting stems from viewing these dimensions as indicating common traits with distinct specificities. Callousness (i.e., lack of empathy, guilt, and remorse), and uncaring (i.e., disregard for consequences, may prompt younger individuals to impulsively engage without considering others' emotions. Conversely, the inability to express or experience emotions may lead to indifference towards romantic and sexual relationships. Therefore, individuals with unemotional characteristics may be less inclined to engage in sexting behaviors due to their overall lack of interest in forming emotional connections, even online. Consequently, young individuals with unemotional traits may avoid sexting altogether, as their reduced emotional responsiveness may extend to online interactions (Frick et al., 2014).

Again, within a personality perspective, our results highlighted important outcomes. Findings revealed a low ICC, suggesting that the variance attributable to differences between groups was small compared to the variance within groups. In other words, individuals within the same group were not substantially more similar than individuals in other groups. In this sense, our primary aim was to investigate the relationships between variables, focusing on a regression-based approach and modeling country membership. We believe that studying these relationships can provide broadly applicable insights across contexts, whereas group differences often reflect country-specific phenomena. Importantly, our findings strongly indicate that the relationships among the study variables hold despite potential differences in objective and contextual factors between countries (e.g., education rates, internet accessibility, GDP,

and similar factors). The results suggest that individual-level factors, rather than country membership, predominantly drive the observed outcomes. This finding enhances the generalizability of our results across diverse contexts. Furthermore, it underscores the crucial role of psychological variables. Notably, despite variations in the prevalence and frequency of sexting across the different countries considered, the relationships between personality factors, such as CU traits, and various forms of aggravated sexting behaviors remain consistent across different countries.

Furthermore, most studies on sexting have been conducted within a single country, limiting the generalizability of their results to other countries. This is the first study to investigate the personality correlates of different aggravated sexting behaviors, providing the opportunity to generalize the findings across countries from different continents.

Interestingly, younger individuals and women are more likely to engage in non-consensual sexting than older individuals and men, which contradicts previous studies suggesting older adolescents are more prone to such behavior (Barroso et al., 2021; Kernsmith et al., 2018). However, some authors propose that this difference could be due to the overall increase in sexual activity and sexting behavior with age (Barroso et al., 2023). Essentially, older teenagers engage in more sexting overall, putting them at greater risk for non-consensual sexting compared to younger individuals.

Younger individuals' tendency towards non-consensual sexting aligns with broader perspectives on psychological and sexual development, as they may exhibit reduced responsibility and future orientation, failing to consider consequences similar to other aspects of life (Clancy et al., 2019; Naezer & van Oosterhout, 2021). Young people often struggle with impulse control and risk assessment, leading to limited awareness of the seriousness of sexting and its consequences. Engagement in non-consensual sexting may stem from seeking attention, enjoyment, or peer acceptance, akin to behaviors like bullying (Barrense-Dias et al., 2020).

The literature on both consensual and non-consensual sexting offers conflicting findings regarding gender prevalence, with some studies indicating higher engagement among adolescent males and others among females (Barroso et al., 2023). Contrary to patriarchal stereotypes, men aren't necessarily more prone to non-consensual sexting. Motivations include misuse, lack of awareness, peer validation, gossip, and entertainment (Barrense-Dias et al., 2020). Women may receive unsolicited sexts, leading to public dissemination as protest or self-protection, with revenge possibly occurring in response to relationship endings or perceived deserving punishment (Naezer & van Oosterhout, 2021).

Two notable interaction effects emerged from the analyses. Firstly, younger individuals with heightened levels of callousness displayed increased likelihood of engaging in non-consensual sexting, suggesting that traits linked to lack of empathy and responsibility may exacerbate risks in younger age groups, potentially due to impulsive tendencies (Blair et al., 2014). Secondly, males with low unemotional traits were more prone to non-consensual sexting, echoing discussions on emotional detachment potentially facilitating abusive behaviors (Frick et al., 2014). However, male predominance in this dimension implies heightened importance of emotional involvement and communication for young males, possibly hindered by societal expectations discouraging emotional

expression, thereby fostering disruptive manifestations. Again, these insights may be useful in developing and providing targeted interventions and education to adolescents who may be at greater risk of this behavior.

This study has several limitations that need to be acknowledged. Firstly, the data collected was cross-sectional, which prevents us from establishing causal relationships among the variables. Additionally, the use of a snowball sampling method may limit the generalizability of our findings. Moreover, relying on self-report questionnaires introduces the possibility of social desirability bias. Furthermore, a more in-depth examination of the role played by recipients of forwarded sexts was lacking. Additionally, the study did not explore the motivations underlying non-consensual sexting, which could have been addressed by directly asking participants about their reasons for forwarding sexts without consent. Understanding these motivations could provide a more nuanced understanding of non-consensual sexting behavior. Future research should address these gaps through more targeted investigations (Barrense-Dias et al., 2020). Utilizing a mixed-method approach could be particularly beneficial, as it allows for the integration of quantitative data, such as the frequency of non-consensual sexting, with qualitative insights into motivations, thus offering a more comprehensive understanding of the phenomenon.

While acknowledging limitations, this study brings significant strengths and practical implications. It addresses the gap in understanding the link between maladaptive personality traits and sexting behaviors across multiple countries, focusing on victim identity. Exploring the unemotional dimension alongside callousness and uncaring, it sheds light on emotional involvement's role in non-consensual sexting. Findings suggest emotional detachment and reduced involvement act as protective factors, offering valuable insights for further research.

Moreover, this study engaged a substantial number of participants from diverse cultural backgrounds, spanning ages 13 to 30. This broader age spectrum facilitated insights into non-consensual sexting across different developmental stages. It is worth noting that previous studies often employed methodologically weak approaches, using single-item measures or general sexting behavior assessments (e.g., Barroso et al., 2023). In contrast, our study employed a comprehensive multi-item questionnaire tailored to explore various aspects of non-consensual sexting, capturing nuances among individuals involved. This robust measurement strategy yielded specific, detailed data with favorable psychometric properties, enhancing its suitability for future research endeavors.

These findings emphasize the importance of prevention interventions concerning non-consensual sexting. Understanding how personality traits influence online behaviors is crucial for designing effective measures. The study highlights a lack of empathy and guilt as predisposing factors for non-consensual sexting. Targeted interventions addressing callousness and uncaring can be developed, such as school programs fostering empathy and emotion management. This aligns with recent research on the role of emotional intelligence in sexting behaviors (Morelli et al., 2023b). The associations between CU traits and non-consensual sexting behaviors emphasize the need for targeted interventions focusing on emotional self-awareness, empathy training, and the promotion of ethical online conduct, particularly among individuals displaying high callous and uncaring traits.

In this regard, the findings underscore the importance of implementing comprehensive emotional and sexual education programs in schools. These programs should prioritize emotional self-awareness, promote gender equality, and challenge gender stereotypes. The primary aim of such initiatives should be to educate young individuals about the importance of refraining from engaging in non-consensual sexting. Educators and psychologists play a crucial role in implementing programs that equip young people with valuable skills for establishing healthy relationships and fostering positive interactions, both online and offline. By increasing awareness of the potential adverse outcomes associated with sexting, individuals can better cope with the negative repercussions that may arise from being involved as either a perpetrator or a victim of aggravated sexting.

#### Author Contributions

**Mara Morelli:** Project administration, Conceptualization, Methodology, Data curation, Writing - Original draft, Writing - review & editing, Supervision. **Fau Rosati:** Writing - Original draft, Data curation, Writing - review & editing. **Elena Cattelino:** Conceptualization, Data curation, Supervision, Writing - review & editing. **Flavio Urbini:** Investigation. **Roberto Baiocco:** Investigation, Writing - review & editing. **Dora Bianchi:** Investigation, Writing - review & editing. **Fiorenzo Laghi:** Investigation, Writing - review & editing. **Maurizio Gasseau:** Data curation. **Piotr Sorokowski:** Data curation. **Michał Misiak:** Data curation. **Martyna Dziekan:** Data curation. **Heather Hudson:** Data curation. **Alexandra Marshall:** Data curation. **Thanh Truc Nguyen:** Data curation. **Lauren Mark:** Data curation. **Kamil Kopecky:** Data curation. **René Szotkowski:** Data curation. **Ezgi Toplu Demirtaş:** Data curation. **Joris Van Ouytsel:** Data curation. **Koen Ponnet:** Data curation. **Michel Walrave:** Data curation. **Tingshao Zhu:** Data curation. **Ya Chen:** Data curation. **Nan Zhao:** Data curation. **Xiaoqian Liu:** Data curation. **Alexander Voiskounsky:** Data curation. **Nataliya Bogacheva:** Data curation. **Maria Ioannou:** Data curation. **John Synnott:** Data curation. **Kalliopi Tzani-Pepelasis:** Data curation. **Vimala Balakrishnan:** Data curation. **Moses Okumu:** Data curation. **Eusebius Small:** Data curation. **Silviya Pavlova Nikolova:** Data curation. **Michelle Drouin:** Data curation. **Antonio Chirumbolo:** Conceptualization, Methodology, Data curation, Writing - Original draft, Writing - review & editing, Supervision.

#### Funding

This work was partially supported by Sapienza University of Rome (Project title: The psychological underpinnings of sexting behaviors: A cross-cultural investigation; Grant number: RG11715C7C530999) and the work of Dr. Joris Van Ouytsel was supported by the the Research Foundation – Flanders (Grant number: 12J8719N).

#### Declaration of Interests

The authors declare that there are no conflicts of interest.

#### Data Availability Statement

Data are available under request to the first author.

#### References

- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Sage Publications.
- Barrense-Dias, Y., Akre, C., Auderset, D., Leeners, B., Morselli, D., & Suris, J. C. (2020). Non-consensual sexting: Characteristics and motives of youths who share received-intimate content without consent. *Sexual Health*, 17(3), 270–278. <https://doi.org/10.1071/SH19201>
- Barroso, R., Marinho, A. R., Figueiredo, P., Ramião, E., & Silva, A. S. (2023). Consensual and non-consensual sexting behaviors in adolescence: A systematic review. *Adolescent Research Review*, 8(1), 1–20. <https://doi.org/10.1007/s40894-022-00199-0>
- Barroso, R., Ramião, E., Figueiredo, P., & Araújo, A. M. (2021). Abusive sexting in adolescence: Prevalence and characteristics of abusers and victims. *Frontiers in Psychology*, 12, Article 610474. <https://doi.org/10.3389/fpsyg.2021.610474>
- Baumgartner, S. E., Sumter, S. R., Peter, J., Valkenburg, P. M., & Livingstone, S. (2014). Does country context matter? Investigating the predictors of teen sexting across Europe. *Computers in Human Behavior*, 34, 157–164. <https://doi.org/10.1016/j.chb.2014.01.041>
- Baumgartner, S. E., Valkenburg, P. M., & Peter, J. (2010). Unwanted online sexual solicitation and risky sexual online behavior across the lifespan. *Journal of Applied Developmental Psychology*, 31(6), 439–447. <https://doi.org/10.1016/j.appdev.2010.07.005>
- Bianchi, D., Morelli, M., Baiocco, R., & Chirumbolo, A. (2017). Sexting as the mirror on the wall: Body-esteem attribution, media models, and objectified-body consciousness. *Journal of Adolescence*, 61(1), 164–172. <https://doi.org/10.1016/j.adolescence.2017.10.006>
- Bianchi, D., Morelli, M., Baiocco, R., Cattelino, E., Laghi, F., & Chirumbolo, A. (2019). Family functioning patterns predict teenage girls' sexting. *International Journal of Behavioral Development*, 43(6), 507–514. <https://doi.org/10.1177/0165025419873037>
- Blair, R. J. R., Leibenluft, E., & Pine, D. S. (2014). Conduct disorder and callous-unemotional traits in youth. *New England Journal of Medicine*, 371(23), 2207–2216. <https://doi.org/10.1056/NEJMc1415936>
- Carlson, M., Oshri, A., & Kwon, J. (2015). Child maltreatment and risk behaviors: The roles of callous/unemotional traits and conscientiousness. *Child Abuse & Neglect*, 50, 234–243. <https://doi.org/10.1016/j.chiabu.2015.07.003>
- Centifanti, L. C. M., Fanti, K. A., Thomson, N. D., Demetriou, V., & Anastassiou-Hadjicharalambous, X. (2015). Types of relational aggression in girls are differentiated by callous-unemotional traits, peers and parental overcontrol. *Behavioral Sciences*, 5(4), 518–536. <https://doi.org/10.3390/bs5040518>
- Chalfen, R. (2009). 'It's only a picture': Sexting, 'smutty' snapshots and felony charges. *Visual Studies*, 24(3), 258–268. <https://doi.org/10.1080/14725860903309203>
- Clancy, E. M., Klettke, B., & Hallford, D. J. (2019). The dark side of sexting: Factors predicting the dissemination of sexts. *Computers in Human Behavior*, 92, 266–272. <https://doi.org/10.1016/j.chb.2018.11.023>
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.
- Dev, P., Medina, J., Agha, Z., De Choudhury, M., Razi, A., & Wisniewski, P. J. (2022). From ignoring strangers' solicitations to mutual sexting with friends: Understanding youth's online sexual risks in Instagram private conversations. In G. Hsieh, A. Tang, M. G. Ames, S. Ding, S. Fussell, V. Liao, A. Monroy-Hernández, S. Munson, I. Shklovski and J. Tang, *Companion Publication of the 2022 Conference on Computer Supported*



- Cooperative Work and Social Computing (pp. 94–97). Association for Computing Machinery.
- Dodaj, A., & Sesar, K. (2020). Sexting categories. *Mediterranean Journal of Clinical Psychology*, 8(2), 1–26. <https://doi.org/10.6092/2282-1619/mjcp-2432>
- Drouin, M., & Landgraff, C. (2012). Texting, sexting, attachment, and intimacy in college students' romantic relationships. *Computers in Human Behavior*, 28(2), 444–449. <https://doi.org/10.1016/j.chb.2011.10.015>
- Drouin, M., Coupe, M., & Temple, J. R. (2017). Is sexting good for your relationship? It depends. *Computers in Human Behavior*, 75, 749–756. <https://doi.org/10.1016/j.chb.2017.06.018>
- Fanti, K. A., Demetriou, C. A., & Kimonis, E. R. (2013). Variants of callous-unemotional conduct problems in a community sample of adolescents. *Journal of Youth and Adolescence*, 42(7), 964–979. <https://doi.org/10.1007/s10964-013-9958-9>
- Fanti, K. A., Frick, P. J., & Georgiou, S. (2009). Linking callous-unemotional traits to instrumental and non-instrumental forms of aggression. *Journal of Psychopathology and Behavioral Assessment*, 31, 285–298. <https://doi.org/10.1007/s10862-008-9111-3>
- Fontaine, N. M. G., Rijdsdijk, F. V., McCrory, E. J. P., & Viding, E. (2010). Etiology of different developmental trajectories of callous-unemotional traits. *Journal of the American Academy of Child and Adolescent Psychiatry*, 49(7), 656–664. <https://doi.org/10.1016/j.jaac.2010.03.014>
- Frick, P. J. (2004). *Inventory of Callous–Unemotional Traits* [Database record]. APA PsycTests. <https://doi.org/10.1037/t62639-000>
- Frick, P. J., & White, S. F. (2008). Research review: The importance of callous-unemotional traits for developmental models of aggressive and antisocial behavior. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 49(4), 359–375. <https://doi.org/10.1111/j.1469-7610.2007.01862.x>
- Frick, P. J., Cornell, A. H., Barry, C. T., Bodin, S. D., & Dane, H. E. (2003). Callous-unemotional traits and conduct problems in the prediction of conduct problem severity, aggression, and self-report of delinquency. *Journal of Abnormal Child Psychology*, 31, 457–470.
- Frick, Paul J., Ray, J. V., Thornton, L. C., & Kahn, R. E. (2014). Annual research review: A developmental psychopathology approach to understanding callous-unemotional traits in children and adolescents with serious conduct problems. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 55(6), 532–548. <https://doi.org/10.1111/jcpp.12152>
- Gallucci, M. (2019). *GAMLj: General analyses for linear models* [Jamovi module]. <https://gamlj.github.io/>.
- Gómez-Guadix, M., & de Santisteban, P. (2018). “sex pics”: Longitudinal predictors of sexting among adolescents. *The Journal of Adolescent Health: Official Publication of the Society for Adolescent Medicine*, 63(5), 608–614. <https://doi.org/10.1016/j.jadohealth.2018.05.032>
- Gassó, A. M., Mueller-Johnson, K., Agustina, J. R., & Gómez-Durán, E. (2021). Mental health correlates of sexting coercion perpetration and victimisation in university students by gender. *The Journal of Sexual Aggression*, 27(2), 247–263. <https://doi.org/10.1080/13552600.2021.1894493>
- Gewirtz-Meydan, A., Mitchell, K. J., & Rothman, E. F. (2018). What do kids think about sexting? *Computers in Human Behavior*, 86, 256–265. <https://doi.org/10.1016/j.chb.2018.04.007>
- Gil-Llario, M. D., Gil-Juliá, B., Morell-Mengual, V., Cárdenas-López, G., & Ballester-Arnal, R. (2021). Analysis of demographic, psychological and cultural aspects associated with the practice of sexting in Mexican and Spanish adolescents. *International Journal of Intercultural Relations*, 82, 197–206. <https://doi.org/10.1016/j.ijintrel.2021.03.013>
- Hare, R. D., & Neumann, C. S. (2008). Psychopathy as a clinical and empirical construct. *Annual Review of Clinical Psychology*, 4(1), 217–246. <https://doi.org/10.1146/annurev.clinpsy.3.022806.091452>
- Helfrit, L. E., & Stanford, M. S. (2006). Personality and psychopathology in an impulsive aggressive college sample. *Aggressive Behavior*, 32(1), 28–37. <https://doi.org/10.1002/ab.20103>
- Kernsmith, P. D., Victor, B. G., & Smith-Darden, J. P. (2018). Online, ofline, and over the line: Coercive sexting among adolescent dating partners. *Youth and Society*, 50(7), 891–904. <https://doi.org/10.1177/0044118X18764040>
- Kimonis, E. R., Frick, P. J., Skeem, J. L., Marsee, M. A., Cruise, K., Munoz, L. C., Aucoin, K. J., & Morris, A. S. (2008). Assessing callous-unemotional traits in adolescent offenders: Validation of the Inventory of Callous-Unemotional Traits. *International Journal of Law and Psychiatry*, 31(3), 241–252. <https://doi.org/10.1016/j.ijlp.2008.04.002>
- Kokkinos, C. M., & Voulgaridou, I. (2017). Relational and cyber aggression among adolescents: Personality and emotion regulation as moderators. *Computers in Human Behavior*, 68, 528–537. <https://doi.org/10.1016/j.chb.2016.11.046>
- Livingstone, S., & Görzig, A. (2014). When adolescents receive sexual messages on the internet: Explaining experiences of risk and harm. *Computers in Human Behavior*, 33, 8–15. <https://doi.org/10.1016/j.chb.2013.12.021>
- Madigan, S., Ly, A., Rash, C. L., Van Ouytsel, J., & Temple, J. R. (2018a). Prevalence of multiple forms of sexting behavior among youth: A systematic review and meta-analysis. *JAMA Pediatrics*, 172(4), 327–335. <https://doi.org/10.1001/jamapediatrics.2017.5314>
- Madigan, S., Van Ouytsel, J., & Temple, J. R. (2018b). Nonconsensual sexting and the role of sex differences-reply. *JAMA Pediatrics*, 172(9), 890–891. <https://doi.org/10.1001/jamapediatrics.2018.1951>
- March, E., Grieve, R., Marrington, J., & Jonason, P. K. (2017). Trolling on Tinder® (and other dating apps): Examining the role of the Dark Tetrad and impulsivity. *Personality and Individual Differences*, 110, 139–143. <https://doi.org/10.1016/j.paid.2017.01.025>
- Marinho, A. R., Figueiredo, P., Ramião, E., Silva, S., & Barroso, R. (2023). Callous-unemotional traits mediate the effect of childhood maltreatment in later non-consensual sexting practices. *Journal of Family Trauma, Child Custody & Child Development*, 20(3), 315–331. <https://doi.org/10.1080/26904586.2022.2152146>
- Morelli, M., Bianchi, D., Baiocco, R., Pezzuti, L., & Chirumbolo, A. (2016). Sexting, psychological distress and dating violence among adolescents and young adults. *Psicothema*, 28(2), 137–142. <http://doi.org/10.7334/psicothema2015.193>
- Morelli, M., Cattelino, E., Baiocco, R., Chirumbolo, A., Crea, G., Longobardi, E., Nappa, M. R., & Graziano, F. (2023b). The relationship between trait emotional intelligence and sexting in adolescence. *Sexuality Research and Social Policy*, 21, 1607–1620. <https://doi.org/10.1007/s13178-023-00913-0>
- Morelli, M., Chirumbolo, A., Bianchi, D., Baiocco, R., Cattelino, E., Laghi, F., Sorokowski, P., Misiak, M., Dziekan, M., Hudson, H., Marshall, A., Nguyen, T. T. T., Mark, L., Kopecky, K., Szotkowski, R., Demirtaş, E. T., Van Ouytsel, J., Ponnet, K., Walrave, M., ... Drouin, M. (2020). The role of HEXACO personality traits in different kinds of sexting: A cross-cultural study in 10 countries. *Computers in Human Behavior*, 113, 106502. <https://doi.org/10.1016/j.chb.2020.106502>
- Morelli, M., Graziano, F., Chirumbolo, A., Longobardi, E., & Cattelino, E. (2023c). Sexting in adolescenza: Profili di intelligenza emotiva di sexters e non sexters. *Sistemi intelligenti*, 35(3), 633–654. <https://doi.org/10.1422/109321>



- Morelli, M., Plata, M. G., Isolani, S., Zabala, M. E. Z., Hoyos, K. P. C., Tirado, L. M. U., ... & Baiocco, R. (2023a). Sexting behaviors before and during COVID-19 in Italian and Colombian young adults. *Sexuality Research and Social Policy*, 20, 1515–1527. <https://doi.org/10.1007/s13178-023-00798-z>
- Morelli, M., Urbini, F., Bianchi, D., Baiocco, R., Cattelino, E., Laghi, F., Sorokowski, P., Misiak, M., Dziekan, M., Hudson, H., Marshall, A., Nguyen, T. T. T., Mark, L., Kopecky, K., Szotkowski, R., Toplu Demirtaş, E., Van Ouytsel, J., Ponnet, K., Walrave, M., ... Chirumbolo, A. (2021). The relationship between Dark Triad Personality Traits and sexting behaviors among adolescents and young adults across 11 countries. *International Journal of Environmental Research and Public Health*, 18(5), 2526. <https://doi.org/10.3390/ijerph18052526>
- Mori, C., Cooke, J. E., Temple, J. R., Ly, A., Lu, Y., Anderson, N., Rash, C., & Madigan, S. (2020). The prevalence of sexting behaviors among emerging adults: A meta-analysis. *Archives of Sexual Behavior*, 49(4), 1103–1119. <https://doi.org/10.1007/s10508-020-01656-4>
- Mori, C., Park, J., Temple, J. R., & Madigan, S. (2022). Are youth sexting rates still on the rise? A meta-analytic update. *Journal of Adolescent Health*, 70(4), 531–539. <https://doi.org/10.1016/j.jadohealth.2021.10.026>
- Mori, C., Temple, J. R., Browne, D., & Madigan, S. (2019). Association of sexting with sexual behaviors and mental health among adolescents: A systematic review and meta-analysis. *JAMA Pediatrics*, 173(8), 770–779. <https://doi.org/10.1001/jamapediatrics.2019.1658>
- Naezer, M., & van Oosterhout, L. (2021). Only sluts love sexting: Youth, sexual norms and non-consensual sharing of digital sexual images. *Journal of Gender Studies*, 30(1), 79–90. <https://doi.org/10.1080/09589236.2020.1799767>
- Pardini, D. A., Lochman, J. E., & Frick, P. J. (2003). Callous/unemotional traits and social-cognitive processes in adjudicated youths. *Journal of the American Academy of Child and Adolescent Psychiatry*, 42(3), 364–371. <https://doi.org/10.1097/00004583-200303000-00018>
- Ringrose, J., Gill, R., Livingstone, S., & Harvey, L. (2013). A qualitative study of children, young people and ‘sexting’: Gendered value in digital image exchange. *Feminist Theory*, 14(3), 305–323. <https://doi.org/10.1177/1464700113499853>
- Salter, M., Crofts, T., & Lee, M. (2013). Beyond criminalisation and responsibilisation: Sexting, gender and young people. *Current Issues in Criminal Justice*, 24(3), 301–316. <https://doi.org/10.1080/10345329.2013.12035963>
- Temple, J. R., & Lu, Y. (2018). “Sexting from a health perspective: Sexting, health, and risky sexual behaviour.” In M. Walrave, J. Van Ouytsel, K. Ponnet & J. R. Temple (Eds.) *Sexting Motives and Risk in Online Sexual Self-Presentation* (pp. 53–61). Springer.
- Van Ouytsel, J., Lu, Y., Shin, Y., Avalos, B. L., & Pettigrew, J. (2021). Sexting, pressured sexting and associations with dating violence among early adolescents. *Computers in Human Behavior*, 125, Article 106969. <https://doi.org/10.1016/j.chb.2021.106969>
- Viding, E., & Kimonis, E. R. (2018). Callous-unemotional traits. In C.J. Patrick (Ed) *Handbook of Psychopathy. Second Edition* (pp. 144–164). The Guilford Press.
- Wachs, S., Wright, M. F., & Wolf, K. D. (2017). Psychological correlates of teen sexting in three countries - direct and indirect associations between self-control, self-esteem, and sexting. *International Journal of Developmental Science*, 11(3–4), 109–120. <https://doi.org/10.3233/DEV-160212>
- Walker, K., & Sleath, E. (2017). A systematic review of the current knowledge regarding revenge pornography and non-consensual sharing of sexually explicit media. *Aggression and Violent Behavior*, 36, 9–24. <https://doi.org/10.1016/j.avb.2017.06.010>
- Waller, R., Gardner, F., Viding, E., Shaw, D. S., Dishion, T. J., Wilson, M. N., & Hyde, L. W. (2014). Bidirectional associations between parental warmth, callous unemotional behavior, and behavior problems in high-risk preschoolers. *Journal of Abnormal Child Psychology*, 42(8), 1275–1285. <https://doi.org/10.1007/s10802-014-9871-z>
- Waller, R., Wagner, N. J., Barstead, M. G., Subar, A., Petersen, J. L., Hyde, J. S., & Hyde, L. W. (2020). A meta-analysis of the associations between callous-unemotional traits and empathy, prosociality, and guilt. *Clinical Psychology Review*, 75, Article 101809. <https://doi.org/10.1016/j.cpr.2019.101809>
- Wolak, J., & Finkelhor, D. (2011). *Sexting: A typology*. Crimes against Children Research Centre. <https://scholars.unh.edu/cgi/viewcontent.cgi?article=1047&context=ccrc>
- Wright, N., Hill, J., Pickles, A., & Sharp, H. (2019). Callous-unemotional traits, low cortisol reactivity and physical aggression in children: Findings from the Wirral Child Health and Development Study. *Translational Psychiatry*, 9(1), 79. <https://doi.org/10.1038/s41398-019-0406-9>

Article

## Assessing Positive Organizational Culture: Psychometric Properties of the POCS

Javier Barriá-González<sup>1</sup> , Jaime García-Fernández<sup>2</sup> , Ricardo Pérez-Luco<sup>1</sup>  and Álvaro Postigo<sup>2</sup> 

<sup>1</sup>Universidad de La Frontera (Chile)

<sup>2</sup>Universidad de Oviedo (Spain)

### ARTICLE INFO

Received: 26/11/2024  
Accepted: 24/02/2025

#### Keywords:

POCS  
Organizational culture  
Work environment  
Psychometric properties  
Measurement invariance

### ABSTRACT

**Background:** The Positive Organizational Culture construct is a set of shared practices, values, and behaviors within an organization that promote healthy and motivating working environments. This study develops a new scale called the Positive Organizational Culture Scale (POCS) to assess how organizational values affect well-being and work performance. **Method:** The sample consisted of 1,420 workers in Chile, with an average age of 39.48 years ( $SD = 11.13$ ). Over half (55.0%) worked in the public sector, 34.5% worked in private organizations, and 10.5% worked in private non-profit organizations. The study examined item descriptions, the scale's internal structure, its measurement invariance regarding sex and organization, and its relationship with other psychological variables (organizational climate, professional burnout, psychosomatic symptomatology). **Results:** The POCS showed a good fit to a correlated two-factor structure (People-Oriented Culture and Results-Oriented Culture;  $CFI = .94$ ;  $RMSEA = 0.08$ ), demonstrating measurement invariance regarding sex and type of organization. The findings show that the POCS has 36 items exhibiting satisfactory psychometric properties and a structure consisting of two first-order factors, which exhibit distinct associations with the other recorded variables. **Conclusions:** The POCS provides relevant information for formulating actions aimed at enhancing the work environment in the Chilean context.

## Evaluación de la Cultura Organizacional Positiva: Propiedades Psicométricas de la POCS

### RESUMEN

**Antecedentes:** La Cultura Organizacional Positiva es un conjunto de prácticas, valores y comportamientos compartidos por una organización que promueven entornos laborales saludables y motivadores. El objetivo del estudio fue desarrollar la Escala de Cultura Organizacional Positiva (ECOP), la cual evalúa cómo los valores organizacionales afectan el bienestar y rendimiento laboral. **Método:** La muestra fueron 1.420 trabajadores de Chile, con una edad media de 39,48 años ( $DT = 11,13$ ). El 55% eran trabajadores del sector público, el 34,5% de organizaciones privadas y el 10,5% de organizaciones privadas sin fines de lucro. Se estudiaron los descriptivos de los ítems, la estructura interna de la escala, su invarianza de medida en términos de sexo y organización y su relación con otras variables psicológicas (clima organizacional, desgaste profesional, sintomatología psicósomática). **Resultados:** La ECOP mostró un buen ajuste a una estructura de dos factores correlacionados (Cultura Orientada a las Personas y Cultura Orientada a los Resultados;  $CFI = .94$ ;  $RMSEA = 0.08$ ), demostrando invarianza de medida en términos de sexo y tipo de organización. Los factores mantienen relaciones diferentes con las otras variables registradas. **Conclusiones:** La ECOP ofrece información relevante para el desarrollo de intervenciones que fortalezcan el ambiente laboral en el contexto chileno.

#### Palabras clave:

POCS  
Cultura organizacional  
Clima laboral  
Propiedades psicométricas  
Invarianza de medida

## Introduction

As complex social systems, organizations exhibit deeply embedded patterns of behavior that shape internal interactions, decisions, and strategies (Ostroff & Schulte, 2014). This culture is defined by collective values and fundamental assumptions that explain organizational behavior and priorities, anchored in its members' common ideas, values, and social norms (Schneider et al., 2017). In turn, these cultural elements provide a framework that guides how members interpret, consider, and react to events within the organization (Schein, 2015).

Organizational culture is a vehicle of cohesion and coordination, fostering a fundamental source of collective identity and commitment. Beyond being a source of cohesion and coordination, it also fosters a shared identity, strengthening the bond between people and the organization and promoting greater commitment to organizational goals. Organizational culture affects employee performance and well-being by creating an atmosphere that either facilitates or impedes the use of personal and professional resources and the satisfaction of job expectations. Maintaining such an environment is crucial for ensuring a safe and effective workplace (Aryani & Widodo, 2020; Bakker & Demerouti, 2018; Prieto-Díez et al., 2022).

Organizational culture significantly impacts workplace stress, performance, and burnout, playing a key role in how employees perceive and manage job-related stress. According to the study by Olynick and Li (2020), an organizational culture that promotes mutual support and recognition can mitigate stress levels and reduce burnout by fostering a positive and cooperative work environment. Conversely, cultures that place excessive value on competitiveness and high-performance demands can increase stress and contribute to employee burnout (Taris, 2023). These cultural dynamics affect workers' mental and physical health and directly impact their effectiveness and efficiency. Di Stefano and Gaudiño (2019) point out that if a culture fails to manage workloads or provide sufficient resources, it can lead to diminished performance and increased absenteeism, adversely affecting organizational outcomes. Therefore, understanding the relationship between organizational culture and job stress is essential to developing effective strategies that promote well-being and sustainable productivity in the workplace (Barría-González et al., 2023; Jacob & Tende, 2022; Rattrie et al., 2020).

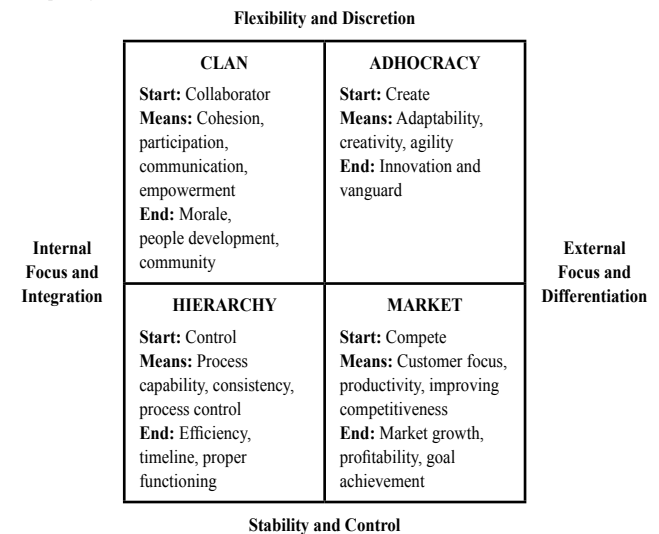
A positive organizational culture promotes respect, integrity, and openness, fostering a healthy work environment and enhancing organizational effectiveness. According to Gelfand et al. (2017), organizations with positive cultures exhibit higher levels of commitment and satisfaction among employees, reducing turnover and improving internal cohesion. In addition, Akpa et al. (2021) note that positive cultures facilitate organizational adaptability, enabling organizations to respond more effectively to market changes and internal crises. The study by Schneider et al. (2017) complements these findings, reporting that positive perceptions of organizational climate are strongly linked to performance and innovation. As research has demonstrated its direct impact on well-being and performance, the concept of positive organizational culture has gained increasing attention (Parent & Lovelace, 2018).

Positive organizational culture is defined as “a set of shared practices, values, and behaviors within an organization that promote a healthy and motivating work environment. It fosters cooperation and support for individual well-being (People-Oriented) while driving efficiency, competitiveness, and the achievement of organizational

goals” (Results-Oriented)”. This culture strives to balance human development and performance, shifting toward ethical and sustainable organizational behavior. It perceives organizational culture not merely as a framework for operational efficiency but also as a catalyst for people's well-being and sustainable organizational advancement (Bal, 2017; Donaldson et al., 2022; Hofstede, 2011; Luthans & Youssef-Morgan, 2016; Parent & Lovelace, 2018; van Zyl et al., 2024).

Similarly, the Competing Values Framework (CVF; Cameron & Quinn, 2006) offers a structured approach to understanding organizational culture. The CVF identifies four types of organizational cultures: clan, adhocracy, market, and hierarchy, each underpinned by specific sets of values and practices that support the achievement of organizational objectives in different ways (see Figure 1). Within this framework, different cultures emphasize distinct organizational priorities. For example, Clan cultures emphasize collaboration and mutual commitment, whereas Adhocracy cultures prioritize innovation and flexibility, which are crucial for organizations operating in dynamic and competitive environments. Likewise, Market and Hierarchy cultures focus on competition and control, respectively, each suitable for contexts where efficiency and consistency are priorities (Cameron et al., 2006).

**Figure 1**  
Competing Values Framework



Note. Taken from Hartnell et al. (2011).

The types of organizational culture and their impact on various organizational dynamics highlight the importance of balancing flexibility and control to enhance organizational performance and well-being. According to the CVF (Clan, Adhocracy, Market, and Hierarchy), Clan and Adhocracy cultures, oriented toward flexibility and mutual support, promote innovation, commitment, and job satisfaction by fostering autonomy and personal development. In contrast, Market and Hierarchy cultures, focused on results, efficiency, and control, drive productivity and operational stability but may limit innovation. Striking a balance between structure and adaptability is essential to address challenges and maintain a resilient and productive work environment (Ehrhart & Kuenzi, 2017; Gregory et al., 2009; Hartnell et al., 2011; Sarros et al., 2008).

Additionally, the importance of a positive organizational culture lies not only in its ability to influence employee well-being and performance but also in the necessity of having accurate tools to evaluate and manage it effectively. In this context, the Positive Organizational Culture Scale (POCS; [Perez-Luco, 2008](#)) emerges as a key instrument for addressing an existing gap in the measurement of organizational culture. This tool aims to fill a gap in measuring organizational culture in complex environments, integrating dimensions such as well-being and performance to strengthen organizational health and sustainability.

The first aim of the present study is to explore the dimensionality of the POCS. Although the original proposal ([Pérez-Luco, 2008](#)) includes six theoretical facets—Skills, Relationships, Branding, Vanguard, Rigor, and Improvisation—these facets require further empirical validation. To achieve this, we propose a model based on the CVF framework, specifically its structural dimension contrasting Flexibility and Stability (see [Figure 1](#), horizontal axis). This approach directly relates the dimensions of People Orientation and Results Orientation, derived from the definition of Positive Organizational Culture and indicated by [Hofstede \(2011\)](#). According to this model, Clan and Adhocracy cultures value collaboration and adaptability and are people-oriented. On the other hand, Market and Hierarchy cultures, which emphasize efficiency, control, and competitiveness, are results-oriented ([Beus et al., 2020](#); [Hartnell et al., 2019](#)). By aligning the POCS dimensions with the CVF, we provide a structured method for evaluating organizational culture within diverse workplace contexts. This is how the Skills and Relationships (Clan) facets emphasize personal development, well-being, internal cohesion, and the importance of personal relationships within the organization. The Vanguard and Improvisation (Adhocracy) facets highlight the importance of innovation, adaptability, and advanced technologies. The Rigor (Hierarchy) facet reflects the importance of organizational structure, process control, regulation, and efficiency. Finally, the Branding (Market) facet focuses on competitiveness and market success.

The creation of a new version of the POCS is proposed to evaluate the positive dimensions of organizational culture that influence the subjective work dynamics of complex organizations, encompassing

both public and private entities. The aim is to systematically analyze how organizational practices and values impact the well-being and productivity of individuals and teams.

In this sense, several instruments have been designed to assess organizational culture based on consolidated theories. These questionnaires, which are widely recognized and used, provide insight into organizational values and practices ([Tadesse & Debela, 2024](#)). In the Spanish-speaking context, instruments to measure organizational culture often present limitations in terms of theoretical consistency and evidence of validity. Many of the questionnaires used are based on models developed by English-speaking authors, such as [Denison \(1990\)](#), [Hartnell et al. \(2019\)](#), [Schein \(2010\)](#), [Cameron & Quinn \(2011\)](#), [Cooke and Lafferty \(1987\)](#), [O'Reilly et al. \(1991\)](#), and [Hofstede \(1991\)](#). Chile is no exception, having created a specific questionnaire for the education field. As in other Latin American countries, several recognized international instruments have been validated. Some of the most relevant questionnaires in English, Spanish, and the Chilean context are in [Table 1](#).

As [Table 1](#) shows, there are instruments in Chile to assess organizational culture; however, none is specifically designed to measure the balance between job demands and resources, focusing on well-being and performance. Most available questionnaires, like [Marcone and Martin del Buey \(2003\)](#) Inventory of Organizational Culture in Education Institutions (ICOE), focus on measuring organizational culture in the education setting without specifically addressing the relationship between demands and resources. The POCS signifies progress in this area, as its dual dimensions—People-Oriented and Results-Oriented—, making it possible to assess the impact of organizational values on well-being and work performance.

In this line, the psychometric properties of this scale will be studied in the Chilean context. The items of the POCS will be analyzed, the reliability of their scores will be explored, and evidence of validity will be collected from different sources, such as those based on internal structure and in relation to other variables such as organizational climate, professional burnout, and psychosomatic symptomatology. POCS will enhance the theoretical

**Table 1**  
*Organizational Culture Questionnaires for the English, Spanish, and the Chilean Contexts*

Questionnaire	Authors
<b>English-language questionnaires</b>	
Organizational Culture Assessment Instrument (OCAI)	<a href="#">Cameron &amp; Quinn (2011)</a>
The FOCUS Questionnaire	<a href="#">van Muijen et al. (1999)</a>
Organization Culture Profile (OCP)	<a href="#">O'Reilly et al. (1991)</a>
Denison Organizational Culture Survey (DOCS).	<a href="#">Denison (1990)</a>
Organizational Culture Inventory (OCI)	<a href="#">Cooke &amp; Lafferty (1987)</a>
<b>Spanish-language questionnaires</b>	
Escala de Diagnóstico de la Cultura Organizacional (EDCO) ( <i>Organizational Culture Diagnostic Scale</i> )	<a href="#">Robles et al. (2018)</a>
Instrumento de cultura organizacional y Competitividad (ICOC) ( <i>Organizational culture and competitiveness instrument</i> )	<a href="#">Hernández et al. (2008)</a>
Brazil's instrument for assessing organizational culture	<a href="#">Ferreira et al. (2002)</a>
Cuestionario Focus 93 ( <i>Focus 93 Questionnaire</i> )	<a href="#">González-Romá et al. (1996)</a>
<b>Chilean-context questionnaires</b>	
Inventory of Organizational Culture in Education Institutions (ICOE)	<a href="#">Marcone &amp; Martin del Buey (2003)</a>



framework of organizational psychology and establish itself as a vital resource for optimizing work dynamics and promoting health within organizations.

## Method

### Participants

The sample comprises 1,420 workers from productive and service organizations, seven public and two private, from different cities in Chile. Fifty-five percent of the sample belongs to public organizations, 34.5% to private organizations, and 10.5% to private non-profit organizations (social development). 97.75% of the sample were full-time workers. The age ranged from 18 to 65 years, with a mean of 39.48 years and a standard deviation of 11.13. Regarding age groups, 325 were classified as young (18 to 30 years), 828 as adults (31 to 50 years), and 241 as older (more than 50 years). Forty-five percent of the sample were women.

### Instruments

#### *Positive Organizational Culture Scale (POCS)*

This is a 41-item questionnaire with Likert-type responses with five response alternatives from 1 (*never*) to 5 (*always*). The scale is used to assess organizational culture. The original version (POCS; Pérez-Luco, 2008) includes six facets (Skills, Relationships, Branding, Vanguard, Rigor, and Improvisation). Evidence of content validity was ensured through a review by organizational psychology experts, who assessed the representativeness and relevance of the items in relation to the construct's facets (Pérez-Luco, 2008). Although this structure has shown good evidence of validity regarding its content, not validity evidence in terms of its internal structure has been reported. Thus, in the present study, the dimensionality of the 41 items will be explored to produce a new version of the POCS. The items can be found in the Supplementary Material.

#### *Subjective Work Environment Climate Scale (SWECS; Barriá-González et al., 2021)*

The SWECS is a questionnaire with 38 items that assesses five dimensions of organizational climate: Organizational Trust, Job Stress, Social Support, Compensation, and Job Satisfaction. The items that make up the questionnaire follow a Likert-type format with five response categories (1 = *never*, 5 = *always*). The scale has adequate psychometric properties to evaluate organizational climate in the Chilean context. The dimension-specific reliability coefficients of the scores ( $\alpha$ ) are: Organizational Trust, .91; Job Stress, .75; Social Support, .82; Compensation, .79; and Job Satisfaction, .78.

#### *Professional Burnout Scale (PBS; Perez-Luco, 2008)*

This scale is composed of 22 items that measure worker burnout. The scale is used to assess the degree of professional burnout and includes three dimensions (Emotional Fatigue, Personal Fulfillment, and Affective Hardening), using a Likert scale from 1 (*never*) to 5 (*always*). The study sample presented reliability coefficients of the scores ( $\alpha$ ) of .86 for Emotional Fatigue, .77 for Personal Fulfillment, and .76 for Affective Hardening.

#### *Psychosomatic Symptomatology Scale (PSS; Pérez-Luco, 2008)*

The scale measures the psychological and somatic symptoms of professional burnout through 22 items, using a dichotomous scale: 0 (*no*) and 1 (*yes*). Reliability coefficients of the scores ( $\alpha$ ) of .87 for Psychological Symptomatology and .78 for Somatic Symptomatology were found in this study sample.

### Procedure

A theoretical matrix of eight fields was defined for the selection of the organizations, considering funding source (public/private), orientation (production/services), and purpose (profit and social development). In each case, different complex organizations (four or more divisions, three or more hierarchical levels or sections, and a minimum of 200 employees) with a presence in two or more regions in Chile were identified, and their managers were contacted through formal and informal channels to invite them to participate in the study. Representation was obtained in seven of the eight types since no representation was obtained from productive for-profit public organizations. The instrument was self-administered and accessible on a website. Informed consent was obtained from each study participant before beginning the application of the instrument to ensure anonymity, confidentiality, and adherence to data protection regulations. The participation agreement encompassed a comprehensive assessment of the subjective work environment, followed by the dissemination of results to the corresponding executives.

### Data Analysis

First, following a cross-validation procedure (Fabrigar et al., 1999; Rey-Sáez, 2022), the sample was divided in two with the SOLOMON algorithm (Lorenzo-Seva, 2021), obtaining two halves of 710 people each. With the first half, the dimensionality of the instrument was explored through an exploratory factor analysis (EFA).

In the EFA, the KMO and Bartlett statistics were used to assess the suitability of the data for the factor analysis. The EFA was performed on the polychoric correlation matrix using diagonally weighted least squares (DWLS) as the estimation method and Promin as the rotation method (Lorenzo-Seva and Ferrando, 2019).

The number of extracted dimensions was determined through the optimal implementation of the parallel analysis (Timmerman and Lorenzo-Seva, 2011) with 500 replicates. The goodness-of-fit index (GFI) and the root mean square root of residuals (RMSR) were used as fit indices, establishing a good fit when the CFI > .95 and the RMSEA < .06 (Hu and Bentler, 1999).

Then, the second half of the sample (710 participants) was used to confirm the internal structure obtained in the exploratory approach. For this, a confirmatory factor analysis (CFA) was performed using DWLS, considering a good fit of the model when the GFI and CFI > .95 and the RMSEA and RMSR < .08 (Hu and Bentler, 1999).

Once the factor structure was clarified, the descriptive statistics (mean, standard deviation, skewness, and kurtosis) and the discrimination indices of the POCS items were examined. The reliability of each dimension was calculated with Cronbach's alpha and McDonald's Omega.



In addition, in light of the importance of studying the factor structure of the construct in different populations (Amérigo et al., 2020; Postigo et al., 2023), measurement invariance was assessed as a function of sex (male-female), type of organization (public-private), and age groups (young [18-30 years], adults [31-50 years], seniors [51-80 years]). The configural, metric, and scalar invariance levels were analyzed by multigroup confirmatory factor analysis (MG-CFA). Given that these are aggregate models, a change in the CFI of less than -.01 and a change in the RMSEA of less than -.015 ( $\Delta\text{CFI} < -.01$ ,  $\Delta\text{RMSEA} < .015$ ; Chen, 2007) makes it possible to accept the measurement invariance.

To analyze the differences in means according to sex and type of organization (public vs. private), the student's *t*-test was applied with Welch's correction, appropriate for unequal variances. In addition, Cohen's *d* was used as an effect size estimator, which makes it possible to interpret the magnitude of the differences observed between the groups. Subsequently, to determine the relationship between the POCS and other psychological variables, a Pearson correlation was calculated between the scale and the scores on climate, professional burnout, and psychosomatic symptomatology (Barria-González et al., 2021).

The analyses were performed with R version 4.3.2. (R Core Team, 2023) and the *haven*, *lavaan* (Rosseel, 2012), *psych* (Revelle, 2024), and *tidyverse* (Wickham et al., 2019) packages. For the EFA, Factor version 12.04.05 was used (Lorenzo-Seva y Ferrando, 2006). Supplementary Material can be accessed at <https://osf.io/wdv75/>

## Results

The parallel analysis with the initial scale (41 items) recommended extracting two factors on the scale (fit of the unidimensional model: CFI = .85, GFI = .89, RMSEA = 0.100, RMSR = 0.132; fit of the bidimensional model: CFI = .99, GFI = 1, RMSEA = 0.027, RMSR = 0.050). In this factor solution, one item from Rigor (8), two from Relationships (21 and 22), and two from Vanguard (25, 26) showed cross and low loadings in both dimensions. After their elimination, a new EFA fitted with the remaining 36 items. These data were adequate to perform a factor analysis (KMO = .92; Bartlett  $p < .001$ ), explaining 42% of the variance. The fit indices of the model were adequate (fit of the final solution: CFI = .99, GFI = 1, RMSEA = 0.025, RMSR = 0.046). The correlation matrix between the battery scores indicated that the two specific dimensions on the POCS are positively related to each other ( $p < .01$ ), with a correlation of .31.

Then, using the second subsample, the factor structure was confirmed by CFA, which showed a good fit to the data (CFI = .94, GFI = .96, RMSEA = 0.080, SRMR = 0.079). The factor loadings of the CFA are in Table 2.

The descriptive statistics (mean, standard deviation, skewness, and kurtosis), as well as the discrimination indices, are in Table 2. The items show adequate values of skewness and kurtosis, as well as adequate discriminative power ( $DI > .30$ ).

The reliability of the scores for each dimension was adequate in both, being  $\alpha = .90$   $\omega = .90$  for the people-oriented culture factor and  $\alpha = .88$ ,  $\omega = .88$  for the results-oriented culture factor.

Table 3 displays the findings concerning the measure's invariance. The measurement invariance of the POCS was confirmed at all levels (configural, metric, and scalar) for sex (male, female), type

**Table 2**  
Descriptive Items, Discrimination Indices, and Factor Loadings

Item	Mean	(SD)	Skew	Kurtosis	DI	$\lambda$	
						F1	F2
1	3.57	(1.21)	-0.64	-0.63	.38	.44	
2	3.74	(1.16)	-0.70	-0.41	.61	.67	
3	3.83	(1.06)	-0.91	0.23	.51	.59	
4	4.06	(0.95)	-1.09	0.98	.62	.70	
5	3.69	(1.04)	-0.70	-0.02	.53	.62	
6	3.62	(1.08)	-0.47	-0.47	.42	.47	
7	4.01	(0.99)	-1.15	1.04	.57	.63	
8*	3.76	(0.99)	-0.70	-0.01	-	-	-
9	3.91	(0.85)	-0.97	1.35	.53	.60	
10	3.85	(0.92)	-0.78	0.52	.52	.64	
11	4.01	(0.95)	-1.09	1.05	.57	.65	
12	3.79	(0.93)	-0.75	0.23	.47	.59	
13	4.07	(0.96)	-1.18	1.28	.61	.68	
14	3.93	(0.87)	-0.95	1.08	.53	.62	
15	4.18	(0.93)	-1.13	0.91	.54	.58	
16	3.91	(1.12)	-1.02	0.26	.46	.53	
17	4.14	(1.02)	-1.29	1.20	.57	.63	
18	3.80	(1.07)	-0.79	-0.04	.55	.66	
19	3.77	(1.04)	-0.71	-0.11	.53	.65	
20	3.88	(1.01)	-0.82	0.17	.46	.51	
21*	3.12	(1.22)	-0.19	-0.96	-	-	-
22*	3.23	(1.09)	-0.11	-0.82	-	-	-
23	3.78	(1.04)	-0.82	-0.01	.34	.41	
24	3.71	(1.06)	-0.71	-0.15	.53	.61	
25*	3.52	(1.09)	-0.51	-0.49	-	-	-
26*	3.41	(1.03)	-0.35	-0.50	-	-	-
27	3.15	(1.13)	-0.06	-1.04	.51		.63
28	2.85	(1.06)	0.26	-0.72	.64		.73
29	3.22	(1.09)	-0.06	-0.93	.55		.62
30	2.82	(1.14)	0.18	-0.85	.52		.56
31	3.39	(1.15)	-0.33	-0.96	.49		.54
32	3.01	(1.05)	0.04	-0.88	.52		.57
33	3.07	(1.04)	-0.01	-0.80	.56		.66
34	2.81	(1.03)	0.15	-0.44	.52		.56
35	2.88	(1.07)	0.18	-0.77	.53		.59
36	2.96	(1.18)	0.16	-0.98	.46		.50
37	3.05	(1.16)	0.01	-0.88	.54		.65
38	3.08	(0.95)	-0.14	-0.04	.42		.54
39	3.43	(1.01)	-0.38	-0.54	.48		.62
40	2.81	(1.09)	0.24	-0.62	.59		.65
41	2.90	(1.08)	0.14	-0.88	.62		.68

Note. SD = Standard Deviation, DI = Discrimination Index.  $\lambda$  = Factor Loadings (CFA), F1 = People-Oriented, F2 = Results-Oriented. Eliminated items are marked with an asterisk (\*).

**Table 3***Invariance of the Measure for POCS by Sex and Type of Organization*

	Sex				Public-Private			
	CFI	RMSEA	$\Delta$ CFI	$\Delta$ RMSEA	CFI	RMSEA	$\Delta$ CFI	$\Delta$ RMSEA
Configural	.936	0.085			.946	0.078		
Metric	.934	0.085	-.002	0	.941	0.081	-0.006	0.003
Scalar	.934	0.081	0	-0.004	.938	0.079	-0.002	-0.002

	Age Groups			
	CFI	RMSEA	$\Delta$ CFI	$\Delta$ RMSEA
Configural	.935	0.086		
Metric	.931	0.087	-.004	0.001
Scalar	.932	0.082	.001	-0.005

Note. CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation.

**Table 4***Differences in Means According to Sex and Type of Organization*

(1-2)	$\bar{X}_1$	$\bar{X}_2$	<i>t</i>	df	<i>p</i>	<i>d</i>
<b>Male/Female</b>						
People-Oriented	81.85	80.53	1.95	1311.1	.051	0.11
Results-Oriented	46.09	44.60	2.84	1368.5	.005	0.15
<b>Public-Private</b>						
People-Oriented	82.05	79.73	3.31	1129.2	< .001	0.19
Results-Oriented	42.95	50.75	-14.96	1073.7	< .001	0.86

Note.  $\bar{X}_1$  = Mean in Men,  $\bar{X}_2$  = Mean in women.

of organization (public, private), and age groups (young [18-30], adults [31-50], seniors [51-80]).

Subsequently, mean differences were analyzed according to sex and type of organization (Table 4). No statistically significant differences were found in the People-Oriented Factor according to sex. The other comparisons were statistically significant, although with small effect sizes, except for the Results-Oriented Factor in the comparison between the public and private sectors, which had a large effect size ( $d = .86$ ).

Finally, the relationships with other variables (organizational climate, professional burnout, and symptomatology) are shown in Table 5. The People-Oriented dimension shows stronger relationships with all the variables than the Results-Oriented dimension.

## Discussion

The assessment of organizational culture is of great relevance for workers' performance and health (e.g., Tadesse & Debela, 2024; Van Zyl et al., 2024). This study sought to examine the psychometric properties of the POCS in the Chilean context, supported by two key dimensions: People-Oriented and Results-Oriented Culture. The development of the POCS marks a breakthrough in the assessment of organizational culture in the Chilean setting.

The POCS is invariant as a function of sex and type of organization, showing that it maintains the same factor structure among different groups at the configural, metric, and scalar levels. This substantiates the need for equitable comparisons among various groups, with any observed discrepancies attributable to genuine disparities.

The People-Oriented dimension measures workers' perception of the organization's interest in their well-being, support, and

**Table 5***Pearson Correlations Between POCS and SWECS, PBS, and PSS*

Scales/Dimensions	People-Oriented	Results-Oriented
SWECS (Organizational Climate)		
Job Satisfaction	.38**	.10**
Organizational Trust	.32**	.02
Job Stress	.17**	-.18**
Social Support	.33**	.05*
Remuneration	.28**	.17**
PBS (Professional Burnout)		
Emotional Fatigue	-.25**	.11**
Personal fulfillment	.41**	-.01
Affective Hardening	-.20**	.27**
PSS (Psychosomatic Symptomatology)		
Somatic	-.19**	-.10**
Psychological	-.20**	-.11**

Note. \*\*  $p < .01$ . \*  $p < .05$ .

development, as reflected in its policies and actions. A high score would indicate that the organization promotes a positive work environment, emphasizing cohesion, satisfaction, and personal growth. A low score would reflect a perception of indifference to employees' well-being. The results show that this dimension is positively associated with job satisfaction, social support, and personal fulfillment and negatively related to emotional burnout and psychosomatic symptomatology. These findings align with Bakker and Demerouti's (2017) Job Demands and Resources theory, which posits that practices prioritizing well-being act as work resources that reduce stress and improve employees' mental health.

The Results-Oriented dimension, on the other hand, assesses the perception of the importance the organization lends to meeting objectives, efficiency, and competitiveness. A high score indicates that the organization is seen as goal-oriented, innovative, and efficient, whereas a low score suggests a lack of focus on productivity and results. The results indicate that this dimension correlates positively with aspects such as social support and pay while also being associated with higher levels of emotional fatigue, affective hardening, job stress, and psychosomatic symptomatology, suggesting that a strong focus on efficiency may result in heightened job demands if inadequately managed.

In the People-Oriented factor, the items “*A good worker adapts to new technologies*” and “*The basis of our success lies in order, planning, and innovation in technology*” suggest that technology and innovation are valued as tools for the development and adaptation of employees in an organized environment focused on well-being. This shows that, in this perspective, the Vanguard promotes the growth and adaptation of individuals within the organization. On the other hand, the items “*To be the best, you must always use the latest technology*” and “*We are the best because we are always the first to incorporate new technologies*” are associated with the Results-Oriented factor. In summary, these items associated with the theoretical dimension “Vanguard” not only align with the Adhocracy Culture of the CVF model but also reflect the organization’s ability to adapt to both the internal well-being of its employees and external market demands. The duality shown by this dimension (Vanguard), through its items, enables the organization to promote an innovative environment that, on the one hand, simultaneously drives personal growth and, on the other hand, competitive positioning, supporting [Cameron & Quinn \(2006\)](#) assertion that balanced organizational cultures are more effective and sustainable ([Shuaib & He, 2021](#); [Suifan, 2021](#)).

According to the definition operationalized by the authors, the Results-Oriented dimension is a valuable aspect for building a positive organizational culture as long as it is kept in balance with the People-Oriented approach. A robust results orientation, while traditionally linked to heightened competitiveness and pressure, can, when balanced appropriately, promote creativity, efficiency, and productivity, which are crucial components for sustainable organizational success. Recent studies, like those by [Bakker and Demerouti \(2018\)](#), suggest that combining job resources with challenging demands allows a results-oriented approach to drive performance and competitiveness without causing excessive professional burnout. Thus, a positive organizational culture can include a strong focus on results, providing it promotes a healthy and equitable environment that supports workers in achieving these goals ([Roll et al., 2019](#); [Schaufeli, 2017](#)). The POCS shows evidence of validity in relation to other variables such as organizational climate, professional burnout, and symptomatology. The connections are more robust within the People-Oriented Culture dimension, wherein an organization that prioritizes cultural care for individuals enhances the organizational climate and mitigates professional burnout and mental health symptomatology (e.g., [van Zyl et al., 2024](#)).

Although the POCS has a solid structure and has proven to be a tool that offers reliable scores with adequate evidence of validity, certain limitations should be considered. First, although representative of different sectors in Chile, the sample is designed specifically for Chilean organizational contexts. This highlights the need to validate the scale in various cultural and organizational settings to determine its factor equivalence and consistency in other national contexts. Another limitation is the cross-sectional design of the study. Although it identifies strong associations between the POCS dimensions and other organizational variables, it does not establish causal inferences. Future longitudinal studies are needed to assess the temporal stability of the measurements and to understand how the People-Oriented and Results-Oriented dimensions dynamically influence each other over time. Furthermore, incorporating evidence of validity of outcome variables would help determine the extent to which POCS assessments can anticipate key outcomes related to organizational performance and workplace

well-being, thus strengthening its practical and theoretical utility as a tool for organizational diagnosis and development in Chile. Thus, future studies should take into account important variables such as possible mental health problems and workers’ work experience.

Adopting the POCS can yield critical insights for formulating interventions to enhance the work environment in Chile. Organizations can use the results to identify areas for improvement and devise strategies that promote a positive culture, balancing a focus on results with the well-being of their employees. In this context, the People-Oriented dimension reflects values that promote social support, cohesion, and personal development, essential organizational resources for alleviating stress and enhancing job satisfaction. The Results-Oriented dimension is related to achieving goals and efficiency, which, when properly managed, drive productivity and work resilience, fostering optimal and sustainable performance over time. Thus, POCS is offered as a tool to assess organizational culture, enabling organizations to identify key areas for intervention and optimize outcomes related to well-being and productivity.

### Author Contributions

**Javier Barría-González:** Data Collection, Conceptualization, Methodology, Formal Analysis, Writing – Original Draft. **Jaime García-Fernández:** Data Curation, Methodology, Investigation, Software, Formal Analysis, Writing – Original Draft. **Ricardo Pérez-Luco:** Data Collection, Data Curation, Validation, Project Administration, Writing – Review & Editing. **Álvaro Postigo:** Supervision, Methodology, Visualization, Software, Writing – Review & Editing.

### Funding

This investigation has been funded (partially) by the Dirección de Investigación, Universidad de La Frontera (Chile).

### Declaration of Interests

The authors declare that there are no conflicts of interest.

### Data Availability Statement

The research data associated with this article is available upon reasonable request to the first author.

### References

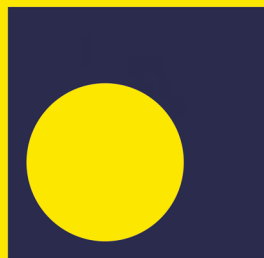
- Akpa, V. O., Asikhia, O. U., & Nneji, N. E. (2021). Organizational culture and organizational performance: A review of literature. *International Journal of Advances in Engineering and Management*, 3(1), 361-372. <https://doi.org/10.35629/5252-0301361372>
- Amérigo, M., García, J. A., Pérez-López, R., Cassullo, G., Ramos, A., Kalyan-Venumbaka, S., & Aragonés, J. I. (2020). Analysis of the structure and factorial invariance of the Multidimensional Environmental Concern Scale (MECS). *Psicothema*, 32(2), 275-283. <https://doi.org/10.7334/psicothema2019.281>
- Aryani, R., & Widodo, W. (2020). The determinant of organizational culture and its impact on organization: A conceptual framework. *International Journal of Higher Education*, 9(3), 64-70. <https://doi.org/10.5430/ijhe.v9n3p64>

- Bakker, A. B., & Demerouti, E. (2017). Job demands–resources theory: Taking stock and looking forward. *Journal of Occupational Health Psychology*, 22(3), 273-285. <https://doi.org/10.1037/ocp0000056>
- Bakker, A. B., & Demerouti, E. (2018). Multiple levels in job demands–resources theory: Implications for employee well-being and performance. In E. Diener, S. Oishi, & L. Tay (Eds.), *Handbook of well-being* (pp. 1–13). DEF Publishers. <https://doi.org/10.1016/j.coldregions.2015.12.009>
- Bal, P. M. (2017). *Dignity in the workplace*. Springer. <https://doi.org/10.1007/978-3-319-55245-3>
- Barría-González, J., Postigo, Á., Pérez-Luco, R., Cuesta, M., & García-Cueto, E. (2021). Evaluación de clima Organizacional: Propiedades psicométricas del ECALS [Assessing organizational climate: Psychometric properties of the ECALS scale]. *Anales de Psicología*, 37(1), 168-177. <https://doi.org/10.6018/analesps.417571>
- Barría-González, J., Postigo, Á., Pérez-Luco, R., Henríquez-Mesa, P., & García-Cueto, E. (2023). Co-Active coping inventory: Development and validation for the Chilean population. *The Spanish Journal of Psychology*, 26, Article e22. <https://doi.org/10.1017/SJP.2023.24>
- Beus, J. M., Solomon, S. J., Taylor, E. C., & Esken, C. A. (2020). Making sense of climate: A meta-analytic extension of the competing values framework. *Organizational Psychology Review*, 10(3-4), 136-168. <https://doi.org/10.1177/2041386620914707>
- Cameron, K. S., & Quinn, R. E. (2006). *Diagnosing and changing organizational culture* (3rd ed.). Jossey-Bass.
- Cameron, K. S., & Quinn, R. E. (2011). *The competing values culture assessment: A tool from the competing values product line*. University Michigan regence.
- Cameron, K. S., Quinn, R. E., DeGraff, J., & Thakor, A. V. (2006). *Competing values leadership: Creating value in organizations*. Edward Elgar Publishing.
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14(3), 464-504. <https://doi.org/10.1080/10705510701301834>
- Cooke, R. A., & Lafferty, J. C. (1987). *Organizational Culture Inventory (OCI)*. Human Synergetics.
- Denison, D. R. (1990). *Corporate culture and organizational effectiveness*. John Wiley.
- Di Stefano, G., & Gaudiino, M. (2019). Workaholism and work engagement: How are they similar? How are they different? A systematic review and meta-analysis. *European Journal of Work and Organizational Psychology*, 28(3), 329-347. <https://doi.org/10.1080/1359432X.2019.1590337>
- Donaldson, S. I., van Zyl, L. E., & Donaldson, S. I. (2022). PERMA+4: A Framework for work-related wellbeing, performance and positive organizational psychology 2.0. *Frontiers in Psychology*, 12, Article 817244. <https://doi.org/10.3389/fpsyg.2021.817244>
- Ehrhart, M. G., & Kuenzi, M. (2017). The impact of organizational climate and culture on employee turnover. In C. Goldstein, H. W. Pulakos, E. D. Passmore, & J. Semedo (Eds.), *The Wiley Blackwell handbook of the psychology of recruitment, selection, and employee retention* (pp. 494–512). John Wiley & Sons Ltd. <https://doi.org/10.1002/9781118972472.ch23>
- Fabrigar, L. R., Wegener, D. T., Maccallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3), 272-299. <https://doi.org/10.1037/1082-989X.4.3.272>
- Ferreira, M. C., Cristina, M., & Helena, C. (2002). Desenvolvimento de um instrumento brasileiro para avaliação da cultura organizacional [Development of a Brazilian instrument for evaluation of organizational culture]. *Estudos de Psicologia*, 7(2), 271-280. <https://doi.org/10.1590/S1413-294X2002000200008>
- Gelfand, M. J., Aycan, Z., Erez, M., & Leung, K. (2017). Cross-cultural industrial organizational psychology and organizational behavior: A hundred-year journey. *Journal of Applied Psychology*, 102(3), 514-529. <https://doi.org/10.1037/apl0000186>
- González-Romá, V., Tomás, I., Peiró, J. M., Lloret, S., Espejo, B., Ferreres, A., & Hernández, A. (1996). Análisis de las propiedades psicométricas del cuestionario de clima organizacional FOCUS-93 [Analysis of the Psychometric Properties of the FOCUS-93 Organizational Climate Questionnaire in a Multiprofessional Sample]. *Revista de Psicología Social Aplicada*, 6(1), 5-22. <https://www.copmadrid.org/webcopm/publicaciones/trabajo/1995/vol1/arti1.htm>
- Gregory, B. T., Harris, S. G., Armenakis, A. A., & Shook, C. L. (2009). Organizational culture and effectiveness: A study of values, attitudes, and organizational outcomes. *Journal of Business Research*, 62(7), 673-679. <https://doi.org/10.1016/j.jbusres.2008.05.021>
- Hartnell, C. A., Ou, A. Y., & Kinicki, A. (2011). Organizational culture and organizational effectiveness: A meta-analytic investigation of the competing values framework's theoretical suppositions. *Journal of Applied Psychology*, 96(4), 677-694. <https://doi.org/10.1037/a0021987>
- Hartnell, C. A., Ou, A. Y., Kinicki, A. J., Choi, D., & Karam, E. P. (2019). A meta-analytic test of organizational culture's association with elements of an organization's system and its relative predictive validity on organizational outcomes. *Journal of Applied Psychology*, 104(6), 832–850. <https://doi.org/10.1037/apl0000380>
- Hernández, M. A., Mendoza, J., & González, L. (2008). Construcción y validez del Instrumento de Cultura Organizacional y Competitividad (ICOC) [Construction and validity of the Organizational Culture and Competitiveness Instrument (ICOC)]. In *Estableciendo puentes en una economía global [Building Bridges in a Global Economy]* (vol. 2, p. 9). Escuela Superior de Gestión Comercial y Marketing (ESIC).
- Hofstede, G. (1991). *Cultures and organizations: Software of the mind*. McGraw-Hill.
- Hofstede, G. (2011). Dimensionalizing cultures: The hofstede model in context. *Online Readings in Psychology and Culture*, 2(1), 1-26. <https://doi.org/10.9707/2307-0919.1014>
- 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>
- Jacob, D. A., & Tende, F. B. (2022). Corporate culture, employee stress, and leadership support: In-route organizational psychotherapy. *Journal of Business Strategy, Finance and Management*, 4(1), 81-90. <https://doi.org/10.12944/jbsfm.04.01.07>
- Lorenzo-Seva, U. (2021). SOLOMON: A method for splitting a sample into equivalent subsamples in factor analysis. *Behavior Research Methods*, 54(6), 2665–2677. <https://doi.org/10.3758/s13428-021-01750-y>
- Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods*, 38(1), 88–91. <https://doi.org/10.3758/BF03192753>
- Lorenzo-Seva, U., & Ferrando, P. J. (2019). Robust Promin: Un método para la rotación de factores de diagonal ponderada [Robust Promin: A method for diagonally weighted factor rotation]. *Liberabit: Revista Peruana de Psicología*, 25(1), 99–106. <https://doi.org/10.24265/liberabit.2019.v25n1.08>
- Luthans, F., & Youssef-Morgan, C. M. (2016). Positive workplaces. In C. R. Snyder, S. J. Lopez, L. M. Edwards, & S. C. Marques (Eds.), *The Oxford*



- handbook of positive psychology* (3<sup>a</sup> ed., pp. 820–831). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199396511.013.47>
- Marcone, R., & Martín del Buey, F. (2003). Construcción y validación de un Inventario de Cultura Organizacional Educativa (ICOE) [Construction and validation of an Educational Organizational Culture Inventory (ICOE)]. *Psicothema*, 15(2), 292–299.
- O'Reilly, C. A., Chatman, J., & Caldwell, D. F. (1991). People and organizational culture: A profile comparison approach to assessing person-organization fit. *The Academy of Management Journal*, 34(3), 487–516. <https://doi.org/10.2307/256404>
- Olynick, J., & Li, H. Z. (2020). Organizational culture and its relationship with employee stress, enjoyment of work and productivity. *International Journal of Psychological Studies*, 12(2), 14. <https://doi.org/10.5539/ijps.v12n2p14>
- Ostroff, C., & Schulte, M. (2014). A configural approach to the study of organizational culture and climate. In B. Schneider & K. M. Barbera (Eds.), *The Oxford handbook of organizational climate and culture* (pp. 532–552). Oxford University Press.
- Parent, J. D., & Lovelace, K. J. (2018). Employee engagement, positive organizational culture and individual adaptability. *On the Horizon*, 26(3), 206–214. <https://doi.org/10.1108/OTH-01-2018-0003>
- Pérez-Luco, R. (2008). *Ambiente laboral subjetivo: Formulación empírica de un constructo* [Subjective work environment: Empirical formulation of a construct] [Doctoral thesis]. Universidad Pontificia de Salamanca. <https://summa.upsa.es/details.vm?q=id:0000030816&lang=es&view=main>
- Postigo, Á., García-Fernández, J., Cuesta, M., Recio, P., Barria-González, J., & Lozano, L. M. (2023). Giving meaning to the dark triad: Comparison of different factor structures of the Dirty Dozen through eight regions of the world. *Assessment*, 31(6), 1218–1232. <https://doi.org/10.1177/10731911231209282>
- Prieto-Díez, F., Postigo, Á., Cuesta, M., & Muñiz, J. (2022). Work engagement: Organizational attribute or personality trait? *Journal of Work and Organizational Psychology*, 38(2), 85–92. <https://doi.org/10.5093/jwop2022a7>
- R Core Team. (2023). R: A language and environment for statistical computing (Versión 4.3.2) [Software]. *Journal of statistical Software, R Foundation for Statistical Computing*. <https://www.R-project.org/>
- Rattrie, L. T. B., Kittler, M. G., & Paul, K. I. (2020). Culture, burnout, and engagement: A meta-analysis on national cultural values as moderators in JD-R theory. *Applied Psychology*, 69(1), 176–220. <https://doi.org/10.1111/apps.12209>
- Revelle, W. (2024). *Psych: Procedures for psychological, psychometric, and personality research*. Northwestern University.
- Rey-Sáez, R. (2022). Réplica a: validación cruzada sobre una misma muestra: Una práctica sin fundamento [Reply to: Cross-validation on the same sample: A practice without foundation]. *R.E.M.A. Revista electrónica de metodología aplicada*, 24(1), 41–44. <https://doi.org/10.17811/rema.24.1.2022.41-44>
- Robles, C., Montes, J., Rodríguez, A., & Ortega, A. O. (2018). Diseño y validación de un instrumento de cultura organizacional para empresas medianas [Validation and design of an organizational culture scale of measurement for medium enterprises]. *Nova Scientia*, 10(21), 552–575. <https://doi.org/10.21640/ns.v10i21.1453>
- Roll, L. C., Van Zyl, L. E., & Griep, Y. (2019). Brief positive psychological interventions within multi-cultural organizational contexts: A systematic literature review. In L. E. Van Zyl & S. Rothmann (Eds.), *Theoretical approaches to multi-cultural positive psychological interventions* (pp. 523–544). Springer International Publishing. <https://doi.org/10.1007/978-3-030-20583-6>
- Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Sarros, J. C., Cooper, B. K., & Santora, J. C. (2008). Through transformational leadership and organizational culture. *Journal of Leadership & Organizational Studies*, 15(2), 145–158. <https://doi.org/10.1177/1548051808324100>
- Schaufeli, W. B. (2017). Applying the job demands-resources model: A 'how to' guide to measuring and tackling work engagement and burnout. *Organizational Dynamics*, 46(2), 120–132. <https://doi.org/10.1016/j.orgdyn.2017.04.008>
- Schein, E. H. (2010). *Organizational culture and leadership* (4<sup>a</sup> ed.). John Wiley & Sons.
- Schein, E. H. (2015). Taking culture seriously in organization development. In W. J. Rothwell, J. M. Stavros, & R. L. Sullivan (Eds.), *Practicing organization development: Leading transformational change* (4<sup>a</sup> ed., pp. 233–244). John Wiley & Sons. <https://doi.org/10.1002/9781119176626.ch14>
- Schneider, B., González-Romá, V., Ostroff, C., & West, M. A. (2017). Organizational climate and culture: Reflections on the history of constructs in JAP. *American Psychological Association*, 102(3), 468–482. <https://doi.org/10.1037/apl0000090>
- Shuaib, K. M., & He, Z. (2021). Impact of organizational culture on quality management and innovation practices among manufacturing SMEs in Nigeria. *Quality Management Journal*, 28(2), 98–114. <https://doi.org/10.1080/10686967.2021.1886023>
- Suifan, T. (2021). How innovativeness mediates the effects of organizational culture and leadership on performance. *International Journal of Innovation Management*, 25(2), Article 2150016. <https://doi.org/10.1142/S136391962150016X>
- Tadesse, A., & Debela, K. L. (2024). Organizational culture: A systematic review. *Cogent Business and Management*, 11(1), Article 2340129. <https://doi.org/10.1080/23311975.2024.2340129>
- Taris, T. W. (2023). Workplace engagement and motivation. In A. Elliott (Ed.), *Advances in motivation science* (1<sup>a</sup> ed., Vol. 10, pp. 1–26). Academic Press.
- Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, 16(2), 209–220. <https://doi.org/10.1037/a0023353>
- van Muijen, J. J., Koopman, P., De Witte, K., De Cock, G., Sušan, Z., Lemoine, C., Bourantas, D., Papalexandris, N., Branyicski, I., Spaltro, E., Jesuino, J., Gonçalves, J., Pitariu, H., Konrad, E., Peiró, J. M., González-Romá, V., & Turnipseed, D. (1999). Organizational culture: The FOCUS questionnaire. *European Journal of Work and Organizational Psychology*, 8(4), 551–568. <https://doi.org/10.1080/135943299398168>
- van Zyl, L. E., Dik, B. J., Donaldson, S. I., Klibert, J. J., di Blasi, Z., van Wingerden, J., & Salanova, M. (2024). Positive organisational psychology 2.0: Embracing the technological revolution. *Journal of Positive Psychology*, 19(4), 699–711. <https://doi.org/10.1080/1743976.2023.2257640>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., ... Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), Article 1686. <https://doi.org/10.21105/joss.01686>





# Psicothema

Volume 37, no. 3

---

**PUBLISHED BY**



**MEMBER OF**



---

**SPONSORED BY**

