METHODOLOGICAL QUALITY OF INTERVENTIONS IN PSYCHOLOGY

EDITED BY: Salvador Chacón-Moscoso, Susana Sanduvete-Chaves and
Jason C. Immekus

PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.

By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

ISSN 1664-8714
ISBN 978-2-88945-249-1
DOI 10.3389/978-2-88945-249-1

METHODOLOGICAL QUALITY OF INTERVENTIONS IN PSYCHOLOGY

Topic Editors:
Salvador Chacón-Moscoso, Universidad de Sevilla, Spain and Universidad Autónoma de Chile, Chile
Susana Sanduvete-Chaves, Universidad de Sevilla, Spain
Jason C. Immekus, University of Louisville, United States

Glacier Perito Moreno (Argentina).
Photo by Salvador Chacón-Moscoso

Evaluations of intervention programs seek to present high-quality design, measures and data to assess their merit and worth. While evaluations differ in their purpose, theoretical framework and methodology, their collective aim is to obtain relevant and meaningful information to inform practice, research, and policy. As such, evaluation findings serve to build a body of knowledge on effective approaches to promote designated psychological outcomes, critical to an individual's overall health and well-being. However, as examined in this e-book, methodological weaknesses directly limit the potential of evaluations of intervention programs. As discussed by

Chacón-Moscoso and Sanduvete-Chaves, methodological weaknesses can be attributed to how to define and measure methodological quality and the contextual dependency of instruments designed to measure this quality.

In response, this e-book provides a collection of studies on methodological approaches to promote the quality of psychological interventions. Specifically, 10 original works published in the Research Topic Methodological Quality of Interventions in Psychology are included. The papers are organized into two chapters. Concretely, Chapter 1 includes studies pertaining to methodological approaches to enhance the quality of psychological intervention, being context independent solutions. Furthermore, Chapter 2 presents original work in different areas (health, education, sport and social welfare) where methodological quality has been better assessed. Collectively, the papers in this e-book serve to expand the awareness of practitioners and researchers interested in psychological interventions of the critical role of methodological quality in this work.

This research was funded by the projects 1150096 (Chilean National Fund of Scientific and Technological Development, FONDECYT); and PSI2015-71947-REDT (Spain's Ministry of Economy and Competitiveness).

Citation: Chacón-Moscoso, S., Sanduvete-Chaves, S., Immekus, J. C., eds. (2017). Methodological Quality of Interventions in Psychology. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-249-1

Table of Contents

Chapter 1: Methodological approaches to enhance the quality of psychological intervention

Chapter 2: Methodological quality in specific areas

Health

Education

Sport

Social welfare

frontiers in Psychology

EDITORIAL
published: 12 June 2017
doi: 10.3389/fpsyg.2017.00975

Editorial: Methodological Quality of Interventions in Psychology

Salvador Chacón-Moscoso [1, 2]* *and Susana Sanduvete-Chaves* [1]

[1] *HUM 649—Innovaciones Metodológicas en Evaluación de Programas, Psicología Experimental, Universidad de Sevilla, Sevilla, Spain,* [2] *Departamento de Psicología, Universidad Autónoma de Chile, Santiago, Chile*

Keywords: methodological quality, interventions, psychology, design, measurement, analysis

Editorial on the Research Topic

Methodological Quality of Interventions in Psychology

The need to evaluate intervention programs rigorously in different areas of psychology (e.g., health, education, sports, or social welfare) is widespread. However, we find clear methodological weaknesses in professional practice when it comes to evaluating intervention programs.

In many cases, fundamental details are not learned, such as how an intervention is framed, how it was implemented, what aspects of it are responsible for the effects, and how effective it is relative to other alternatives. Such absences hinder the replicability of interventions, learning what program aspects could be improved and how the knowledge from a single intervention can be integrated with other findings. All this prevents the growth of cumulative knowledge, the ability to use research to inform policy, and even the advancement of science.

According to previous research, much of this methodological weakness can be attributed to two factors: disagreement about how to conceptualize and measure methodological quality in evaluation, and the context dependency of existing instruments that claim to measure such quality.

The concept *quality* is complex and multidimensional. It has been defined from different theoretical perspectives that variously emphasize individual concepts or sets of concepts dealing with, for example, internal, external, and construct validity. This theoretical diversity leads to different approaches to measuring research quality, such as scales (tools where at least content, construct, and criterion validity evidence was tested), checklists (tools that have not been tested through an extensive validation process), and general recommendations (taking the form of advice).

The second methodological weakness stems from the context dependency of the instruments used that reduces the chance of the information they generate to be general. Indeed, many tools are used on just one occasion, and so dependable knowledge about its psychometric properties, including reliability and validity, are rarely available.

In this Research Topic, some works present methodological approaches to enhance the quality of psychological intervention, being context independent solutions. Thus, Chacón-Moscoso et al. (a) systematize and summarize the available literature about methodological quality in primary studies to describe the state of the art in assessing the methodological quality of interventions; (b) propose a specific, parsimonious, context independent, 12-items checklist to empirically define the methodological quality of primary studies based on a content validity study; and (c) present an inter-coder reliability study for the resulting 12 items.

Holgado-Tello et al. use Structural Equation Modeling (SEM) as a first approximation to operationalize the analytical implications of threats to validity in quasi-experimental designs. The study presents this empirical solution to the existing weak link between design features, measurement issues, and concrete impact estimation analyses. Finally, Manolov et al. make practitioners and applied researchers aware of the available appropriate options for extracting

OPEN ACCESS

Edited and reviewed by:
Pietro Cipresso,
IRCCS Istituto Auxologico Italiano, Italy

****Correspondence:***
Salvador Chacón-Moscoso
schacon@us.es

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *07 April 2017*
Accepted: *26 May 2017*
Published: *12 June 2017*

Citation:
Chacón-Moscoso S and Sanduvete-Chaves S (2017) Editorial: Methodological Quality of Interventions in Psychology. Front. Psychol. 8:975. doi: 10.3389/fpsyg.2017.00975

maximum information from the data. Concretely, they suggest that the evaluation of behavioral change should include visual and quantitative analyses, complementing the substantive criteria regarding the practical importance of the behavioral change.

In a complementary way, this Research Topic also presents original work in different areas where methodological quality has been better assessed in order to estimate unbiased effect sizes and study possible moderator variables influencing the results obtained.

In health area, Cano-García et al. evaluate formatively (before, during and after the intervention), a program of multicomponent psychological intervention for patients with chronic pain implemented: (a) based on techniques with empirical evidence, but developed in Spain; (b) at a public primary care center; (c) among patients with limited financial resources and lower education; (d) by a novice psychologist; and (e) taking measures of all domains of painful experience using the instruments recommended by the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT).

Additionally, Moreno et al. use the adversity level associated with family functioning and the positive adaptation level, as measures of a global health score, to distinguish four groups within adolescents: maladaptive, resilient, competent, and vulnerable. Such groups are compared in a number of demographic, school context, peer context, lifestyles, psychological, and socioeconomic variables, which can facilitate or inhibit positive adaptation in each context. In this way, they offer very valuable information for optimizing design and assessment of interventions and policies aimed at fostering adolescent health.

Furthermore, Vargas et al. use animal models of mental illness as a useful tool to characterize indicators of possible cognitive dysfunctions in humans. In this way, the subjectivity of the classical psychological evaluation processes where the patient must calibrate the magnitude of his/her symptoms and therefore the severity of his/her disorder, is overcome.

In education, Liu et al. extend the measurement part of latent transition analysis to the growth mixture model to examine the reading ability development of children. They found that the new model fitted the data well. Results also revealed that most of the children stayed in the same ability group with few cross-level changes in their classes. Finally, after adding the environmental factors as predictors, analyses showed that children receiving higher teachers' ratings, with higher socioeconomic status, and of above average poverty status, would have higher probability to transit into the higher ability group.

In sports area, Liu et al. examine relevant randomized controlled trials (RCTs) published in the past 20 years (1996–2015) for methodological concerns arise from Lord's paradox. Their analysis revealed that RCTs supporting the positive effect of exercise on cognition are likely to include Type I Error(s). This result can be attributed to the use of gain score analysis on pretest-posttest data as well as the presence of control group superiority over the exercise group on baseline cognitive measures. To improve accuracy of causal inferences in this area, analysis of covariance on pretest-posttest data is recommended under the assumption of group equivalence.

Finally, referring to social welfare, Izquierdo-Sotorrío et al. explore the informant effect and incremental validity to examine the relationships between perceived parental acceptance and children's behavioral problems (externalizing and internalizing) from a multi-informant perspective.

AUTHOR CONTRIBUTIONS

The two authors contributed to documenting, designing, drafting, and writing the manuscript, and revised it for important theoretical and intellectual content. Additionally, both authors provided final approval of the version to be published and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

FUNDING

This research was funded by the projects 1150096 (Chilean National Fund of Scientific and Technological Development, FONDECYT); and PSI2015-71947-REDT (Spain's Ministry of Economy and Competitiveness).

ACKNOWLEDGMENTS

The editors greatly appreciate the contributions received from the authors in this research topic.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 18 November 2016
doi: 10.3389/fpsyg.2016.01811

The Development of a Checklist to Enhance Methodological Quality in Intervention Programs

Salvador Chacón-Moscoso[1,2]*, *Susana Sanduvete-Chaves*[1] *and Milagrosa Sánchez-Martín*[3]

[1] HUM-649 Innovaciones Metodológicas en Evaluación de Programas, Departamento de Psicología Experimental, Facultad de Psicología, Universidad de Sevilla, Sevilla, Spain, [2] Universidad Autónoma de Chile, Santiago de Chile, Chile, [3] Department of Psychology, Universidad Loyola Andalucia, Sevilla, Spain

OPEN ACCESS

Edited by:
Pietro Cipresso,
IRCCS Istituto Auxologico Italiano, Italy

Reviewed by:
Michael Bosnjak,
Free University of Bozen-Bolzano, Italy
Vassilis Barkoukis,
Aristotle University of Thessaloniki, Greece

***Correspondence:**
Salvador Chacón-Moscoso
schacon@us.es

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *08 October 2015*
Accepted: *02 November 2016*
Published: *18 November 2016*

Citation:
Chacón-Moscoso S, Sanduvete-Chaves S and Sánchez-Martín M (2016) The Development of a Checklist to Enhance Methodological Quality in Intervention Programs. Front. Psychol. 7:1811. doi: 10.3389/fpsyg.2016.01811

The methodological quality of primary studies is an important issue when performing meta-analyses or systematic reviews. Nevertheless, there are no clear criteria for how methodological quality should be analyzed. Controversies emerge when considering the various theoretical and empirical definitions, especially in relation to three interrelated problems: the lack of representativeness, utility, and feasibility. In this article, we (a) systematize and summarize the available literature about methodological quality in primary studies; (b) propose a specific, parsimonious, 12-items checklist to empirically define the methodological quality of primary studies based on a content validity study; and (c) present an inter-coder reliability study for the resulting 12-items. This paper provides a precise and rigorous description of the development of this checklist, highlighting the clearly specified criteria for the inclusion of items and a substantial inter-coder agreement in the different items. Rather than simply proposing another checklist, however, it then argues that the list constitutes an assessment tool with respect to the representativeness, utility, and feasibility of the most frequent methodological quality items in the literature, one that provides practitioners and researchers with clear criteria for choosing items that may be adequate to their needs. We propose individual methodological features as indicators of quality, arguing that these need to be taken into account when designing, implementing, or evaluating an intervention program. This enhances methodological quality of intervention programs and fosters the cumulative knowledge based on meta-analyses of these interventions. Future development of the checklist is discussed.

Keywords: checklist, methodological quality, content validity, inter-coder reliability, primary studies

INTRODUCTION

Meta-analyses and systematic reviews aim to summarize the literature and generalize the results from a series of different studies about a given area of interest (Cheung, 2015). To avoid biased or erroneous conclusions, this requires clear criteria regarding the methodological quality of the primary studies and how to combine or analyze studies of different methodological quality (Jüni et al., 2001). Although, there is a general consensus about this need (Moher et al., 1996; Altman et al., 2001), a number of controversies arise when studying methodological quality in practice. For example, is it possible to give a one-dimensional answer to what is probably a multidimensional problem? Do we have clear criteria for deciding which specific and differently weighted methodological quality items should be considered? Which criteria should be used to decide between methodological quality indexes based on scores obtained from just one item or from a global assessment of several weighted items? Is it worthwhile trying to study a general construct that might not be equally applicable to all the contexts in which it might be used?

Despite this complexity, the extensive literature on these issues is testament to the importance of considering the methodological quality of primary studies. The present paper reviews the work in this area until July 2015. We begin by summarizing the relevant literature and then introduce the main problems derived from the state of the art.

Theoretical and Empirical Definition of *Methodological Quality*

The concept of *methodological quality* is complex and multidimensional. It has been defined theoretically from different perspectives, such as (a) internal validity (Moher et al., 1996); (b) external validity (Rubinstein et al., 2007); (c) both internal and external validity (Jüni et al., 2001); (d) internal, external, statistical, and construct validity (Valentine and Cooper, 2008); (e) precision of the study report (Moher et al., 1998; Altman et al., 2001; Efficace et al., 2006; Hopewell et al., 2006; Rutjes et al., 2006; Cornelius et al., 2009; Li et al., 2009); (f) appropriate statistical analysis (Minelli et al., 2007); (g) ethical implications (Jüni et al., 1999); (h) relevance for the intervention area (Sargeant et al., 2006; Jefferson et al., 2009; Jiménez-Requena et al., 2009); or (i) publication status (Moher et al., 2009).

This theoretical diversity of the concept of *methodological quality* leads to different approaches to measuring it empirically. The main approaches described in the literature are:

- Scales. These can be defined as validated tools used to measure the construct. At least the content, construct, and criterion validity evidence should be tested (Crocker and Algina, 1986; American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999; Abad et al., 2011). They are usually structured into different dimensions comprising differently weighted items (Sanderson et al., 2007). These items are either summed to obtain a global index (Jadad et al., 1996; Classen et al., 2008) or yield various indexes based on the dimensions considered (Jefferson et al., 2009).
- Checklists. The main difference between these tools and scales is that checklists have not been tested through an extensive validation process. Partial validity evidence may be presented, for example, based only on content or construct validity evidence. Checklists may also propose a final global index (Effective Public Health Practice Project, 1998; Efficace et al., 2003; Sanderson et al., 2007; Pluye et al., 2009); just one individual component (Gilbody et al., 2007); or several components (Bossuyt et al., 2003; Taji et al., 2006; Schulz et al., 2010).
- General recommendations. These take the form of advice, including general aspects to consider when assessing methodological quality. They may sometimes describe just a few examples of possible items, without specifying a whole list of proposed items. In sum, recommendations refer to those approaches that do not fulfill the criteria required by the previous two categories (Ford and Moayyedi, 2009; Linde, 2009; Wilson, 2009).

At this point, it is interesting to mention the difference between *quality in primary studies* and *quality of the report of primary studies* (Leonardi, 2006). It is very important to study the quality of the report of primary studies because the study of quality in primary studies is mostly based on reports given by authors. Indeed, this is usually the only source to obtain information about primary studies (Altman et al., 2001; Grimshaw et al., 2006; Cornelius et al., 2009). Nevertheless, we base our study on quality of primary studies (instead of the report) to (a) give researchers guidelines to check the methodological quality of studies included in a meta-analysis, to facilitate conclusions about possible risk of bias in the conclusions; (b) provide practitioners with a checklist to enhance methodological quality when designing, implementing, and evaluating their interventions; and (c) make explicit the criteria for why we included some concrete items and excluded others from an available extensive list. This information can be useful in case researchers or practitioners are interested in including different items from the extensive list based on their aims and specific contexts.

Problems Derived from the Dispersion in the Definition of *Methodological Quality*

The abovementioned characteristics of the concept of *methodological quality*, that is, the diversity in its theoretical and empirical definition (Linde, 2009), imply three interrelated and specific problems:

Lack of *representativeness* (R), the extent to which the specific item represents the methodological quality domain to which it is assigned. There are no clear criteria for choosing the optimal tool to measure methodological quality. This occurs especially since it is common to use non-randomized studies in social sciences (Shadish et al., 2005). This is due to a shortage of instruments that (a) are rigorously developed and (b) have reliability and/or validity evidence with tested R (Crowe and Sheppard, 2011). Their use is based on criteria that have no empirical support (Valentine and Cooper, 2008). For example, some authors opt to use individual components (Field et al., 2014; Eken, 2015). Other authors apply scales that provide a global value, even when they are strongly criticized for the lack of a bias estimation (Crowe and Sheppard, 2011). In spite of this, many scales are available and used nowadays (Dechartres et al., 2011). As a consequence, different scales applied to the same group of studies may indicate different levels of methodological quality (Greenland and O'Rourke, 2001; Jüni et al., 2001). Furthermore, some tools might be labeled as scales but without providing information about their construction process (Taji et al., 2006; Jefferson et al., 2009).

Lack of *utility* (U), the extent to which the specific item is useful for assessing the methodological quality of the study with respect to the assigned domain. In practice, scales usually include many items susceptible to omission because they are not relevant or essential for measuring the construct. Therefore, they could be shortened (Jüni et al., 2001; Conn and Rantz, 2003).

Lack of *feasibility* (F), the extent to which data codification is viable because data are available and can be gathered. Tools to measure methodological quality are usually complex and their items lack operational specificity. As a consequence, they are hard to understand and require previous training for coders. Additionally, the information needed is in most cases unavailable (Classen et al., 2008; Valentine and Cooper, 2008).

Objectives

To resolve the aforementioned problems when measuring methodological quality, the objectives of this paper are (a) to systematize and summarize the available literature about methodological quality in primary studies published until July 2015 (Study 1: systematic review); (b) to propose a specific, parsimonious checklist to empirically define the methodological quality of primary studies in meta-analyses and systematic reviews. This tool offers evidence of good R, U, and F based on expert judges (Study 2: content validity); and (c) to present evidence of adequate inter-coder reliability in the items that form the checklist (Study 3).

Contributions of this Study Compared to Other Studies Available in the Literature

The most popular tools to measure methodological quality present some of these problems. For example, the study Design and Implementation Assessment Device (DIAD) (Valentine and Cooper, 2008) was systematically developed. Nevertheless, it did not present reliability and validity evidence (weak R), and its application was complex (weak F).

Another example is the Cochrane Collaboration's tool for assessing risk of bias in randomized trials. It focuses on individual biases (Higgins et al., 2011). In this case, we did not find reliability and validity evidence (weak R). Furthermore, there was lack of U in social sciences because it is only applicable for randomized control trials (Shadish et al., 2005). Finally, at least two of the items (incomplete outcome data and selective reporting) are difficult to assess (weak F).

The Physiotherapy Evidence Database quality scale for randomized control trials —the PEDro scale— (Sherrington et al., 2000) presents reliability (Maher et al., 2003) and validity (Macedo et al., 2010) evidence (good in R). A website[1] offers access to the tool and a training program for raters (good in F). Nevertheless, it lacks U for our proposal because it is an adequate tool only for randomized control trials and only in the context of physiotherapy.

The checklist for the assessment of methodological quality presented by Downs and Black (1998) is good in U because it can be applied to randomized and non-randomized studies. Nevertheless, it partially presents weaknesses in R because, although it presents validity evidence, it attains poor reliability in a subscale and some specific items. Furthermore, practitioners who are not experts in methodology might experience some problems in its application (weak F).

The Newcastle–Ottawa Scale (NOS) for assessing the quality of non-randomized studies in a meta-analysis (Wells et al., 2009) presents good F: the tool and its manual are freely accessible through the Internet. Nevertheless, its R is medium because it presents intra-rater reliability and content and criterion validity but its construct validity has not been established yet. In addition, its U can be considered medium because it has been tested exclusively to be applied to non-randomized studies, but we do not know how it works for randomized studies.

There are quite well-developed tools that measure the quality of the report of primary studies, indicating the aspects to be made explicit when reporting a study, but without valuing the actions to improve the methodological quality of a study or intervention. Some of them are (Portell et al., 2015) (a) the Consolidated Standards of Reporting Trials (CONSORT) statement (Schulz et al., 2010) for randomized control trials; (b) the STrengthening the Reporting of OBservational Studies in Epidemiology (STROBE) statement (von Elm et al., 2007); (c) Guidelines for Reporting Momentary Studies (Stone and Shiffman, 2002) for intensive repeated measurements in naturalistic settings; (d) Guidelines for Qualitative Research Methodologies (Blignault and Ritchie, 2009); (e) Guidelines for Conducting and Reporting Mixed Research for Counselor Researchers (Leech and Onwuegbuzie, 2010); and (f) Guidelines for Reporting Evaluations Based on Observational Methodology (Portell et al., 2015) for low intervention designs. Our proposal is to measure the methodological quality of primary studies instead of the report of these studies. Consequently, our aim and the aim of the previously mentioned tools are clearly different. They both can be considered complementary because the methodological quality of a study cannot be valued when the aspects to evaluate are not reported.

Literature reviews about methodological quality have already been done (e.g., Donegan et al., 2010). Furthermore, tools to measure methodological quality with good results in inter-rater reliability and content validity already exist (e.g., Wells et al., 2009). This paper integrates both contributions: it updates the literature reviews until July 2015 exhaustively providing a list of the most frequent quality items; and based on the results, proposes a tool to enhance methodological quality with content validity (R, U, and F of items) and inter-rater reliability evidence.

In sum, our proposed 12-items checklist addresses the limitations that the other proposals present in total or partially. First, it presents R, U, and F evidence for each of its items based on a systematic literature review and content validity study. Second, appropriate results in reliability can be considered an additional evidence of R and F. In that case, we can describe our items as operationally specified, easy to be applied, and understandable. Third, additional U evidence of the tool is its applicability in different designs (randomized and non-randomized) and different contexts (it can be applied in the

[1] www.pedro.org.au

design, intervention, and/or evaluation of any program). Forth, additional F evidence is the transparency in procedure and results (presented objectively, thoughtfully, and in detail). We made explicit (a) the inclusion and exclusion criteria applied in each stage of the development of the tool; (b) the papers, tools, and items found in the literature; (c) the values obtained in the content validity study in R, U, and F for the most frequently used items to measure quality; and (d) the reliability coefficients. Finally, the proposed tool measures methodological quality instead of the quality of the report in methodological aspects.

STUDY 1. SYSTEMATIC REVIEW TO SEARCH FOR METHODOLOGICAL QUALITY INDICATORS

Method

Inclusion and Exclusion Criteria

We searched for papers published up to July 2015. Four inclusion criteria were applied: (a) methodological quality in primary studies was measured, (b) the full text was available, (c) it was written in English or Spanish, and (d) the instrument used to measure methodological quality was not previously included (was original, not repeated).

Information to Code

Tools to measure methodological quality in primary studies were identified. After that, they were assigned to the previously defined categories regarding the empirical definition of methodological quality: scales, checklists, and general recommendations.

Subsequently, the most frequently used items in the previously identified tools were compiled by two independent researchers. This item gathering was exhaustive but not necessarily mutually exclusive; that is, different items could refer to the same methodological quality content but define it with different degrees of detail/accuracy. Any redundancies in this regard would be removed in the content validity study (Study 2).

Finally, items were assigned to different dimensions and sub-dimensions based on a categorization of moderator variables in meta-analyses (Lipsey, 1994; Sánchez-Meca, 1997; Sánchez-Meca et al., 1998; Merrett et al., 2013): (a) substantive characteristics, pertinent to characterizing the phenomenon under study and referring to three aspects: subject characteristics (description of participants such as gender, age, or cultural status), the setting in which the intervention was implemented (e.g., geographical, cultural, temporal, or political context), and the nature of the intervention provided (e.g., modality, underlying theory, duration or number of sessions); (b) methodological or procedural aspects, referring to the manner in which the study was conducted (i.e., variations in the design, research procedures, quality of measures, and forms of data analysis); and (c) characteristics extrinsic to both the substantive phenomenon and the research methods. This includes characteristics of the researcher(s) (e.g., gender or affiliation), research circumstances (e.g., sponsorship), or reporting (e.g., form of publication or accuracy of the reporting). It has been reported that these variables are correlated with the magnitude of the effect in many meta-analyses (Lipsey, 1994).

Search Strategies

The search was carried out in 12 databases that were of interest due to their content. Specifically, these were Web of Science,

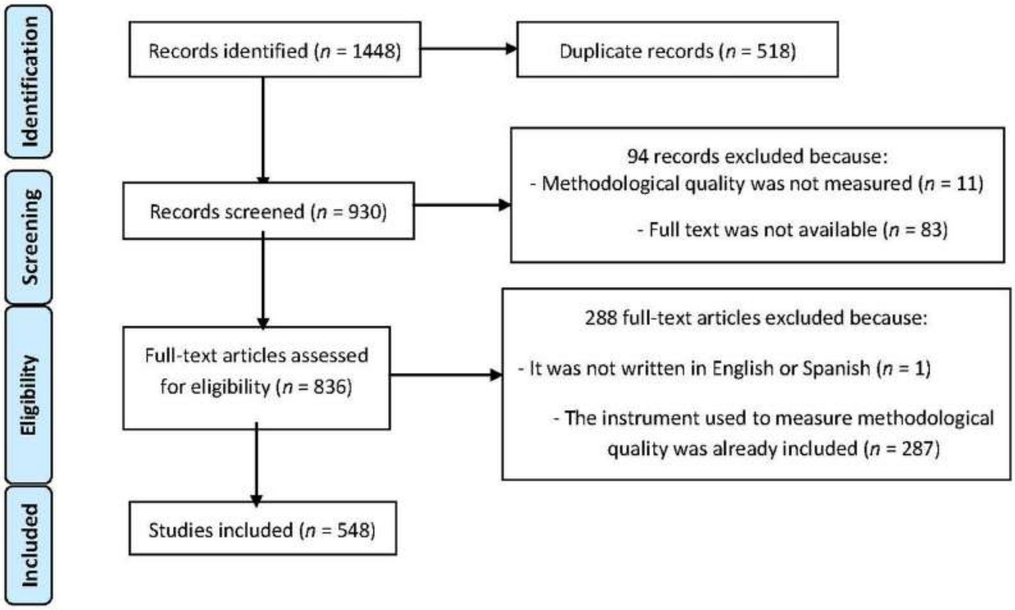

Identification
Screening
Eligibility
Included
Records identified (n = 1448)
Duplicate records (n = 518)
Records screened (n = 930)
94 records excluded because:
- Methodological quality was not measured (n = 11)
- Full text was not available (n = 83)
Full-text articles assessed for eligibility (n = 836)
288 full-text articles excluded because:
- It was not written in English or Spanish (n = 1)
- The instrument used to measure methodological quality was already included (n = 287)
Studies included (n = 548)

FIGURE 1 | Flow chart in the search for papers (Moher et al., 2009).

Scopus, Springer, EBSCO Online, Medline, CINAHL, Econlit, MathSci Net, Current Contents, Humanities Index, ERIC, and PsycINFO.

The keywords were "methodological quality" AND "meta-analysis" AND "primary studies." Title, abstract, keywords, and full text were examined. In addition, the reference lists of studies found were checked to identify other studies of interest. This procedure was repeated until no further relevant studies were discovered.

Coding Procedures

Inter-coder reliability (Nimon et al., 2012; Stolarova et al., 2014) was studied. The degree of agreement between two researchers (two of the authors, CM and SC) was calculated using Cohen's κ coefficient. Any disagreements were resolved by consensus.

Results

Figure 1 presents the flow chart based on the PRISMA statement (Moher et al., 2009). A total of 930 abstracts were initially screened. Considering full-text availability and exclusion criteria, the final sample comprised 548 full texts that referred to the measurement of methodological quality in primary studies, using different procedures (Supplementary Data 1). Four were scales, 425 checklists, and 119 sets of general recommendations (Supplementary Table S1). The inter-rater reliability gave a $\kappa = 0.874$ ($p < 0.001$), 95% CI [0.827, 0.921].

We gathered a list of the most frequent 43-items to measure methodological quality. Supplementary Tables S2 and S3 list these items, along with the corresponding original references from Supplementary Data 1. The inter-rater reliability coefficient was $\kappa = 0.924$ ($p < 0.001$), 95% CI [0.918, 0.93]. This was considered an adequate level of agreement between the two researchers.

Finally, the 43-items identified were assigned to the previously defined dimensions and sub-dimensions according to their content (see Supplementary Table S4). Specifically, six items were assigned to extrinsic characteristics, 14 to substantive characteristics (five referred to the sample, three to the setting, and six to the intervention), and 23 to methodological characteristics. The degree of consensus across items assigned to different dimensions yielded a good agreement with a $\kappa = 0.842$ ($p < 0.001$), 95% CI [0.695, 0.989].

STUDY 2. CONTENT VALIDITY STUDY

Method

Sample

Thirty judges participated in the content validity study. They were experts in design, systematic reviews, quality measurement, program evaluation, and/or applied psychology (social, educational, developmental, or clinical). They were all members of the Methods Group of the Campbell Collaboration and/or European Association of Methodology. Specifically, they consisted of 12 women and 18 men, 20 from Europe and 10 from the USA. Their mean age was 42 years. They had an average of 14 years of experience on these issues.

Instruments

The 43-items previously obtained and structured by the dimensions were presented as a questionnaire (see Supplementary Table S4). Experts had to score each item by taking into account the three previously mentioned problems: R, U, and F (Chacón-Moscoso et al., 2001; Martínez-Arias et al., 2006). This was done using a three-point rating scale (Osterlind, 1998): −1 was the lowest, 0 the medium, and +1 the highest score. The experts could also offer suggestions (such as including another item not currently considered, modifying or eliminating existing items, or changing the dimension to which an item was assigned).

Procedure

Tool distribution and gathering

The questionnaire was sent by e-mail to 52 experts. After the third request, a total of 30 questionnaires were completed and returned. Anonymity was assured in all cases.

Data analysis

The Osterlind index of congruence (1998) was used to quantify the consensus between experts in their judgments of each item and issue (Glück et al., 2015). The formula used was

$$I_{ik} = \frac{(N-1)\sum_{j-1}^{n} X_{ijk} + N\sum_{j=1}^{n} X_{ijk} - \sum_{j=1}^{n} X_{ijk}}{2(N-1)n}$$

where N = number of dimensions; X_{ijk} = score given by each expert to each item (between −1 and +1); and n = number of experts.

The results could range from −1 to +1. A score of −1 meant that all the experts awarded the most negative rating to the item in question. A score of +1 indicated that they all considered that the item in question merited the highest rating.

Inclusion criterion

Items that obtained a score of 0.5 or more on at least two of the three issues studied (R, U, and F) were included as important indicators to take into account when studying methodological quality in primary studies (Osterlind, 1998).

Results

Table 1 shows the Osterlind index obtained for each item on the three issues studied: R, U, and F. Fourteen methodological items fulfilled the inclusion criterion. A total of 18-items obtained scores equal to or higher than 0.5 on R, whereas 15-items obtained this score on U and 16 on F.

Item 22 was omitted because of its redundant content and suggestions by the experts (it shared redundant information with items 21 and 36). Furthermore, items 26 and 27 were combined into a single item. Consequently, the final proposed checklist contained 12-items focused on *methodological* characteristics. Definitions of items and their coding criteria can be found in the Appendix.

TABLE 1 | Osterlind indexes of representativeness (R), utility (U), and feasibility (F) obtained for the 43 items.

Extrinsic characteristics ($N = 30$)	**R**	**U**	**F**
(1) Type of publication	−0.2	0.4	**0.6**
(2) Year of publication	−0.4	−0.6	**0.6**
(3) Citation impact factor for the journal	−0.4	−0.2	0
(4) Raw data from the study available	−0.8	0	**0.8**
(5) Training of treatment implementers	0.4	**0.8**	0
(6) APA format	−0.2	−0.4	−0.2
Substantive characteristics ($N = 30$)			
Sample			
(7) Age (range)	0.4	0	0.4
(8). Age (mean)	**0.6**	0.467	0.4
(9) Age (standard deviation)	−0.2	−0.4	0
(10) Cultural origin	−0.2	0.2	0.2
(11) Socioeconomic level	−0.4	0	−0.2
Setting			
(12) Implementation context	−0.8	−0.2	0.4
(13) Intervention field	−0.2	−0.4	**0.8**
(14) Country in which study was conducted	0.2	0.4	**0.8**
Treatment			
(15) Theoretical orientation	0.2	−0.2	0.2
(16) Previous empirical evidence	0	−0.2	0.4
(17) Period of treatment	0.467	0.467	**1**
(18) Degree of treatment intensity	0.4	0.467	**1**
(19) Units	**0.737**	0.433	0.467
(20) Strengths and weaknesses of treatment are discussed	0.4	−0.2	0.4
Methodological characteristics ($N = 30$)			
(21) Inclusion and exclusion criteria for units provided	**0.6**	**0.8**	0.4
(22) Random assignment of units	**0.8**	**1**	**0.8**
(23) Methodology or design	**0.8**	**1**	**0.8**
(24) Sample size	0.367	0.467	**1**
(25) Analysis to calculate sample size	0.4	0.4	−0.4
(26) Attrition	**0.8**	**1**	0
(27) No attrition occurred	**0.6**	**0.6**	**0.6**
(28) Attrition between groups	**1**	**1**	**0.6**
(29) Exclusions after randomization	**0.8**	**1**	0.4
(30) Units studied before treatment implementation	0	0.4	0.2
(31) Follow-up period	**0.5**	**0.6**	0.2
(32) Occasions of measurement on each variable	**0.8**	**1**	**1**
(33) Measures in pre-test appear in post-test	**0.6**	**0.8**	0.4
(34) Standardized dependent variables	**0.5**	**0.8**	0.357
(35) Intervention context homogeneity	**0.6**	0.433	0.2
(36) Control techniques	**0.6**	**0.6**	−0.2
(37) Construct definition of outcome	**1**	0.**6**	−0.2
(38) Statistical methods for imputing missing data	**0.6**	**0.6**	0.4
(39) Specification of confidence intervals in statistical analysis	0.2	0.2	**0.6**
(40) Effect size value	0.2	0.4	**0.8**
(41) Effectiveness of treatment	0	0.4	**0.8**
(42) Interpretation of results	−0.2	−0.4	0.2
(43) Discussion of bias and limitations	**0.6**	0	0.4

Items appear in abbreviated form; the whole version can be consulted in Supplemental Material 4. Scores of 0.5 or higher are printed in bold.

STUDY 3. INTER-CODER RELIABILITY STUDY

Method

Sample

Four coders participated in the study. Two of them (C1 and C2) were coauthors of this study (SC and SM) and two others (C3 and C4) were not. Each coder had a high level of understanding of written English and received prior training on the coding task by an expert in the topic, also a coauthor of this article (CM).

Instruments

The 12-items checklist resulting from the previous Studies 1 and 2 was applied. The Appendix presents the final version of the coding scheme after including the changes derived from the pilot study described in this Study 3.

Papers were found by searching 11 computerized databases to locate training programs: EBSCO Online, Medline, Serfile, CABHealth, CINAHL, PsycINFO, Econlit, ERIC, MathSci, Current Contents, and Humanities Index. Finally, we used SPSS 17.0 to calculate Cohen's κ coefficient.

Procedure

First, we conducted a bibliographic search to collect articles published in the training program field. The issue was chosen by research interest. The keywords used were "evaluation," "training programs," and "work." From the resulting 1,399 published journal articles, we obtained 124 after discarding (a) the duplicates ($n = 223$); (b) those that were not written in English or Spanish ($n = 46$); (c) those for which the complete text was not available ($n = 421$); or (d) where the training program was not aimed at employees to improve their professional skills ($n = 585$). Twenty-five studies (20% of the total) were randomly selected to be used in the pilot study.

C1 and C2 were trained under the supervision of one of the authors of this article (CM), an expert on the topic. The three researchers revised the coding scheme to be sure that they understood each item in the same way (Bennett et al., 1991). CM solved the questions that C1 and C2 asked. Later, as a test, C1 and C2 jointly coded one study that was not included in this research. This task was useful to clarify some discrepancies between the coders about the items and their meaning and the way to locate the information in the papers. Then, independently, they applied the checklist to the 25 studies selected. Each study was coded in an average of 15 min.

To analyze the degree of agreement on each item, Cohen's κ (Cohen, 1960; Bechger et al., 2003; Engelhard, 2006; Nimon et al., 2012) was used for categorical items. For quantitative items (items 3–6), a correlation coefficient was calculated. When assumptions were accepted (normality Kolmogorov–Smirnov z with $p > 0.05$ and independence of errors Durbin–Watson d between 1.5 and 2.5), the Pearson correlation coefficient (r) was calculated; when at least one of the assumptions was violated, the Spearman correlation coefficient (ρ) was calculated.

This reliability study was replicated twice: (a) C1 and C2 applied the scale to 20 new studies (20% of the total, randomly chosen after excluding the 25 papers previously analyzed). After analyzing the results, the wording of some definitions and alternatives of the items that might have caused coding discrepancies were modified to achieve greater clarity and simplicity in the instrument; (b) C3 and C4 applied the scale to the same 20 studies. C3 and C4 received information about the research, its main characteristics, the topic it covered, the task to do, and guidelines to codify the studies. In both replications, reliability was analyzed using the same coefficients that were used in the pilot study. In addition, the reliability among the four coders in the replication phase was analyzed. For that, we calculated Cohen's κ for categorical items and Krippendorff's α coefficient for quantitative items 3–6 (Hayes and Krippendorff, 2007).

Results

Testing Assumptions for Quantitative Items 3–6

Table 2 presents the results obtained on the normality (Kolmogorov–Smirnov z) and independence of errors (Durbin–Watson d) assumptions for the quantitative items 3–6.

Normality and independence of errors assumptions were accepted for item 4 in the pilot study and items 3 and 4 in the replication carried out by C3 and C4. In these cases, Pearson's r coefficient was calculated as inter-coder agreement value. For the rest of the situations (when at least one assumption was violated), Spearman's ρ coefficient was obtained.

Inter-coder Reliability

Table 3 shows the results obtained for each item individually. In the pilot study, we obtained a significant agreement value for seven items; only items 4 and 10 obtained an agreement value higher than 0.7; and, in general, the 95% CI amplitudes were wide, ranging from 0.376 in item 4, [0.994, −0.618] to 1.422 in item 5, [0.551, −0.871].

In the replication of the reliability study carried out by C1 and C2, we obtained a significant κ value for nine items. Four of them obtained an agreement value higher than 0.8, seven of them an agreement value higher than 0.7. The highest agreement value was 1 for item 5, *Exclusions after randomization*. The lowest agreement value was 0.5 for item 12, *Statistical methods for imputing missing data*. Compared to the results in the pilot study, the level of agreement improved substantially for most of the items except for items 4, 9, 10, and 12, where it fell slightly; 95% CIs were, in general, narrower than in the pilot study but still wide, ranging in amplitude from 0.045 (item 6, [0.994, −0.949]) to 1.168 (items 2 and 11, both [1.445, −0.277]).

In the second reliability study replication, performed by C3 and C4, the agreement value was significant for all the items. Ten items obtained an agreement value higher than 0.8. The lowest value was equal to 0.744, obtained for item 10 (*Control techniques*). Five items obtained the highest agreement value (1). Compared to the results in the replication study carried out by C1 and C2, the level of agreement was higher for C3 and C4 in all the items except for item 11, where it fell slightly, although it maintained significance and had an agreement value close to 0.8. 95% CIs were in general narrower than in the pilot study but still wide in some occasions, ranging in amplitude from 0 (items 2, 7, 8, and 12, in all cases [1-1]) to 0.998 (item 11, [1.269, −0.271]).

TABLE 2 | Testing assumptions for quantitative items.

	Pilot study			Replication					
Item	**C1 *z***	**C2 *z***	***d***	**C1 *z***	**C2 *z***	***d***	**C3 *z***	**C4 *z***	***d***
(3) Attrition	0.449	**1.696****	1.587	0.767	0.683	**1.289**	0.683	0.757	1.633
(4) Attrition between	0.77	0.873	2.31	0.667	0.536	**2.799**	0.49	0.595	2.244
(5) Exclusions after	1.335	0.57	**0.692**	0.451	0.513	**2.974**	0.38	0.506	**2.974**
(6) Follow-up	**1.661****	**1.919****	1.768	**1.639****	**1.478***	2.276	**1.532***	**1.478***	1.742

Items appear in abbreviated form; the whole and final version (after including improvements derived from the pilot study) can be consulted in the Appendix. C1–C4 = coder 1–4, respectively. z = Kolmogorov–Smirnov z to study normality assumption (accepted when p > 0.05). d = Durbin–Watson d to study independence of errors assumption (accepted when 1.5 < d < 2.5). Results that imply an assumption violation are in boldface.
$^{*}p < 0.05$; $^{**}p < 0.01$.

The results obtained in reliability across the four coders were positive, with significant values in all the items, ranging in agreement values between 0.73 and 0.931; whereas some 95% CIs remained too wide, ranging in amplitude from 0.248 (item 8, [0.854, −0.606]) to 1.15 (item 10, [1.342, −0.192]).

DISCUSSION

In this paper, we propose a simple 12-items checklist that, when used, can contribute to enhance the methodological quality of interventions. This checklist is formed by individual methodological features that serve as indicators of quality to be taken into account when designing, implementing, or evaluating an intervention. Thus, its use does not imply obtaining a single methodological quality measure by summing the evaluation of several indicators, which is a highly criticized approach due to the inconsistent results when measuring the same studies with different methodological quality scales (Greenland and O'Rourke, 2001).

It must be asked what this checklist adds to the state of the art. Why and how is our measurement tool any different from other proposed measures that are routinely used? The first advantage is its clear, careful, and explicit process of development. First, we made an extensively updated review of all available papers referring to the measurement of methodological quality in primary studies. Second, we carried out a content validity study through expert judges. Thus, we obtained results about the congruence between checklist items with respect to their R, U, and F in relation to the dimensions they were assigned to Osterlind (1998). Third, we carried out an inter-coder reliability pilot study and multiple replication studies. As a result, we obtained appropriate coefficients in all the items, comparing the degree of agreement in pairs and with four coders joined.

In this sense, lack of R can be considered solved. In contrast to existing publications, we have clarified to the reader how and why the checklist was developed, setting up the criteria for the inclusion of items. In this regard, the appraisal made by each item on the complete checklist can be consulted with respect to its R, U, and F; as well as in relation to the categorization of the moderator variables (i.e., substantive —about subjects, setting, and intervention—, methodological and extrinsic characteristics) usually used in a meta-analysis (Lipsey, 1994). The following information has also been made available as supplementary material: the complete list of 548 reviewed papers referring to the measurement of methodological quality in primary studies and published until July 2015 (Supplementary Data 1); the list of references classified according to different and specific approaches to the empirical definition of methodological quality (Supplementary Table S1); the 43-items chosen and the original references in which they were found (Supplementary Tables S2 and S3); and the content validity questionnaire given to experts (Supplementary Table S4).

Referring to the lack of U, some issues have been solved. The proposed 12-items checklist can be useful, not just for improving the reporting of studies. First, it can assess the methodological quality of studies that have already been carried out. It gives researchers guidelines regarding inclusion–exclusion criteria in a systematic review or meta-analysis. It also checks the methodological quality of included studies to facilitate conclusions about possible risk of bias in the conclusions. Additionally, the checklist items can be used as potential moderator variables in a meta-analysis (Conn and Rantz, 2003). Second, the checklist can enhance the methodological quality in ongoing interventions that are being planned, designed, or implemented. It is extensively useful because it can be applied to experimental and non-experimental studies (interventions with random assignment of participants to the different groups or without random assignment). This is a critical issue for practitioners and in practical systematic reviews and meta-analyses because the latter type of design is frequently used in the social sciences (Shadish et al., 2005; Mayer et al., 2014).

One advantage of focusing on methodological characteristics is that it enables the tool to be extrapolated and generalized to different areas of intervention rather than being linked to one specific context. It is therefore interesting to use a common methodological framework through which one can obtain and analyze differences and communalities both within and between different intervention contexts. Logically, conclusions obtained with the same checklist would be modulated, depending on the area of intervention.

In a parallel way, we made explicit the criteria by which we included some concrete items and excluded others. Thus, we provided practitioners and researchers with clear criteria for choosing items that may be adequate to their needs. As a consequence, some of the 43-items categorized in the extrinsic,

TABLE 3 | Results of inter-coder reliability.

	Pilot study				Replication			
	C1–C2		C1–C2		C3–C4		4C	
Items	**Agreement**	**95% CI**	**Agreement**	**95% CI**	**Agreement**	**95% CI**	**Agreement**	**95% CI**
(1) Inclusion/exclusion criteria	[a]0.684**	[0.292, 1]	[a]0.798**	[0.533, 1]	[a]0.9**	[0.71, 1]	[a]0.851**	[0.707, 0.995]
(2) Methodology/design	[a]0.252	[−0.062, 0.566]	[a]0.861**	[0.277, 1]	[a]1**	[1, 1]	[a]0.931**	[0.639, 1]
(3) Attrition	[b]0.505	[0.078, 0.898]	[b]0.653*	[0.463, 0.962]	[c]0.943**	[0.772, 0.986]	[d]0.79**	[0.617, 0.963]
(4) Attrition between groups	[c]0.952**	[0.618, 0.994]	[b]0.866	[0.326, 1]	[c]0.991**	[0.629, 1]	[d]0.849**	[0.478, 1]
(5) Exclusions after	[b]−0.206	[−0.871, 0.551]	[b]1**	[0.137, 0.998]	[b]1**	[0.476, 1]	[d]0.775**	[0.306, 1]
(6) Follow-up	[b]0.522	[−0.133, 0.802]	[b]0.783**	[0.949, 0.994]	[b]0.963**	[0.976, 0.997]	[d]0.76**	[0.5, 1]
(7) Occasions of measurement	[a]0.486*	[0.131, 0.841]	[a]0.653**	[0.32, 0.986]	[a]1**	[1, 1]	[a]0.827**	[0.66, 0.994]
(8) Pre/post measures	[a]0.592**	[0.173, 1]	[a]0.714*	[0.212, 1]	[a]1**	[1, 1]	[a]0.73**	[0.606, 0.854]
(9) Dependent variables	[a]0.577**	[0.25, 0.904]	[a]0.512**	[0.038, 0.986]	[a]0.857**	[0.588, 1]	[a]0.745**	[0.313, 1]
(10) Control techniques	[a]0.706**	[0.323, 1]	[a]0.667	[0.104, 1]	[a]0.744*	[0.281, 1]	[a]0.767**	[0.192, 1]
(11) Construct definition	[a]0.047	[−0.18, 0.274]	[a]0.861**	[0.277, 1]	[a]0.77**	[0.271, 1]	[a]0.772**	[0.438, 1]
(12) Imputing missing data	[a]0.571*	[0.081, 1]	[a]0.5	[0.014, 0.986]	[a]1**	[1, 1]	[a]0.841**	[0.581, 1]

Items appear in abbreviated form; the whole and final version (after including improvements derived from the pilot study) can be found in the Appendix. C1–C2 = reliability results between coders 1 and 2; C3–C4 = reliability results between coders 3 and 4; 4C = reliability results across the four coders combined.

[a]*Cohen's κ coefficient;* [b]*Spearman's ρ coefficient;* [c]*Pearson's r coefficient;* [d]*Krippendorff's α coefficient.*

$*p < 0.05$; $**p < 0.01$.

substantive, and methodological characteristics (available in Supplementary Table S4), which were obtained from the search described in Study 1, can be selected in case researchers and practitioners are interested in including different characteristics based on their aims and specific contexts.

Referring to the lack of F, we also made advances due to the acceptable results yielded in the inter-coder reliability study (Study 3), that is, few discrepancies when different professionals coded the same studies, and because the average time needed to apply the checklist was 15 min per primary study. These facts can be interpreted in that the checklist is relatively easy to apply by having the definitions of the 12-items and their coding criteria for the final proposed checklist (Appendix).

Although this is not particularly relevant for reliability studies, the performance in Study 3 in only one intervention area is another possible limitation. Nevertheless, we are certain that the results can be generalized to other areas. We applied previous versions of the final proposed checklist in a number of pilot studies, systematic reviews, and meta-analyses. The topic was varied: psychological interventions in general, for elderly people, and for children with attention deficit hyper-activity disorder (e.g., see Supplementary Table S5). In all these cases, results obtained in inter-coder reliability were adequate.

Some of the research is ongoing or being planned. We will carry out another inter-coder reliability study enlarging the sample size to improve the accuracy of the results found in Study 3. Furthermore, we will conduct pilot studies to analyze the psychometric properties of the 12 previously obtained items. Thus, for example, we will calculate their capacity for discrimination by using the mean discrimination index and item reliability according to classical test theory (Holgado-Tello et al., 2006). Finally, the inter-coder reliability obtained was adequate but could be improved. This is why we will constantly review the definition of the 12-items of the checklist based on comments obtained from different professionals who use this tool.

CONCLUSION

There is no single approach for the issue of methodological quality, and this paper was not intended to give a definitive answer. However, we do offer a justified response to the question. For that, we summarized our continuous and collaborative research over the past 15 years, which began with our first pilot applications in Baltimore in 2002 (Methods Campbell Collaboration Meeting). Furthermore, we do not merely argue the case for our own 12-items approach but also encourage other possible answers by researchers and practitioners, based on the R, U, and F assessment of the 43 most used methodological quality items in a meta-analysis.

In sum, this paper describes the rigorous process of methodological quality index selection for meta-analyses and systematic reviews and for designing, implementing, and evaluating interventions. To achieve this, we carry out an updated review on an ongoing basis. Instead of partial reviews, with poorly specified criteria for the inclusion of items, we present a checklist that has been and is being reviewed periodically. This checklist is based on the literature, experts' opinion, applications, and feedback from related professional meetings, mainly from the *Campbell Collaboration* group (C2), the *Society for Research Synthesis Methodology* (SRSM), the *European Association of Methodology* (EAM) and the *Spanish Association of Methodology in Behavioral Sciences* (AEMCCO). The most recent comments on this work were received from the last editions of some of these meetings: the VI European Congress of Methodology in Utrecht, Netherlands (July 2014), and the XIV Congress of Methodology in Health and Social Sciences in Palma de Mallorca, Spain (July 2015).

Finally, we would like to invite any interested readers who design, implement, and/or evaluate interventions to collaborate with this project, so that we can share comments or results regarding the application of the proposed checklist. We also invite collaborations from those who are able and willing to assess the methodological quality of primary studies in meta-analyses and systematic reviews.

AUTHOR CONTRIBUTIONS

SC-M developed the initial idea and design of the work and performed the analysis. SS-C, and MS-M performed the analyses and interpreted the data. SC-M and SS-C were in charge of drafting the manuscript. MS-M revised the manuscript critically for important intellectual content. All three authors (SC-M, SS-C, and MS-M) provided final approval of the version to be published and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work were appropriately investigated and resolved.

FUNDING

This research was funded by the Projects with Reference PSI2011-29587 (Spanish Ministry of Science and Innovation); 1150096 (Chilean National Fund of Scientific and Technological Development, FONDECYT); and PSI2015-71947-REDT (Spanish Ministry of Economy and Competitiveness).

ACKNOWLEDGMENTS

The authors greatly appreciate all the comments received from William R. Shadish (recently passed away), the reviewers, and the English language editor. We believe that the quality of this paper has been substantially enhanced as a result.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.01811/full#supplementary-material

REFERENCES

Abad, F. J., Olea, J., Ponsoda, V., and García, C. (2011). *Medición en Ciencias Sociales y de la Salud [Measurement in Health and Social Sciences]*. Madrid: Síntesis.

Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., et al. (2001). The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann. Intern. Med.* 134, 663–694. doi: 10.7326/0003-4819-134-8-200104170-00012

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Bechger, T. M., Maris, G., Verstralen, H. H. F. M., and Béguin, A. A. (2003). Using classical test theory in combination with Item Response Theory. *Appl. Psychol. Meas.* 27, 319–334. doi: 10.1177/0146621603257518

Bennett, R. E., Sebrechts, M. M., and Rock, D. A. (1991). Expert-system scores for complex constructed-response quantitative items: a study of convergent validity. *Appl. Psychol. Meas.* 15, 227–239. doi: 10.1177/014662169101500302

Blignault, I., and Ritchie, J. (2009). Revealing the wood and the trees: reporting qualitative research. *Health Promot. J. Austr.* 20, 140–145.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., et al. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Radiology* 226, 24–28. doi: 10.1148/radiol.2261021292

Chacón-Moscoso, S., Pérez-Gil, J. A., Holgado-Tello, F. P., and Lara, A. (2001). Evaluation of quality in higher education: content validity. *Psicothema* 13, 294–301.

Cheung, M. W. L. (2015). MetaSEM: an R package for meta-analysis using structural equation modeling. *Front. Psychol.* 5:1521. doi: 10.3389/fpsyg.2014.01521

Classen, S., Winter, S., Awadzi, K. D., Garvan, C. W., Lopez, E. D. S., and Sundaram, S. (2008). Psychometric testing of SPIDER: data capture tool for systematic literature reviews. *Am. J. Occup. Ther.* 62, 335–348. doi: 10.5014/ajot.62.3.335

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46. doi: 10.1177/001316446002000104

Conn, V. S., and Rantz, M. J. (2003). Research methods: managing primary study quality in meta-analyses. *Res. Nurs. Health* 26, 322–333. doi: 10.1002/nur.10092

Cornelius, V. R., Perrio, M. J., Shakir, S. A. W., and Smith, L. A. (2009). Systematic reviews of adverse effects of drug interventions: a survey of their conduct and reporting quality. *Pharmacoepidemiol. Drug Saf.* 18, 1223–1231. doi: 10.1002/pds.1844

Crocker, L., and Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York, NY: Holt, Rinehart and Winston.

Crowe, M., and Sheppard, L. (2011). A review of critical appraisal tools shows they lack rigor: alternative tool structure is proposed. *J. Clin. Epidemiol.* 64, 79–89. doi: 10.1016/j.jclinepi.2010.02.008

Dechartres, A. C. P., Hopewell, S., Ravaud, P., and Altman, D. G. (2011). Reviews assessing the quality or the reporting of randomized controlled trials are increasing over time but raised questions about how quality is assessed. *J. Clin. Epidemiol.* 64, 136–144. doi: 10.1016/j.jclinepi.2010.04.015

Donegan, S., Williamson, P., Gamble, C., and Tudur-Smith, C. (2010). Indirect comparisons: a review of reporting and methodological quality. *PLoS ONE* 5:e11054. doi: 10.1371/journal.pone.0011054

Downs, S. H., and Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J. Epidemiol. Commun. Health* 52, 377–384. doi: 10.1136/jech.52.6.377

Effective Public Health Practice Project (1998). *Quality Assessment Tool for Quantitative Studies*. Available at: http://www.ephpp.ca/tools.html

Efficace, F., Bottomley, A., Osoba, D., Gotay, C., Flechtner, H., D'haese, S., et al. (2003). Beyond the development of health-related quality-of-life (HRQOL) measures: a checklist for evaluating HRQOL outcomes in cancer clinical trials–does HRQOL evaluation in prostate cancer research inform clinical decision making? *J. Clin. Oncol.* 21, 3502–3511. doi: 10.1200/JCO.2003.12.121

Efficace, F., Horneber, M., Lejeune, S., Van Dam, F., Leering, S., Rottmann, M., et al. (2006). Methodological quality of patient-reported outcome research was low in complementary and alternative medicine in oncology. *J. Clin. Epidemiol.* 59, 1257–1265. doi: 10.1016/j.jclinepi.2006.03.006

Eken, C. (2015). Critical reappraisal of intravenous metoclopramide in migraine attack: a systematic review and meta-analysis. *Am. J. Emerg. Med.* 33, 331–337. doi: 10.1016/j.ajem.2014.11.013

Engelhard, G. (2006). Book review: analyzing rater agreement: manifest variable methods. *Appl. Psychol. Meas.* 30, 154–156. doi: 10.1177/0146621605277030

Field, N., Cohen, T., Struelens, M. J., Palm, D., Cookson, B., Glynn, J. R., et al. (2014). Strengthening the reporting of molecular epidemiology for infectious diseases (STROME-ID): an extension of the STROBE statement. *Lancet Infect. Dis.* 14, 341–352. doi: 10.1016/S1473-3099(13)70324-4

Ford, A. C., and Moayyedi, P. (2009). Redundant data in the meta-analysis on *Helicobacter pylori* eradication. *Ann. Intern. Med.* 151, 513–514. doi: 10.7326/0003-4819-151-7-200910060-00015

Gilbody, S., Richards, D., Brealey, S., and Hewitt, C. (2007). Screening for depression in medical settings with the patient health questionnaire (PHQ): a diagnostic meta-analysis. *J. Gen. Intern. Med.* 22, 1596–1602. doi: 10.1007/s11606-007-0333-y

Glück, J., König, S., Naschenweng, K., Redzanowski, U., Dorner, L., Straßer, I., et al. (2015). How to measure wisdom: content, reliability, and validity of five measures. *Front. Psychol.* 4:405. doi: 10.3389/fpsyg.2013.00405

Greenland, S., and O'Rourke, K. (2001). On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics* 2, 463–471. doi: 10.1093/biostatistics/2.4.463

Grimshaw, J., Eccles, M., Thomas, R., MacLennan, G., Ramsay, C., Fraser, C., et al. (2006). Toward evidence-based quality improvement. Evidence (and its limitations) of the effectiveness of guideline dissemination and implementation strategies 1966–1998. *J. Gen. Intern. Med.* 21(Suppl. 2), 14–20. doi: 10.1007/s11606-006-0269-7

Hayes, A. F., and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Commun. Methods Meas.* 1, 77–89. doi: 10.1080/19312450709336664

Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., et al. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 343:d5928. doi: 10.1136/bmj.d5928

Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero-García, M. I., and Sanduvete-Chaves. (2006). Training satisfaction rating scale: development of a measurement model using polychoric correlations. *Eur. J. Psychol. Assess.* 22, 268–279. doi: 10.1027/1015-5759.22.4.268

Hopewell, S., Clarke, M., and Askie, L. (2006). Reporting of trials presented in conference abstracts needs to be improved. *J. Clin. Epidemiol.* 59, 681–684. doi: 10.1016/j.jclinepi.2005.09.016

Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J. M., Gavaghan, D. J., et al. (1996). Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control. Clin. Trials* 17, 1–12. doi: 10.1016/0197-2456(95)00134-4

Jefferson, T., Di Pietrantonj, C., Debalini, M. G., Rivetti, A., and Demicheli, V. (2009). Relation of study quality, concordance, take home message, funding, and impact in studies of influenza vaccines: systematic review. *BMJ* 338:b354. doi: 10.1136/bmj.b354

Jiménez-Requena, F., Delgado-Bolton, R. C., Fernández-Pérez, C., Gambhir, S. S., Schwimmer, J., Pérez-Vázquez, J. M., et al. (2009). Meta-analysis of the performance of F-FDG PET in cutaneous melanoma. *Eur. J. Nucl. Med. Mol. Imaging* 37, 284–300. doi: 10.1007/s00259-009-1224-8

Jüni, P., Altman, D. G., and Egger, M. (2001). "Assessing the quality of randomised controlled trials," in *Systematic Reviews in Health Care*, eds M. Egger, G. D. Smith, and D. G. Altman (London: BMJ), 87–108.

Jüni, P., Witschi, A., Bloch, R., and Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 282, 1054–1060. doi: 10.1001/jama.282.11.1054

Leech, N. L., and Onwuegbuzie, A. J. (2010). Guidelines for conducting and reporting mixed research in the field of counseling and beyond. *J. Couns. Dev.* 88, 61–69. doi: 10.1002/j.1556-6678.2010.tb00151.x

Leonardi, M. (2006). Public health support to policies in the fields of headache. Different ways of producing data and modalities of reading them with the aid of the meta-analytic approach. *J. Headache Pain* 7, 157–159. doi: 10.1007/s10194-006-0298-y

Li, L. C., Moja, L., Romero, A., Sayre, E. C., and Grimshaw, J. M. (2009). Nonrandomized quality improvement intervention trials might overstate the strength of causal inference of their findings. *J. Clin. Epidemiol.* 62, 959–966. doi: 10.1016/j.jclinepi.2008.10.008

Linde, K. (2009). Can you trust systematic reviews of complementary and alternative therapies? *Eur. J. Integr. Med.* 1, 117–123. doi: 10.1016/j.eujim.2009.09.002

Lipsey, M. W. (1994). "Identifying potentially interesting variables and analysis opportunities," in *The Handbook of Research Synthesis*, eds H. M. Cooper and L. V. Hedges (New York, NY: Sage), 111–123.

Macedo, L. G., Elkins, M. R., Maher, C. G., Moseley, A. M., Herbert, R. D., and Sherrington, C. (2010). There was evidence of convergent and construct validity of physiotherapy evidence database quality scale for physiotherapy trials. *J. Clin. Epidemiol.* 63, 920–925. doi: 10.1016/j.jclinepi.2009.10.005

Maher, C. G., Sherrington, C., Herbert, R. D., Moseley, A. M., and Elkins, M. (2003). Reliability of the PEDro Scale for rating quality of randomized controlled trials. *Phys. Ther.* 83, 713–721.

Martínez-Arias, M. R., Hernández-Lloreda, M. J., and Hernández-Lloreda, M. V. (2006). *Psicometría [Psychometrics]*. Madrid: Alianza.

Mayer, A., Nagengast, B., Fletcher, J., and Steyer, R. (2014). Analyzing average and conditional effects with multigroup multilevel structural equation models. *Front. Psychol.* 5:304. doi: 10.3389/fpsyg.2014.00304

Merrett, D. L., Peretz, I., and Wilson, S. J. (2013). Moderating variables of music training-induced neuroplasticity: a review and discussion. *Front. Psychol.* 4:606. doi: 10.3389/fpsyg.2013.00606

Minelli, C., Thompson, J. R., Abrams, K. R., Thakkinstian, A., and Attia, J. (2007). How should we use information about HWE in the meta-analyses of genetic association studies? *Int. J. Epidemiol.* 37, 136–146. doi: 10.1093/ije/dym234

Moher, D., Jadad, A. R., and Tugwell, P. (1996). Assessing the quality of randomized controlled trials: current issues and future directions. *Int. J. Technol. Assess. Health Care* 12, 195–208. doi: 10.1017/S0266462300009570

Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. G. (2009). Preferred reporting items for systematic review and meta-analyses: the PRISMA statement. *BMJ* 339, 332–336. doi: 10.1136/bmj.b2535

Moher, D., Pham, B., Jones, A., Cook, D. J., Jadad, A. R., Moher, M., et al. (1998). Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 352, 609–613. doi: 10.1016/S0140-6736(05)60370-4

Nimon, K., Zientek, L. R., and Henson, R. K. (2012). The assumption of a reliable instrument and other pitfalls to avoid when considering the reliability of data. *Front. Psychol.* 3:102. doi: 10.3389/fpsyg.2012.00102

Olivares, J., Rosa, A. I., and Sánchez-Meca, J. (2000). Meta-análisis de la eficacia de las habilidades de afrontamiento en problemas clínicos y de salud en España [Meta-analysis of the effectiveness of coping skills in clinical and health problems in Spain]. *Anuario Psicol.* 31, 43–61.

Osterlind, S. J. (1998). *Constructing Tests Items*. Boston, MA: Kluwer Academic Publishers.

Pluye, P., Gagnon, M. P., Griffiths, F., and Johnson-Lafleur, J. (2009). A scoring system for appraising mixed methods research, and concomitantly appraising qualitative, quantitative and mixed methods primary studies in mixed studies reviews. *Int. J. Nurs. Stud.* 46, 529–546. doi: 10.1016/j.ijnurstu.2009.01.009

Portell, M., Anguera, M. T., Chacón-Moscoso, S., and Sanduvete-Chaves, S. (2015). Guidelines for reporting evaluations based on observational methodology. *Psicothema* 27, 283–289. doi: 10.7334/psicothema2014.276

Rubinstein, S. M., Pool, J. J. M., van Tulder, M. W., Riphagen, I. I., and De Vet, H. C. W. (2007). A systematic review of the diagnostic accuracy of provocative tests of the neck for diagnosing cervical radiculopathy. *Eur. Spine J.* 16, 307–319. doi: 10.1007/s00586-006-0225-6

Rutjes, A. W. S., Reitsma, J. B., Di Nisio, M., Smidt, N., van Rijn, J. C., and Bossuyt, P. M. M. (2006). Evidence of bias and variation in diagnostic accuracy studies. *Can. Med. Assoc. J.* 174, 469–476. doi: 10.1503/cmaj.050090

Sánchez-Meca, J. (1997). "Methodological issues in the meta-evaluation of correctional treatment," in *Advances in Psychology and Law: International Contributions*, eds S. Redondo, V. Garrido, J. Pérez, and R. Barberet (New York, NY: Walter de Gruyter), 486–498.

Sánchez-Meca, J., Rosa, A. I., and Olivares, J. (1998). Cognitive-behavioral techniques in clinic and healthy disorders. Meta-analysis of Spanish literature. *Psicothema* 11, 641–654.

Sanderson, S., Tatt, I. D., and Higgins, J. P. T. (2007). Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *Int. J. Epidemiol.* 36, 666–676. doi: 10.1093/ije/dym018

Sargeant, J. M., Torrence, M. E., Rajic, A., O'Connor, A. M., and Williams, J. (2006). Methodological quality assessment of review articles evaluating interventions to improve microbial food safety. *Foodborne Pathog. Dis.* 3, 447–456. doi: 10.1089/fpd.2006.3.447

Schulz, K. F., Altman, D. G., and Moher, D. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 340, 698–702. doi: 10.1136/bmj.c332

Shadish, W. R., Chacón-Moscoso, S., and Sánchez-Meca, J. (2005). Evidence-based decision making: enhancing systematic reviews of program evaluation results in Europe. *Evaluation* 11, 95–109. doi: 10.1177/1356389005053196

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York, NY: Houghton Mifflin Company.

Sherrington, C., Herbert, R. D., Maher, C. G., and Moseley, A. M. (2000). PEDro. A database of randomized trials and systematic reviews in physiotherapy. *Man. Ther.* 5, 223–226. doi: 10.1054/math.2000.0372

Stolarova, M., Wolf, C., Rinker, T., and Brielmann, A. (2014). How to assess and compare inter-rater reliability, agreement and correlation of ratings: an exemplary analysis of mother–father and parent–teacher expressive vocabulary rating pairs. *Front. Psychol.* 5:509. doi: 10.3389/fpsyg.2014.00509

Stone, A. A., and Shiffman, S. (2002). Capturing momentary, self-report data: a proposal for reporting guidelines. *Ann. Behav. Med.* 24, 236–243. doi: 10.1207/S15324796ABM2403_09

Taji, Y., Kuwahara, T., Shikata, S., and Morimoto, T. (2006). Meta-analysis of antiplatelet therapy for IgA nephropathy. *Clin. Exp. Nephrol.* 10, 268–273. doi: 10.1007/s10157-006-0433-8

Valentine, J. C., and Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: the study design and implementation assessment device (study DIAD). *Psychol. Methods* 13, 130–149. doi: 10.1037/1082-989X.13.2.130

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., and Vandenbroucke, J. P. (2007). The strengthening the reporting of observational studies in epidemiology (STROBE) statement. Guidelines for reporting observational studies. *Epidemiology* 18, 800–804.

Wells, G., Shea, B., O'Connell, D., Robertson, J., Peterson, J., Welch, V., et al. (2009). *The Newcastle-Ottawa Scale (NOS) for Assessing the Quality of Nonrandomized Studies in Meta-Analysis*. Available at: http://www.ohri.ca/programs/clinical$_e$pidemiology/oxford.htm

Wilson, D. B. (2009). Missing a critical piece of the pie: simple document search strategies inadequate for systematic reviews. *J. Exp. Criminol.* 5, 429–440. doi: 10.1007/s11292-009-9085-5

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

A Simulation Study of Threats to Validity in Quasi-Experimental Designs: Interrelationship between Design, Measurement, and Analysis

Fco. P. Holgado-Tello[1], *Salvador Chacón-Moscoso*[2,3]*, *Susana Sanduvete-Chaves*[2] and *José A. Pérez-Gil*[2]

[1] *Metodología de las Ciencias del Comportamiento, Universidad Nacional de Educación a Distancia, Madrid, Spain,* [2] *HUM-649, Innoevalua, Psicología Experimental, Universidad de Sevilla, Sevilla, Spain,* [3] *Universidad Autónoma de Chile, Santiago, Chile*

OPEN ACCESS

Edited by:
Pietro Cipresso,
IRCCS Istituto Auxologico Italiano, Italy

Reviewed by:
Timothy R. Brick,
Pennsylvania State University, USA
Adam R. Ferguson,
University of California, San Francisco, USA

****Correspondence:***
Salvador Chacón-Moscoso
schacon@us.es

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *13 October 2015*
Accepted: *31 May 2016*
Published: *16 June 2016*

Citation:
Holgado-Tello FP, Chacón-Moscoso S, Sanduvete-Chaves S and Pérez-Gil JA (2016) A Simulation Study of Threats to Validity in Quasi-Experimental Designs: Interrelationship between Design, Measurement, and Analysis. Front. Psychol. 7:897. doi: 10.3389/fpsyg.2016.00897

The Campbellian tradition provides a conceptual framework to assess threats to validity. On the other hand, different models of causal analysis have been developed to control estimation biases in different research designs. However, the link between design features, measurement issues, and concrete impact estimation analyses is weak. In order to provide an empirical solution to this problem, we use Structural Equation Modeling (SEM) as a first approximation to operationalize the analytical implications of threats to validity in quasi-experimental designs. Based on the analogies established between the Classical Test Theory (CTT) and causal analysis, we describe an empirical study based on SEM in which range restriction and statistical power have been simulated in two different models: (1) A multistate model in the control condition (pre-test); and (2) A single-trait-multistate model in the control condition (post-test), adding a new mediator latent exogenous (independent) variable that represents a threat to validity. Results show, empirically, how the differences between both the models could be partially or totally attributed to these threats. Therefore, SEM provides a useful tool to analyze the influence of potential threats to validity.

Keywords: threats to validity, quasi-experimental designs, Structural Equation Modeling, causal analysis, Classical Test Theory

THREATS TO VALIDITY: THEORETICAL AND ANALYTICAL PERSPECTIVES

The unstable social and political conditions of most contexts to which evaluation programs are applied (Gorard and Cook, 2007) imply that, *a priori*, there are no standardized evaluation design structures (Chacón-Moscoso et al., 2013). Due to this fact and because random assignment of participants to different groups is not always possible (Colli et al., 2014), quasi-experimental designs are more commonly used in social sciences than experimental ones (Shadish et al., 2005). Quasi-experiments lack control over extraneous variables generated by random allocation; therefore, it is extremely important that the evaluation process is conducted in a manner that provides reliable and valid results based on the analysis of the influence of potential threats to validity (Reichardt and Coleman, 1995).

There are conditions other than the intervention program itself that could be responsible for the outcomes. These conditions are called threats to validity which, unless controlled, limit the confidence of causal findings (Weiss, 1998).

This evaluation research context presents two main problems. On the one hand, as a conceptual-theoretical framework, the Campbellian tradition presents a series of threats to validity that can affect four different kinds of validity (Campbell, 1957; Campbell and Stanley, 1963; Cook and Campbell, 1979; Shadish et al., 2002): (a) statistical conclusion validity (García-Pérez, 2012) can be affected by a low statistical power (Tressoldi and Giofré, 2015) and a restricted range (Vaci et al., 2014); (b) internal validity can be affected by selection, history, maturation, and regression; (c) construct validity can be affected by construct confounding, treatment-sensitive factorial structure, and inadequate explication of constructs; and (d) external validity can be affected by interaction of the causal relationship with units or outcomes. Although Campbell's approach provides a conceptual framework for evaluating the main threats to four types of validity (Shadish et al., 2002) and some guidelines (design features) to enhance validity were presented, there is not an empirical, systematic approach to check and control the influence of threats to validity on the treatment effect estimations in program evaluation practice (e.g., Stocké, 2007; Krause, 2009; Johnson et al., 2015).

On the other hand, from an analytical point of view, procedures have been developed to assess some construct validity threats, such as inadequate explication of constructs, confounding of constructs operations, mono-operationalization, and mono-method bias. Some of these procedures include the multimethod-multitrait approach (Eid et al., 2008) and factor retention analysis, through the study of systematic pattern in the error covariance (Brown, 2015). The apparently useful analytical proposal weakens because it is not based on any theoretical framework.

In sum, there is a small connection between design features, measurement issues, and concrete impact estimation analyses. In order to provide an empirical solution to this problem, we use Structural Equation Modeling (SEM) as a first approximation to operationalize the analytical implications of threats to validity in quasi-experimental designs.

STRUCTURAL EQUATION MODELING (SEM): AN INTEGRATED APPROACH

Based on Steyer (2005), who draws analogies between the Classical Test Theory (CTT)— measurement point of view (Trafimow, 2014) —and causal analysis and Rubin's Causal Model— design and analysis points of view, which determine the concepts of statistical inference for causal effects in experiments and observational studies (West, 2011) — we assume that SEM can be used to systematize the model assumptions and test the model fit likelihood statistically, and empirically check the way threats to validity affect data and how different threats to validity influence each other.

If we focus on the participant-level scores in each experimental condition, we can establish an analogy between causal analysis and CTT, so that the measurement, design, and analysis aspects would be linked. The participants' expected value in causal analysis is similar to the true score defined by CTT. That is, it would be the expected score obtained after an infinite number of independent administrations of a measurement, under some assumptions (Lord and Novick, 1968). Based on the concept of parallel test, we can assume that across a set of scores, the variance of the observed score is composed of the sum of the true scores and the error variance. If we consider an experimental or quasi-experimental design with two conditions (control and experimental), then we can expect two true scores for each unit—one for the control condition and another for the experimental condition (Steyer, 2005)—and, therefore, two observed variances (one for the control condition and another for the experimental condition), and one covariance (control-experimental).

Taking into account the number of groups, and the number of measurement occasions, this theoretical framework could be translated into any experimental or quasi-experimental design. For example, if we combine the measurement occasion [pre-test (f0) and post-test (f1)] and the expected value or true score of each group (control: $X = 0$ and experimental: $X = 1$), **Table 1** presents the variance/covariance matrix in a non-equivalent control group design.

Variances are in the diagonal in boldface; e.g., $S^2[f0/X = 0]$ is the control group variance for the pre-test measurement and $S^2[f1/X = 1]$ is the experimental group variance for the post-test measurement. Covariances are out of the diagonal; e.g., $S[f0,f0/X = 1, X = 0]$ is the covariance between the control and experimental groups at pre-test; and $S[f1,f0/X = 0]$ is the pre-testpost-test covariance for the control group.

From the implied variance-covariance matrix, we can establish the model derivations for the non-equivalent control group design: (a) the control-experimental group covariance in the pre-test ($S[f0,f0/X = 1, X = 0]$) should be equal to the control and experimental variance in the pre-test ($S^2[f0/X = 0]$; and $S^2[f0/X = 1]$); equal to the control variance in the post-test ($S^2[f1/X = 0]$); and equal to the pre-testpost-test control group covariance $S[f1,f0/X = 0]$); and (b) the pre-testpost-test experimental group covariance ($S[f1,f0/X = 1]$) should be equal to the control-experimental group covariance in the post-test ($S[f1,f1/X = 0, X = 1]$); and equal to the pretestcontrol posttest-experimental covariance ($S[f1,f0/X = 1, X = 0]$).

These assumptions are shown in **Figure 1** over a non-equivalent control group design scheme.

The variance-covariance derivations could be operationalized via SEM through restriction of models, including more latent variables or testing the error terms, and therefore statistically tested. For example, the true scores (expected values or μ, the means of the population) of the control and experimental groups should be significantly equivalent in the pre-test (f0) and significantly different in the post-test (f1); in the control group, true scores should be significantly equivalent between the pre- (f0) and post-test (f1); and in the experimental group, these expected values should be significantly different between the pre- (f0) and post-test (f1). We can design these restrictions via

TABLE 1 | Implied variance/covariance matrix in a non-equivalent control group design.

		Pre (f0)		**Post (f1)**	
		Control (X0)	**Exptal.(X1)**	**Control (X0)**	**Exptal. (X1)**
Pre (f0)	Control (X0)	$\mathbf{S^2}$ **[f0/*X* = 0]**			
	Exptal. (X1)	*S* [f0, f0/*X* = 1, *X* = 0]	$\mathbf{S^2}$ **[f0/X = 1]**		
Post (f1)	Control (X0)	*S* [f1, f0/*X* = 0]	*S* [f0, f1/*X* = 0, *X* = 1]	$\mathbf{S^2}$ **[f1/*X* = 0]**	
	Exptal. (X1)	*S* [f1, f0/*X* = 1, *X* = 0]	*S* [f1, f0/*X* = 1]	*S* [f1, f1/*X* = 1, *X* = 0]	$\mathbf{S^2}$ **[f1/*X* = 1]**

Pre (f0), pre-treatment time point; Post (f1), post-treatment time point; Control (X0), Control group; Exptal. (X1), Experimental group; S^2*, variance (in boldface); and S, covariance.*

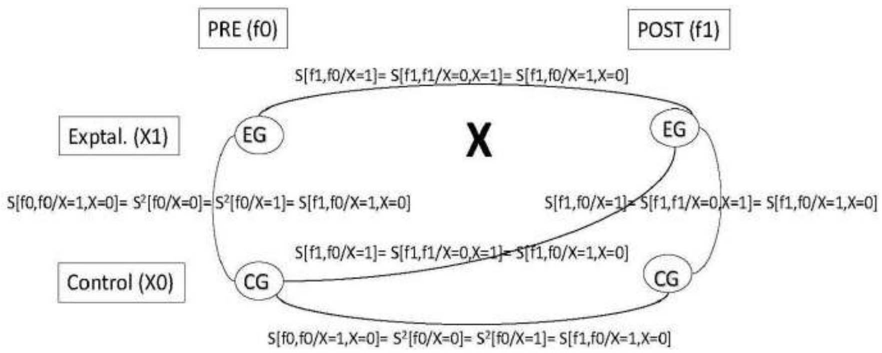

PRE (f0)
POST (f1)
S[f1,f0/X=1]= S[f1,f1/X=0,X=1]= S[f1,f0/X=1,X=0]
Exptal. (X1)
EG
X
EG
S[f0,f0/X=1,X=0]= S²[f0/X=0]= S²[f0/X=1]= S[f1,f0/X=1,X=0]
S[f1,f0/X=1]= S[f1,f1/X=0,X=1]= S[f1,f0/X=1,X=0]
S[f1,f0/X=1]= S[f1,f1/X=0,X=1]= S[f1,f0/X=1,X=0]
Control (X0)
CG
CG
S[f0,f0/X=1,X=0]= S²[f0/X=0]= S²[f0/X=1]= S[f1,f0/X=1,X=0]

FIGURE 1 | Covariances in a non- equivalent control group design. Pre (f0), pre-treatment measurement occasion; Post (fl), post-treatment measurement occasion; Control (X0), Control group; Exptal. (Xl), Experimental group; S^2, variance; *S,* covariance.

SEM, whether working with one or more groups or measurement occasions (Bollen and Curran, 2006). If the above conditions are not satisfied, it may be due to any validity threats that could be tested in an SEM framework. At this point, it is important to emphasize that this approach is only applicable in cases when the intervention aims to change the level of the dependent variable/s. However, when the aim is to maintain it, then the logic would be different (the expected changes would be in the control group across the pre-test and the post-test, rather than in the experimental group).

Once we have established the theoretical assumptions, the next step is to try to draw a non-equivalent control group into the SEM framework. As an example, we opt to use only one group (control group) in a simple design (pre-test and post-test) because the conditions are more easily simulated (a more complex design and model with control and experimental groups would require two pre-tests and two post-tests). In this sense, **Figure 2** presents a multistate model where four different endogenous latent variables are measured at the same time (in the pre-test).

Believable inferences are based on the assumption that all changes between the pre-test and post-test are caused by the treatment, and this assumption can only be true if we do not find systematic changes between the pre-test and post-test in the control group.

Then, we can suppose that $X = 1$ in the variance-covariance matrix is not an experimental group (i.e., a group that participated in a treatment), but a group affected by a threat to validity that can modify the data and generate systematic changes. In a parallel way, the latent variable *T* represented in **Figure 3** is not a treatment, but a threat to validity, so this figure would represent the control group in a concrete time point (the post-test), where an odd element, such as history, for example, can be affecting the results in two of the four endogenous latent variables, i.e., η_3 and $\eta_{4.}$ Let's suppose that, to measure the effectiveness of an intervention program to improve attitudes toward immigration in a developed country with an aging population, participants from an experimental group and a control group filled in a questionnaire at an early stage (pre-test). A year later, after the implementation of the intervention program, the post-test was completed. This questionnaire was formed by four dimensions: public safety (η_1), education (η_2), economy (η_3) and public health (η_4). It was not expected to obtain significant differences between pre- and post-test measures in the control group. However, a wave of young immigrants (*T*) occurred concurrently with the study and, according to research, promoted an increase in economic activity and an improvement in public health by increasing the number of taxpayers. Thus, in the control group, there was a significant improvement in attitudes toward immigration in economy (η_3) and public health (η_4), while attitudes in public safety (η_1) and education (η_2) did not vary significantly.

At this point, it is important to clarify that this is just a hypothetical situation; the model could have been defined with

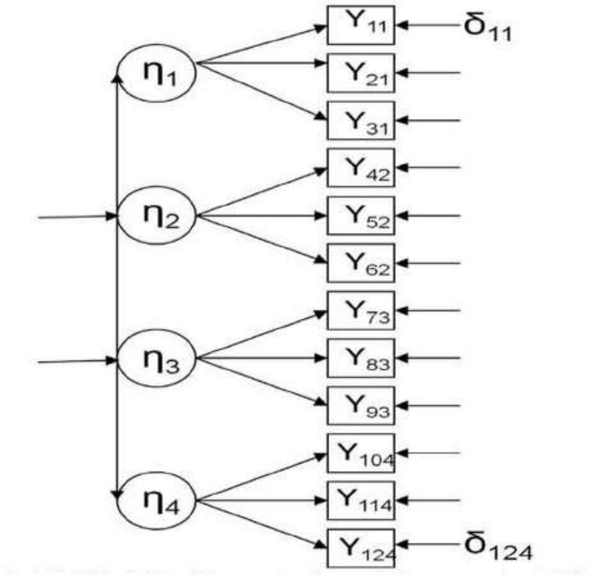


FIGURE 2 | Multistate model in the control condition, pre-test (Model 1). η, latent endogenous (dependent) variable; *Y*, observed endogenous (dependent) variable; δ, error.

the possible influence of *T* over three endogenous latent variables instead of two, over only one, over η_2 and η_3 instead of over η_3 and η_4, and so on.

In this case, we expect the same results in all the variances and covariances presented in **Table 1**. As **Figures 2** and **3** represent, respectively, pre- and post-test in the control group, any systematic change found could be attributed to the influence of a threat to validity (*T*).

ADVANTAGES OF SEM OVER OTHER METHODS

Scarce research in psychology was aimed to empirically detect the influence of threats to validity in interventions based on a theoretical framework. In this regard, Crutzen et al. (2015) used meta-analysis in order to study the relationship between differential attrition and several moderator variables; nevertheless, they could not study the relationship between the differential attrition and the effect size owing to technical problems. Furthermore, Damen et al. (2015) carried out a longitudinal study to finally conclude that a possible attrition bias occurred in a percutaneous coronary intervention, as drop-outs and completers differed systematically on some socio-demographic, clinical, and psychological baseline characteristics; nevertheless, as drop-outs did not receive the complete intervention, the authors could not study the difference across both groups (drop-outs and completers) in the results obtained in the post-test. Mixed-effects regression is useful to study the difference between the pre- and post-test across experimental and control groups. Nevertheless, this option is based on a pure analytical perspective and is restricted to include only directly observed variables; whereas, SEM is not just based on analysis, but on the integration of design, measurement, and analysis. Thus, it provides the possibility of obtaining concrete data about the relationship between latent and observed variables used to measure them and the associated measurement error for each one (measurement model), and the relationship between different latent variables (structural model), as shown in Duncan et al. (2006). Additionally, when the design presented includes two groups, the degree of equivalence between them can be defined depending on the restrictions imposed: we can assume equivalence between experimental and control groups in both the measurement and the structural model, or only in one of them. As a consequence of these differences, regression tends to obtain less sensitive results compared with SEM (Nusair and Hua, 2010).

Therefore, the SEM framework presents several advantages compared with other procedures. This approach includes a measurement-design-analysis point of view, so it is more complete than the traditional approaches based on a single aspect. Moreover, it allows to (a) take conclusions about the relationship across multiple latent variables between them (structural model) and each latent variable with the observed variables that measure it (measurement model); (b) define the degree of equivalence between the experimental and the control group; and (c) obtain inferences about the influence of threats to validity in the results; (d) be generalized to any threat to any kind of validity; and (e) study the concrete conditions under which different threats to validity could be influencing the results.

A similar methodology has been already found as useful to study the influence of threats to validity in other fields; e.g., (a) in forest research, Ficko and Boncina (2014) operationalized the influence of response style bias and the robustness of statistical methods in the results using simulations and including latent variables in the models representing those threats to validity; and (b) in medical research, Mickenautsch et al. (2014) studied the inflation of effect size owing to selection bias using simulations. In the current study, we show the application and usefulness of simulations and the SEM framework in social sciences, specifically in psychology, to detect the influence of other different threats to validity.

OBJECTIVE

The objective of this study is to illustrate conceptual problems of threats to validity through causal analyses using SEM, under the framework of design. Concretely, based on the multistate and single-trait-multistate models, we carry out a simulation study where two threats to statistical conclusion validity are manipulated (restriction of range and low statistical power) in order to analyze the way in which a third threat to validity named *T* (unspecified, it could be potentially any of them) could be affecting the measures in the post-test of a non-equivalent control group in a quasi-experimental design.

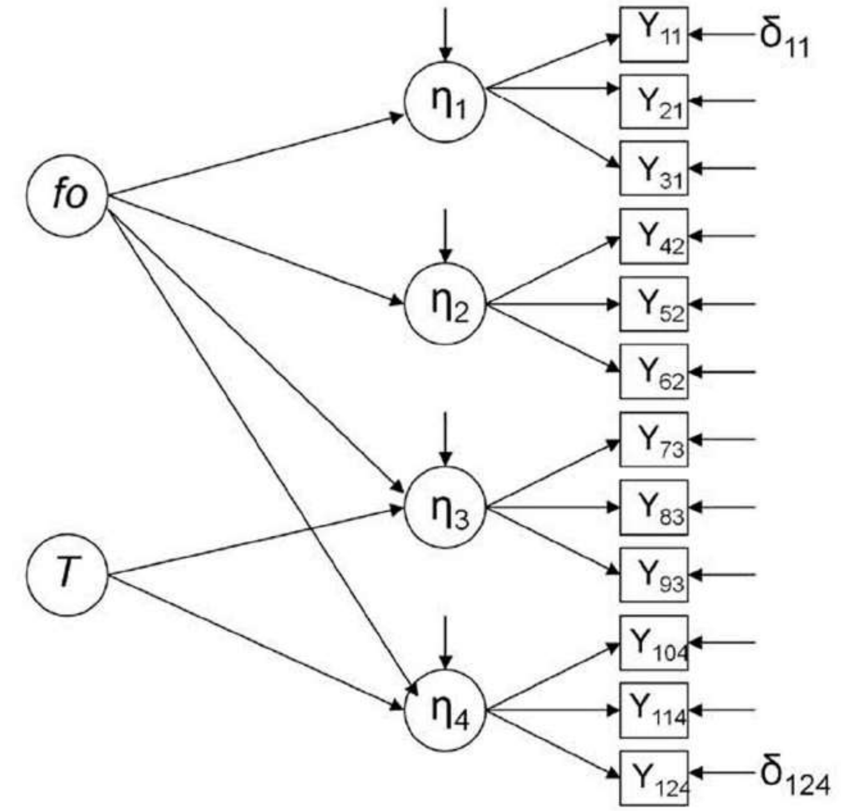


FIGURE 3 | Singletrait-multistate Model in the control condition, post-test (Model 2; Steyer, 2005). *fo*, latent exogenous (independent) variable representing the expected outcome under control condition; *T,* latent exogenous (independent) variable representing a threat to validity; η, latent endogenous (dependent) variable; *Y*, observed endogenous (dependent) variable; and δ, error.

MATERIALS AND METHODS

Data was generated using two different models. However, in both the cases, we considered only the control condition: (a) **Figure 2** represents the multistate model in the control condition (Model 1), where four latent endogenous (dependent) variables (η) are measured through three observed endogenous variables (Y) in a concrete time point (pre-test); (b) **Figure 3** represents the single-trait-multistate model in the control condition (Model 2), where the same four latent endogenous variables are measured through the same three observed endogenous variables in another concrete time point (post-test) (Steyer, 2005; Dumenci and Windle, 2010; Pohl and Steyer, 2010). In this case, a new mediator latent exogenous (independent) variable that represents a threat to validity (T) was added in order to detect its possible influence in the model fit; f_0 is another latent exogenous variable that represents the expected outcome under the control condition (Steyer, 2005). Both Models 1 and 2 assume that all effects are linear (Kline, 2012).

When the multistate model (Model 1, pre-test in control group; **Figure 2** that does not include T) is rejected and the single-trait-multistate model (Model 2, post-test in control group; **Figure 3** that includes T) is accepted, we can conclude that the T variable could be affecting the data in the post-test; thus, differences found between the pre-test and the post-test could be partially or totally attributed to threats to validity. Under these circumstances, further analysis would not provide valid inferences about the effectiveness of treatment. In that case, the T variable could be operationalized in a SEM (Ryu, 2014).

Additionally, two previously mentioned threats to statistical conclusion validity are manipulated in order to study the possible interaction with the threat to validity named T in **Figure 3**: (a) the *low statistical power* implies obtaining non-significant relationship between the treatment and outcome because the experiment has insufficient power (probability of finding an effect when the effect exists). This threat to validity was manipulated by varying the sample size, with 5 conditions: 100, 500, 750, 1000, and 5000 participants; and (b) the *restricted range* implies that reduced range on a variable usually weakens the relationship between this variable and another (Coenders and Saris, 1995; DiStefano, 2002; Holgado-Tello et al., 2010; Yang-Wallentin et al., 2010; Williams and Vogt, 2011; Bollen, 2014). This threat to validity was manipulated by varying the number of levels in the dependent observed variables (Y), with four conditions: 3, 5, and 7 discrete categories, and as continuous variables.

For each latent endogenous variable, three observed variables were simulated. The factor loadings of the observed variables were always the same in all factors (0.8). The simulated factor loadings were high in order to avoid doubts about the specification in the estimation stage of the parameters. Observed variables were generated according to a normal distribution *N(0,1)*. Then, these answers were categorized according to 3, 5, and 7 discrete categories, that is, were categorized in Likert scales with different numbers of possible responses to restrict the range of variation, or remained as continuous variables. The responses to all observed variables remained symmetrical in order to avoid the influence of skewness. To categorize the Likert scales, as stated by Bollen and Barb (1981), the continuum was divided into equal intervals from $z = -3$ to $z = 3$ in order to calculate the thresholds of the condition in which the response distribution to all items is symmetrical (skewness = 0). Finally, the sample size had five experimental values (100, 500, 750, 1000, and 5000).

The combination of the two experimental factors produced 20 experimental conditions (4 × 5) which were replicated 1000 times. These replications were performed using R version 2.0.0, which invoked PRELIS successively (Jöreskog and Sörbom, 1996b) to generate the corresponding data matrices according to the specifications resulting from the combination of the experimental conditions. Thus, for each data generated matrix, correlation matrices were obtained. After obtaining correlation matrices for each replication, the corresponding Confirmatory Factor Analysis was performed successively. The instrumental problem of underidentification in Model 2 (**Figure 3**) was solved by constraining four model components as equal to one: two beta parameters (concretely, β_{11} and β_{32}) and the variances of F_0 and T.

As in the previous case, these replications were performed using R version 2.0.0, which invoked LISREL 8.8 successively (Jöreskog and Sörbom, 1996a).

RESULTS

Table 2 presents the results obtained in the multistate model in the control condition, pre-test (Model 1) and the single-trait-multistate model in the control condition, post-test (Model 2) in the different experimental conditions.

We found, in general, that: (a) in none of occasions Model 1 fitted better than Model 2; (b) increase in chi-square ($\Lambda\chi^2$) was significant from Model 1 to 2; therefore, Model 2 fitted significantly better than Model 1 in all the experimental conditions in most of the replications (in 100% of replications when there were 500 participants or more); (c) with 100 participants, both models were rejected, regardless of the categorization of the observed dependent variables (Y).

Taking into account the percentage of replications where *RMSEA* was lower than 0.08 we found the following results: (a) with 500 participants or more, both Models 1 and 2 fitted when the observed dependent variables (Y) were continuous; and (b) Model 2 fitted better than Model 1 when the observed dependent variables (Y) had 5 or 7 categories; with 750 participants or more

TABLE 2 | Results obtained in Models 1 and 2 in different conditions.

n	**Categories**	**% Accepted Ho**		**% $\Lambda\chi^2$ is significant ($\Lambda df = 3$)**	**% *RMSEA* < 0.08**	
		Model 1	**Model 2**		**Model 1**	**Model 2**
100	**3**	0.3	0.3	**72.6**	0.4	1.3
	5	0.0	0.9	**100**	0.9	17
	7	0.0	0.0	**99.9**	0.0	1.5
	Con.	0.0	0.0	**99.1**	1.7	7.5
500	**3**	0.0	0.0	**100**	0.0	29
	5	0.0	5.4	**100**	**4.1**	**100**
	7	0.0	0.0	**100**	**0.1**	**94.7**
	Con.	**3.5**	**95.6**	**100**	100	100
750	**3**	0.0	0.0	**100**	**0.0**	**84.5**
	5	0.0	6.7	**100**	**4.0**	**100**
	7	0.0	0.1	**100**	**0.4**	**99.8**
	Con.	**0.9**	**96.4**	**100**	100	100
1000	**3**	0.0	0.0	**100**	**0.2**	**99.3**
	5	0.0	6.3	**100**	**4.7**	**100**
	7	0.0	0.0	**100**	**0.6**	**100**
	Con.	**0.0**	**93.9**	**100**	100	100
5000	**3**	0.0	0.0	**100**	**0.0**	**99.9**
	5	0.0	6.5	**100**	**0.6**	**100**
	7	0.0	0.1	**100**	**0.4**	**100**
	Con.	**0.0**	**96.4**	**100**	100	100

Model 1, the pre-test in control group, which does not include a possible threat to validity influence (T); Model 2, the post-test in control group, which includes a possible threat to validity influence (T); n, sample size; % accepted Ho, percentage of null hypothesis accepted (i.e., the model fits) in Models 1 and 2 using χ^2; % $\Lambda\chi^2$ is significant, percentage of significant increase in χ^2 between Models 1 and 2; Λdf, increase in the degrees of freedom between Models 1 and 2; % RMSEA < 0.08, percentage of occasions in which the Root Mean Square Error of Approximation is under 0.08 (i.e., the model fits); Con., the dependent variables (Y) are continuous. Values marked in bold are the results that suggest a better fit in Model 2 than in Model 1.

as well, the same result was found in the case that the observed dependent variables (Y) had 3 categories.

Taking into account the percentage of accepted null hypothesis considering χ^2, an index more sensible than *RMSEA*, the only model that fitted was Model 2 in the case where there were 500 participants or more and the observed dependent variables (Y) were continuous.

Whether the multistate model (Model 1) is rejected and the single-trait-multistate model (Model 2) is accepted, following the results obtained, T could be affecting the results: (a) in all the experimental conditions, if we consider the percentage of significant increase of chi squared values (% $\Lambda\chi^2$); (b) when there were 500 participants or more and the observed dependent variables (Y) had 5 or 7 categories, if we consider the percentage of *RMSEA* lower than 0.08 (% RMSEA < 0.08); and (c) when there were 500 participants or more and the observed dependent variables (Y) were continuous, if we consider the percentage of accepted null hypothesis considering χ^2.

Following the same logic, we can conclude that the possible effect of the threat to validity (T) was only annulled in the case that we had at least 500 participants and the observed dependent variables (Y) were continuous, if we consider the percentage of *RMSEA* lower than 0.08 (% RMSEA < 0.08).

DISCUSSION

We would like to remark that the current study is a very preliminary approximation to study the threats to validity in quasi-experimental designs under the Campbellian tradition. We have attempted to present the basic aspects of the conceptual foundations to approach the study of threats to validity from an empirical perspective. The combination of design, CTT, and SEM has enabled us to obtain the design models derivations expressed in a variance-covariance matrix whose likelihood could be tested via SEM. Finally, from a pragmatic perspective, we have attempted to empirically illustrate the proposals presented via a simulation study. This study is an attempt to open slightly a door to develop vast empirical research for the solution of the problem regarding the threats to validity. From this perspective, we suggest potential future research to analyze other types of validity taking into account many possible designs.

Overall, we conclude that the single-trait-multistate model in the control condition, post-test (Model 2, including T), presented a better fit than the multistate model in the control condition, pre-test (Model 1, without T), across the experimental conditions. As the number of categories and sample size increase, the results showed that Model 1 was rejected in favor of Model 2.

The key findings obtained based on the simulation study of threats to validity using SEM applied to causal analysis are as follows: (a) a general view including measurement, design, and analysis aspects can be provided, bridging design issues and analytical implications, by analytically studying the consequences of threats to validity; this would give a necessary insight for practitioners when considering the consequences of design features on impact analysis; (b) it is useful to include several variables in the analysis using SEM representing any threat to any kind of validity in order to explain the inter-individual differences in the individual causal effects of the treatment variable on the response variable; with SEM, a test of measurement invariance using a concrete set of data could be carried out in a complementary way to study the possible differences between models, obtaining conclusions for specific situations in specific conditions (Muthén and Asparouhov, 2013); the advantage of using simulations is that conclusions about possible threats to validity can be easily generalized to any situation and to different conditions due to the high number of replications obtained (1000 in this study) and the possibility of manipulating different variables (e.g., number of possible responses and sample size in this study); thus, based on this study, we can conclude that (c) it is recommended not to categorize the dependent variables and, when done, try to have as many categories as possible; with continuous dependent variables, the possible negative effect of the threat to validity included in Model 2 (named T) tended to be neutralized (Model 1, without T, also obtained a good fit considering *RMSEA*); and (d) using small sample sizes (less than 100 participants) is not adequate (Models, including T or not, did not present an acceptable fit).

For future research: (a) we shall apply the procedure presented in the current study using real data obtained from a real situation in order to show practitioners how this proposal can increase the control over extraneous variables and, consequently, the quality of the interventions; (b) it will be necessary to work under the logic of multigroup analysis. This perspective would enable us to consider the control and experimental groups at the same time, and the pre- and post-test measures; then, it would be possible to impose the restrictions of the variance-covariance matrix of **Table 1**. In this way, some weaknesses of the present proposal would be solved, such as the fact that extraneous variables do not necessarily imply a threat to validity because, when provoking the same effect in the treatment and control groups, this effect is neutralized (for example, the effect of maturation in children); in this sense, we would find a positive change in the control group when comparing pre- and post-test (instead of the same true score), but the change would be significantly lower than in the treatment group (if the treatment were effective). In sum, the control group does not need to have an identical level of X in pre and post-test, but this possible level difference does not need to be statistically significant compared to the treatment group. These differences can only be detected when comparing both groups (control and experimental) and both measurement occasions (pre and post-test); (c) when working with control and experimental groups, we shall manipulate the degree of equivalence between both to study the changes in the model fit when control and experimental groups are strictly equivalents (strong equivalence; i.e., equal structural and measurement model), or only the structural model is equal between control and experimental groups, or only the measurement model is equal between control and experimental groups (weak equivalence). Thus, it has completely different consequences on possible inferences to be made from the quasi-experimental designs. If a strong equivalence is achieved, then we would have empirical evidence of a “high degree of validity” in the results obtained. However, when the equivalence found is only weak, we could suspect that some threats to construct validity could be affecting the results when the equivalence is found only in the measurement model, and it is possible that some threats to internal validity could be working if the equivalence is achieved only in the structural model; and (d) we shall manipulate other threats to validity (Shadish et al., 2002) using the same approach; e.g., violated assumptions of statistical tests (a threat to statistical conclusion validity) can be studied by simulating data with and without normal distribution and checking if the same model fits under both conditions; regression (a threat to internal validity) can be studied by simulating data sets with and without extreme values and checking if the same model fits under both conditions; treatment-sensitive factorial structure (a threat to construct validity) can be studied by simulating possible changes in data when comparing pre- and post-test results and checking if the factorial structure obtained in the pre-test is maintained equal in the post-test; inadequate explication of constructs (another threat to construct validity) can be studied by taking real data obtained from questionnaires and, before checking the possible relationships between constructs (structural model), confirming that items measure adequately each construct (measurement model); or interaction of the causal relationship with units (a threat to external validity) can be studied by checking the measurement invariance of a model across different groups, such as male and female.

AUTHOR CONTRIBUTIONS

FH-T and SC-M came up with the initial idea and design of the work and interpreted the results. FH-T, SC-M, and JP-G made critical revisions to the manuscript for important intellectual content. FH-T and JP–G performed the analyses. SS-C helped out in the interpretation of data and was in charge of drafting the manuscript. All the four authors (FH-T, SC-M, SS-C, and JP-G) provided the final approval of the version to be published, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

FUNDING

This study forms part of the results obtained in research project PSI2011-29587, funded by Spain's Ministry of Science and Innovation; and in research project number 1150096, funded by Chilean National Fund of Scientific and Technological Development - FONDECYT).

ACKNOWLEDGMENTS

We would like to dedicate this article to our fellow researchers and very good friends José A. Pérez-Gil, who passed away while the article was being written and William R. Shadish, an example of honesty and hard work, who is a referent in quasi-experimental designs. May they rest in peace. The authors highly appreciate all the comments received from the reviewers and the English language editor. We believe that the quality of this paper has been substantially enhanced as a result.

REFERENCES

Bollen, K. A. (2014). *Structural Equations with Latent Variables*. New York, NY: Wiley-Interscience Publication.

Bollen, K. A., and Barb, K. H. (1981). Pearson's r and coarsely categorized measures. *Am. Sociol. Rev.* 46, 232–239. doi: 10.2307/2094981

Bollen, K. A., and Curran, P. (2006). *Latent Curve Models: A Structural Equation Perspective*. Hoboken, NJ: Wiley-Interscience Publication.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research*. New York, NY: Guilford Publications.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychol. Bull.* 54, 297–312. doi: 10.1037/h0040950

Campbell, D. T., and Stanley, J. C. (1963). *Experimental and Quasiexperimental Designs for Research*. Chicago, IL: RandMcnally.

Chacón-Moscoso, S., Sanduvete-Chaves, S., Portell-Vidal, M., and Anguera, M. T. (2013). Reporting a program evaluation: needs, program plan, intervention, and decisions. *Int. J. Clin. Health Psychol.* 13, 58–66. doi: 10.1016/S1697-2600(13)70008-5

Coenders, G., and Saris, W. E. (1995). "Categorization and measurement quality. The choice between Pearson and polychoric correlations," in *The Multitrait-Multimethod Approach to Evaluate Measurement Instruments*, eds W. E. Saris and A. Münnich (Budapest: Eötvös University Press), 125–144.

Colli, A., Pagliaro, L., and Duca, P. (2014). The ethical problem of randomization. *Int. Emerg. Med.* 9, 799–804. doi: 10.1007/s11739-014-1118-z

Cook, T. D., and Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston, MA: Houghton Mifflin.

Crutzen, R., Viechtbauer, W., Spigt, M., and Kotz, D. (2015). Differential attrition in health behaviour change trials: a systematic review and meta-analysis. *Psychol. Health* 30, 122–134. doi: 10.1080/08870446.2014.953526

Damen, N. L., Versteeg, H., Serruys, P. W., van Geuns, R.-J. M., van Domburg, R. T., Pedersen, S. S., et al. (2015). Cardiac patients who completed a longitudinal psychosocial study had a different clinical and psychosocial baseline profile than patients who dropped out prematurely. *Eur. J. Prev. Cardiol.* 22, 196–199. doi: 10.1177/2047487313506548

DiStefano, C. (2002). The impact of categorization with Confirmatory Factor Analysis. *Struct. Equ. Modeling* 9, 327–346. doi: 10.1207/S15328007SEM0903_2

Dumenci, L., and Windle, M. (2010). A latent trait-state model of adolescent depression using the center for epidemiologic studies-depression scale. *Multivariate Behav. Res.* 31, 313–330. doi: 10.1207/s15327906mbr3103_3

Duncan, T. E., Duncan, S. C., and Strycker, L. A. (2006). *An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues, and Application*. Mahwah, NJ: Lawrence Erlbaum Associates.

Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., and Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: different models for different types of methods. *Psychol. Methods* 13, 230–253. doi: 10.1037/a0013219

Ficko, A., and Boncina, A. (2014). Ensuring the validity of private forest owner typologies by controlling for response style bias and the robustness of statistical methods. *Scand. J. For. Res.* 29, 210–223. doi: 10.1080/02827581.2013.837194

García-Pérez, M. A. (2012). Statistical conclusion validity: some common threats and simple remedies. *Front. Psychol.* 3:325. doi: 10.3389/fpsyg.2012.00325

Gorard, S., and Cook, T. D. (2007). Where does good evidence come from? *Int. J. Res. Methods Educ.* 30, 307–323. doi: 10.1080/17437270701614790

Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero-García, M. I., and Vila-Abad, E. (2010). Polychoric versus Pearson correlations in Exploratory and Confirmatory Factor Analysis with ordinal variables. *Qual. Quant.* 44, 153–166. doi: 10.1007/s11135-008-9190-y

Johnson, S. P., Malay, S., and Chung, K. C. (2015). The quality of control groups in nonrandomized studies published in the Journal of Hand Surgery. *J. Hand Surg.* 40, 133–139. doi: 10.1016/j.jhsa.2014.09.021

Jöreskog, K., and Sörbom, D. (1996a). *LISREL 8: User's Reference Guide*. Chicago, IL: Scientific Software International.

Jöreskog, K., and Sörbom, D. (1996b). *PRELIS 2: User's Reference Guide*. Chicago, IL: Scientific Software International.

Kline, R. B. (2012). "Assumptions of structural equation modeling," in *Handbook of Structural Equation Modeling*, ed. R. Hoyle (New York, NY: Guilford Press), 111–125.

Krause, M. S. (2009). Reversion toward the mean independently of regression toward the mean. *Methodology* 5, 3–6. doi: 10.1027/1614-2241.5.1.3

Lord, F. M., and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. New York, NY: Addison Wesley.

Mickenautsch, S., Fu, B., Gudehithlu, S., and Berger, V. W. (2014). Accuracy of the Berger-Exner test for detecting third-order selection bias in randomised controlled trials: a simulation-based investigation. *BMC Med. Res. Methodol.* 14:114. doi: 10.1186/1471-2288-14-114

Muthén, B., and Asparouhov, T. (2013). *New Methods for the Study of Measurement Invariance with Many Groups*. Available at: https://www.statmodel.com/download/PolAn.pdf

Nusair, K., and Hua, N. (2010). Comparative assessment of structural equation modeling and multiple regression research methodologies: e-commerce context. *Tourism Manage.* 31, 314–324. doi: 10.1016/j.tourman.2009.03.010

Pohl, S., and Steyer, R. (2010). Modeling common traits and method effects in multitrait-multimethod analysis. *Multivariate Behav. Res.* 45, 45–72. doi: 10.1080/00273170903504729

Reichardt, C. S., and Coleman, S. C. (1995). The criteria for convergent and discriminant validity in a multitrait-multimethod matrix. *Multivariate Behav. Res.* 30, 513–538. doi: 10.1207/s15327906mbr3004_3

Ryu, E. (2014). Model fit evaluation in multilevel structural equation models. *Front. Psychol.* 5:81. doi: 10.3389/fpsyg.2014.00081

Shadish, W. R., Chacón, S., and Sánchez-Meca, J. (2005). Evidence-based decision-making: enhancing systematic reviews of program evaluation results in Europe. *Evaluation* 11, 95–109.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton-Mifflin.

Steyer, R. (2005). Analyzing individual and average causal effects via Structural Equation Models. *Methodology* 1, 39–54. doi: 10.1027/1614-1881.1.1.39

Stocké, V. (2007). Determinants and consequences of survey respondents' social desirability beliefs about racial attitudes. *Methodology* 3, 125–138. doi: 10.1027/1614-2241.3.3.125

Trafimow, D. (2014). Estimating true standard deviations. *Front. Psychol.* 5:235. doi: 10.3389/fpsyg.2014.00235

Tressoldi, P., and Giofré, D. (2015). The pervasive avoidance of prospective statistical power: major consequences and practical solutions. *Front. Psychol.* 6:726. doi: 10.3389/fpsyg.2015.00726

Vaci, N., Gula, B., and Bilalic, M. (2014). Restricting range restricts conclusions. *Front. Psychol.* 5:569. doi: 10.3389/fpsyg.2014.00569

Weiss, C. H. (1998). *Evaluation*. Saddle River, NJ: Prentice Hall.

West, S. G. (2011). Editorial: introduction to the special section on causal inference in cross-sectional and longitudinal mediational models. *Multivariate Behav. Res.* 46, 812–815. doi: 10.1080/00273171.2011.606710

Williams, M., and Vogt, W. P. (2011). *The SAGE Handbook of Innovation in Social Research Methods*. London: SAGE Publications.

Yang-Wallentin, F., Jöreskog, K. G., and Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Struct. Equ. Model.* 17, 392–423. doi: 10.1080/10705511.2010.489003

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

METHODS
published: 21 January 2016
doi: 10.3389/fpsyg.2016.00032

Analyzing Two-Phase Single-Case Data with Non-overlap and Mean Difference Indices: Illustration, Software Tools, and Alternatives

Rumen Manolov[1]*, *José L. Losada*[1], *Salvador Chacón-Moscoso*[2,3]* and *Susana Sanduvete-Chaves*[2]

[1] *Departamento de Metodología de las Ciencias del Comportamiento, Facultad de Psicología, Universidad de Barcelona, Barcelona, Spain,* [2] *Psicología Experimental, Universidad de Sevilla, Seville, Spain,* [3] *Universidad Autónoma de Chile, Santiago, Chile*

Two-phase single-case designs, including baseline evaluation followed by an intervention, represent the most clinically straightforward option for combining professional practice and research. However, unless they are part of a multiple-baseline schedule, such designs do not allow demonstrating a causal relation between the intervention and the behavior. Although the statistical options reviewed here cannot help overcoming this methodological limitation, we aim to make practitioners and applied researchers aware of the available appropriate options for extracting maximum information from the data. In the current paper, we suggest that the evaluation of behavioral change should include visual and quantitative analyses, complementing the substantive criteria regarding the practical importance of the behavioral change. Specifically, we emphasize the need to use structured criteria for visual analysis, such as the ones summarized in the What Works Clearinghouse *Standards*, especially if such criteria are complemented by visual aids, as illustrated here. For quantitative analysis, we focus on the non-overlap of all pairs and the slope and level change procedure, as they offer straightforward information and have shown reasonable performance. An illustration is provided of the use of these three pieces of information: visual, quantitative, and substantive. To make the use of visual and quantitative analysis feasible, open source software is referred to and demonstrated. In order to provide practitioners and applied researchers with a more complete guide, several analytical alternatives are commented on pointing out the situations (aims, data patterns) for which these are potentially useful.

Keywords: non-experimental, single-case, data analysis, guidelines, methodological quality

OPEN ACCESS

Edited by:
Pietro Cipresso,
IRCCS Istituto Auxologico Italiano,
Italy

Reviewed by:
John G. Holden,
University of Cincinnati, USA
Shana Cornelis,
Ghent University, Belgium

Correspondence:
Rumen Manolov
rrumenov13@ub.edu;
Salvador Chacón-Moscoso
schacon@us.es

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: *22 October 2015*
Accepted: *08 January 2016*
Published: *21 January 2016*

Citation:
Manolov R, Losada JL,
Chacón-Moscoso S
and Sanduvete-Chaves S (2016)
Analyzing Two-Phase Single-Case
Data with Non-overlap and Mean
Difference Indices: Illustration,
Software Tools, and Alternatives.
Front. Psychol. 7:32.
doi: 10.3389/fpsyg.2016.00032

INTRODUCTION

The evidence-based practices movement aims to provide guidelines for carrying out methodologically sound research in fields such as psychology (Apa Presidential Task Force on Evidence-Based Practice, 2006) and special education (Odom et al., 2005). According to this movement, the studies providing solid evidence need to meet a series of criteria related to how an experimental effect is documented and how generality can be established (Maggin et al., 2014). The first of these aspects refers, among other features of the study, to its design and analysis. In the current work, we focus on two-phase designs that do not meet the criteria established by the What Works Clearinghouse *Standards* (Kratochwill et al., 2010), unless they are part of a within-study replication, as in a multiple-baseline design. Two-phase designs may be weaker, from the perspective of internal validity, but they are still used (e.g., Cordery et al., 2010; O'Neill et al., 2013; Finn and McDonald, 2014; Winkens et al., 2014) and can be useful as pilot studies and also due to the fact that establishing the evidence basis of interventions is related to the replication of results and their integration via systematic reviews and meta-analyses (Jenson et al., 2007). Such reviews can offer a comprehensive summary of findings while trying to avoid publication bias, which would take place when excluding studies on the basis of the design. In that sense, it is potentially useful to report the results of all studies and, afterward, consider whether some studies show no differences or negative results (Kratochwill et al., 2001) or whether there are differences according to the design used or the methodological quality of the study. Actually, Gage and Lewis (2014) suggest that experimental control can be used as a moderator variable in meta-analyses.

In this context, the present paper arises from our conviction that practitioners' professional practice, mainly aimed to help individual clients, can also contribute to informing fellow professionals about the results of applying certain interventions. In order to make this contribution possible and in order to be able to translate practice into research certain design and analysis considerations are necessary. The current paper mainly aims to answer two specific questions "What can be done to improve the data analysis in my practice so that its results are more useful to the discipline, despite using a sub-optimal design?" and "How can I easily implement some appropriate analytical techniques?" However, design and data analysis should be considered jointly (Brossart et al., 2014) and this is why we first review some aspects related to how the study is conducted.

Regarding the ways in which a study can be considered as providing evidence, a design implemented as a randomized controlled trial is one option, but it is not always feasible. Another alternative is single-case designs, also referred to as N-of-1 trials (Howick et al., 2011). For this latter option, there are several guidelines on how the studies should be carried out (see Smith, 2012; Maggin et al., 2014, for a review). Two of these guidelines are What Works Clearinghouse *Standards* (Kratochwill et al., 2010) and the Risk of Bias in N-of-1 Trials (RoBiNT) scale by Tate et al. (2013). In brief, the optimal features of a single-case study contributing solid evidence are: to use a design allowing for at least three comparisons between conditions (as in multiple baseline, alternating treatments, and ABAB designs; Barlow et al., 2009); to include randomization in the design when assigning measurement times to conditions (Kratochwill and Levin, 2010); to include blinding of the patient, therapist, and assessor; to show high inter-rater reliability when recording the data (especially useful when by means of observation, Cohen, 1960); to apply the intervention as planned (see also Ledford and Gast, 2014, for a discussion on procedural fidelity); the use a repeatable measure for the target behavior; to use an appropriate data analysis procedure; to assess generalization across other behaviors and settings; and to replicate the results.

These requirements reflect the aspects of a study or a professional practice that moderate the extent to which its findings are "solid evidence" and also affect the practitioner's confidence in the conclusions regarding intervention effectiveness. Accordingly, using a sub-optimal two-phase design such as AB (referred to as "pre-experimental," Kazdin, 1982, or "quasi-experimental," Campbell and Stanley, 1966) is a drawback, but it does not necessarily preclude a study from being useful[1], as there are other characteristics that can increase the credibility in the obtained results. In the present work, we focus on one of these aspects – data analysis – showing how to meet the condition for an appropriate data analysis.

The structure of this article is as follows. First, we comment on the characteristics of non-experimental studies in order to frame a context, where improvements are required (Institute of Education Sciences, 2013). Second, we present an analytical method meeting the criterion for appropriate data analysis; we refer to its strengths, limitations, and alternatives. Third, we apply the analytical method to a real data set. Fourth, we point out several analytically challenging situations and present our own advice to practitioners and applied researchers. With the justification and illustration of the analytical method and the software, we aim to offer practitioners and applied researchers a useful tool, and indications about its alternatives.

NON-EXPERIMENTAL STUDIES

Demonstration of causal relations via experimental designs is considered optimal for building the evidence basis of interventions (Kratochwill et al., 2010; Tate et al., 2013), but everyday practice cannot always meet this requirement (e.g., due to time pressure or to the unethical withholding or removal of a potentially beneficial intervention). However, non-experimental studies can still contribute via in-depth assessment of effects, taking into consideration different sources of information (e.g., visual and numerical analyses of the data gathered, the interpretation of the client, his/her significant ones, and the practitioner) and relying on replication.

Non-experimental studies consisting only of a pre-intervention and post-intervention condition resemble "natural experiments," such as disasters or legislation changes, and they also resemble observational studies in which continuous recording of a single individual is taking place (see **Figure 1** representing the taxonomy of observation studies by Anguera et al., 2001, used in Jonsson et al., 2006). Moreover, an experimental multiple-baseline design across behaviors is similar to an observational plan in which several behaviors of the same participant are recorded each time that a video-taped situation is seen by the observers (i.e., a multidimensional observational

[1] Actually, even pre–post designs with a single measurement before and after an intervention can provide useful evidence (e.g., Pazzagli et al., 2014), especially if clinical significance is assessed, for instance using the Reliable Change Index (Jacobson and Truax, 1991).

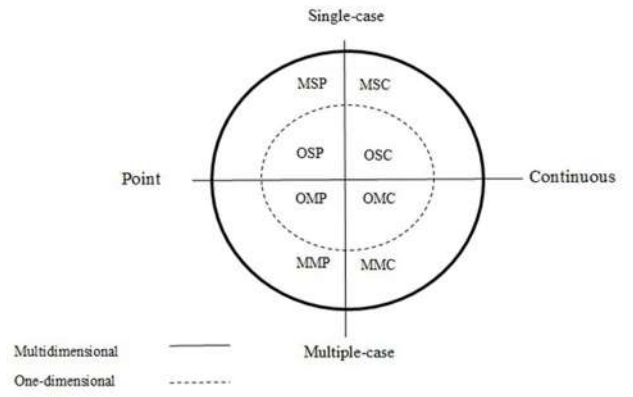


FIGURE 1 | A classification system for gathering data via observation. The acronyms of the figure correspond to the initials of the levels of the three components: behavior (multidimensional or one-dimensional), participant (single-case or multiple-case), and time (point or continuous), respectively.

recording according to Anguera et al., 2001). Another similarity can be seen between a multiple-baseline design across subjects and a multiple-case one-dimensional continuous recording observational plan. However, observational (or non-experimental, in general) and experimental methodology allow reaching different conclusions. Regarding experimental control, the main differences are in: (a) the use of randomization to decide when to introduce and withdraw an intervention, (b) the staggered introduction of the intervention and (c) the replication of effects. Accordingly, in the absence of staggered introduction of the intervention, in an observational study there is less control over alternative explanations of potential behavioral change and the demonstration of intervention effectiveness is not so strong (Kazdin, 1984). Thus, multidimensional single-case continuous observation is not equivalent to multiple-baseline design across behaviors. Moreover, in a natural setting it is usually not possible to choose *at random* when to intervene in order to support internal and conclusion validity (Kratochwill and Levin, 2010). Thus, the conclusions made need to refer to the existence and amount of change in the behavior, but not to the cause for such a change.

THE ANALYTICAL METHOD EXPLAINED

The analytical method is grounded on the “data analysis” item of the RoBiNT scale: controversy remains about whether the appropriate method of analysis in single-case reports is visual or statistical. Nonetheless, two points are awarded if systematic visual analysis is used according to steps specified by Kratochwill et al. (2010, 2013), or visual analysis is aided by quasi-statistical techniques, or statistical methods are used where a rationale is provided for their suitability (Tate et al., 2013, p. 629).

Our proposal is to use the option of “visual analysis aided by quasi-statistical techniques,” where the latter are understood as descriptive measures that do not intend to yield statistical significance values due to various reasons. First, visual analysis is not only frequently used, but it is apparently the only kind of single-case data analysis that researchers seem to agree that is necessary (e.g., Parker et al., 2006; Gast and Spriggs, 2010; Kratochwill et al., 2010; Davis et al., 2013; Fisher and Lerman, 2014). Second, the evidence on visual analysis suggests that its exclusive use is potentially problematic (i.e., visual analysis is not sufficient) and techniques increasing the reliability of visual analysis are necessary (Maggin et al., 2013). Third, we consider that certain quasi-statistical techniques with favorable evidence for their performance can be used as natural complements of the commonly used visual analysis, as they share the emphasis on the same main data features (overlap, level, and trend), whereas the visual aids also take data variability into account and allow comparing projected and actual data. Fourth, applied researchers may not be willing to use the more complex statistical techniques whose results are more easily misinterpreted, in case of incomplete understanding of what exactly is being done with the data. Fifth, the use of inferential statistical procedures may not be fully justified in the absence of random sampling (Edgington and Onghena, 2007). Moreover, an inference to a population is not necessarily an aim of idiographic research (Johnston and Pennypacker, 2008) that focuses on the needs and the improvement of the individual clients. Sixth, easy to use software is available for the descriptive statistical procedures recommended here.

SYSTEMATIC VISUAL ANALYSIS

Rationale

Visual analysis has been and still is popular among professionals in their everyday psychological practice (Robey et al., 1999; Parker and Brossart, 2003) and is still advocated for (Lane and Gast, 2014) and used as a gold standard for assessing quantitative procedures (Wolery et al., 2010). Visual analysis has been considered both appropriate and sufficient for data gathered longitudinally (Michael, 1974). However, this sufficiency has been defended only for experimental studies (Sidman, 1960), which points at the need for complementing it with a quantitative procedure.

Tate et al. (2013) advise for systematic visual analysis and it necessarily starts with assessing the baseline, specifically, whether the intervention can be introduced or it should be postponed until stability is reached (Barlow et al., 2009). Alternatively, deterioration in the behavior of interest would suggest even more clearly the need for intervention. In that sense, deterioration is not expected to interfere with subsequent conclusions about intervention effectiveness (Kazdin, 1978), given that it allows exploring whether an intervention reverts the situation. Nonetheless, it is possible to assess intervention effectiveness even when the behavior is already improving before the intervention itself, as it will be shown later.

The specific data aspects, which are foci of attention, are the amount of overlap between data in the different conditions, within- and between-phase variability, slope and level change (SLC; Kratochwill et al., 2010; Lane and Gast, 2014). A more objective assessment of the degree to which data share the same values (i.e., overlap), whether levels and trends are similar across conditions, and whether data become more stable or more variable after the intervention can be done using visual aids instead of relying on naked-eye impressions. Finally, visual analysis focuses on the whole data pattern (Parker et al., 2006) in order to assess whether it resembles the expected one, that is, a consistent improvement only during intervention. Kratochwill et al. (2010) summarize the overall assessment as a comparison between projected and actually obtained measurements. Specifically, in two-phase designs, it is relevant to project the baseline (in case it is stable or presents trend stability) into the intervention phase and compare this projection with the real treatment phase data.

Potentially Useful Tools

The assessment of overlap can be done using visual aids, such as range lines, as provided by the SCDA plug-in (Bulté and Onghena, 2012[2]) for R-Commander. The upper left panel of **Figure 2** shows an example with the data reported by Taylor and Weems (2011) for a participant called Elizabeth. This graph suggests a minor overlap between the observations. Regarding the assessment of changes in level, the same software can be used to superimpose, for instance, the median of the behavioral observations in the pre-intervention and post-intervention conditions.The upper right panel of **Figure 2** shows an example with the same data and suggests that there has been a reduction in the level of target behavior. However, the median is not very useful for the post-intervention observations in which there is a clear downward trend.

Regarding the assessment of changes in slope, two situations should be considered: when pre-intervention data are stable and when baseline data show an upward or downward trend. In case of stability, it is possible to use the stability envelope (Lane and Gast, 2014) or the two-standard deviations band used in statistical process control (Callahan and Barisa, 2005). The two-standard deviations band implies computing the average of the data for a specific condition and representing it with a solid line. The standard deviation of the same data is also computed and two dashed lines are represented: one located two standard deviations below the mean and the other two standard deviations above. The basis of this procedure is that, for a normally distributed variable, few points (less than 5%) are expected to be out of these limits in case there is no change in the behavior with the introduction of the intervention. However, we suggest using it only as visual aid and not as a formal statistical procedure, as the data cannot be reasonably assumed to be normal, continuous, or independent. This visual aid is implemented in R Core Team (2013) code[3] that only requires inputting the data and specifying the number of pre-intervention observations. As an example see the lower left panel of **Figure 2**, indicating that the reduction in behavior is beyond what is expected only by random variability as there are multiple observations with values smaller than the lower limit.

In case the pre-intervention data show a trend, it is necessary to compare the projection of this trend and the actually obtained measurements (Kratochwill et al., 2010). For that purpose, there is another potentially useful R code[4] which allows applying the stability envelope to the trend line: (a) estimating split-middle trend (Miller, 1985), (b) projecting it into the next phase, and (c) constructing an envelope around it. The envelope can be constructed on the basis of the baseline median[5], so that the lower limit is located 25% of the median below the estimated split-middle trend and the upper limit at the same distance above it (Lane and Gast, 2014). In case 80% of the data are within those limits, this would indicate trend stability, that is, it would suggest that no change in slope has been produced with the introduction of the intervention. For using this code only data input is required before copy-pasting it in R. The lower right panel of **Figure 2** shows an example with Elizabeth's data. Given that the projected trend and its stability envelope are lower than the actual observations, this is the only piece of graphical information that does not suggest improvement in the behavior, but practitioners should be cautious when trend is estimated from as few as four observations and when it is projected farther away in time into values that are out of the range of possible measurements (Parker et al., 2011b).

Another aspect assessed is whether the introduction of the intervention has led to an immediate change in the behavior. Moreover, the duration of the change (maintained or transitory) is also taken into account in order to evaluate the strength of the intervention. A structured guide on visual analysis is offered by the What Works Clearinghouse *Standards* (Kratochwill et al., 2010; see also the application and a scoring procedure by Maggin et al., 2013) and by Lane and Gast (2014).

Limitations

Despite these guidelines on visual analysis, there are still no soundly based formal decision rules for all data aspects that are visually assessed (Kazdin, 1982) and objective and replicable outcomes are also missing (Robey et al., 1999). These two drawbacks might be among the reasons for the frequently reported inadequate performance of visual analysts (Gibson and Ottenbacher, 1988; Ottenbacher, 1990; Danov and Symons, 2008; Ximenes et al., 2009; see also Ninci et al., 2015, for a recent meta-analysis reporting insuficient interrater agreement, especially among single-case experts). Moreover, the visual analysts' decisions are not directly useful for documentation or for meta-analysis (Busse et al., 1995), which would allow establishing the evidence basis for interventions (Jenson et al., 2007), especially as generalization in single-case studies depends on replication[6] rather than on random sampling and statistical inference. As a result of these limitations, there is a consensus that visual and quantitative analyses should be used jointly (Franklin et al., 1996; Fisch, 2001; Houle, 2009; Harrington and Velicer, 2015).

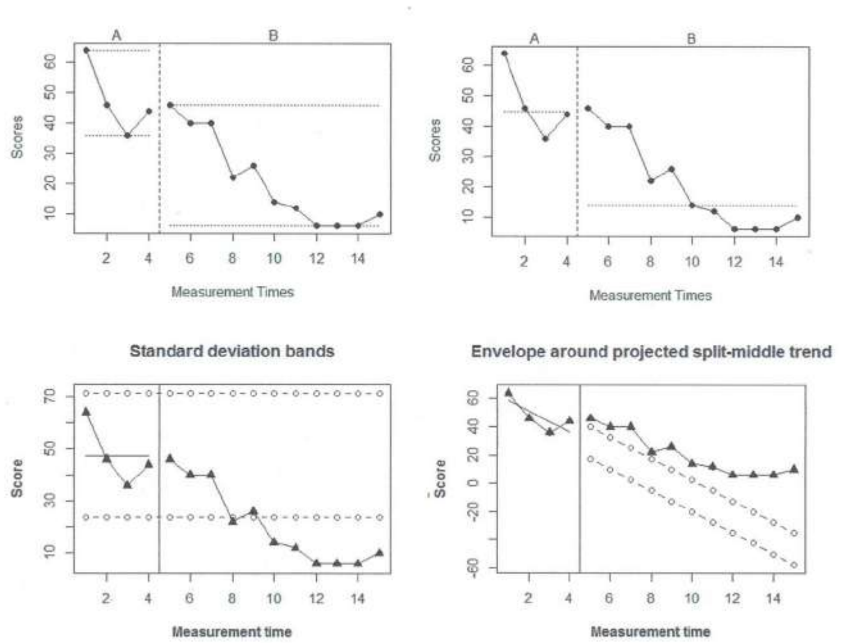
[2] http://cran.r-project.org/web/packages/RcmdrPlugin.SCDA/index.html

[3] https://dl.dropboxusercontent.com/s/elhy454ldf8pij6/SD_band.R

[4] https://dl.dropboxusercontent.com/s/5z9p5362bwlbj7d/ProjectTrend.R

[5] Another option is to take into account the baseline data variability, operationally defined as the interquarile range, when constructing the trend stability envelope (Manolov et al., 2014).

[6] Kratochwill et al. (2013) recommend that the findings be replicated in at least five different studies, conducted by at least three different research teams on a total of 20 participants or more (i.e., the 5-3-20 rule).



Standard deviation bands

Envelope around projected split-middle trend

FIGURE 2 | An illustration of visual sides using Taylor and Weems (2011) data on a participant called Elizabeth. Upper left panel—range bars. Upper right panel—medians. Lower left panel—2-standard deviation bands. Lower right panel—stability envelope around split middle trend.

QUANTITATIVE ANALYSES RECOMMENDED

Our choice of procedures [non-overlap of all pairs (NAPs); Parker and Vannest, 2009 and SLC; Solanas et al., 2010a] is based on the six criteria detailed below, although alternative quantifications are provided later in this article.

Criterion 1: Simple to Compute

The techniques are relatively simple to compute and offer straightforward interpretations for practitioners who are not experts in statistics (as the Institute of Education Sciences, 2013, suggests). The calculation does not entail statistical decisions about the likelihood of obtaining such a large difference under the null hypothesis. This criterion also relates to the need for easily trainable procedures (Fisher et al., 2003).

Criterion 2: Complementary to Visual Analysis

This criterion is related to the popularity of visual analysis among practitioners (Parker and Brossart, 2003), which makes necessary to develop and promote suitable complements to it. NAP and SLC are actually based on relevant visual criteria (i.e., data overlap, change in slope and in level) and thus potentially useful as complements[7]. Specifically, visual inspection can be used to assess the adequacy of the baseline as a reference for comparison. The change identified visually can then be quantified in an objective manner. The numerical values also offer information that can be communicated among researchers and professionals and used for further analyses with different analytical techniques or as part of research synthesis (e.g., NAP was used in the meta-analysis by Jamieson et al., 2014, whereas the new developments on SLC make possible its comparability across studies; Manolov and Rochat, 2015).

Criterion 3: Synergic Application

Wolery et al. (2010) criticized non-overlap methods for omitting relevant data aspects such as level, trend, and stability or variability: SLC partially addresses this issue and it also responds to Beretvas and Chung's (2008) suggestion for quantifying separately level and slope change. Moreover, SLC yields unstandardized results, which help assessing the practical importance of the behavior change when using meaningful measures (Grissom and Kim, 2012) such as the number of tantrums or the number of self-injurious behaviors. In contrast, NAP is bounded, which allows comparisons and quantitative integrations. Thus, NAP and SLC can be used jointly as they provide different information. Specifically, NAP is an ordinal measure (Solomon et al., 2015) that does not distinguish between conditions once complete overlap is achieved. In contrast, SLC can be used even in absence of overlap to quantify how different the measurements belonging to different phases are.

[7] Wolery et al. (2010) found that no overlap technique had highest agreement with visual analysts for both data with and without a change. However, they did not include NAP or Tau-U (Brossart et al., 2014) in their study, and these two non-overlap indices are considered to be superior, given their more solid statistical basis and greater statistical power according to the review performed by Parker et al. (2011a).

Criterion 4: Absence of Assumptions and Restrictions of Use

The procedures used here do not make explicit *a priori* assumptions about independence or homoscedasticity of the data, as serial dependence is likely to present in data obtained from the same individual (Matyas and Greenwood, 1996). There are also no specific design requirements.

Criterion 5: Appropriate Performance

In relation to the previous point, there is evidence that their performance is appropriate for a variety of single-case data patterns (Manolov et al., 2011). NAP is a suitable indicator when data is stable and even when data is variable. In contrast, in such situations visual analysis is more difficult to perform and means and medians are not informative and trends are not estimated with precision. On the other hand, NAP is not suitable when the data show improving trend, but SLC can be applied in such a situation – this complementarity relates to Criterion 3 "Synergic application." SLC is useful for separately quantifying the change in level and the change in slope in potentially meaningful terms. In relation to this criterion, it is important to discourage the use of methods for comparing conditions that have been shown not to perform appropriately, such as the binomial test applied after the split-middle method (Crosbie, 1987) which does not control for Type I error rates, ITSACORR which presents modeling flaws (Huitema et al., 2007), or the C-statistic (Young, 1941; Tryon, 1982; used by Fabio et al., 2013), which is actually an estimator of autocorrelation (DeCarlo and Tryon, 1993).

Criterion 6: Reduced Likelihood of Misinterpretation

Using descriptive measures like the ones provided by NAP and SLC makes it less likely for applied researchers to make inferences, which would be statistically incorrect in absence of random sampling of the participant or of the behavior of interest (Barlow et al., 2009). We consider that inferential statistical techniques are more susceptible to being misunderstood and to prompt researchers to make dichotomous decisions (Cohen, 1994) about intervention effectiveness or behavioral change. In case inference is desired, we recommend causal inference, instead of population inference, in line with the recommendations by Heyvaert et al. (2015).

Non-overlap of All Pair

Non-overlap of all pairs is an improvement of the Percent of non-overlapping data commonly used for quantifying the degree to which the measurements pertaining to each phase share the same values (Scruggs and Mastropieri, 2013). It represents the number of non-overlapping data relative to all possible comparisons and it is actually identical to the non-parametric version of the probability of superiority (Grissom, 1994), which is related to the common language effect size (McGraw and Wong, 1992). When a decrease in the behavior is expected, as in the example provided later, the formula for this indicator can be written as $(\#(X_{pre(i)} > X_{post(j)}) + 0.5\#(X_{pre(i)} = X_{post(j)}))/n_{pre}n_{post}$ where X_{pre} and X_{post}, which represent the values of the pre-intervention and post-intervention phases, respectively, with $i = 1, 2, \cdots, n_{pre}$ and $j = 1, 2, \cdots, n_{post}$, and # denotes the number of times that the inequality or the equality is true. Given that each data point of the pre-intervention phase is compared to a data point from the post-intervention phase there is a total of $n_{pre}n_{post}$ comparisons, where n_{pre} and n_{post} denote the number of measurements in the first and second phase, respectively. In each of these comparisons, a non-overlap occurs when a post-intervention measurement represents an improvement over a pre-intervention measurement, with ties counting as half a non-overlap. To obtain the index value, the number of non-overlapping pairs is divided by number of comparisons. This value can be interpreted in two different ways. One the one hand, it represents the proportion of comparisons for which intervention phase data improve baseline data. On the other hand, it can be conceptualized as the probability that a randomly selected post-intervention data point will improve (here, be smaller than) a randomly selected pre-intervention data point. The NAP can be computed via the online calculator http://www.singlecaseresearch.org/calculators/nap by Vannest et al. (2011), where it is only necessary to enter the data from the different conditions in separate columns. It is also part of the output ("A vs. B" comparison) of the R code for Tau-U https://dl.dropboxusercontent.com/u/2842869/Tau_U.R (Brossart et al., 2014), which requires loading a data file with a single comma-separated column including "Time" (1, 2, . . ., $n_{pre}+n_{post}$), "Score" (denoting the measurements) and "Phase" denoting the condition (n_{pre} times the value of 0 followed by n_{post} times the value of 1).

Slope and Level Change

Slope and level change quantifies two aspects of behavior's evolution after a change in the conditions: change in slope and change in level. Actually, this procedure first estimates pre-intervention linear trend ($\widehat{\beta_A}$) as the average of the differenced first phase measurements, that is, $\widehat{\beta_A} = \sum_{i=1}^{n_{pre}-1}(X_{i+1} - X_i)/(n_{pre} - 1)$. Baseline trend is thus the average increase (or, if negative, decrease) from one baseline measurement occasion to the next one. This estimation can inform about the characteristics of the data before an intervention is introduced. Moreover, baseline trend is removed from the whole data series so that it does not affect the quantification of the effects of the intervention. Technically, each data point is corrected according to its position in the series of observational sessions. This initial step allows for applying an intervention even when the theoretically undesirable linear improvement is present already during the assessment period. Thus, SLC would show whether there is an effect of the intervention beyond the initial improvement. After the correction it is assumed that the pre-intervention phase shows zero trend (i.e., stable data) and thus the trend present in the post-intervention phase actually represents an effect (i.e., a change in slope). This effect is estimated in the same manner as in the initial step, that is, as the average of the differenced (and already detrended) post-intervention measurements: $\widehat{SC} = \sum_{j=1}^{n_{post}-1} (\widetilde{X}_{j+1} - \widetilde{X}_j)/(n_{post} - 1)$, where $\widetilde{X}$ represent detrended values (i.e., after eliminating pre-intervention trend), instead of the original measurements. Therefore, the intervention phase estimate of trend presents the average increase (or, if negative, decrease) from one intervention phase measurement occasion to the next one, after controlling for baseline linear trend. For instance, the slope change estimate reflects the average decrease in the number of tantrums in a child with each successive post-intervention measurement, that is, a progressive change.

Once slope change is estimated, post-intervention trend is removed in order to obtain a net estimate of the change in level. This way of proceeding is similar to what is done in ARIMA models, before obtaining a quantification of change in level (see Harrington and Velicer, 2015). Net change in level is estimated as the difference between the average of the corrected post-intervention measurements and the average of the corrected pre-intervention measurements. The expression for this step is $\widehat{LC} = \sum_{j=1}^{n_{post}} \tilde{X}_j/n_{post} - \sum_{i=1}^{n_{pre}} \tilde{X}_i/n_{pre}$, where $\tilde{X}$ represents post-intervention measurements with both pre-intervention trend and post-intervention trend (i.e., slope change) removed and $\tilde{X}$ represents pre-intervention measurements with pre-intervention trend removed. The net level change estimate quantifies, for instance, the average decrease of tantrums in a child after the intervention, once slope change has been taken into account. Thus, it can be conceptualized as a quantification of an abrupt and maintained effect. The SLC can be computed using R code https://dl.dropboxusercontent.com/s/ltlyowy2ds5h3oi/SLC.Ror via the R-Commander Plug-in offering point-and-click menus, available at http://cran.r-project.org/web/packages/RcmdrPlugin.SLC/index.html. For obtaining the numerical results and a graphical representation of the original and detrended data, both options only require inputting the values of the observations and specifying the pre-intervention phase length.

ALTERNATIVES FOR QUANTITATIVE ANALYSIS

There is currently no consensus on which the optimal quantitative procedure for single-case designs is (Kratochwill et al., 2010; Smith, 2012), as the RoBiNT scale also reflects (Tate et al., 2013). For a comprehensive review of most currently available techniques the interested reader should consult the state-of-the-art information provided in the Special Issues of the *Journal of School Psychology* in 2014, volume 52, issue 2 (e.g., Shadish et al., 2014; Swaminathan et al., 2014) and of *Neuropsychological Rehabilitation* also in 2014, volume 24, issues 3-4 (e.g., Borckardt and Nash, 2014; Brossart et al., 2014; Heyvaert and Onghena, 2014). Here, we provide brief comments on the strengths and limitations of several analytical alternatives, which in some cases may be more appropriate than NAP and SLC included in the analytical method suggested.

Considering specifically observational studies in which data is recorded continuously within a session, it is possible to follow an analytical approach different from the one used in single-case designs, namely, to apply sequential analysis to explore whether the occurrence of some behaviors make more or less probable that other behaviors take place (Bakeman and Quera, 2011). Additionally, longer series of data gathered across time can be analyzed using Markov chains or analyses of rhythm, according to the aims of the study (Suen and Ary, 1989).

Starting our discussion from procedures similar to the ones included in the analytical method, Tau-U (Parker et al., 2011b) is closely related to NAP and it is preferable when pre-intervention trend is present in the data. For both Tau-U and NAP p-values have been offered, although their basis has not clearly been explained in the presence of autocorrelation. However, Tau-U is interpretatively and computationally less straightforward than NAP (i.e., Criterion 2 “Complementary to visual analysis” is met to a lesser extent). For instance, even in case a baseline trend is generally deteriorating, if there is a single improving value in the baseline phase, as compared to a previous baseline data point, this would reduce the value of the non-overlap index. Thus, in case trend is not reasonably clear, Tau-U can be an excessively conservative procedure (i.e., it would overcorrect). Furthermore, more evidence is required on its performance (thus the abovementioned Criterion 5 “Appropriate performance” is not fully met, as Parker et al., 2011a,b, offer only applications to real data, but no simulation study).

Regarding procedures quantifying average differences, similar to the SLC, the d-statistic (Shadish et al., 2014) has to be mentioned. We highlight here the d-statistic developed by Shadish et al. (2014), which has been created specifically for single-case designs rather than the d-statistic described by Busk and Serlin (1992; approach one[8]), recommended by Beeson and Robey (2006), for two reasons: (a) the latter is an adaptation of the group designs indicator and does not take into account autocorrelation, while it has been shown to be somewhat affected by autocorrelation (Manolov and Solanas, 2008); and (b) its sampling distribution in single-case studies is unknown (Beretvas and Chung, 2008). In contrast, the d-statistic developed by Shadish et al. (2014), offers a standardized measure of the mean difference with a solid statistical basis offering the possibility to estimate the index variance for future meta-analyses. So far, it has been developed for AB, reversal (e.g., ABAB) and multiple-baseline designs and assuming that pre-intervention data is stable, assuming that within-case residuals and between-case variation do not change over time. Thus, this procedure fails in terms of Criterion 4 “Absence of assumptions and restrictions of use.” Some potential drawbacks include: (a) its computation requires several cases per study; and (b) the calculations are potentially difficult to understand by applied researchers with less statistical knowledge and require the use of software, such as the R code provided in the appendix of the Shadish et al. (2014) paper. Hence, the d-statistic is preferable to SLC when there is more than one participant per study and the aim is to obtain a standardized

[8] This indicator is equivalent to Glass' Δ (Glass et al., 1981), as it divides the mean difference by the standard deviation of the pre-intervention phase data.

measure, but it is not suitable when pre-intervention trend is present and when the focus on a specific client.

Generalized least squares regression analysis (Swaminathan et al., 2014) also enables computing an effect size index. Its strengths include the fact that it can take into account changes in level and in slope (although they are quantified as part of the same overall indicator, unlike SLC), the versatility in modeling (e.g., controlling for linear and non-linear trends), and that it deals explicitly with autocorrelation. However, autocorrelation estimation has been shown to be problematic (Solanas et al., 2010b) and the analytical procedure requires several steps, some of them taking place iteratively (i.e., Criterion 1 "Simple to compute" is not met). This procedure is applicable to longer data series for which autocorrelation can be estimated with greater precision. Moreover, we recommend that practitioners work together with a statistician, so that the analysis can be properly run. Brossart et al. (2006) compared the agreement between visual analysis and several regression-based approaches and the best performer in this terms (related to Criterion 2 "Complementary to visual analysis") was Allison and Gorman's (1993) method, which is however affected by autocorrelation (Manolov and Solanas, 2008). The generalized least squares approach was not yet proposed by the time Brossart et al. (2006) conducted their study and more evidence is necessary to assess its performance.

Multilevel models are an extension of piecewise regression and can be used to model several data aspects (e.g., trend, autocorrelation, heterogeneous data variability across phases) and they yield estimates of the change in the same measurement units as the target behavior and their statistical significance (Moeyaert et al., 2014a). The main drawbacks of multilevel models are the problematic estimation of variance (Ferron et al., 2009), their relative complexity for applied researchers with less statistical knowledge and the fact that they the replication of the intervention in several participants. Actually, such a complex procedure is more suitable for more complex design structures that the two-phase AB (Moeyaert et al., 2014b). Finally, most implementations of this analytical procedure have been done in commercial software (e.g., Moeyaert et al., 2014a include SAS code in their article).

An effect size index can also be computed from interrupted time series analysis via ARIMA (autoregressive integrate moving average) models, which allow controlling for trend and autocorrelation (Simonton, 1977). The main difficulties of this option are the need for long data series and the problematic initial model identification step. However, there have been suggestions for using some general models that make model identification unnecessary (Harrop and Velicer, 1985). A recent application of ARIMA models has shown that these can be applied to two-phase data, but there might be convergence problems and, more importantly, the agreement with visual analysis is low (Harrington and Velicer, 2015). We consider that this latter drawback and the relative complexity of the technique make it less attractive to applied researchers with no statistical expertise.

Statistical significance (i.e., *p*-values) can be estimated for *d* and the generalized least squares procedure on the basis of the comparison between the test statistic and a theoretical reference (the sampling distribution) and allows making inference about the population from which the individual was drawn. In contrast, randomization tests (Heyvaert and Onghena, 2014) yield a *p*-value on the basis of a comparison between the test statistic and an empirical reference –the randomization distribution. In the current context of two-phase studies, this reference is the distribution of the test statistic values quantifying the difference between the two conditions for each possible intervention start point (i.e., for each possible way in which the data series can be split into two; Edgington, 1980). For this analytical option the inference is restricted to the case studied, referring to the likelihood of obtaining such a large difference in case the intervention was ineffective. Randomization tests are versatile in terms of test statistic to use (e.g., it can be an effect size such as a non-overlap index) and offer flexible options for dealing with different situations (e.g., Levin et al., 2012). However, the necessary randomization as part of the data collection process is both a strength (Kratochwill and Levin, 2010) and a limiting characteristic (Fisher and Lerman, 2014) in a clinical setting (i.e., Criterion 4 "Absence of assumptions and restrictions of use" is not met). Moreover, in certain conditions Type I error rates are not controlled (Manolov et al., 2010). Randomization tests can be recommended when the aim is to obtain statistical significance and the point(s) of change in the conditions can be chosen at random. Randomization tests are also accompanied by freely available software (Bulté and Onghena, 2013; Levin et al., 2014).

Another procedure using an empirical reference distribution is simulation modeling analysis (SMA; Borckardt and Nash, 2014). In SMA, data are generated with the same autocorrelation as estimated from the data, but with no difference between the conditions, thus representing the null hypothesis of identical behavioral level across conditions. The *p*-value represents the likelihood of the outcome, computed as a point biserial correlation between the measurements and a dummy variable representing the condition (0 = without intervention, 1 = with intervention). This approach is intuitive, takes autocorrelation into account, and it can be implemented via the software available freely at http://clinicalresearcher.org/software.htm. However, so far the evidence on its performance (i.e., Criterion 5 "Appropriate performance") is not sufficient. Finally, as the focus of is put on the *p*-value, which may enter in conflict with Criterion 6 "Reduced likelihood of misinterpretation."

Whereas SMA uses Monte Carlo methods or bootstrap for generating samples and estimating the likelihood of the value of test statistic in case there is not difference between conditions, bootstrap has also been suggested for single-case as a way of reducing bias and estimating standard errors (McKnight et al., 2000) and specifically for estimating confidence intervals of regression-based *R*-squared values (Parker, 2006). This option has not received much attention lately and it is unclear whether applied researchers would be willing to use it.

Another computer-intensive option could be the Monte Carlo based method for modeling non-linearity proposed by Theiler et al. (1992). However, modeling non-linear patterns can also be achieved without prior knowledge and without the need to specify a model, by using local regression (LOESS; Jacoby, 2000; Solmi et al., 2014). We consider LOESS to be more practical for applied researchers than the Theiler et al. proposal. Moreover, randomization tests are also more parsimonious as they require no assumptions about the process generating the data or about random sampling. Actually, Theiler et al. (1992) mention this option as rank statistic approach for obtaining *p*-values. Randomization test offer the advantage of not only mimicking the preserved data features (such as mean and standard deviation), as expressed by Theiler et al. (1992), but they actually preserve the whole data series and its order, taking advantage of the different possible moments of change in phase, when such moments are determined at random.

A simplified summary of these general recommendations regarding the use of the analytical techniques can be found in **Figure 3**.

INTERVENTION EFFECTIVENESS IS NOT ONLY DATA ANALYSIS

Assessing the relevance of an intervention cannot be constrained solely to visual and descriptive or inferential statistical analyses. It is important to assess aspects such as quality of life (Kendall, 1999), whether the behavior has moved from dysfunctional to functional ranges (Kazdin, 1999), without forgetting subjective evaluation (Hugdahl and Ost, 1981). Regarding the latter, Kratochwill and Levin (2010) highlight the need to get to know the perceptions of the client and of significant others. According to the specific context being studied, these significant others would be the family members (parents, siblings, marital partner), the teacher, the coach, or the boss (as figure with a higher hierarchical role), and friends, classmates, or colleagues (at the level of "peers"). Kazdin (1984) has referred to these groups of people as "paraprofessionals," as they help detecting the behavior that requires intervention and they can also be the agents reinforcing the behavior of interest (e.g., a mother reinforcing a child's disruptive behavior by paying attention to it) or producing stimuli for discriminating conditions in which certain types of behavior are desirable (e.g., a boss may encourage jokes with one type of clients and more distant behavior with others).

THE ANALYTICAL METHOD APPLIED

In the present section, we will illustrate the application of the analytical method and the information that can be obtained via visual and quantitative analyses, while also considering substantive criteria. This application focuses on the family context, where it is common to gather data before and after an intervention (Crane, 1985). One of the empirically supported interventions in this context is the Parent Child Interaction Therapy (PCIT; Eyberg et al., 2008), which has been reported to increase positive parent behavior and reduce child behavior problems (Borrego et al., 2006). For the current example, the data gathered by Bagner et al. (2009) will be used. The participants are a 23-months-old premature-born child displaying difficult behaviors and his mother. The application of the PCIT focuses on teaching parenting skills in order to improve the interaction with the child and to decrease his externalizing behavior. Teaching takes place in two phases. First, child-directed intervention (CDI) takes place. It is similar to play therapy: the child is the leader and the parent has to learn how to act positively (e.g., praising the child, imitating the child's play). Second, parent-directed intervention (PDI) phase occurs. It is similar to clinical behavior therapy: the parent is more directive and has to improve her way of disciplining so that a greater compliance is achieved. In order to assess intervention effectiveness, several sources of information are used: parent reports provided via inventories, observation of the parent–child interaction, and physiological measurements. In the running example, we focus on the parent weekly reports obtained via the Intensity scale of the Eyberg Child Behavior Inventory (ECBI; Eyberg and Pincus, 1999) on disruptive behavior, although a complete assessment entails exploring whether all available information converges to the same conclusion. The Bagner et al. (2009) ECBI data were chosen here given that there is a cut-off point at a T-score of 60 which indicates clinically significant results and eases the interpretation in substantive terms. The data gathered[9] on the ECBI scale are represented on **Figure 4**. The upper panel contains ordinary least squares trend lines provided by the SCDA plug-in for R, the middle panel contains split-middle trend for the first phase, and the lower panel represents the application of the two-standard deviations band fit to the first condition's data and projected into the second one.

Firstly, when visually inspecting the data, it has to be kept in mind that both phases are treatment phases and thus in both some reduction in child's behavior is expected and desired. Moreover, it has to be taken into account that the pre-treatment (i.e., actual baseline) value is 82, equal to the first CDI phase measurement. At the beginning of the first phase there is actually a reduction, but then a new increment starts. Considering this alternating pattern the CDI does not seem especially effective. Given the amount of variability in the first phase, neither the central tendency measure (mean represented on the lower panel of **Figure 4**), nor the different types of trend fitted (upper and middle panel) seem to represent the data well-enough. This can hamper the comparison between this condition and the subsequent one.

Once the intervention is introduced, there is apparently a decrease in the ECBI score on disruptive behavior. The downward trend is stable, as shown by the good fit of the ordinary least squares regression line to the data (upper panel of **Figure 4**). For such data it is not meaningful to discuss level or variability around a mean or a median level; actually variability is only assessed looking at the (small) distance of the measurements from the fitted trend line.

Comparing the two phases in terms of overlap, the values in the beginning of the PDI-phase are similar to the ones in the CDI-phase, but not so in the end. Comparing levels is not meaningful. Comparing trends is hindered by the lack of fit of the trend lines to the CDI data, but if we focus on the last four (out of five)

[9] We would like to thank Dr. Daniel Bagner for kindly offering the raw data for re-constructing their original figure.

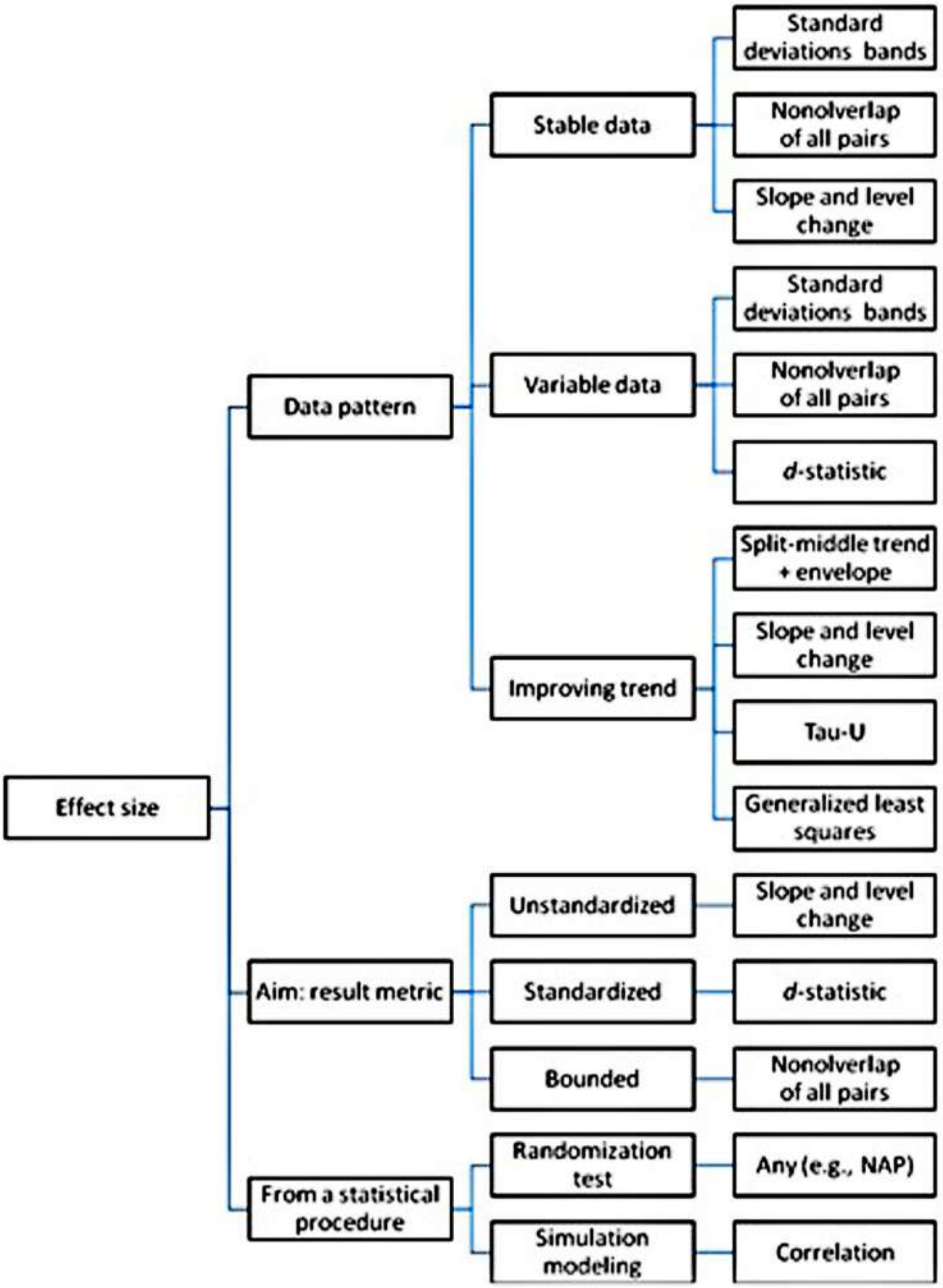


FIGURE 3 | Graphical (simplified) summary of the recommendations regarding the use of several analytical techniques for single-case experimental and pre-experimental designs.

CDI measurements, there is a deterioration that is reverted with the introduction of the PDI: thus a change in slope has taken place. The comparison between projected and actual data is done in two ways, projecting the baseline mean with limits based on the baseline standard deviation and projecting the split-middle trend line with limits based on 25% of the baseline median. In this case, both approaches lead to a very similar graphical representation, which is well-aligned with the conclusion that the last PDI data points are clearly lower that what would be expected (i.e., values within the limits) in case there was no difference between the two interventions. Additionally, we should consider that Bagner et al. (2009) collected a post-treatment measurement equal to 38 – a value even lower than the last PDI-phase measurement and so the downward trend seems to continue, which could be interpreted as maintenance of the effect.

Secondly, regarding quantitative analyses, the NAP performs 50 comparisons, given that $n_{pre} = 5$ and $n_{post} = 10$, in which there are 19 full overlaps, that is, 19 cases in which a CDI datum is better (here, lower) than a PDI measurement, 0 ties, and 31 cases in which a PDI measurement is better than a CDI data point. (Lower rather than greater values are considered as overlaps, given that the aim is to reduce the disruptive behavior and thus also the ECBI T-score.) The value yielded by NAP is 62.00%, which can be interpreted as the percentage of PDI measurements that improve the CDI measurements. Therefore, the index does not suggest that the change is especially salient, given that the value is only slightly higher than the one expected by chance (50%) and it is within the range of values (0–65%) denoting small effect according to Parker and Vannest (2009). However, it has to be considered that this may be due to the fact that the effect is delayed. The data pattern is not specifically easily analyzed by the SLC either. The procedure estimates the CDI-phase trend as −2.25, which represents an average of approximately two T-score units reduction for each CDI measurement time. However, this value does not reflect the visual impression, provided that this phase shows a specific kind of variability (i.e., an alternating pattern). Correcting for this initial phase trend, the slope change estimate is −1.64, that is, nearly two T-score points average reduction for each PDI measurement time. This quantification reflects to some extent the visual impression of slope change. SLC's estimate of the net change in level is positive, 18.15, which contrasts with the visual impression of the graphed data.

Thirdly, focusing on substantive criteria, Bagner et al. (2009) summarize their results in terms of improved parent practice and increased child compliance. In fact, while the former result stems from observation and evaluation by the authors, the latter is based in reports from the parents (i.e., the paraprofessionals). Regarding the ECBI scores, the last three scores during the PDI phase fall out of the clinical range, indicating that a practically significant change in behavior of the child has taken place. Interestingly, these same three scores also fall out of the two-standard deviations band and out of the split-middle trend stability envelope represented in the middle and lower panels of **Figure 4**. To complement this assessment, the authors report that at a 4-months follow-up the results of the ECBI remained in the normal range (the value was 47), which increases the confidence in the importance of the behavioral change. Finally, it should be noted that Bagner et al. (2009) comment explicitly the "inability to conduct statistical analyses" (p. 475), which suggests that informing applied researchers about analytical options for two-phase single-case designs, as we intend with the current paper, is a timely endeavor.

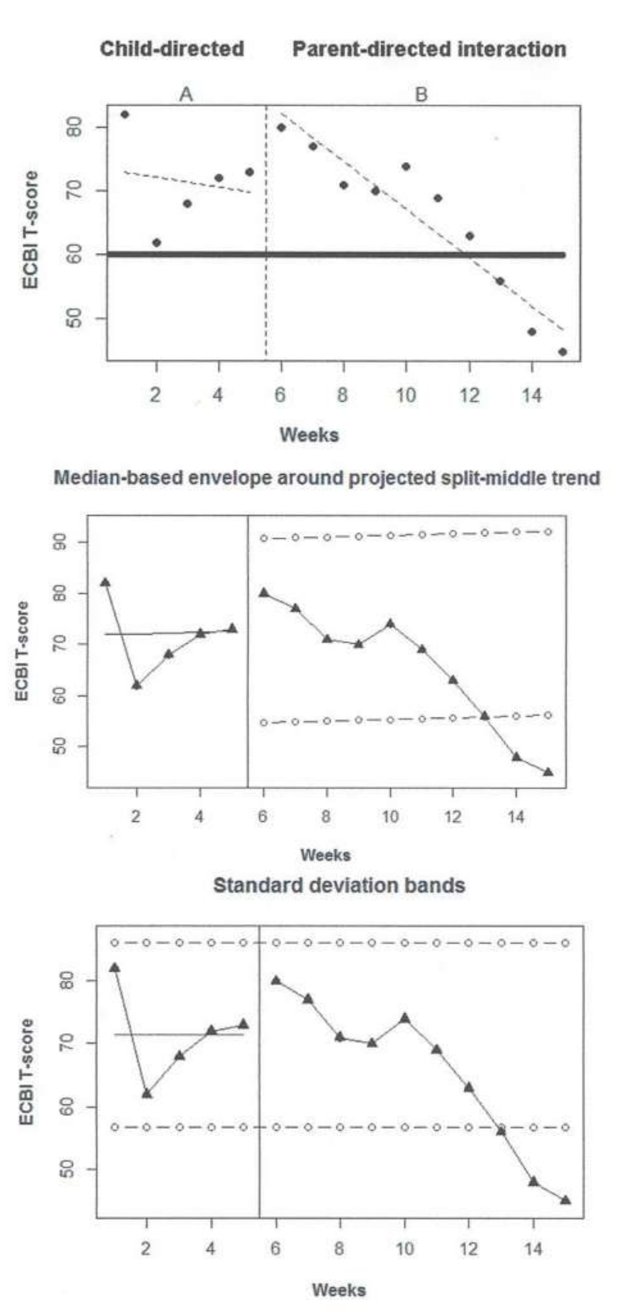


FIGURE 4 | Graphical representations of the Bagner et al. (2009) data gathered through observation in the family context: upper panel – trend lines; middle panel – split middle and trend envelope; lower panel – standard deviation bands.

The main conclusion of this application of the analytical method is that visual analysis is necessary for focusing at different aspects of the data, such as an unstable baseline which is not well-represent by mean or trend lines, a somewhat delayed slope change, and a considerable amount of overlap only in the beginning of the second condition but not at the end. The variability and relative shortness of the first phase (although it meets the current standards of five measurements; Kratochwill et al., 2010) have to be kept in mind when comparing it to the measurements obtained in the subsequent condition. In the current case, the visual aids reflected this variability and suggested a similar conclusion as the one based on substantive criterion expressed as a cut-off point. All this information is critical for interpreting correctly the numerical yielded by descriptive statistical procedures. Actually, we preferred to use a data set that is challenging for the quantitative analyses in order to alert applied researchers on the need to interpret numerical values with caution and to use all information available; we also wanted to avoid doubts about the data being picked up only to show the quantification in a positive way (Fisher and Lerman, 2014). Finally, the follow-up measures, the parent-report and the physiological measures recorded by Bagner et al. (2009) also contribute to building solid conclusions. The two-phase design may not be sufficient for establishing a causal effect in a scientifically sound way, but there is enough information pointing at the clinically important reduction of problematic behavior.

DISCUSSION

The present work focused on the question of what can be done to improve the data analysis in studies/practices using sub-optimal designs in such a way that results are more useful to the discipline. We recommended an analytical method consisting of structured visual analysis complemented with descriptive statistical procedures, while also keeping in mind substantive criteria (i.e., the opinion of the individuals involved in the process: family members, teachers, peers, coworkers, or supervisors). On the one hand, quantifications are useful for summarizing different aspects of the data and making the results available for subsequent meta-analysis. On the other hand, visual analysis is required for gaining an in-depth knowledge of the data and for assessing the adequacy of any specific quantitative procedures, due to the lack of consensus regarding the most appropriate technique (Tate et al., 2013).

A second question concerned the availability of tools for implementing the procedures proposed as part of the analytical method. We have mentioned, referenced, and illustrated the output of several tools implemented in the freeware R. Some of them are based on clickable menus, whereas others only require inputting the data before copying and pasting the code. The availability of software is crucial for eliminating the errors in obtaining the numerical and graphical results and in terms of time efficiency, both for short and relatively straightforward data series (e.g., Bunn et al., 2005) and for longer series with and less visually clear data patterns (e.g., Abney et al., 2014).

One potential issue with the analytical method is that it is possible that, in some instances, the three components do not coincide. A cautious approach would be to gather follow-up data after a certain period of time in order to check whether the initial ambiguous result of the assessment still holds. In case the unclear change is maintained and perceived as a change by the participants, then there would be evidence in favor of its practical importance. If there is disagreement between the substantive criterion and the other two components, we think that if the clients' well-being, quality of life, functionality, performance, etc. is improved according to their own opinion, then the substantive criterion should prevail, regardless of its numerical expression. In any case, the general effectiveness of an intervention depends on replications (Pashler and Wagenmakers, 2012) and not on the numerical result in a single study. Finally, if there is a divergence between the visual and quantitative information, it is important to know: (a) whether there is any data feature (e.g., pre-intervention trend, outliers) that might affect the performance of the quantitative analysis – in such case visual inspection should prevail; or (b) whether the data pattern prevents from getting a clear visual impression (e.g., due to highly variable data and/or a complex design structure) – in such case the quantitative summary is potentially more useful.

Another issue with the analytical method is that it might fail in certain situations such as the ones described in this paragraph (the list is not necessarily comprehensive). First, it is possible that the pre-intervention phase is too short or the measurements too variable for estimating trend with precision: the SLC quantifications would be less useful, but if there is no clear evidence of trend, then the NAP can be used as main quantification. Second, if there is complete non-overlap between the observations of the two conditions, the NAP will not be very informative, but the SLC can be used as an unstandardized quantification of the amount of difference and the *d*-statistic as a standardized quantification if more than one participant is being studied. Third, there might be a non-linear trend present in data, which is not an optimal situation for applying the SLC. In such case running medians (Tukey, 1977) can be used as a visual aid via the SCDA plug-in for R, while data modeling via the generalized least squares approach and LOESS is also possible. Fourth, there might be a delayed change in the behavior, not occurring simultaneously with the change in conditions (an issue that has remained practically unstudied except for Lieberman et al., 2010). In such case, the descriptive statistics will reflect the delay with lower quantifications of the effect, but it would be crucial to explore the cause of the change among the external uncontrolled factors (i.e., the solution is not an analytical one), given that the immediacy of the effect is one of the cornerstones for demonstrating causality (Kratochwill et al., 2010).

We hope that the discussion presented here would help practitioners and applied researchers to apply a systematic approach to data analysis and take a step toward partially improving the methodological quality of the studies. However, this would only be *one* step and studies would also need to meet the recommendations about the assessment and measurement of the target behavior, the implementation of the intervention, and the use of blinding to ensure objectivity, and also about reporting the results of the study (Tate et al., in press). Finally, it should always be considered whether what is assessed can be considered an "intervention effect" (in causal terms) or only a "behavioral change," which after several replications might point at the possible effectiveness of the intervention. In that sense, the analytical method was described in the context of studies with less-than-optimal designs in which causal relations cannot be readily established. Nonetheless, it is possible to extrapolate the method to experimental situations (e.g., multiple-baseline designs in which it is crucial to assess whether the behavioral change coincides with the staggered introduction of the intervention).

As a limitation of the quasi-statistical component of the analytical method, it is debatable whether the numerical results can be presented confidently in absence of a conventionally accepted optimal procedure, i.e., when all analytical techniques can be criticized. Considering the analytical method as a whole, further discussion is necessary on how to proceed when practitioners are faced with data that cannot be easily analyzed visually or quantitatively (e.g., short series, great data variability). One option would be to use the substantive criteria as basis for the conclusions and label the study as "practice" but not as "research." In contrast, when all three pieces of information (visual, quantitative, and substantive) coincide, it still has to be kept in mind that not meeting current *Standards* (Kratochwill et al., 2010) could render two-phase studies only a "pilot" status and, when included in meta-analysis, they are likely to be assigned lower weights and have less influence on the summary measures obtained.

AUTHOR CONTRIBUTIONS

The initial idea was due to JL and it was subsequently complemented and further developed by RM. The manuscript was written by JL (observational, non-experimental conceptual part in the Introduction) and RM (analytical part in the Analytical Method Explained, Analytical Method Applied, and Discussion). SC-M and SS-C made substantial contribution to the design of the work. All four authors (RM, JL, SC-M, and SS-C) participated in several revisions during the process of creating, discussing, and improving the manuscript, with RM leading all revisions and guiding the continuous improvement of the manuscript; gave their consent that this final version is submitted for publication; and agreed in their co-responsibility regarding all aspects of the work, such as the accuracy of the data and the integrity of the research.

ACKNOWLEDGMENTS

This study forms part of the results obtained in research project PSI2011-29587, funded by Spain's Ministry of Science and Innovation; and in research project number 1150096, funded by Chilean National Fund of Scientific and Technological Development -FONDECYT).

REFERENCES

Abney, D. H., Warlaumont, A. S., Haussman, A., Ross, J. M., and Wallot, S. (2014). Using nonlinear methods to quantify changes in infant limb movements and vocalizations. *Front. Psychol.* 5:771. doi: 10.3389/fpsyg.2014.00771

Allison, D. B., and Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: the case of the single case. *Behav. Res. Ther.* 31, 621–631. doi: 10.1016/0005-7967(93)90115-B

Anguera, M. T., Blanco-Villaseñor, Á, and Losada, J. L. (2001). Diseños observacionales, cuestión clave en el proceso de la Metodología Observacional. [Observational designs, a critical question in the process of Observational Methodology]. *Methodol. Behav. Sci.* 3, 135–160.

Apa Presidential Task Force on Evidence-Based Practice (2006). Evidence-based practice in psychology. *Am. Psychol.* 61, 271–285. doi: 10.1037/0003-066X.61.4.271

Bagner, D. M., Steinkopf, S. J., Miller-Loncar, C. L., Vohr, B. R., Hinckley, M., Eyberg, S. M., et al. (2009). Parent-Child Interaction Therapy for children born premature: a case study and illustration of vagal tone as a physiological measure of treatment outcome. *Cogn. Behav. Pract.* 16, 468–477. doi: 10.1016/j.cbpra.2009.05.002

Bakeman, R., and Quera, V. (2011). *Sequential Analysis and Observational Methods for the Behavioral Sciences.* Cambridge: Cambridge University Press.

Barlow, D. H., Nock, M. K., and Hersen, M. (Eds) (2009). *Single Case Experimental Designs: Strategies for Studying Behavior Change*, 3rd Edn. Boston, MA: Pearson.

Beeson, P. M., and Robey, R. R. (2006). Evaluating single-subject treatment research: lessons learned from the aphasia literature. *Neuropsychol. Rev.* 16, 161–169. doi: 10.1007/s11065-006-9013-7

Beretvas, S. N., and Chung, H. (2008). A review of meta-analyses of single-subject experimental designs: methodological issues and practice. *Evid. Based Commun. Assess. Interv.* 2, 129–141. doi: 10.1080/17489530802446302

Borckardt, J., and Nash, M. (2014). Simulation modelling analysis for small sets of single-subject data collected over time. *Neuropsychol. Rehabil.* 24, 492–506. doi: 10.1080/09602011.2014.895390

Borrego, J. Jr., Anhalt, K., Terao, S. Y., Vargas, E. C., and Urquiza, A. J. (2006). Parent-child interaction therapy with a Spanish-speaking family. *Cogn. Behav. Pract.* 13, 121–133. doi: 10.1016/j.cbpra.2005.09.001

Brossart, D. F., Parker, R. I., Olson, E. A., and Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behav. Modif.* 30, 531–563. doi: 10.1177/0145445503261167

Brossart, D. F., Vannest, K., Davis, J., and Patience, M. (2014). Incorporating nonoverlap indices with visual analysis for quantifying intervention effectiveness in single-case experimental designs. *Neuropsychol. Rehabil.* 24, 464–491. doi: 10.1080/09602011.2013.868361

Bulté, I., and Onghena, P. (2012). When the truth hits you between the eyes: a software tool for the visual analysis of single-case experimental data. *Methodology* 8, 104–114. doi: 10.1027/1614-2241/a000042

Bulté, I., and Onghena, P. (2013). The single-case data analysis package: analysing single-case experiments with R software. *J. Mod. Appl. Stat. Methods* 12, 450–478.

Bunn, R., Burns, M. K., Hoffman, H. H., and Newman, C. L. (2005). Using incremental rehearsal to teach letter identification with a preschool-aged child. *J. Evid. Based Pract. Schools* 6, 124–134.

Busk, P. L., and Serlin, R. (1992). "Meta-analysis for single case research," in *Single-Case Research Design and Analysis: New Directions for Psychology and Education*, eds T. R. Kratochwill and J. R. Levin (Hillsdale, NJ: Lawrence Erlbaum), 187–212.

Busse, R. T., Kratochwill, T. R., and Elliott, S. N. (1995). Meta-analysis for single-case consultation outcomes: applications to research and practice. *J. Sch. Psychol.* 33, 269–285. doi: 10.1016/0022-4405(95)00014-D

Callahan, C. D., and Barisa, M. T. (2005). Statistical process control and rehabilitation outcome: the single-subject design reconsidered. *Rehabil. Psychol.* 50, 24–33. doi: 10.1037/0090-5550.50.1.24

Campbell, D. T., and Stanley, J. C. (1966). *Experimental and Quasi-experimental Designs for Research*. Chicago, IL: Rand McNally.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46. doi: 10.1177/001316446002000104

Cohen, J. (1994). The earth is round (p < .05). *Am. Psychol.* 49, 997–1003. doi: 10.1037/0003-066X.49.12.997

Cordery, J. L., Morrisson, D., Wright, B. M., and Wall, T. B. (2010). The impact of autonomy and task uncertainty on team performance: a longitudinal field study. *J. Organ. Behav.* 31, 240–258. doi: 10.1002/job.657

Crane, D. R. (1985). Single-Case experimental designs in family therapy research: limitations and considerations. *Fam. Process* 24, 69–77. doi: 10.1111/j.1545-5300.1985.00069.x

Crosbie, J. (1987). The inability of the binomial test to control Type I error with single-subject data. *Behav. Assess.* 9, 141–150.

Danov, S. E., and Symons, F. J. (2008). A survey evaluation of the reliability of visual inspection and functional analysis graphs. *Behav. Modif.* 32, 828–839. doi: 10.1177/0145445508318606

Davis, D. H., Gagné, P., Fredrick, L. D., Alberto, P. A., Waugh, R. E., and Haardörfer, R. (2013). Augmenting visual analysis in single-case research with hierarchical linear modeling. *Behav. Modif.* 37, 62–89. doi: 10.1177/0145445512453734

DeCarlo, L. T., and Tryon, W. W. (1993). Estimating and testing correlation with small samples: a comparison of the C-statistic to modified estimator. *Behav. Res. Ther.* 31, 781–788. doi: 10.1016/0005-7967(93)90009-J

Edgington, E. S. (1980). Validity of randomization tests for one-subject experiments. *J. Educ. Stat.* 5, 235–251. doi: 10.3102/10769986005003235

Edgington, E. S., and Onghena, P. (2007). *Randomization Tests*, 4th Edn. London, UK: Chapman & Hall/CRC.

Eyberg, S. M., Nelson, M. M., and Boggs, S. R. (2008). Evidence-based psychosocial treatments for children and adolescents with disruptive behavior. *J. Clin. Child Adolesc. Psychol.* 37, 1–23. doi: 10.1080/15374410701820117

Eyberg, S. M., and Pincus, D. (1999). *Eyberg Child Behavior Inventory and Sutter-Eyberg Student Behavior Inventory: Professional Manual*. Odessa, FL: Psychological Assessment Resources.

Fabio, R. A., Castelli, I., Marchetti, A., and Antonietti, A. (2013). Training communication abilities in Rett Syndrome through reading and writing. *Front. Psychol.* 4:911. doi: 10.3389/fpsyg.2013.00911

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., and Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: the utility of multilevel modeling approaches. *Behav. Res. Methods* 41, 372–384. doi: 10.3758/BRM.41.2.372

Finn, M., and McDonald, S. (2014). A single case study of computerised cognitive training for older persons with mild cognitive impairment. *NeuroRehabilitation* 35, 261–270. doi: 10.3233/NRE-141121

Fisch, G. S. (2001). Evaluating data from behavioral analysis: visual inspection or statistical models? *Behav. Processes* 54, 137–154. doi: 10.1016/S0376-6357(01)00155-3

Fisher, W. W., Kelley, M. E., and Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *J. Appl. Behav. Anal.* 36, 387–406. doi: 10.1901/jaba.2003.36-387

Fisher, W. W., and Lerman, D. C. (2014). It has been said that, "There are three degrees of falsehoods: lies, damn lies, and statistics". *J. School Psychol.* 52, 243–248. doi: 10.1016/j.jsp.2014.01.001

Franklin, R. D., Gorman, B. S., Beasley, T. M., and Allison, D. B. (1996). "Graphical display and visual analysis," in *Design and Analysis of Single-Case Research*, eds R. D. Franklin, D. B. Allison, and B. S. Gorman (Mahwah, NJ: Lawrence Erlbaum), 119–158.

Gage, N. A., and Lewis, T. J. (2014). Hierarchical linear modeling meta-analysis of single-subject design research. *J. Spec. Educ.* 48, 3–16. doi: 10.1177/0022466912443894

Gast, D. L., and Spriggs, A. D. (2010). "Visual analysis of graphic data," in *Single Subject Research Methodology in Behavioral Sciences*, ed. D. L. Gast (London, UK: Routledge), 199–233.

Gibson, G., and Ottenbacher, K. (1988). Characteristics influencing the visual analysis of single-subject data: an empirical analysis. *J. Appl. Behav. Sci.* 24, 298–314. doi: 10.1177/0021886388243007

Glass, G. V., McGaw, B., and Smith, M. L. (1981). *Meta-analysis in Social Research*. Beverly Hills, CA: Sage.

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *J. Appl. Psychol.* 79, 314–316. doi: 10.1037/0021-9010.79.2.314

Grissom, R. J., and Kim, J. J. (2012). *Effect Size for Research: Univariate and Multivariate Applications*, 2nd Edn. London, UK: Routledge.

Harrington, M., and Velicer, W. F. (2015). Comparing visual and statistical analysis in single-case studies using published studies. *Multivariate Behav. Res.* 50, 162–183. doi: 10.1080/00273171.2014.973989

Harrop, J. W., and Velicer, W. F. (1985). A comparison of three alternative methods of time series model identification. *Multivariate Behav. Res.* 20, 27–44. doi: 10.1207/s15327906mbr2001_2

Heyvaert, M., and Onghena, P. (2014). Randomization tests for single-case experiments: state of the art, state of the science, and state of the application. *J. Contextual Behav. Sci.* 3, 51–64. doi: 10.1016/j.jcbs.2013.10.002

Heyvaert, M., Wendt, O., Van Den Noortgate, W., and Onghena, P. (2015). Randomization and data-analysis items in quality standards for single-case experimental studies. *J. Spec. Educ.* 49, 146–156. doi: 10.1177/0022466914525239

Houle, T. T. (2009). "Statistical analyses for single-case experimental designs," in *Single Case Experimental Designs: Strategies for Studying Behavior Change*, 3rd Edn, eds D. H. Barlow, M. K. Nock, and M. Hersen (Boston, MA: Pearson), 271–305.

Howick, J., Chalmers, I., Glasziou, P., Greenhaigh, T., Heneghan, C., Liberati, A., et al. (2011). *The 2011 Oxford CEBM Evidence Table (Introductory Document)*. Oxford: Centre for Evidence-Based Medicine.

Hugdahl, K., and Ŏst, L.-G. (1981). On the difference between statistical and clinical significance. *Behav. Assess.* 3, 289–295.

Huitema, B. E., McKean, J. W., and Laraway, S. (2007). Time series intervention analysis using ITSACORR: fatal flaws. *J. Mod. Appl. Stat. Methods* 6, 367–379.

Institute of Education Sciences (2013). *Request for Applications: Statistical and Research Methodology in Education*. Available at: http://ies.ed.gov/funding/pdf/2014_84305D.pdf

Jacobson, N. S., and Truax, P. (1991). Clinical significance: a statistical approach to meaningful change in psychotherapy research. *J. Consult. Clin. Psychol.* 59, 12–19. doi: 10.1037/0022-006X.59.1.12

Jacoby, W. G. (2000). Loess: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Stud.* 19, 577–613. doi: 10.1016/S0261-3794(99)00028-1

Jamieson, M., Cullen, B., McGee-Lennon, M., Brewster, S., and Evans, J. J. (2014). The efficacy of cognitive prosthetic technology for people with memory impairments: a systematic review and meta-analysis. *Neuropsychol. Rehabil.* 24, 419–444. doi: 10.1080/09602011.2013.825632

Jenson, W. R., Clark, E., Kircher, J. C., and Kristjansson, S. D. (2007). Statistical reform: evidence-based practice, meta-analyses, and single subject designs. *Psychol. Schools* 44, 483–493. doi: 10.1002/pits.20240

Johnston, J. M., and Pennypacker, H. S. (2008). *Strategies and Tactics of Behavioral Research*, 3rd Edn. New York, NY: Routledge.

Jonsson, G. K., Anguera, M. T., Blanco-Villaseñor, Á, Losada, J. L., Hernández-Mendo, A., Ardá, T., et al. (2006). Hidden patterns of play interaction in soccer using SOF-CODER. *Behav. Res. Methods* 38, 372–381. doi: 10.3758/BF03192790

Kazdin, A. E. (1978). Methodological and interpretive problems of single-case experimental designs. *J. Consult. Clin. Psychol.* 46, 629–642. doi: 10.1037/0022-006X.46.4.629

Kazdin, A. E. (1982). *Single-Case Research Designs: Methods for Clinical and Applied Settings*. New York, NY: Oxford University Press.

Kazdin, A. E. (1984). *Behavior Modification in Applied Settings*, 3rd Edn. Homewood, IL: The Dorsey Press.

Kazdin, A. E. (1999). The meanings and measurements of clinical significance. *J. Consult. Clin. Psychol.* 67, 332–339. doi: 10.1037/0022-006X.67.3.332

Kendall, P. C. (1999). Clinical significance. *J. Consult. Clin. Psychol.* 67, 283–284. doi: 10.1037/0022-006X.67.3.283

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., et al. (2010). *Single Case Designs Technical Documentation. In What Works Clearinghouse: Procedures and Standards Handbook (Version 2.0)*. Available at: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., et al. (2013). Single-Case intervention research design standards. *Remedial Spec. Educ.* 34, 26–38. doi: 10.1177/0741932512452794

Kratochwill, T. R., and Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: randomization to the rescue. *Psychol. Methods* 15, 124–144. doi: 10.1037/a0017736

Kratochwill, T. R., Stoiber, K. C., and Gutkin, T. B. (2001). Empirically supported interventions in school psychology: the role of negative results in outcome research. *Psychol. Schools* 37, 399–413. doi: 10.1177/0741932512452794

Lane, J. D., and Gast, D. L. (2014). Visual analysis in single case experimental design studies: brief review and guidelines. *Neuropsychol. Rehabil.* 24, 445–463. doi: 10.1080/09602011.2013.815636

Ledford, J., and Gast, D. L. (2014). Measuring procedural fidelity in behavioural research. *Neuropsychol. Rehabil.* 24, 332–348. doi: 10.1080/09602011.2013.861352

Levin, J. R., Evmenova, A. S., and Gafurov, B. S. (2014). "The single-case data-analysis ExPRT (Excel Package of Randomization Tests)," in *Single-Case Intervention Research: Methodological and Statistical Advances*, eds T. R. Kratochwill and J. R. Levin (Washington, DC: American Psychological Association), 185–219.

Levin, J. R., Ferron, J. M., and Kratochwill, T. R. (2012). Nonparametric statistical tests for single-case systematic and randomized ABAB. . .AB and alternating treatment intervention designs: new developments, new directions. *J. Sch. Psychol.* 50, 599–624. doi: 10.1016/j.jsp.2012.05.001

Lieberman, R. G., Yoder, P. J., Reichow, B., and Wolery, M. (2010). Visual analysis of multiple baseline across participants graphs when change is delayed. *Sch. Psychol. Q.* 25, 28–44. doi: 10.1037/a0018600

Maggin, D. M., Briesch, A. M., and Chafouleas, S. M. (2013). An application of the What Works Clearinghouse standards for evaluating single-subject research: synthesis of the self-management literature base. *Remedial Spec. Educ.* 34, 44–58. doi: 10.1177/0741932511435176

Maggin, D. M., Briesch, A. M., Chafouleas, S. M., Ferguson, T. D., and Clark, C. (2014). A comparison of rubrics for identifying empirically supported practices with single-case research. *J. Behav. Educ.* 23, 287–311. doi: 10.1007/s10864-013-9187-z

Manolov, R., and Rochat, L. (2015). Further developments in summarising and meta-analysing single-case data: an illustration with neurobehavioural interventions in acquired brain injury. *Neuropsychol. Rehabil.* 25, 637–662. doi: 10.1080/09602011.2015.1064452

Manolov, R., Sierra, V., Solanas, A., and Botella, J. (2014). Assessing functional relations in single-case designs: quantitative proposals in the context of the evidence-based movement. *Behav. Modif.* 38, 878–913. doi: 10.1177/0145445514545679

Manolov, R., and Solanas, A. (2008). Comparing N = 1 effect size indices in presence of autocorrelation. *Behav. Modif.* 32, 860–875. doi: 10.1177/0145445508318866

Manolov, R., Solanas, A., Bulté, I., and Onghena, P. (2010). Data-division-specific robustness and power for ABAB designs. *J. Exp. Educ.* 78, 191–214. doi: 10.1080/00220970903292827

Manolov, R., Solanas, A., Sierra, V., and Evans, J. J. (2011). Choosing among techniques for quantifying single-case intervention effectiveness. *Behav. Ther.* 42, 533–545. doi: 10.1016/j.beth.2010.12.003

Matyas, T. A., and Greenwood, K. M. (1996). "Serial dependency in single-case time series," in *Design and Analysis of Single-Case Research*, eds R. D. Franklin, D. B. Allison, and B. S. Gorman (Mahwah, NJ: Lawrence Erlbaum), 215–243.

McGraw, K. O., and Wong, S. P. (1992). A common language effect size statistic. *Psychol. Bull.* 111, 361–365. doi: 10.1037/0033-2909.111.2.361

McKnight, S. D., McKean, J. W., and Huitema, B. E. (2000). A double bootstrap method to analyze linear models with autoregressive error terms. *Psychol. Methods* 3, 87–101. doi: 10.1037/1082-989X.5.1.87

Michael, J. (1974). Statistical inference for individual organism research: mixed blessing or curse? *J. Appl. Behav. Anal.* 7, 647–653. doi: 10.1901/jaba.1974.7-647

Miller, M. J. (1985). Analyzing client change graphically. *J. Couns. Dev.* 63, 491–494. doi: 10.1002/j.1556-6676.1985.tb02743.x

Moeyaert, M., Ferron, J., Beretvas, S., and Van Den Noortgate, W. (2014a). From a single-level analysis to a multilevel analysis of since-case experimental designs. *J. Sch. Psychol.* 52, 191–211. doi: 10.1016/j.jsp.2013.11.003

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S. N., and Van den Noortgate, W. (2014b). The influence of the design matrix on treatment effect estimates in the quantitative analyses of single-case experimental designs research. *Behav. Modif.* 38, 665–704. doi: 10.1177/0145445514535243

Ninci, J., Vannest, K. J., Willson, V., and Zhang, N. (2015). Interrater agreement between visual analysts of single-case data: a meta-analysis. *Behav. Modif.* 39, 510–541. doi: 10.1177/0145445515581327

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., and Harris, K. R. (2005). Research in special education: scientific methods and evidence-based practices. *Except. Child.* 71, 137–148. doi: 10.1177/0014402905071 00201

O'Neill, B., Best, C., Gillespie, A., and O'Neill, L. (2013). Automated prompting technologies in rehabilitation and at home. *Soc. Care Neurodisability* 4, 17–28. doi: 10.1108/20420911311302281

Ottenbacher, K. J. (1990). Visual inspection of single-subject data: an empirical analysis. *Ment. Retard.* 28, 283–290.

Parker, R. I. (2006). Increased reliability for single-case research results: is bootstrap the answer? *Behav. Ther.* 37, 326–338. doi: 10.1016/j.beth.2006.01.007

Parker, R. I., and Brossart, D. F. (2003). Evaluating single-case research data: a comparison of seven statistical methods. *Behav. Ther.* 34, 189–211. doi: 10.1016/S0005-7894(03)980013-8

Parker, R. I., Cryer, J., and Byrns, G. (2006). Controlling baseline trend in single-case research. *Sch. Psychol. Q.* 21, 418–443. doi: 10.1037/h0084131

Parker, R. I., and Vannest, K. J. (2009). An improved effect size for single-case research: nonoverlap of all pairs. *Behav. Ther.* 40, 357–367. doi: 10.1016/j.beth.2008.10.006

Parker, R. I., Vannest, K. J., and Davis, J. L. (2011a). Effect size in single-case research: a review of nine nonoverlap techniques. *Behav. Modif.* 35, 303–322. doi: 10.1177/0145445511399147

Parker, R. I., Vannest, K. J., Davis, J. L., and Sauber, S. B. (2011b). Combining nonoverlap and trend for single-case research: Tau-U. *Behav. Ther.* 42, 284–299. doi: 10.1016/j.beth.2010.08.006

Pashler, H., and Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? *Perspect. Psychol. Sci.* 7, 528–530. doi: 10.1177/1745691612465253

Pazzagli, C., Laghezza, L., Manaresi, F., Mazzeschi, C., and Powell, B. (2014). The circle of security parenting and parental conflict: a single case study. *Front. Psychol.* 5:887. doi: 10.3389/fpsyg.2014.00887

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Robey, R. R., Schultz, M. C., Crawford, A. B., and Sinner, C. A. (1999). Single-subject clinical outcome research: designs, data, effect sizes, and analysis. *Aphasiology* 13, 445–473. doi: 10.1080/026870399402028

Scruggs, T. E., and Mastropieri, M. A. (2013). PND at 25: past, present, and future trends in summarizing single-subject research. *Remedial Spec. Educ.* 34, 9–19. doi: 10.1177/0741932512440730

Shadish, W. R., Hedges, L. V., and Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: a primer and applications. *J. Sch. Psychol.* 52, 123–147. doi: 10.1016/j.jsp.2013.11.005

Sidman, M. (1960). *Tactics of Scientific Research: Evaluating Experimental Data in Psychology*. New York, NY: Basic Books.

Simonton, D. K. (1977). Cross-sectional time-series experiments: some suggested statistical analyses. *Psychol. Bull.* 84, 489–502. doi: 10.1037/0033-2909.84.3.489

Smith, J. D. (2012). Single-Case experimental designs: a systematic review of published research and current standards. *Psychol. Methods* 17, 510–550. doi: 10.1037/a0029312

Solanas, A., Manolov, R., and Onghena, P. (2010a). Estimating slope and level change in N = 1 designs. *Behav. Modif.* 34, 195–218. doi: 10.1177/0145445510363306

Solanas, A., Manolov, R., and Sierra, V. (2010b). Lag-one autocorrelation in short series: estimation and hypothesis testing. *Psicológica* 31, 357–381.

Solmi, F., Onghena, P., Salmaso, L., and Bulté, I. (2014). A permutation solution to test for treatment effects in alternation design single-case experiments. *Commun. Stat.* 43, 1094–1111. doi: 10.1080/03610918.2012.725295

Solomon, B. G., Howard, T. K., and Stein, B. L. (2015). Critical assumptions and distribution features pertaining to contemporary single-case effect sizes. *J. Behav. Educ.* 24, 438–458. doi: 10.1007/s10864-015-9221-4

Suen, H. K., and Ary, D. (1989). *Analyzing Quantitative Behavioral Data*. Hillsdale, NJ: Lawrence Erlbaum.

Swaminathan, H., Rogers, H. J., and Horner, R. H. (2014). An effect size measure and Bayesian analysis of single-case designs. *J. Sch. Psychol.* 52, 213–230. doi: 10.1016/j.jsp.2013.12.002

Tate, R. L., Perdices, M., Rosenkoetter, U., Shadish, W., Vohra, S., and Wilson, B. (in press). The single-case Reporting guideline in behavioural interventions (SCRIBE) 2015 statement. *Arch. Sci. Psychol*.

Tate, R. L., Perdices, M., Rosenkoetter, U., Wakima, D., Godbee, K., Togher, L., et al. (2013). Revision of a method quality rating scale for single-case experimental designs and n-of-1 trials: the 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale. *Neuropsychol. Rehabil.* 23, 619–638. doi: 10.1080/09602011.2013.824383

Taylor, L. K., and Weems, C. F. (2011). Cognitive-behavior therapy for disaster-exposed youth with posttraumatic stress: results from a multiple-baseline examination. *Behav. Ther.* 42, 349–363. doi: 10.1016/j.beth.2010.09.001

Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., and Farmer, J. D. (1992). Testing for nonlinearity in time series: the method of surrogate data. *Physica D* 58, 77–94. doi: 10.1016/0167-2789(92)90102-S

Tryon, W. W. (1982). A simplified time-series analysis for evaluating treatment interventions. *J. Appl. Behav. Anal.* 15, 423–429. doi: 10.1901/jaba.1982.15-423

Tukey, J. W. (1977). *Exploratory Data Analysis*. London, UK: Addison-Wesley.

Vannest, K. J., Parker, R. I., and Gonen, O. (2011). *Single Case Research: Web Based Calculators for SCR Analysis. (Version 1.0) [Web-Based Application]*. College Station, TX: Texas A&M University.

Winkens, I., Ponds, R., Pouwels-van den Nieuwenhof, C., Eilander, H., and van Heugten, C. (2014). Using single-case experimental design methodology to evaluate the effects of the ABC method for nursing staff on verbal aggressive behaviour after acquired brain injury. *Neuropsychol. Rehabil.* 24, 349–364. doi: 10.1080/09602011.2014.901229

Wolery, M., Busick, M., Reichow, B., and Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *J. Spec. Educ.* 44, 18–29. doi: 10.1177/0022466908328009

Ximenes, V. M., Manolov, R., Solanas, A., and Quera, V. (2009). Factors affecting visual inference in single-case designs. *Span. J. Psychol.* 12, 823–832. doi: 10.1017/S1138741600002195

Young, L. C. (1941). On the randomness in ordered sequences. *Ann. Math. Stat.* 12, 293–300. doi: 10.1214/aoms/1177731711

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 23 March 2017
doi: 10.3389/fpsyg.2017.00435

Evaluation of a Psychological Intervention for Patients with Chronic Pain in Primary Care

Francisco J. Cano-García[1]*, *María del Carmen González-Ortega*[1], *Susana Sanduvete-Chaves*[2], *Salvador Chacón-Moscoso*[2,3] and *Roberto Moreno-Borrego*[4]

[1] *Departamento de Personalidad, Evaluación y Tratamiento Psicológicos, Universidad de Sevilla, Seville, Spain,* [2] *Departamento de Psicología Experimental, Facultad de Psicología, Universidad de Sevilla, Seville, Spain,* [3] *Departamento de Psicología, Universidad Autónoma de Chile, Santiago, Chile,* [4] *Centro de Atención Primaria Príncipe de Asturias, Servicio Andaluz de Salud, Utrera, Spain*

According to evidence from recent decades, multicomponent programs of psychological intervention in people with chronic pain have reached the highest levels of efficacy. However, there are still many questions left to answer since efficacy has mainly been shown among upper-middle class patients in English-speaking countries and in controlled studies, with expert professionals guiding the intervention and with a limited number of domains of painful experience evaluated. For this study, a program of multicomponent psychological intervention was implemented: (a) based on techniques with empirical evidence, but developed in Spain; (b) at a public primary care center; (c) among patients with limited financial resources and lower education; (d) by a novice psychologist; and (e) evaluating all domains of painful experience using the instruments recommended by the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT). The aim of this study was to evaluate this program. We selected a consecutive sample of 40 patients treated for chronic non-cancer pain at a primary care center in Utrera (Seville, Spain), adults who were not in any employment dispute, not suffering from psychopathology, and not receiving psychological treatment. The patients participated in 10 psychological intervention sessions, one per week, in groups of 13–14 people, which addressed psychoeducation for pain; breathing and relaxation; attention management; cognitive restructuring; problem-solving; emotional management; social skills; life values and goal setting; time organization and behavioral activation; physical exercise promotion; postural and sleep hygiene; and relapse prevention. In addition to the initial assessment, measures were taken after the intervention and at a 6-month follow-up. We assessed the program throughout the process: before, during and after the implementation. Results were analyzed statistically (significance and effect size) and from a clinical perspective (clinical significance according to IMMPACT standards). According to this analysis, the intervention was successful, although improvement tended to decline at follow-up, and the detailed design gave the program assessment a high degree of standardization and specification. Finally, suggestions for improvement are presented for upcoming applications of the program.

Keywords: formative evaluation, clinical effectiveness, chronic pain, Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT), methodological quality, primary care

OPEN ACCESS

Edited by:
Pietro Cipresso,
IRCCS Istituto Auxologico Italiano,
Italy

Reviewed by:
Daniel Saverio John Costa,
University of Sydney, Australia
Bárbara Oliván Blázquez,
University of Zaragoza, Spain
Deepak Kumar,
Institute of Human Behaviour
and Allied Sciences, India

****Correspondence:***
Francisco J. Cano-García
fjcano@us.es

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: *10 December 2016*
Accepted: *08 March 2017*
Published: *23 March 2017*

Citation:
Cano-García FJ,
González-Ortega MC,
Sanduvete-Chaves S,
Chacón-Moscoso S and
Moreno-Borrego R (2017) Evaluation
of a Psychological Intervention
for Patients with Chronic Pain
in Primary Care.
Front. Psychol. 8:435.
doi: 10.3389/fpsyg.2017.00435

INTRODUCTION

Pain is an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage (Merskey, 1994). Pain becomes chronic when it loses its adaptive function, lasts longer than expected (3–6 months), and does not respond to the prescribed medical treatments. Pain and chronic pain are global, complex experiences for human beings, and interdisciplinary theoretical models have been developed to study them. One such model is the gate control theory (Melzack and Wall, 1967) and its more recent version, the neuromatrix theory (Melzack, 1999). Essentially, painful experience is defined at different levels here, including the sensory, behavioral, emotional and cognitive level, all of which are integrated in a more comprehensive framework of stress processes (for a more detailed description, see Gatchel et al., 2007). For this reason, psychology's contribution to the study and treatment of chronic pain has been critically important for the past few decades.

Chronic pain is a public health issue in the developed world. In an aging population like that of Europe, 19% of the population suffers from chronic pain; in Spain, where this study was conducted, chronic pain stands at 11%. A recent study by Andrew et al. (2014) estimated the costs associated with chronic pain. In the work world, for every dollar lost by the average person, the costs associated with a person suffering from chronic pain are between $3.60 and $12.50 for absenteeism, between $2.50 and $3.00 for loss of productivity, and between $1.90 and $2.60 in paid unemployment. In terms of healthcare costs, for every dollar spent on other patients, the costs associated with a person suffering from chronic pain are between $2.50 and $3.00 in visits to primary care centers, between $3.30 and $7.60 in hospital stays, $4.00 in medicine and $3.00 in emergency care.

The gateway for patients with chronic pain in healthcare systems is usually the primary care center, as seen in Europe, where 70% of these patients saw a general practitioner (Breivik et al., 2006). Patients with chronic pain are seen as a challenging but low-priority customer similar to those suffering from mental health disorders, in contrast to high-priority patients like those suffering from cardiovascular disease (Johnson et al., 2013). Although professionals who see such patients usually have clinical practice guidelines, they tend not to use them to either evaluate or treat such patients because they are overwhelmed by the quantity and complexity of the demand. In most cases, such physicians limit themselves to prescribing drugs or referring the patient to a specialist.

There is unquestionable evidence on the efficacy of psychological intervention in chronic pain. According to the Society of Clinical Psychology (APA, 2016), evidence is particularly strong for two types of psychological intervention: cognitive-behavioral therapy (Morley et al., 1999; Huguet et al., 2014; Cherkin et al., 2016; Kroner et al., 2016) and Acceptance and Commitment Therapy (Veehof et al., 2011, 2016; Hann and McCracken, 2014). Other treatment options like relaxation therapy (Meeus et al., 2014), guided meditation and hypnosis have yielded moderate efficacy levels. Finally, evidence of efficacy has been growing for more recent treatment options such as eye movement desensitization and reprocessing (EMDR) (Tesarz et al., 2014) and particularly, mindfulness (Lauche et al., 2013). Given the current state of knowledge, multicomponent psychological treatments could be considered more efficacious than others and represent a viable alternative for healthcare when applied in small groups (APA, 2016). However, identifying efficacious treatment is one thing and getting the general population to benefit from such treatment is quite another. A good example of this is an epidemiological study conducted among 2,596 fibromyalgia patients in the USA: only 8% had received cognitive-behavioral therapy (Bennett et al., 2007).

In the scientific study of pain, the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) began in 2002 to improve the quality of assessments in clinical trials, bringing together scholars, regulatory bodies and public healthcare institutions, consumer and patient associations, and representatives from the pharmaceutical industry. Various scientific disciplines within healthcare like anesthesiology, clinical pharmacology, internal medicine, law, neurology, nursing, oncology, psychology, rheumatology and surgery are part of IMMPACT. The initiative has yielded three main results: the identification of the basic and complementary areas of the pain experience that must be evaluated (Turk et al., 2003; McGrath et al., 2008); the identification, development and validation of instruments to assess them (Dworkin et al., 2005; Turk et al., 2006; McGrath et al., 2008); and the determination of clinical importance standards to assess treatment outcomes (Dworkin et al., 2008, 2009; Turk et al., 2008).

The evidence presented above regarding both psychological treatment and the IMMPACT initiative is generally produced by studies conducted in ideal conditions, with the funding necessary for an adequate selection of participants: expert psychologists, patients with middle-high educational levels who are motivated to participate and do not leave the study, etc. In conditions such as these, many doubts regarding the efficacy of psychological intervention go unanswered. However, little information is available on clinical efficacy in real healthcare contexts: what if the studies focused on patients from a rural area in the south of Spain with different educational levels and from a different sociodemographic? What happens when they visit a primary care facility and are seen by a novice psychologist?

Ehde et al. (2014) addressed these challenges in an interesting review on cognitive-behavioral therapy for patients with chronic pain. The authors found only one study with rural and low literacy samples (Thorn et al., 2011). Worse still, they found no study that considered the level of experience of the therapist, but indicated that this variable might be relevant, since cognitive-behavioral therapy is more effective when performed by psychologists than other care providers (Nicholas et al., 2011).

These questions are what motivated us to assess a multicomponent cognitive-behavioral program specifically designed for patients with chronic pain, applied in a public primary care center located in the south of Spain, with participants from different socioeconomic and educational backgrounds and implemented by an inexperienced psychologist.

MATERIALS AND METHODS

Participants

Patients at the Príncipe de Asturias primary care center participated in the study. The primary care center is located in Utrera, a small rural town in the province of Seville, Spain.

The inclusion criteria were the following: (a) to be at least 18 years old; (b) to have visited primary care due to difficulties handling chronic pain during the recruitment period (present maladaptive adjustment to pain); (c) to not be in the middle of an employment dispute or waiting for approval on a disability pension; (d) to not have a primary psychopathologic disorder; (e) to not be in psychiatric or psychological treatment, but could be taking psychotropic drugs; (f) the ability to follow group sessions, thus excluding conditions such as deafness, blindness, or dementia; (g) willingness to sign an agreement to attend the sessions (group and/or individual); and (h) not be hospitalized.

Design

This study presents a quasi-experimental one-group pre-test – post-test – follow-up design (Shadish et al., 2002; Chacón-Moscoso et al., 2008). This means that there are three measurement instances: one before the intervention and two after the intervention (specifically, one immediately after the intervention and another 6 months later). Additionally, this design lacks a control group. As we are interested in studying the change over time in only one group, this is a within-subject design (APA, 2010).

Variables and Measures

Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials recommendations were used to assess the pain experience in terms of both procedures and instruments (Turk et al., 2003; Dworkin et al., 2005). The assessment covered pain, physical functioning, emotional functioning, and the patient's rating of change. Although IMMPACT recommendations do not establish the main assessment variables, pain, specifically pain intensity, is usually considered a primary outcome. As a result, the remaining areas and variables would be considered secondary in this study, but also extremely important as indicators of possible improvements in the patients' quality of life. To evaluate pain, the patient was asked to describe the intensity of perceived pain in the 24 h period preceding the interview and at the time of the interview, using a numerical scale with 0 meaning "No pain" and 10 meaning "Pain as bad as you can imagine" (Dworkin et al., 2005).

Physical functioning was evaluated through (1) the items *How much has pain interfered in your daily life during the last 24 h?* and *How much is pain interfering right now?*, with a four-point rating scale where 0 is nothing and 3, totally; and (2) the Spanish language version of the pain interference subscale (Ferrer et al., 1993) of the West Haven-Yale Multidimensional Pain Inventory (WHYMPI) (Kerns et al., 1985). The WHYMPI is the first psychometric instrument for multidimensional pain evaluation. The 11 items interference subscale consists of a seven-point Likert scale (0–6) to rate pain interference in daily life; the total points are then divided by the number of items. The psychometric properties of the original scale have been clearly demonstrated internationally (Haythornthwaite, 2003). Cronbach's α was 0.68 for the Spanish language version of the interference scale (Ferrer et al., 1993).

Two instruments were used to evaluate emotional functioning: (1) the Profile of Mood States (POMS) (Haythornthwaite, 2004). This psychometric instrument assesses, using 58 adjectives rated from 0 (not at all) to 4 (extremely) on a five-point Likert scale, six mood states: Fatigue (0–28), Depression (0–60), Tension (0–36), Hostility (0–48), Confusion (0–28), and Vigor (0–32). In addition to six partial scores, it provides a global score on Total Mood Disturbance that ranges from −32 to 200 after adding the scores obtained in Fatigue, Depression, Tension, Hostility and Confusion, and subtracting the score obtained in Vigor. The POMS properties have also been demonstrated within the framework of IMMPACT with an internal consistency of the different scales between 0.63 (Confusion) and 0.96 (Depression) (Dworkin et al., 2005); and (2) the Beck Depression Inventory (BDI) (Beck et al., 1961). This instrument is comprised of 21 items that are answered on a four-point Likert scale (0–3). A total score is obtained by adding the values given for the 21 items ranging from 0 to 63. Higher values mean higher levels of depression. Specifically, 0–9 indicates none or minimal depression; 10–18 indicates mild to moderate depression; 19–29 indicates moderate to severe depression; and 30–63 indicates severe depression. This tool presents evidence of reliability and validity in the assessment of symptoms of depression and emotional distress (Dworkin et al., 2005).

The expected rating of change (pre-test) and the rating of change (post-test and follow-up) were evaluated using the Patient Global Impression of Change Scale (PGIC) (Guy, 1976). This measure is a single-item rating of a patient's rating of improvement as the result of treatment on a seven-point scale that ranges from 1 "very much worse" to 7 "very much improved" with no change at the middle of the scale. Due to its simplicity, validity and reliability, the PGIC was included as a scale recommended by IMMPACT (Farrar, 2003).

Patient willingness was evaluated using the CONSORT (Consolidated Standards of Reporting Trials) guideline (Altman et al., 2001; Moher et al., 2001) which provides information on recruitment processes; the number of candidates excluded and the reasons for exclusion; the number of candidates who did not start treatment and the reasons; and the number of participants who abandon treatment and the reasons.

Psychological Intervention

Psychological intervention consisted in a multicomponent protocol developed and published in Spain by a group of professionals and scholars, including one of the authors of this work (FJC). This protocol incorporates the principal cognitive-behavioral techniques with evidence of efficacy in pain treatment and combines them with a few others inspired by Acceptance and Commitment Therapy. A description of the program can be found in Moix and Casado (2011) and the full program is available at Kovacs and Moix (2011).

The program is structured in 10 weekly sessions, each lasting an hour and a half, that approach the following topics sequentially: (1) introduction to cognitive-behavioral intervention; (2) breathing and relaxation; (3) attention management; (4) cognitive restructuring I; (5) cognitive restructuring II; (6) problem-solving; (7) emotional management and assertiveness; (8) life values and goal setting; (9) time management and reinforcement activities; and (10) exercise, postural and sleep hygiene and relapse prevention.

Each session consists of three parts: first, a review of doubts and the tasks presented in the previous session; second, a discussion of the contents corresponding to the current session; and third, an overview of the tasks for the following session.

In addition to providing a handbook for the therapist, the program provides each patient with a dossier that includes a summary of the sessions and the tasks to accomplish as well as a CD audio guide on the breathing and relaxation exercises done in session 2.

The 40 patients assigned to the intervention were divided into three groups based on age and gender variables that will be detailed in Section "Results." The first consisted of 14 women ages 33–55, the second of 13 men ages 33–55, and the last of 13 patients (eight women and five men) between ages 55 and 69. The total compliance rates for the full sessions were 78% in group 1 and 69% in groups 2 and 3. In groups 2 and 3, the intervention was not applied to two patients and in group 1, it was not applied to one patient; one patient from group 1, three from group 2 and two from group 3 discontinued.

Procedure

This study was carried out in accordance with the recommendations of the Ethics Committee of the Southern Seville Health District (Andalusian Health Service) with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the South Seville Health District (Andalusian Health Service).

This study was carried out as part of a scientific-technical agreement with Southern Seville Primary Care. As part of this agreement, the second author of this study, MCG, then a post-graduate student, was selected through Ícaro[1], a blog to manage practices in business and employment, as the psychologist who would carry out the study. She was selected because of an impressive academic record and after receiving a positive evaluation in a personal interview. The first author, FJC, informed her of the aim of the intervention and the task she was going to carry out; gave her all the materials (slides, handbook, dossier for the patients and CDs to be used during relaxation techniques); and provided her with training in a 4-h session.

The first step was to get the healthcare personnel, doctors and nurses involved in patient information and recruitment. This task that was handled by the last author of this study, RM. Recruitment relied on the inclusion criteria specified in Section "Participants."

Following patient recruitment by the healthcare personnel, MCG informed the patients what the study entailed. The patients then signed the informed consent form and their first appointment was scheduled. During that appointment, each patient had a one-on-one interview with an undergraduate psychology student instructed in the application of the measures to be used in the study. Next, they participated in the group intervention with MCG. The sessions were held in a meeting room in the center with audiovisual equipment and mats for the participants to do the breathing and relaxation exercises.

Formative assessments (Chacón-Moscoso et al., 2013) were done throughout the process (before, during and after the implementation of the program). Immediately after the program ended and 6 months later, another assessment session with a one-on-one interview like the one described above was held.

All the data collected before the intervention, immediately afterward and 6 months later were anonymously added to a database by interning students from the authors' departments and supervised by two of the authors, SS and SC, who also did the statistical analysis using the SPSS 22.

Statistical Analyses

Cronbach's (1951) α was used to test the reliability of the measures gauged with psychological tests and comprising more than one item, specifically the pain interference subscale of WHYMPI, POMS (subscales and global score), and BDI. Additionally, given the small sample size, in order to obtain a more precise reliability coefficient the unbiased estimator of Cronbach's α was calculated (Feldt et al., 1987); and the significance of each unbiased estimator was calculated using the procedure of Kristof (1963) and Feldt et al. (1987). Following criteria established by George and Mallery (2003), values above 0.9 were considered excellent; between 0.8 (excluded) and 0.9 (included), good; between 0.7 (excluded) and 0.8 (included), acceptable; and between 0.6 (excluded) and 0.7 (included), questionable. Following criteria by Huh et al. (2006), values equal or higher than 0.7 were considered appropriate.

To study the changes to the different dependent variables across the three measurement instances (pre-test, post-test and follow-up), we first checked the normality assumption using Shapiro–Wilk's test –W– (Shapiro and Wilk, 1965), adequate for small samples ($N \leq 50$). When normal distribution was rejected ($p \leq 0.05$), we used a non-parametric test (Friedman test); when this assumption was accepted ($p > 0.05$), we calculated a parametric test (ANOVA for repeated measures). In the case of ANOVA, Mauchly's test of sphericity was calculated. When sphericity was assumed ($p > 0.05$), no correction of degrees of freedom *(df)* of F distribution was made; when it was rejected ($p \leq 0.05$), *df* were multiplied by Greenhouse–Geisser's epsilon to correct them.

Additionally, linear and quadratic trend contrasts were used to compare the three levels (pre-test, post-test, and follow-up). ANOVA trend analysis was used as a parametric test and showed results to be statistically significant when $p < 0.05$. As a non-parametric test, *post hoc* comparisons for trends were used (Marascuilo and McSweeney, 1967); here results were statistically significant when zero was not included in the interval

[1] https://icaro.ual.es/

obtained with a confidence level of 0.95. A significant linear trend would be interpreted as an increase, or at least maintenance, of changes detected in post-test during follow-up. In our case, this would be ideal. A significant quadratic trend would be interpreted as a reversal of the change detected in post-test during follow-up.

To calculate the effect size in the case of ANOVA, the partial eta or omega squared index can be overestimated in repeated measure designs (Olejnik and Algina, 2003). For this reason, we proceeded to calculate r^2 by dividing the sum of squares of the intra-subject by the addition of the sum of squares of the intra-subject, the sum of squares of the intra-subject error and the sum of squares of the within-subject error. To calculate the effect size in the case of Friedman test, we calculated Kendall's *W* coefficient of concordance, considered a strength-of-relationship index. It ranges from 0 to 1. Higher values indicate a stronger relationship (Green and Salkind, 2010). To interpret the effect size, we follow the conventional levels (Cohen, 1992) of effect size: small (0.01), medium (0.06), and large (0.16).

Finally, we used the IMMPACT clinical importance criteria (Dworkin et al., 2008). In terms of pain intensity, score drops (mean differences) between 1 and 2.9 were considered scarcely important; 3–4.9, moderately important; and above 5, substantial. In terms of the WHYMPI interference subscale, score drops equal to or higher than 0.6 are considered clinically important. For the POMS subscales, a reduction (or increase in the case of Vigor) of the score equal to or higher than two points is considered clinically important. In the case of the scale total, the required reduction is at least 10 points. Finally, in terms of the patient's perception of improvement (PGIC), *minimally improved* (category 5) suggests a minor change, *much improved* (category 6), a moderately important change, and *very much improved* (category 7), a substantial change. In all cases, we compared the score obtained in pre-test with post-test and pre-test with follow-up.

RESULTS

Forty patients participated in the study. The age range was 33–69, with an average age of 47.9 and a standard deviation of 8.68. Twenty-two patients (55%) were women and 18 (45%) were men; 38 (80%) were married or lived with a partner; four (10%) were separated or divorced; three (7.5%) were single; and one (2.5%) was widowed. Eighteen (45%) had finished only elementary school and 10 (25%) had not; 11 (27.5%) had received their high school degree; and only one (2.5%) had attended college. In terms of employment, nine (22.5%) were unemployed; nine (22.5%) were housewives; 12 (30%) worked; eight (20%) had received early retirement for illness; one (2.5%) had retired after reaching retirement age; and one (2.5%) was laid off. According to their diagnoses, 22 (55%) were suffering from chronic low back pain, 12 (30%) from fibromyalgia and the remaining six (15%) from chronic headaches. They had been dealing with chronic pain for 2–30 years, with an average of 16.75 years and a standard deviation of 9.14 years. In 22 (55%) of the cases, the patient's support person was their partner or spouse; in 13 (32.5%) of the cases, their father or mother; and in the remaining five (12.5%), other people. In 30 (75%) of the cases, the support person lived with the patient.

One important advantage of this intervention program is its high degree of standardization and specificity, aspects that facilitate its assessment and its replication and, as a consequence, allow its results to be generalized. Next we present the evaluation of the intervention program before, during and after the implementation.

Before the Intervention: Needs Assessment, and Evaluation of Objectives and Design

In general, as this stage was based on IMMPACT recommendations, the objectives, design and instruments used to measure the aspects that the intervention aims to improve were all based on empirical evidence and a theoretical framework.

In order to facilitate the comparison with the results (measures before and after the intervention), information about the scores obtained by the sample before the intervention and its reliability are presented in Section "After the Intervention: Evaluation of Outcomes."

The study of the internal coherence of the program yielded adequate results: all the needs had an associated objective, and at least one activity was included for each objective. Specifically, sessions 2 (training in breathing and relaxation) and 10 (physical activity, sleep and postural hygiene, and relapse prevention) were developed to reduce perceived pain; sessions 3 (attention management), 6 (problem solving), 8 (life values and goal setting), 9 (time management and reinforcement activities) and 10 were implemented to reduce the degree to which pain interferes in a patient's life; sessions 4 and 5 (cognitive restructuring), 6, and 7 (emotions management and assertiveness) were developed in order to improve mood; and all activities (from session 1, the introduction to cognitive-behavioral intervention, through session 10) had a positive influence on patient's perceived satisfaction. Additionally, the timeframe was realistic and the materials available for each activity were made explicit.

During the Intervention: Evaluation of Implementation

As a measure of participant willingness, **Figure 1** presents a participant flow chart in keeping with CONSORT recommendations (Moher et al., 2001).

After the Intervention: Evaluation of Outcomes

Reliability

Table 1 presents the reliability results. All were significant at 95% CI. Considering the unbiased estimator of Cronbach's α, six (22.2%) were excellent, 10 (37%) were good, nine (33.3%) were acceptable, and two (7.5%) were questionable (the subscale

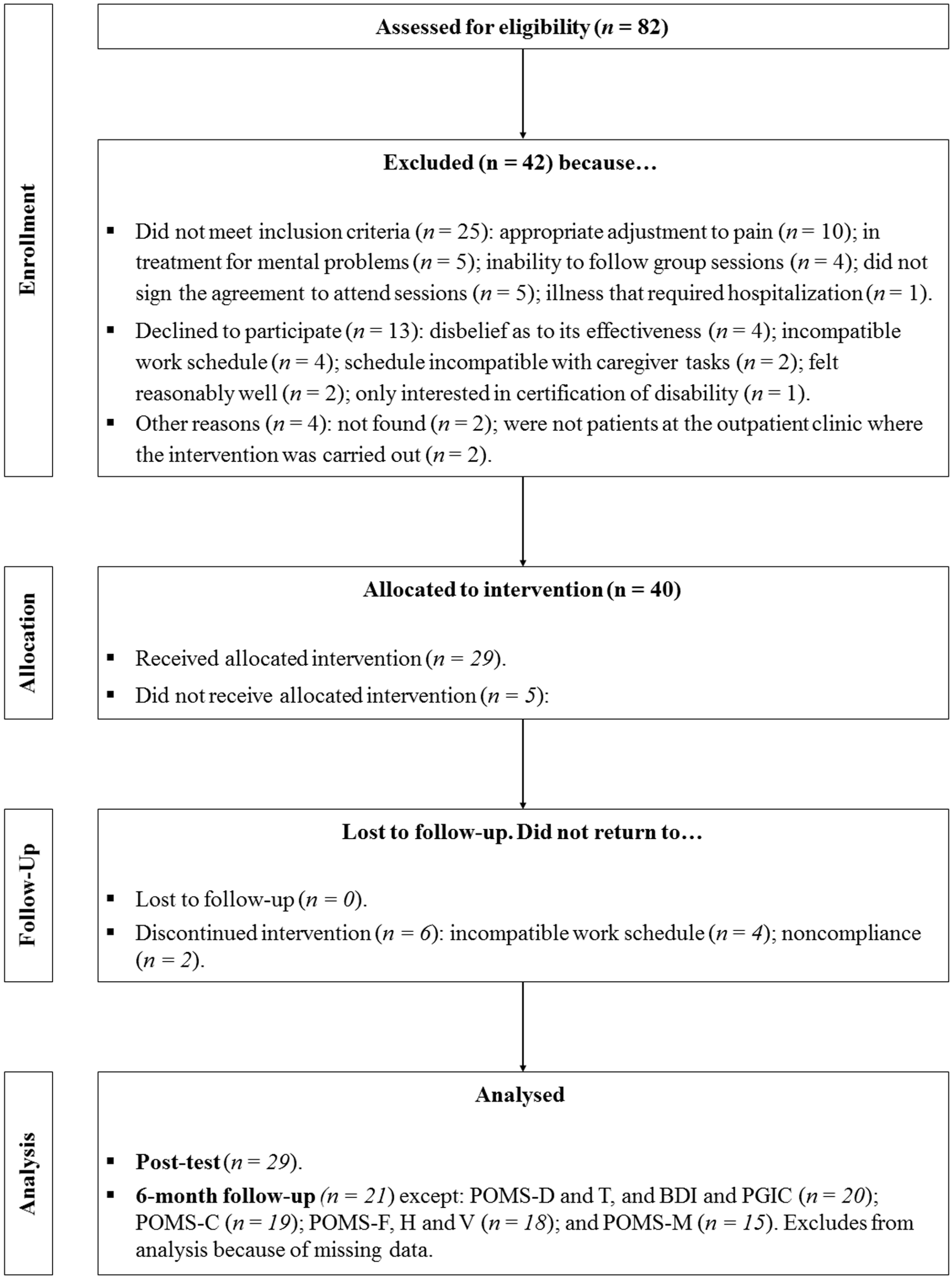

Enrollment
Assessed for eligibility (*n* = 82)
Excluded (n = 42) because…
▪ Did not meet inclusion criteria (*n* = 25): appropriate adjustment to pain (*n* = 10); in treatment for mental problems (*n* = 5); inability to follow group sessions (*n* = 4); did not sign the agreement to attend sessions (*n* = 5); illness that required hospitalization (*n* = 1).
▪ Declined to participate (*n* = 13): disbelief as to its effectiveness (*n* = 4); incompatible work schedule (*n* = 4); schedule incompatible with caregiver tasks (*n* = 2); felt reasonably well (*n* = 2); only interested in certification of disability (*n* = 1).
▪ Other reasons (*n* = 4): not found (*n* = 2); were not patients at the outpatient clinic where the intervention was carried out (*n* = 2).
Allocation
Allocated to intervention (n = 40)
▪ Received allocated intervention (*n = 29*).
▪ Did not receive allocated intervention (*n = 5*):
Follow-Up
Lost to follow-up. Did not return to…
▪ Lost to follow-up (*n = 0*).
▪ Discontinued intervention (*n = 6*): incompatible work schedule (*n = 4*); noncompliance (*n = 2*).
Analysis
Analysed
▪ **Post-test** (*n = 29*).
▪ **6-month follow-up** *(n = 21)* except: POMS-D and T, and BDI and PGIC (*n = 20*); POMS-C (*n = 19*); POMS-F, H and V (*n = 18*); and POMS-M (*n = 15*). Excludes from analysis because of missing data.

FIGURE 1 | Consolidated Standards of Reporting Trials (CONSORT) flow chart of participants through the study (Moher et al., 2001). WHYMPI, West Haven-Yale Multidimensional Pain Inventory; POMS, Profile of Mood States; F, fatigue; D, depression; T, tension; H, hostility; C, confusion; V, vigor; M, total mood disturbance; BDI, Beck Depression Inventory; PGIC, Patient Global Impression of Change Scale.

TABLE 1 | Reliability.

	Pre-test					Post-test					Follow-up				
	α	*N*	$\bar{\alpha}$	*F*	*p*	α	*N*	$\bar{\alpha}$	*F*	*p*	α	*N*	$\bar{\alpha}$	*F*	*p*
WHYMPI	0.728	40	0.796	4.902	<0.001	0.744	29	0.795	4.883	<0.001	0.765	21	0.922	12.766	<0.001
POMS-F	0.717	38	0.732	3.735	<0.001	0.800	29	0.814	5.385	<0.001	0.889	18	0.902	10.210	<0.001
POMS-D	0.845	38	0.853	6.820	<0.001	0.907	29	0.914	11.580	<0.001	0.878	20	0.891	9.161	<0.001
POMS-T	**0.666**	38	**0.684**	3.165	<0.001	0.782	29	0.798	4.940	<0.001	**0.630**	20	**0.669**	3.021	<0.001
POMS-H	0.829	38	0.838	6.182	<0.001	0.886	29	0.894	9.447	<0.001	0.705	18	0.740	3.842	<0.001
POMS-C	0.765	38	0.778	4.498	<0.001	0.766	29	0.783	4.602	<0.001	0.819	19	0.839	6.215	<0.001
POMS-V	0.701	38	0.717	3.536	<0.001	0.804	29	0.818	5.495	<0.001	0.739	18	0.770	4.342	<0.001
POMS-M	0.932	38	0.936	15.546	<0.001	0.960	29	0.963	26.923	<0.001	0.978	15	0.981	53.030	<0.001
BDI	0.876	38	0.883	8.525	<0.001	0.852	29	0.863	7.277	<0.001	0.860	20	0.875	7.983	<0.001

WHYMPI, West Haven-Yale Multidimensional Pain Inventory; POMS, Profile of Mood States; F, fatigue; D, depression; T, tension; H, hostility; C, confusion; V, vigor; M, total mood disturbance; BDI, Beck Depression Inventory; $\bar{\alpha}$= unbiased estimator of Cronbach's α. α and $\bar{\alpha}$ lower than 0.7 are marked in bold.

Tension of POMS in the pre-test and the follow-up). Overall, 25 (92.6%) of the results reached at least appropriate values (above 0.7) and the remaining two (7.5%) were close to 0.7 (concretely, 0.684 and 0.669).

Normality

Considering the 14 variables and the three instances separately ($14 \times 3 = 42$ combinations), the normality assumption using Shapiro–Wilk (W) was accepted on all occasions but nine: 24-h intensity, follow up ($W = 0.857$, $p = 0.027$); 24-h interference, pre-test ($W = 0.639$, $p < 0.001$) and follow-up ($W = 0.798$, $p = 0.005$); present interference pre-test ($W = 0.639$, $p < 0.001$) and follow-up ($W = 0.849$, $p = 0.022$); POMS-V, follow-up ($W = 0.841$, $p = 0.017$); BDI, pre-test ($W = 0.834$, $p = 0.003$); and PGIC post-test ($W = 0.816$, $p = 0.008$) and follow up ($W = 0.851$, $p = 0.023$).

As a result, the calculations for the six variables affected by normality rejection in at least one instance (24-h intensity, 24-h interference, present interference, POMS-V, BDI and PGIC), were done using non-parametric tests.

Effectiveness of the Psychological Intervention

Pain

Table 2 presents the results. In terms of pain, both the pain intensity present at the time of the interview and the pain experienced in the 24 h beforehand diminished in a statistically significant manner after the intervention, with a large effect size.

In present intensity, the clinical significance was minimally important and both linear and quadratic trends were significant. The quadratic trend was stronger, however, with a large effect size, while the effect size for the linear trend was medium. This can be interpreted as a slight maintenance of results obtained in post-test at follow-up.

On pain intensity in the previous 24 h, we found a minimally important change when pre and post-test results were compared, and no change in the pre-test and follow-up comparison. The quadratic trend was statistically significant. This suggests that after the intervention, there was a decrease in 24-h pain intensity, but an increase 6 months later.

Physical functioning

The 24-h and present pain interference and the WHYMPI interference score diminished in a statistically significant manner after the intervention with a large effect size.

The clinical significance in WHYMPI was substantial in the pre–post comparison and moderately important when comparing pre-test and follow-up. The significant linear and quadratic trends with medium effect size revealed that, although there was a slight deterioration, the improvement continued in the follow-up period.

There was a statistically significant deterioration with regards to 24 h-interference in the follow-up period (significant quadratic trend). Nevertheless, the improvement in present interference continued in the follow-up period (significant linear trend).

Emotional functioning

In general, we can say that there was a statistically significant improvement in POMS and BDI. The effect size was medium/large in all the variables. In all cases, the clinical significance implied a substantial change when comparing pre-test and post-test. Additionally, the quadratic trend was statistically significant in all cases. This can be interpreted as an important deterioration in a comparison of the post-test and follow-up. Comparing the clinical significance at pre-test and follow-up, we find that the deterioration does not represent a return to the starting point in all the variables studied, because there is a substantial change in POMS-T, POMS-H, and POMS-V (with this last variable also showing a significant linear trend), and a moderately important change in POMS-F, POMS-D and POMS-M. Moreover, BDI also yielded a significant linear trend in favor of a possible maintenance of the results obtained.

Improvement perceived by the patient

Patient Global Impression of Change Scale shows that the improvement participants expected before the intervention was statistically lower than the subjective improvement perceived by the participants after the intervention, with a large effect size and a moderately important clinical change. This variable presents a statistically significant trend both linearly and quadratically,

TABLE 2 | Global comparison at the different instances of measurement, trend contrasts and clinical significance when comparing pre-test to post-test and to follow-up.

Variable	Pre-test		Post-test		Follow-up			Global		^{f}Clinical significance		Linear trend		Quadratic trend	
	M	*SD*	*M*	*SD*	*M*	*SD*	*N*	Statistic	ES	Pre–post	Pre-follow-up	Statistic	ES	Statistic	ES
24-h Inten	7.52	1.03	4.95	1.72	6.76	2.36	21	^{b}19.9***	^{d}0.47^{+++}	2.57 ♠	0.76	^{h}[−0.0140, 1.054]	—	^{h}**[1.215, 3.065]**	—
Pr Inten	6.71	1.23	4.05	1.56	5	2.88	21	^{a,c}10.37**	^{e}0.3^{+++}	2.66♠	1.71 ♠	^{c}5.6*	^{e}0.13^{++}	^{c}24.24***	^{e}0.25^{+++}
24-h Inter	2.43	0.55	1.52	0.91	2.24	0.89	21	^{b}10.26**	^{d}0.24^{+++}	—	—	^{h}[−0.224, 0.844]	—	^{h}**[0.425, 2.275]**	—
Pr Inter	2.33	0.66	1.52	0.87	1.52	1.08	21	^{b}14.06**	^{d}0.34^{+++}	—	—	^{h}**[0.366, 1.434]**	—	^{h}[−0.205, 2.275]	—
WHYMPI	4.12	0.73	3.33	0.74	3.7	0.86	21	^{c}9.62***	^{e}0.15^{+++}	0.79♠♠♠	0.42♠♠	^{c}5.08*	^{e}0.06^{++}	^{c}14.62**	^{e}0.13^{++}
POMS-F	15.33	5.76	10	5.29	13.63	7.28	18	^{c}4.59*	^{e}0.1^{++}	5.33♠♠♠	1.7♠♠	^{c}0.75	^{e}0.01^{+}	^{c}7.6*	^{e}0.11^{++}
POMS-D	22.65	10.96	14.40	11.05	20.70	16.46	20	^{c}4.60	^{e}0.07^{++}	8.25♠♠♠	1.95♠♠	^{c}0.428	^{e}0.01^{+}	^{c}9.673**	^{e}0.08^{++}
POMS-T	19.15	4.74	15.20	7.01	17.05	9.15	20	^{c}3.26	^{e}0.05^{++}	3.95♠♠♠	2.10♠♠♠	^{c}1.63	^{e}0.02^{+}	^{c}0.355*	^{e}0.04^{+}
POMS-H	20.06	8.01	14	9.41	16.50	12.55	18	^{c}3.34	^{e}0.06^{++}	6.06♠♠♠	3.56♠♠♠	^{c}2.241	^{e}0.02^{+}	^{c}4.476*	^{e}0.05^{++}
POMS-C	13.47	5.89	9.89	5.42	11.95	6.75	19	^{c}4.03*	^{e}0.06^{++}	3.58♠♠♠	1.52♠	^{c}1.678	^{e}0.01^{+}	^{c}5.818*	^{e}0.05^{++}
POMS-V	10.89	4.78	16.37	6.49	13.42	4.83	18	^{b}12.10**	^{d}0.32^{+++}	−5.48♠♠♠	−2.53♠♠♠	^{h}**[0.026, 1.094]**	—	^{h}**[0.575, 2.425]**	—
POMS-M	81.40	36.81	47.33	41.17	73.33	52.77	15	^{c}5.03*	^{e}0.1^{++}	34.07♠♠♠	8.07♠♠	^{c}0.583	^{e}0.01^{+}	^{c}8.548*	^{e}0.14^{++}
BDI	17.95	10.36	10.15	7.62	33.45	8.61	20	^{b}31.3***	^{d}0.78^{+++}	—	—	^{h}**[−1.634,−0.569]**	—	^{h}**[1.475, 3.325]**	—
PGIC	5.00	0	6.20	0.70	5.85	0.93	20	^{b}22.07***	^{d}0.55^{+++}	^{g}♠♠	^{g}♠♠	^{h}**[0.386, 1.454]**	—	^{h}**[0.655, 2.505]**	—

24-h, 24-hour; Inten, intensity; Pr, present; Inter, interference; WHYMPI, West Haven-Yale Multidimensional Pain Inventory; POMS, Profile of Mood States; F, fatigue; D, depression; T, tension; H, hostility; C, confusion; V, vigor; M, total mood disturbance; BDI, Beck Depression Inventory; PGIC, Patient Global Impression of Change Scale. $^{+}$*, small effect size (around 0.01);* $^{++}$*, medium effect size (around 0.06); and* $^{+++}$*, large effect size (around 0.16). Following Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) recommendations, ♠, minimally important change; ♠♠, moderately important change; and ♠♠♠, substantial change.*

^{a}*Sphericity was not assumed (Mauchly's test of sphericity* $p < 0.001$*).* ^{b}*Friedman test.* ^{c}*ANOVA F for repeated measures.* ^{d}*Kendall's coefficient of concordance W.* $^{e}r^{2}$*.* ^{f}*Mean difference.* ^{g}*More detailed information can be consulted on* ***Table 3****.* ^{h}*Non-parametric post hoc comparisons for trend, 95% confidence interval (statistically significant trends –zero excluded from the confidence interval are in bold).*

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

so it can be concluded that participants maintain their positive assessment when comparing post-test and follow-up.

In more detail, **Table 3** shows that, at post-test, all patients noted improvement, with more than half reporting a moderately important change and around one-third reporting substantial change (the maximum). At the 6-month follow-up, two patients reported that their chronic pain was similar to what it had been before the intervention. However, approximately half noted a moderately important improvement and one-fourth, substantial improvement. Overall, 90% of the patients stated that they had improved 6 months after the intervention.

DISCUSSION

This study has provided additional evidence on the generalization of multicomponent interventions that have been already shown in other contexts (Morley et al., 1999; Veehof et al., 2011; Hann and McCracken, 2014; Huguet et al., 2014). While such interventions are usually implemented in English-speaking contexts, this paper presents an implementation in a Spanish rural area. While reported interventions are generally performed in a very controlled context, the sample of this study was selected from among users of a public health center who came in for a consultation. Participants in most studies are usually upper-middle class with a high educational level; 70% of participants in this intervention had a low educational level (complete or incomplete elementary) and 45% had no paid work and, as a result, low income. Finally, it is usual to find a limited number of domains of painful experience to evaluate interventions; in this case, we evaluated all the domains of chronic pain using instruments recommended by IMMPACT, i.e., to measure pain, intensity of perceived pain the previous 24 h and at the time of the interview (Dworkin et al., 2005). To measure physical functioning, we utilized the items referring to pain interference in daily life in the previous 24 h and at the time of the interview, and WHYMPI (Kerns et al., 1985). POMS and BDI were used to gauge emotional functioning. To measure perceived improvement after the treatment, PGIC was used.

Patient flow data were similar to those of other studies. Wetherell et al. (2011) carried out a randomized controlled trial comparing acceptance and compromise therapy with cognitive behavioral therapy in patients with chronic pain. They reported that 66% of patients were excluded from the recruitment, 12% of patients did not receive the intervention, and 16% of patients dropped out. Our percentages were 51, 12.5, and 21%, respectively. The principal reasons for exclusion and drop out of our study were similar to those reported by Wetherell et al. (2011): schedule incompatibilities, adverse life events and non-compliance.

In spite of the variants our study introduced to the standard intervention, the program assessment showed a high degree of standardization and specification owed to its highly detailed design (Kovacs and Moix, 2011; Moix and Casado, 2011). Moreover, the evaluation followed the IMMPACT recommendations, using instruments with tested psychometric properties. This facilitates the replication of the intervention and reinforces the results obtained. Second, there was a high degree of internal coherence. The same measures taken before the intervention were repeated immediately after and again 6 months later, using the same instruments. This comparison of the three instances facilitated the analysis of the change and provided evidence not only of the program's effectiveness but also of the duration of the effects for a longer period of time. Each assessed need had at least an associated objective to be covered and each objective had at least one activity to be reached, and fitted timeframe and resources. Third, explicit selection criteria for participants were applied to all potential participants (Chacón-Moscoso et al., 2016). Forth, the measures presented sufficient reliability coefficients. Fifth, we found evidence of effectiveness, as there was a statistically significant improvement after the intervention or at least a medium effect size in all the variables measured and all the domains taken into account; and substantial clinical change in 75% of the variables measured.

From our point of view, the main contributions of the study is to demonstrate that cognitive-behavioral therapy can be effective even if performed by an inexperienced therapist to groups of low-literacy patients with a low socioeconomic status. As for therapist experience, although common sense suggests that it should improve the effectiveness of therapy, the first longitudinal study that addresses this question, with data from 170 psychotherapists and 6,591 patients (Goldberg et al., 2016), did not endorse this. In our opinion, the highly structured intervention program and the wealth of resources and material available to the therapist minimize the possible impact of their inexperience. In terms of the second aspect, literacy and socioeconomic resources are considered a barrier for the efficacy of cognitive behavioral treatment of chronic pain (Campbell, 2011) and this led to the

TABLE 3 | Improvement perceived by patients after the intervention in the Patient Global Impression of Change (PGIC) scale.

	Category	Post (*N* = 29)		Follow-up (*N* = 20)	
		f	%	*f*	%
No change	4	0	0	2	10
Minimally improved♠	5	3	10.3	4	20
Much improved♠♠	6	17	58.6	9	45
Very much improved♠♠♠	7	9	31	5	25

Following Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) recommendations, ♠, minimally important change; ♠♠, moderately important change; and ♠♠♠, substantial change.

creation of personalization initiatives for these patients (Thorn et al., 2011; Eyer and Thorn, 2016). Even so, in the first study with personalized treatment (Thorn et al., 2011), 26.5% of patients did not complete the intervention, which is 5.5% more than in our study. This could be explained by the therapist's familiarity with the patients and by the effort that she carried out to make the program contents understandable for the patients.

The improvement observed just after the intervention worsened in approximately two-thirds of the variables measured (only the quadratic trend was statistically significant), though the measures did not return to their starting points. The ostensibly mild deterioration is still strong enough to be statistically significant. Maintaining the long-term effects of these programs is another major challenge, considering the high chronicity of these patients (in our study, patients had been suffering from chronic pain for over 16 years on average). A possible moderating factor could be the quantity and quality of homework, a neglected aspect of cognitive behavioral therapy research, the importance of which has been revealed in a recent meta-analysis (Kazantzis et al., 2016). Anyway, it would be highly advisable to add some sessions after the intervention, one every 4 months, to maintain the improvements patients have obtained.

On the other hand, the principal limitation was the absence of a control group that would have enhanced the design and increased evidence of the intervention's effectiveness. Nevertheless, a control group would not have been feasible in this study, because we were ethically obliged to offer the intervention program to every patient in a public primary care setting. In any case, we were less interested in the program efficacy than in identifying who could benefit from the intervention.

Further research is going to take two directions. First, we are going to adapt the intervention to a broader potential population. People with a disability such as deafness, blindness or dementia were excluded from the initial intervention, but we trust that it is possible to adapt the intervention to cases such as these. Second, in order to increase the evidence of the efficacy of the intervention applied in this study (Moix and Casado, 2011), a meta-analysis will be developed. This will assist us in obtaining a global effect size after a statistical synthesis of the results obtained in the different interventions while also allowing us to detect possible moderator variables that influence the effectiveness of these interventions. From this study, we would be able to establish practical recommendations for psychologists to increase the likelihood of success of this kind of programs.

AUTHOR CONTRIBUTIONS

FC-G came up with the initial idea and design. RM-B recruited the sample. MG-O carried out the intervention. SS-C and SC-M performed the analyses and interpreted the data. FC-F, SS-C, and SC-M were entrusted with drafting the manuscript. All the authors reviewed the manuscript, approved the final version to be published, and agree to be accountable for all aspects of the work, ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

FUNDING

This research was funded by the following project grants: PSI2011-29587 (the Spanish Ministry of Science and Innovation); 1150096 (the Scientific and Technological Development Fund of Chile -FONDECYT-); and PSI2015-71947-REDT (the Spanish Ministry of Economy and Competitiveness).

ACKNOWLEDGMENTS

This work would not have been possible without the collaboration of Southern Seville Primary Care, which opened its door to us, and of the professionals at the Príncipe de Asturias Health Center (Utrera, Spain), led by Dr. José Sánchez-Blanco.

The authors greatly appreciate all the comments received from the reviewers and the English language editor Wendy Gosselin. We believe that the quality of this paper has been substantially enhanced as a result.

REFERENCES

Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., et al. (2001). The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann. Intern. Med.* 134, 663–694. doi: 10.7326/0003-4819-134-8-200104170-00012

Andrew, R., Derry, S., Taylor, R. S., Straube, S., and Phillips, C. J. (2014). The costs and consequences of adequately managed chronic non-cancer pain and chronic neuropathic pain. *Pain Pract.* 14, 79–94. doi: 10.1111/papr.12050

APA (2010). *Publication Manual of The American Psychological Association*, 6th Edn. Washington, DC: American Psychological Association.

APA (2016). *Treatments for Chronic Or Persistent Pain*. Available at: http://www.div12.org/psychological-treatments/disorders/chronic-or-persistent-pain/ [accessed September 25, 2016].

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., and Erbaugh, J. (1961). An inventory for measuring depression. *Arch. Gen. Psychiatry* 4, 561–571. doi: 10.1001/archpsyc.1961.01710120031004

Bennett, R. M., Jones, J., Turk, D. C., Russell, I. J., and Matallana, L. (2007). An internet survey of 2,596 people with fibromyalgia. *BMC Musculoskelet. Disord.* 8:27. doi: 10.1186/1471-2474-8-27

Breivik, H., Collett, B., Ventafridda, V., Cohen, R., and Gallacher, D. (2006). Survey of chronic pain in Europe: prevalence, impact on daily life, and treatment. *Eur. J. Pain* 10, 287–287. doi: 10.1016/j.ejpain.2005.06.009

Campbell, L. C. (2011). Addressing literacy as a barrier in delivery and evaluation of cognitive-behavioral therapy for pain management. *Pain* 152, 2679–2680. doi: 10.1016/j.pain.2011.09.004

Chacón-Moscoso, S., Sanduvete-Chaves, S., Portell-Vidal, M., and Anguera-Argilaga, M. T. (2013). Reporting a program evaluation: needs, program plan, intervention, and decisions. *Int. J. Clin. Health Psychol.* 13, 58–66. doi: 10.1016/S1697-2600(13)70008-5

Chacón-Moscoso, S., Sanduvete-Chaves, S., and Sánchez-Martín, M. (2016). The development of a checklist to enhance methodological quality in intervention programs. *Front. Psychol.* 7:1811. doi: 10.3389/fpsyg.2016.01811

Chacón-Moscoso, S., Shadish, W. R., and Cook, T. D. (2008). "Diseños de intervención media," in *Evaluación de Programas Sociales y Sanitarios*, eds M. T.

Anguera, S. Chacón-Moscoso, and A. Blanco-Villaseñor (Madrid: Sïntesis), 185–218.

Cherkin, D. C., Sherman, K. J., Balderson, B. H., Cook, A. J., Anderson, M. L., Hawkes, R. J., et al. (2016). Effect of mindfulness-based stress reduction vs cognitive behavioral therapy or usual care on back pain and functional limitations in adults with chronic low back pain: a randomized clinical trial. *JAMA* 315, 1240–1249. doi: 10.1001/jama.2016.2323

Cohen, J. (1992). A power primer. *Psychol. Bull.* 112, 155. doi: 10.1037/0033-2909.112.1.155

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334. doi: 10.1007/BF02310555

Dworkin, R. H., Turk, D. C., Farrar, J. T., Haythornthwaite, J. A., Jensen, M. P., Katz, N. P., et al. (2005). Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. *Pain* 113, 9–19. doi: 10.1016/j.pain.2004.09.012

Dworkin, R. H., Turk, D. C., Mcdermott, M. P., Peirce-Sandner, S., Burke, L. B., Cowan, P., et al. (2009). Interpreting the clinical importance of group differences in chronic pain clinical trials: IMMPACT recommendations. *Pain* 146, 238–244. doi: 10.1016/j.pain.2009.08.019

Dworkin, R. H., Turk, D. C., Wyrwich, K. W., Beaton, D., Cleeland, C. S., Farrar, J. T., et al. (2008). Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *J. Pain* 9, 105–121. doi: 10.1016/j.jpain.2007.09.005

Ehde, D. M., Dillworth, T. M., and Turner, J. A. (2014). Cognitive-behavioral therapy for individuals with chronic pain: efficacy, innovations, and directions for research. *Am. Psychol.* 69, 153–165. doi: 10.1037/a0035747

Eyer, J. C., and Thorn, B. E. (2016). The learning about my pain study protocol: reducing disparities with literacy-adapted psychosocial treatments for chronic pain, a comparative behavioral trial. *J. Health Psychol.* 21, 2063–2074. doi: 10.1177/1359105315570985

Farrar, J. T. (2003). "Participant ratings of global improvement," in *Proceedings of the Second Meeting of the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT-II)*, Washington, DC.

Feldt, L. S., Woodruff, D. J., and Salih, F. A. (1987). Statistical inference for coefficient alpha. *Appl. Psychol. Meas.* 11, 93–103. doi: 10.1177/014662168701100107

Ferrer, V. A., González, R., and Manassero, M. A. (1993). El west haven yale multidimensional pain questionnaire: un instrumento para evaluar al paciente con dolor crónico. *Dolor* 8, 153–160.

Gatchel, R. J., Peng, Y. B., Peters, M. L., Fuchs, P. N., and Turk, D. C. (2007). The biopsychosocial approach to chronic pain: scientific advances and future directions. *Psychol. Bull.* 133, 581. doi: 10.1037/0033-2909.133.4.581

George, D., and Mallery, P. (2003). *SPSS for Windows Step by Step: A Simple Guide and Reference, 11.0 Update*, 4th Edn. Boston, MA: Allyn & Bacon.

Goldberg, S. B., Rousmaniere, T., Miller, S. D., Whipple, J., Nielsen, S. L., Hoyt, W. T., et al. (2016). Do psychotherapists improve with time and experience? A longitudinal analysis of outcomes in a clinical setting. *J. Couns. Psychol.* 63, 1–11. doi: 10.1037/cou0000131

Green, S. B., and Salkind, N. J. (2010). *Using SPSS for Windows and Macintosh: Analyzing and Understanding Data*. Upper Saddle River, NJ: Prentice Hall Press.

Guy, W. (ed.). (1976). "Clinical global impression scale," in *ECDEU Assessment Manual for Psychopharmacology-Revised* (Rockville, MD: US Department of Health, Education and Welfare), 218–222.

Hann, K. E., and McCracken, L. M. (2014). A systematic review of randomized controlled trials of acceptance and commitment therapy for adults with chronic pain: outcome domains, design quality, and efficacy. *J. Contextual Behav. Sci.* 3, 217–227. doi: 10.1016/j.jcbs.2014.10.001

Haythornthwaite, J. (2003). "The assessment of pain-related physical function for clinical trials in chronic pain," in *Proceedings of the Second Meeting of the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT-II)*, Washington, DC.

Haythornthwaite, J. (2004). "Profile of Mood States," in *Proceedings of the Fourth Meeting of the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT-IV)*, Washington, DC.

Huguet, A., Mcgrath, P. J., Stinson, J., Tougas, M. E., and Doucette, S. (2014). Efficacy of psychological treatment for headaches: an overview of systematic reviews and analysis of potential modifiers of treatment efficacy. *Clin. J. Pain* 30, 353–369. doi: 10.1097/AJP.0b013e318298dd8b

Huh, J., Delorme, D. E., and Reid, L. N. (2006). Perceived third-person effects and consumer attitudes on prevetting and banning DTC advertising. *J. Consum. Aff.* 40, 90–116. doi: 10.1111/j.1745-6606.2006.00047.x

Johnson, M., Collett, B., and Castro-Lopes, J. M. (2013). The challenges of pain management in primary care: a pan-European survey. *J. Pain Res.* 6, 393. doi: 10.2147/JPR.S41883

Kazantzis, N., Whittington, C., Zelencich, L., Kyrios, M., Norton, P. J., and Hofmann, S. G. (2016). Quantity and quality of homework compliance: a meta-analysis of relations with outcome in cognitive behavior therapy. *Behav. Ther.* 47, 755–772. doi: 10.1016/j.beth.2016.05.002

Kerns, R. D., Turk, D. C., and Rudy, T. E. (1985). The west haven-yale multidimensional pain inventory (WHYMPI). *Pain* 23, 345–356. doi: 10.1016/0304-3959(85)90004-1

Kovacs, F., and Moix, J. (2011). *Manual Del Dolor: Tratamiento Cognitivo Conductual Del Dolor Crónico*. Barcelona: Grupo Planeta.

Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika* 28, 221–238. doi: 10.1007/BF02289571

Kroner, J. W., Hershey, A. D., Kashikar-Zuck, S. M., Lecates, S. L., Allen, J. R., Slater, S. K., et al. (2016). Cognitive behavioral therapy plus amitriptyline for children and adolescents with chronic migraine reduces headache days to ≤ 4 per month. *Headache* 56, 711–716. doi: 10.1111/head.12795

Lauche, R., Cramer, H., Dobos, G., Langhorst, J., and Schmidt, S. (2013). A systematic review and meta-analysis of mindfulness-based stress reduction for the fibromyalgia syndrome. *J. Psychosom. Res.* 75, 500–510. doi: 10.1016/j.jpsychores.2013.10.010

Marascuilo, L. A., and McSweeney, M. (1967). Nonparametric post hoc comparisons for trend. *Psychol. Bull.* 67, 401–412. doi: 10.1037/h0020421

McGrath, P. J., Walco, G. A., Turk, D. C., Dworkin, R. H., Brown, M. T., Davidson, K., et al. (2008). Core outcome domains and measures for pediatric acute and chronic/recurrent pain clinical trials: PedIMMPACT recommendations. *J. Pain* 9, 771–783. doi: 10.1016/j.jpain.2008.04.007

Meeus, M., Nijs, J., Vanderheiden, T., Baert, I., Descheemaeker, F., and Struyf, F. (2014). The effect of relaxation therapy on autonomic functioning, symptoms and daily functioning, in patients with chronic fatigue syndrome or fibromyalgia: a systematic review. *Clin. Rehabil.* 29, 221–233. doi: 10.1177/0269215514542635

Melzack, R. (1999). From the gate to the neuromatrix. *Pain* 82, S121–S126. doi: 10.1016/S0304-3959(99)00145-1

Melzack, R., and Wall, P. D. (1967). Pain mechanisms: a new theory. *Surv. Anesthesiol.* 11, 89–90. doi: 10.1097/00132586-196704000-00002

Merskey, H. (1994). "Part III: pain terms, a current list with definitions and notes on usage," in *Classification of Chronic Pain, IASP Task Force on Taxonomy*, eds H. Merskey and N. Bogduk (Seattle, WA: IASP press), 209–214.

Moher, D., Schulz, K. F., Altman, D. G., and Group, C. (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 357, 1191–1194. doi: 10.1016/S0140-6736(00)04337-3

Moix, J., and Casado, M. (2011). Terapias psicológicas para el tratamiento del dolor crónico. *Clín. Salud* 22, 41–50. doi: 10.5093/cl2011v22n1a3

Morley, S., Eccleston, C., and Williams, A. (1999). Systematic review and meta-analysis of randomized controlled trials of cognitive behaviour therapy and behaviour therapy for chronic pain in adults, excluding headache. *Pain* 80, 1–13. doi: 10.1016/S0304-3959(98)00255-3

Nicholas, M. K., Linton, S. J., Watson, P. J., and Main, C. J. (2011). Early identification and management of psychological risk factors ("yellow flags") in patients with low back pain: a reappraisal. *Phys. Ther.* 91, 737–753. doi: 10.2522/ptj.20100224

Olejnik, S., and Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychol. Methods* 8:434. doi: 10.1037/1082-989x.8.4.434

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Wadsworth Cengage learning.

Shapiro, S. S., and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika* 52, 591–611. doi: 10.1093/biomet/52.3-4.591

Tesarz, J., Leisner, S., Gerhardt, A., Janke, S., Seidler, G. H., Eich, W., et al. (2014). Effects of eye movement desensitization and reprocessing (EMDR) treatment in chronic pain patients: a systematic review. *Pain Med.* 15, 247–263. doi: 10.1111/pme.12303

Thorn, B. E., Day, M. A., Burns, J., Kuhajda, M. C., Gaskins, S. W., Sweeney, K., et al. (2011). Randomized trial of group cognitive behavioral therapy compared with a pain education control for low-literacy rural people with chronic pain. *Pain* 152, 2710–2720. doi: 10.1016/j.pain.2011.07.007

Turk, D. C., Dworkin, R. H., Allen, R. R., Bellamy, N., Brandenburg, N., Carr, D. B., et al. (2003). Core outcome domains for chronic pain clinical trials: IMMPACT recommendations. *Pain* 106, 337–345. doi: 10.1016/j.pain.2003.08.001

Turk, D. C., Dworkin, R. H., Burke, L. B., Gershon, R., Rothman, M., Scott, J., et al. (2006). Developing patient-reported outcome measures for pain clinical trials: IMMPACT recommendations. *Pain* 125, 208–215. doi: 10.1016/j.pain.2006.09.028

Turk, D. C., Dworkin, R. H., Mcdermott, M. P., Bellamy, N., Burke, L. B., Chandler, J. M., et al. (2008). Analyzing multiple endpoints in clinical trials of pain treatments: IMMPACT recommendations. *Pain* 139, 485–493. doi: 10.1016/j.pain.2008.06.025

Veehof, M. M., Oskam, M.-J., Schreurs, K. M., and Bohlmeijer, E. T. (2011). Acceptance-based interventions for the treatment of chronic pain: a systematic review and meta-analysis. *PAIN* 152, 533–542. doi: 10.1016/j.pain.2010.11.002

Veehof, M. M., Trompetter, H. R., Bohlmeijer, E. T., and Schreurs, K. M. G. (2016). Acceptance- and mindfulness-based interventions for the treatment of chronic pain: a meta-analytic review. *Cogn. Behav. Ther.* 45, 5–31. doi: 10.1080/16506073.2015.1098724

Wetherell, J. L., Afari, N., Rutledge, T., Sorrell, J. T., Stoddard, J. A., Petkus, A. J., et al. (2011). A randomized, controlled trial of acceptance and commitment therapy and cognitive-behavioral therapy for chronic pain. *Pain* 152, 2098–2107. doi: 10.1016/j.pain.2011.05.016

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 04 July 2016
doi: 10.3389/fpsyg.2016.00983

Characterization of Vulnerable and Resilient Spanish Adolescents in Their Developmental Contexts

*Carmen Moreno[1], Irene García-Moya[1], Francisco Rivera[2] and Pilar Ramos[1]**

[1] Developmental and Educational Psychology, University of Seville, Sevilla, Spain, [2] Department of Behavioral Sciences, University of Huelva, Huelva, Spain

Research on resilience and vulnerability can offer very valuable information for optimizing design and assessment of interventions and policies aimed at fostering adolescent health. This paper used the adversity level associated with family functioning and the positive adaptation level, as measured by means of a global health score, to distinguish four groups within a representative sample of Spanish adolescents aged 13–16 years: maladaptive, resilient, competent and vulnerable. The aforementioned groups were compared in a number of demographic, school context, peer context, lifestyles, psychological and socioeconomic variables, which can facilitate or inhibit positive adaptation in each context. In addition, the degree to which each factor tended to associate with resilience and vulnerability was examined. The majority of the factors operated by increasing the likelihood of good adaptation in resilient adolescents and diminishing it in vulnerable ones. Overall, more similarities than differences were found in the factors contributing to explaining resilience or vulnerability. However, results also revealed some differential aspects: psychological variables showed a larger explicative capacity in vulnerable adolescents, whereas factors related to school and peer contexts, especially the second, showed a stronger association with resilience. In addition, perceived family wealth, satisfaction with friendships and breakfast frequency only made a significant contribution to the explanation of resilience. The current study provides a highly useful characterization of resilience and vulnerability phenomena in adolescence.

Keywords: adolescence, resilience, vulnerability, family functioning, global health score

OPEN ACCESS

Edited by:
Jason C. Immekus,
University of Louisville, USA

Reviewed by:
Evgueni Borokhovski,
Concordia University, Canada
Eleonora Riva,
University of Milan, Italy

****Correspondence:***
Pilar Ramos
pilarramos@us.es

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *21 December 2015*
Accepted: *14 June 2016*
Published: *04 July 2016*

Citation:
Moreno C, García-Moya I, Rivera F and Ramos P (2016) Characterization of Vulnerable and Resilient Spanish Adolescents in Their Developmental Contexts. Front. Psychol. 7:983. doi: 10.3389/fpsyg.2016.00983

INTRODUCTION

Fostering wellbeing is one of the current priorities of international agendas in health promotion (WHO, 2012, 2014), and adolescence has been considered to be a key developmental stage for this objective (WHO, 2014). Scientific evidence on factors that help mitigate risk or promote good adjustment despite adversity is crucial to governments and international agencies, which need to efficiently and effectively invest their resources. Positive and negative factors for wellbeing accumulate throughout life and health promotion interventions, which maximize protective factors while minimizing risks, can be successful in achieving wellbeing gains (Marmot, 2010). Resilience research, which analyses risk and protective factors to understand positive development under adverse circumstances, therefore presents itself as a particularly valuable approach that can provide the foundations for the design of effective health promotion and preventive interventions (Roosa, 2000).

More specifically, the value of resilience studies for the design and evaluation of health promotion interventions is apparent for the following reasons. First, resilience research provides critical information about key factors that help reduce potential harm and encourage positive adaptation (Masten, 2014). Each identified protective or vulnerability factor offers a possible focus of intervention (Olsson et al., 2003). Furthermore, the advantage of these studies is that they not only provide a list of intervention targets, but also emphasize the most relevant factors for different population groups and adversity levels (Luthar and Cicchetti, 2000).

Additionally, in highlighting an individual's positive adaptation resilience studies facilitate a change of approach (Luthar and Cicchetti, 2000; Olsson et al., 2003; Fergus and Zimmerman, 2005). Thus, resilience is in line with the perspective shift which has gradually taken place in different disciplines, including psychology, in the last decades: from the reduction of existing problems and exclusive emphasis on deficit and risk, to a focus on the development and promotion of health resources and assets (Morgan et al., 2010).

Lastly, it is important to bear in mind that the utility of resilience research goes further than merely understanding the processes linked to adversity. According to existing evidence, protective factors (as vulnerability ones) are not specific to situations of adversity, but they are the manifestation of basic adaptational systems that come into play in a variety of situations (Masten and Coatsworth, 1998; Masten, 2001). Therefore, increasing our knowledge about resilience and vulnerability phenomena provides useful evidence for intervention and evaluation in adversity contexts and helps to better understand and promote positive development in the general population.

In order for scientific research to make a significant contribution to the design and evaluation of interventions and policies, it is fundamental that studies on resilience (as well as those on vulnerability) include a clear definition and operationalization of the terminology involved (Luthar and Cicchetti, 2000; Masten, 2014; Luthar et al., 2015). In this regard, resilience is defined as "a dynamic process encompassing positive adaptation within the context of significant adversity" (Luthar et al., 2000, p. 543). There is a wide consensus that the two criteria implicit in this definition must be met in order to identify resilience. Indeed, exposure to adversity and some evidence of positive adaptation have been referred to as the two "judgements," "dimensions," "sides" or "coexisting conditions" of resilience (Masten and Coatsworth, 1998; Luthar et al., 2000, 2015; Luthar and Cicchetti, 2000; Masten, 2001; Rutter, 2006).

The adversity element has been defined by characteristics as diverse as: an experience of war or catastrophe (Masten and Narayan, 2012), low economic status (Buckner et al., 2003), belonging to minority groups (Sandín-Esteban and Sánchez-Martín, 2015), living in disadvantaged neighborhoods (Tiêt and Huizinga, 2002) and an individual's or caregiver's disorders or illnesses (Werner and Smith, 1982). Nevertheless, the key defining characteristic of adversity is that a significant threat to development or demonstrable risk must be present (Luthar and Cicchetti, 2000; Masten, 2001). More specifically, adversity is defined by "current or past hazards judged to have the potential to derail normative development" (Masten, 2001, p. 228) and it "typically encompasses negative life circumstances that are known to be statistically associated with adjustment difficulties" (Luthar and Cicchetti, 2000, p. 858).

In this regard, putting key adaptational systems in danger, including the relationship with loving and competent adult caregivers in a family context, is amongst the principal hazards to human development (Masten, 2001). Extant evidence has documented the fundamental links between the quality of parent-child relationships and adolescent development and adjustment (Steinberg and Silk, 2002; Clarke-Stewart and Dunn, 2006). In this sense, family context has a very strong influence on the person from the beginning of life and through multiple channels. No wonder, therefore, that family is the center of many adaptation and human development studies in this field (Masten and Shaffer, 2006). Hence, low-quality parent-child relationships (García-Moya et al., 2013b) or the existence of problems in the family (Fergusson and Linskey, 1996) have been considered to be key elements in defining an adverse situation. Accordingly, low scores in a composite factorial measure of the quality of parent-child relationships (García-Moya et al., 2013a) will be used as the indicator of adversity in the present study.

In defining positive adaptation, resilience research is especially varied. Luthar et al. (2000, 2015) concluded that a single criterion to establish the best adaptation indicator for any given study does not exist. External criteria such as behavioral adjustment and social competence have tended to predominate (Olsson et al., 2003) but internal criteria including emotional health, life satisfaction or absence of emotional distress are increasingly seen as similarly important indicators of positive adaptation (Masten and Reed, 2005). Furthermore, some revealing studies show that individuals showing positive adaptation according to external competence criteria can still experience internalizing symptoms and health problems (e.g., Luthar et al., 1993). Drawing on this evidence, we selected a global health score, which encompasses self-rated health, psychosomatic complaints, health-related quality of life and life satisfaction, as the indicator of positive adaptation in the present study. This is not to say that positive adaptation is synonymous to health or wellbeing, but we made the conceptually-informed decision to give priority to the aforementioned internal dimensions of health to define positive adaptation. More specifically, the global health score (Ramos et al., 2010) was selected because of its relevance for the kind of adversity examined (Karademas et al., 2008; Jiménez-Iglesias et al., 2015), as well as being a sound composite factorial score that encompasses multiple domains of health and has shown good psychometric properties in adolescents (Ramos et al., 2012). Specifically, using the global health score as the criterion for positive adaptation fits with one of the approaches mentioned in a seminal chapter about measurement issues in the empirical study of resilience, underlining that the assessment of positive adaptation "must be tied in to the particular risk domain being studied" and "rests on multiple-item instruments, typically with well-documented psychometric properties, that provide assessments on the continuum between adjustment and maladjustment" (Luthar and Cushing, 1999, pp. 139–140). Furthering the definition of the constructs related to resilience and adaptation, some authors (Tiêt and Huizinga, 2002) have proposed an interesting classification of individuals based on their level of exposure to adversity and the resulting adaptation shown, which divides the population into four large groups. Two of the groups show expected results in accordance with their level of exposure to adversity: low-risk—good adaptation (*competent* or *unchallenged*) and high-risk—bad adaptation (*maladaptive*). The paradox occurs in the remaining two groups: those that are exposed to high-risk but show good adaptation and those that, despite being exposed to low levels of risk, exhibit low competence levels. The first of these latter two groups constitutes the sample of interest in resilience studies whereas the second group, although rarely studied, could offer interesting information about vulnerability factors in the normative population.

After establishing the group or groups of interest, the next step is to identify which factors facilitate (protective factors) or inhibit (vulnerability factors) positive adaptation in the given context. Research has tended to classify these factors using a theoretical framework which distinguishes three fundamental levels: individual-level, family-level, and extrafamily-level factors (Masten and Coatsworth, 1998; Luthar and Cicchetti, 2000; Olsson et al., 2003).

On the individual level, self-esteem, self-efficacy, and intellectual capacity have been extensively studied in classic literature on resilience as determinant factors on the individual level (Masten and Coatsworth, 1998; Dumont and Provost, 1999; Hamill, 2003). Nevertheless, the claim that positive self-perception along with confidence in one's efficacy and motivation to engage in the environment are fundamental for successful adaptation (Masten, 2001) justifies the need to explore the role of other constructs with clear links to the aforementioned description. Regarding positive self-perception, satisfaction with body image is one aspect that has been considered especially influential in adolescence (Tiggemann, 2005). Confidence in one's efficacy and motivation to engage in the environment are linked to some novel constructs in positive psychology that are likely to play a significant role in explaining positive adaptation, such as sense of coherence (Antonovsky, 1987) and curiosity and exploration (Kashdan et al., 2009). Finally, another fundamental factor is emotional regulation. This skill, which is closely related to intellectual functioning, is currently receiving special scientific attention since it seems to be fundamental for successful coping and good behavioral, emotional and social adjustment (Lengua, 2002; Buckner et al., 2003). The present study will try to further the understanding of individual-level factors by exploring the aforementioned constructs that, despite having connections with well-known individual factors in resilience studies, have not usually been included in previous resilience research.

Along with them, we will analyse the role of lifestyles that, despite their significant contribution to health and wellbeing, have also received little attention to date (Elliot, 1993; Ramos, 2010). Regarding tobacco, alcohol and cannabis use, the absence of these risk behaviors has been predominantly used as criteria for defining adaptation (for a review, see Fergus and Zimmerman, 2005) or its presence has been analyzed as a risk indicator (Anteghini et al., 2001). The associations between resilience and healthy lifestyles, such as eating habits, dental hygiene and physical activity, has also been rarely explored in classic studies. Nonetheless, physical activity, for example, has been highlighted as a relevant factor when explaining resilience due to its protective effects on health in stress situations (Gerber and Pühse, 2009; Silverman and Deuster, 2014) or the fact that it tends to be incompatible with some health-threatening activities or risk behaviors, such as alcohol and other substances abuse (Pate et al., 1996). Consequently, examining the associations between lifestyles and resilience is of unquestionable interest.

On the family level, besides aspects related to the aforementioned quality of relationships and processes in the family context (which will be used to define adversity in the present study), it is worth exploring the contribution of the families' socioeconomic status (Masten and Coatsworth, 1998). A good socioeconomic position is associated with access to material, cultural and educational resources, making it a significant source of social capital (Bornstein and Bradley, 2003), whereas low family affluence limits access to the aforesaid resources and could become a significant source of stress, having negative consequences on children's development (Conger et al., 2000). Unlike objective indicators, subjective measures of socioeconomic status have not generally been analyzed in resilience studies. However, the study of socioeconomic inequalities in health indicates that subjective perceptions of wealth have a strong predictive capacity regarding adolescent health (Goodman et al., 2001) and their significant effects on health remain even after controlling for objective measures such as educational level, parents' occupation and family affluence (Elgar et al., 2016).

Lastly, on the extrafamily level, experiences of belonging and efficacy, such as a positive school climate and experiences of academic achievement, can significantly contribute to positive adaptation outcomes (Masten and Coatsworth, 1998), whereas bullying episodes can hamper them (McVie, 2014). Significant extrafamily relationships with important adults, including teachers (DuBois et al., 1992; Masten and Coatsworth, 1998), as well as the contribution of peer support and the degree in which peers provide positive or adjusted models of behavior (e.g., Jain et al., 2012) have also been emphasized. The present study will consider all the aforementioned aspects.

Therefore, the selection of variables in the present study is supported by an ample consensus on the need to analyse factors from individual, family and extrafamily levels in order to obtain a detailed view of the factors associated with resilience and vulnerability (Masten and Coatsworth, 1998; Luthar and Cicchetti, 2000; Olsson et al., 2003). In addition, the selection of variables is guided by an explicit effort to explore relevant content from those levels that have not been sufficiently examined in resilience research so far. Thus, the present study will try to further the understanding of individual-level factors by exploring emotional regulation along with other constructs such as satisfaction with body image, sense of coherence and curiosity and exploration that, despite having connections with well-known individual factors in resilience studies, have not usually been included in previous resilience research. Similarly, because lifestyles contribute significantly to wellbeing, the selection of variables included breakfast frequency, physical activity and substance use, which have also received little attention in the study of resilience. On the family level, a similar rationale motivated the selection of perceived family wealth as the measure of socioeconomic status, instead of the objective indicators which have dominated previous resilience research. Finally, on the extrafamily level, the selected variables (including academic achievement, classmate and teacher support, bullying victimization, peer support, and models of behavior in the peer group) ensure simultaneous consideration of the most frequently mentioned factors on this level.

Accordingly, this paper starts by using the criteria on adversity and positive adaptation described above to identify two reference groups within a representative sample of adolescents: those that showed good global health despite having a low-quality family environment (*resilient*), and those that showed poor health even with high-quality parent-child relationships (*vulnerable*). Afterwards, drawing on the classification by Tiêt and Huizinga (2002), the phenomena of resilience and vulnerability are characterized by comparing them to groups of *maladaptive* (high risk, poor adaptation) and *competent* (low risk, good adaptation) adolescents, respectively.

The aim of the paper is to characterize resilience and vulnerability in adolescents, considering an ample number of potential protective and vulnerability factors that were selected from the three main levels described in scientific literature (individual, family, and extrafamily). The selection of the specific factors used in this study is also intended to initiate a new direction by exploring relevant constructs for positive adaptation in adolescence which had not received sufficient attention in classic resilience research, amongst others, satisfaction with body image, sense of coherence, curiosity and exploration, and diverse lifestyles.

In short, after conducting preliminary analyses on the differences among resilient, vulnerable, competent, and maladaptative adolescents in individual factors (including psychological variables and lifestyles), family socioeconomic status and extrafamily factors (including those from the school context and the peer context), the ability of those factors (as independent variables) to explain the dependent variables resilience (vs. maladaptation) and vulnerability (vs. competence) is examined. A detailed list of research question is presented in **Table 1**.

This approach is designed to identify important factors for adaptation in adverse and non-adverse contexts respectively, but it may also provide valuable findings that contribute to informing the debate on whether some factors contribute to positive development in the face of adversity but have little impact in the absence of it or whether there are some common protective and risk factors associated with positive adaptation irrespective of the level of adversity exposure (Roosa, 2000). Also, on the potential implications and contributions offered by the present study, it has been stated that although "this kind of epidemiological research does not unpack the processes by which each individual is impacted by contextual experience, it does document the multiple factors in the environment that are candidates for more specific analyses (Sameroff, 2010, p. 14)." The aforementioned factors and levels do not operate independently, rather they relate amongst themselves in people's lives (Fergus and Zimmerman, 2005). For this reason, approaches which provide an ample characterization of resilience and vulnerability phenomena while taking into account a significant number of the aforementioned factors (usually referred as person-focused approaches) provide a very valuable complementary approach (Masten, 2001).

METHOD

Participants

Data were obtained from the Health Behavior in School-aged Children (HBSC) cross-sectional survey. The HBSC study is an international network supported by the World Health Organization that collects data in more than 40 countries in Europe and North America. The survey is conducted every 4 years with the aim of increasing knowledge about health-related behaviors, lifestyles and developmental contexts of young people.

Participants of the present study come from a representative sample of school-aged children aged 13–16 years residing in Spain, who were selected for the 2014 edition of the HBSC study using a random multi-stage sampling stratified by conglomerates, representative by age, area of residence (rural or urban), type of school (public or private) and region (Spain has 19 regions) (Moreno et al., 2016). Participants were recruited from a database of schools published by the Spanish Ministry of Education. Those centers that refused to participate in the study were substituted for other centers, also selected randomly within the same stratum. The final student participation rate was 87%.

For the purpose of this article, terciles were used to identify adolescents scoring high (upper tercile) and low (lower tercile) in the scales for Global Health Score (GHS) and the Quality of the Parent-Child Relationship (QPCR) (described later in the section on instruments).

Despite the limitations of categorizing quantitative variables (Preacher et al., 2005), dividing them into three groups in order to identify their extremes is supported by three reasons: firstly, by the essence of the construct itself, since "resilience is never directly measured, but instead is indirectly inferred based on evidence of the two subsumed constructs" ("adversity" and "positive adaptation"; Luthar et al., 2015, p. 248); secondly, it is consistent with literature that identifies both resilient and vulnerable subjects as extreme groups in unfavorable and favorable circumstances, respectively, but whose results in adjustment indicators are not consistent with their circumstances (Luthar et al., 2000; Masten, 2014); and lastly, from a purely methodological perspective, because, as DeCoster et al. (2009) argues, categorization is advised when focusing on the extreme groups since it allows for identification of groups of subjects based on conceptual definitions.

Based on the four combinations resulting from this division 1753 adolescents were selected from the total of 3845 studied (see **Table 2**). In the selected sample, 45.8% are boys and 54.2% are girls, with a mean age of 14.62 years (*SD* = 1.11). Additionally, 62.7% attended public schools and 37.3% private, with 54.1% living in urban areas and 54.9% in rural areas.

TABLE 1 | Research questions in the present study.

Research question 1
How do the four groups of adolescents analyzed in this paper (maladaptative, resilient, competent, and vulnerable) characterize and differentiate from each other in relation to the three sets of variables analyzed: individual factors (including psychological variables and lifestyles), family socioeconomic status and extrafamily factors (including those from the school context and the peer context)?

Research question 2
Which factors (individual, familial, and extrafamiliar) are useful to understand adaptation in high adversity contexts? In other words: which factors (individual, familial, and extrafamilial) are useful to distinguish between resilient and maladaptative adolescents?

The following *specific questions* will be answered before addressing research question 2:

2a. Which psychological factors (sense of coherence, emotional regulation, curiosity and exploration, perceived body image and satisfaction with body image) distinguish between resilient and maladaptative adolescents?
2b. Which factors related to lifestyles (breakfast frequency, fruit consumption, physical activity, dental hygiene, tobacco, alcohol and cannabis use) distinguish between resilient and maladaptative adolescents?
2c. Which family socioeconomic factors (father's educational level, mother's educational level and perceived family wealth) distinguish between resilient and maladaptative adolescents?
2d. Which factors referring to school (perceived academic achievement, feelings toward school and perceived teacher support) distinguish between resilient and maladaptative adolescents?
2e. Which factors referring to peer group (perceived peer support, models of behavior, satisfaction with friendships, having been bullied and having bullied others) distinguish between resilient and maladaptative adolescents?

Research question 3
Which factors (individual, familial, and extrafamilial) are useful to understand adaptation in low adversity contexts? In other words: which factors (individual, familial, and extrafamilial) are useful to distinguish between vulnerable and competent adolescents?

The following *specific questions* will be answered before addressing research question 3:

3a. Which psychological factors (sense of coherence, emotional regulation, curiosity and exploration, perceived body image and satisfaction with body image) distinguish between vulnerable and competent adolescents?
3b. Which factors related to lifestyles (breakfast frequency, fruit consumption, physical activity, dental hygiene, tobacco, alcohol, and cannabis use) distinguish between vulnerable and competent adolescents?
3c. Which family socioeconomic factors (father's educational level, mother's educational level and perceived family wealth) distinguish between vulnerable and competent adolescents?
3d. Which factors referring to school (perceived academic achievement, feelings toward school and perceived teacher support) distinguish between vulnerable and competent adolescents?
3e. Which factors referring to peer group (perceived peer support, models of behavior, satisfaction with friendships, having been bullied and having bullied others) distinguish between vulnerable and competent adolescents?

TABLE 2 | Sample subgroups according to their tercile position in the global health and the quality of parent–child relationship scores (the four groups examined in the present study are highlighted in bold).

		Global Health Score (GHS)		
		Tercile 1 (low)	**Tercile 2 (medium)**	**Tercile 3 (upper)**
Quality of Parent-Child Relationships (QPCR)	**Tercile 1 (low)**	**726**	386	**172**
	Tercile 2 (medium)	402	505	398
	Tercile 3 (upper)	**150**	401	**705**

Therefore, following the classification criteria for adaptation status developed by classic research (Tiêt and Huizinga, 2002), the sample was classified in four groups, defined as follows: resilient adolescents (tercil 1 in QPCR and tercil 3 in GHS), maladaptative adolescents (tercil 1 in CRPF and tercil 1 in GHS), vulnerable adolescents (tercil 3 in CRPF and tercil 1 in GHS) and competent adolescents (tercil 3 in QPCR and tercil 3 in GS).

Instruments

The variables were assessed using the 2014 Spanish HBSC Questionnaire, which included questions about lifestyles, positive health and characteristics of the principal developmental contexts (family, peers, and school) in adolescence. The instrument is comprised of an extensive series of mandatory questions, optional packages and questions that cover specific national interests (Roberts et al., 2009). The complete questionnaire is revised and improved for each edition of the study (for the last edition, see Inchley et al., 2016). For the present paper, key measures of quality of parent-child relationship and health, as well as sociodemographic, school and peer contexts, lifestyle, and psychological and socioeconomic variables were selected from the Spanish version of the 2014 HBSC survey.

Firstly, the following two measures were used to derive the classification in groups (maladaptative, resilient, vulnerable, and competent) that acts as the dependent variable.

1. Global Health Score (GHS). This measure is based on 20 items related to the variables: life satisfaction, self-rated health, health-related quality of life and psychosomatic complaints.

The details of these instruments can be consulted in **Table 3**. The GHS is a score with a mean of 50 and standard deviation of 10 that has shown good fit indices (NNFI = 0.98, CFI = 0.99, RMSEA = 0.03), as well as good reliability and validity (Ramos et al., 2010). This measure assesses the adolescent's physical, psychological and social wellbeing, following the most widely used and currently accepted definition of health, i.e., the definition proposed by the World Health Organization (WHO, 1948). As previously described, terciles were used in the present study to classify the adolescents in three groups according to this measure of global health.

2. Factorial score on the Quality of Parent-Child Relationship (QPCR), with a mean of 5 and a standard deviation of 2. This score is an adaptation of the measure developed by García-Moya et al. (2013a), that consists of the following three indicators: perceived affection, ease of communication with parents and satisfaction with family relations. The details of these instruments can be consulted in **Table 3**. The factorial score on the QPCR showed good fit indices (NNFI = 0.99, CFI = 0.99, RMSEA = 0.02) has been considered a useful tool in global assessments of the relationships between parents and children according to the adolescents' perception (García-Moya et al., 2013a). As previously mentioned, terciles were used in the present study to classify adolescents in three groups according to the quality of their parent-child relationship.

In addition, the independent variables were selected in line with the aims of this study and assessed by means of several instruments that were part of the 2014 HBSC Questionnaire, explained above. The details of these instruments are presented in **Table 4**.

Procedure

New information and communication technologies (ICT), based on a CAWI (Computer-Assisted Web Interviewing) model, were used in the data collection process. The data was always collected in the school setting, under the supervision of teachers. In those schools with internet-connection problems or problems with the condition or number of computers, members of the research team traveled personally to those schools to collect data using tablets. Ultimately, the guided computerized procedure has the advantage of immediately receiving and incorporating the students' responses in the database, thus reducing the possible errors from the data entry process, as well as helping to maintain the anonymity of the responses.

In all of the schools, after contacting via telephone with the head teacher, deputy head teacher or school counselor, instructions were given to the teachers who would be supervising the classroom when the adolescents responded to the questionnaires. On the other hand, instructions for the students were included at the beginning of the questionnaire to guarantee homogeneity amongst all the participants.

Ultimately, data collection complied with the three requirements dictated by the HBSC international protocol (Roberts et al., 2009): students themselves answered the questionnaires; anonymity was guaranteed; and the questionnaires were completed at school under the supervision of instructed staff.

Statistical Analysis

Firstly, bivariate analyses including chi-square and ANOVA (with Bonferroni test for multiple comparisons) were used to compare the four groups of adolescents (maladaptative, resilient, competent, and vulnerable) in each one of the independent variables (sociodemographic, school context, peer context, lifestyle, psychological, and socioeconomic variables). This analysis corresponds to the research question 1. Also, Crammer's V and Cohen's d were calculated to determine the effect size, with the following cut-off points: 0–0.19 = negligible, 0.20–0.49 = small, 0.50–0.79 = medium, 0.80 and above = high (Cohen, 1988).

Secondly, separate binary logistic regression analyses were carried out for resilience and vulnerability, with adaptation status (resilient vs. maladaptative -research question 2- and vulnerable vs. competent -research question 3-, respectively) as the dependent variables, and the different sets of variables analyzed (demographic, school context, peer context, lifestyle, psychological, and socioeconomic variables) as predictor variables. The predictive capacity of each set of variables (controlling for significant demographic variables) was calculated using the Nagelkerke R^2. Afterwards, a final model including only significant variables in previous analysis was estimated. The odds ratio (OR) and its confidence interval at the 95% level (95% CI) was calculated for each examined predictor, establishing the statistical significance as $p < 0.05$ for each variable.

Statistical analyses were conducted using the IBM SPSS Statistics 22.0 software.

RESULTS

Research Question 1. Comparisons Among the Four Adaptation Groups: Maladaptative, Resilient, Competent, and Vulnerable Adolescents

This first subsection focuses on the comparisons among maladaptative, resilient, competent and vulnerable adolescents in all variables of this study. The comparisons of these groups show significant differences ($p < 0.001$, $V = 0.231$, medium effect size) in the distribution of gender. **Table 5** shows that the maladaptative and vulnerable groups have a greater proportion of girls. However, comparisons between the four groups are not significant neither for educational center ($p = 0.067$, $V = 0.087$, negligible effect size) nor habitat ($p = 0.145$, $V = 0.051$, negligible effect size).

Table 6 shows the distribution of the maladaptative, resilient, competent and vulnerable groups in the age, school context, peer context, lifestyle, psychological, and socioeconomic variables. The mean comparisons of the contrasts between pairs of groups can be consulted in **Table 7**.

Regarding age, older adolescents fell into the maladaptive and vulnerable categories, followed by the resilient adolescents

TABLE 3 | Dependent variables and instruments used for their assessment in the present study.

Global Health Score (GHS)	Life satisfaction	It was measured by the Cantril's Ladder (Cantril, 1965), with the question: "Here is a picture of a ladder. The top of the ladder '10' is the best possible life for you and the bottom '0' is the worst possible life for you. In general, where on the ladder do you feel you stand at the moment? Tick the box next to the number that best describes where you stand." This variable represents the global perception adolescents have of their lives, *from 0 to 10*. Level of measurement: quantitative variable.
	Self-reported health	A single item asked the adolescent to consider their health at that moment, with their response fitting to one of the following four options: *excellent, good, passable,* or *poor* (Idler and Benyamini, 1997). This measure has been validated for quantitative use (Silventoinen et al., 2007). Level of measurement: ordinal variable.
	Health-related quality of life	It was measured with the Kidscreen instrument designed for a population between the ages of 8 to 18. Specifically the Kidscreen-10 version was used, which provides a global, health-related quality of life index with 10 items covering physical, psychological and social aspects (Ravens-Sieberer et al., 2001). These items, which show a Cronbach's alpha of 0.83, refer to feeling well and fit, full of energy, sad, lonely, having enough time for themselves, doing things they want in their free time, receiving fair treatment from their parents, having a good time with friends, getting on well at school and being able to pay attention/concentrate. Items were answered on a 5-point Likert scale, from 1, *never*, to 5*, always.* Level of measurement: continuous variable.
	Psycho-somatic complaint	It was measured with the HBSC-symptom checklist. It measures two aspects (Ravens-Sieberer et al., 2008): psychological complaint (nervousness, feeling low, irritability and sleeping problems) and somatic manifestations (headache, stomach-ache, back ache, and feeling dizzy), with a Cronbach's alpha of 0.83. These 8 items were answered on a 5-point Likert scale: *about every day, more than once a week, about every week, about every month,* and *rarely or never.* Level of measurement: continuous variable.
Factorial Score on the Quality of Parent-Child Relationship (QPCR)	Perceived affection	This variable was assessed by means of the 4-item subscale of the Parental Bonding Inventory-Brief Current form (PBI-BC; Klimidis et al., 1992), with the aim of determining if the parents showed to be warm and supportive toward their children. This dimension includes the following items repeated for the mother and the father: "helps me as much as I need," "is loving," "understand my problems and worries," and "makes me feel better when I'm upset." An average score from 0, *never*, to 2, *almost always*, was obtained from this scale, with a Cronbach's alpha of 0.85. Level of measurement: continuous variable.
	Ease of communication with parents	Participants were asked: "how easy is it for you to talk to your father about things that really bother you?" and "how easy is it for you to talk to your mother about things that really bother you?" (these questions were created by the HBSC study). An average score on ease of communication with parents was obtained that ranged from 1, *very difficult*, to 4, *very easy*. Level of measurement: ordinal variable.
	Satisfaction with family relations	This variable was measured by means of an item based on Cantril's Ladder (1965): "in general, how satisfied are you with the relationships in your family?" A quantitative score was obtained that ranged *from 0 "we have very bad relationships in our family" to 10 "We have very good relationships in our family*." Level of measurement: quantitative variable.

and finally, the youngest fell into the category of competent adolescents.

With respect to school, the competent adolescents show higher perception of academic achievement than the resilient adolescents, who in turn have a higher perception than the maladaptative and vulnerable adolescents. In relation to feeling toward school, the competent adolescents have the most positive feelings toward school and the highest perception of teacher support, followed by the resilient and vulnerable adolescents and, finally, the maladaptative adolescents.

In their peer relationships, the competent and resilient adolescents show the highest perception of social support, followed by the vulnerable adolescents and, finally, the maladaptative adolescents. The resilient and competent adolescents have a higher rate of positive models of behavior in their peer group than the maladaptative adolescents, with the vulnerable adolescents falling in the middle. Likewise, resilient and competent adolescents have higher satisfaction with their friendships than the vulnerable adolescents, and this group shows more satisfaction than maladaptative adolescents. In relation to bullying, the maladaptative adolescents show a higher likelihood to have been bullied and to have bullied others than the other groups (resilient, competent, and vulnerable adolescents).

Regarding lifestyles, the competent and resilient adolescents eat breakfast more days a week, followed by the vulnerable adolescents and, finally, the maladaptative adolescents. The resilient and competent adolescents eat fruit more frequently than the maladaptative adolescents do (the vulnerable adolescents show an intermediate score between the maladaptive and resilient adolescents). Likewise, resilient and competent adolescents do more physical activity (moderate to vigorous and vigorous) than the maladaptative and vulnerable adolescents. The competent adolescents brush their teeth more frequently than the maladaptative and resilient adolescents (the vulnerable adolescents show an intermediate score between the competent and resilient adolescents). In relation to tobacco, the maladaptative adolescents show higher use than the other three

TABLE 4 | Independent variables and instruments used for their assessment in the present study.

Sociodemographic variables	Sex		Boy and girl. Level of measurement: categorical variable.
	Age		13–16 years. Level of measurement: continuous variable.
	Type of educational center		Public and private. Level of measurement: categorical variable.
	Habitat		Urban and rural. Level of measurement: categorical variable.
School context variables	Perceived academic achievement		They were asked: "in your opinion, what does your teacher think about your school performance compared to your classmates" (this question was created by the HBSC study). This question is answered on a 4-point Likert scale, ranging from 1, *below average*, to 4, *very good*. Level of measurement: quantitative variable.
	Feelings toward school		The following question: "how do you feel about school at the present?" (this question was created by the HBSC study). Four response options were available on a 4-point Likert scale from 1, *I don't like it at all*, to 4, *I like it a lot*. Level of measurement: continuous variable.
	Teacher support		It was assessed by means of the following three items: "I feel that my teachers accept me as I am," "I feel that my teachers care about me as a person," and "I feel a lot of trust in my teacher," with a Cronbach's alpha of 0.84. Items were answered on a 5-point Likert scale, from 1, *I completely disagree*, to 5*, I completely agree* (Torsheim et al., 2000). Level of measurement: continuous variable.
Peer context variables	Perceived social support		It was assessed by means of the Multidimensional Scale of Perceived Social Support (MSPSS; Zimet et al., 1988). This scale consists of the following four items: "my friends really try to help me," "I can count on my friends when things go wrong," "I have friends with whom I can share my joys and sorrows," and "I can talk about my problems with my friends," Items are answered on a 7-point Likert scale, from *completely disagree* (1) to *completely agree* (7), with a Cronbach's alpha of 0.98. Level of measurement: continuous variable.
	Models of behavior in the peer group		It was assessed by means on a scale developed by the HBSC study network and validated by Gaspar de Matos et al. (unpublished manuscript). Adolescents were asked about the frequency of 8 different behaviors in their group of friends, including both positive (such as "do well in school," "participate in sports activities with other kids," "participate in cultural activities other than sports" and "get along well with parents") and negative (such as "smoke cigarettes," "drink alcohol," "get drunk," and "consume drugs to get high") behaviors. Items were answered on a Likert scale from 1, *never or almost never*, to 3, *often*, with a Cronbach's alpha of 0.70. The items corresponding to negative behaviors were reverse-coded, so that a higher score on this scale represents a higher presence of positive models of behavior in the peer group. Level of measurement: continuous variable.
	Satisfaction with friendships		Measure adapted by the HBSC network from the Cantril's Ladder on life satisfaction scaled *from 0 to 10* (Cantril, 1965), but referring specifically to satisfaction with friendships. Level of measurement: quantitative variable.
	Having been bullied		It was assessed by means of the Revised Bully/Victim Questionnaire (Olweus, 1996). The response options ranged from 1, *I haven't been bullied in school in the past 2 months*, to 5*, multiple times a week.* Level of measurement: quantitative variable.
	Having bullied others		Also assessed by means of the Olweus (1996) questionnaire and with similar response options. Level of measurement: quantitative variable.
Lifestyle variables	Eating habits	Breakfast frequency	Adolescents were asked how many days a week they typically ate breakfast (something more than a glass of milk or juice), with the corresponding response values ranging from 1 to 7 days. In addition, they also answered questions on how many times a week they typically ate two specific types of foods: fruits and snacks. These questions were created by the HBSC study. The response options varied from 1*, never*, to 7, *every day, more than once.* Level of measurement: quantitative variable.
		Fruit consumption	
		Snack consumption	
	Physical activity	MVPA	Adolescents were asked about their level of Moderate to Vigorous Physical Activity (MVPA), as indicated by the number of days in which they felt physically active during a total of at least 60 min a day over the last 7 days. The response options ranged from 0 to 7 days (Prochaska et al., 2001). In addition, they were asked about their level of Vigorous Physical Activity (VFA), assessed in the HBSC study by the frequency with which the adolescents, in their free time outside of school hours, engaged in some type of physical activity that made them sweat or out of breath. The response options on a Likert scale ranged from 1, *never*, to 7*, every day.* Level of measurement: quantitative variable.
		VFA	

(Continued)

TABLE 4 | Continued

	Dental hygiene		Adolescents were asked how often they brushed their teeth (these questions were created by the HBSC study), with the following response options: *never; less than once a week; at least once a week but not daily; once a day;* and *more than once a day.* Level of measurement: quantitative variable.
	Substance use	Tobacco use	Three questions referring to the frequency of substance use over the past 30 days were included. These items have been adapted from the questions included in the European School Survey Project on Alcohol and Other Drugs (Hibell et al., 2000). Specifically, adolescents were asked about the number of days, out of past 30 days, in which they had smoked cigarettes, had drank alcohol and had smoked cannabis (hash or marijuana, "joints"). These items were answered on a 7-point Likert scale, from 1, *never,* to 7*, 30 days.* Level of measurement: quantitative variable.
		Alcohol use	
		Cannabis use	
Psychological variables	Sense of coherence		This construct was assessed by means of the SOC-13 scale (Antonovsky, 1987). It consists of 13 items, such as "has it happened in the past that you were surprised by the behavior of people whom you thought you knew well?," and "how often do you have the feeling that there's little meaning in the things you do in your daily life?." Questions are answered on a 7-point Likert scale, with a Cronbach's alpha of 0.77. The SOC-13 scale has shown good reliability and validity in different countries (Lindström and Eriksson, 2010). Level of measurement: continuous variable.
	Emotional regulation		It was assessed by means of the impulsiveness/emotion-control subscale from the reduced version of the Emotion Regulation Index for Children and Adolescents scale (ERICA; MacDermott et al., 2010). This subscale comprises 8 items (for example, "I have angry outbursts," "I have trouble waiting for something I want") and it is answered on a 5-point Likert scale, from 1, *totally agree*, to 5, *totally disagree.* The Cronbach's alpha was of 0.84. Level of measurement: continuous variable.
	Curiosity and Exploration		It was assessed by means of the Curiosity and Exploration Inventory-II (Kashdan et al., 2009). It is a scale with 10 items (some examples are: "I am at my best when doing something that is complex or challenging," "I am the kind of person who embraces unfamiliar people, events, and places," or "I like to do things that are a little frightening") with 5 response options on a Likert scale, from 1, *a little or none*, to 5, *a lot.* The Cronbach's alpha was of 0.87. Level of measurement: continuous variable.
	Perceived body image		It was assessed with an item created for the HBSC study. Specifically, they are asked "do you think your body is...?" and the response options on a 5-point Likert scale ranged from 1, *much too fat*, to 5*, much too thin*. Level of measurement: ordinal variable.
	Satisfaction with body image		It was assessed by means of the subscale of feelings and attitudes toward the body of the Body Investment Scale (BIS; Orbach and Mikulincer, 1998). This subscale consists of 6 items ("I am frustrated with my physical appearance," "I am satisfied with my appearance," "I hate my body," "I feel comfortable with my body," "I feel anger toward my body," and "I like my appearance in spite of its imperfections"), and is answered on a 5-point Likert scale, from 1, *totally agree*, to 5*, totally disagree.* The Cronbach's alpha was of 0.89. Level of measurement: continuous variable.
Socioeconomic variables	Father educational level		Father's and mother's educational level and perceived family wealth were assessed with three questions created by the HBSC study. Educational level was measured on a 4-point Likert scale, from 1, *never studied (does not know how to read nor write, or does so with difficulty*) to 4, *university studies, either finished or unfinished*. The level of perceived family wealth was assessed by asking "how well off do you think your family is?." The question was answered on a 5-point Likert scale, from 1, *not at all well off*, to 5, *very well off.* Level of measurement: quantitative variable.
	Mother educational level		
	Perceived family wealth		

groups. However, the competent adolescents show lower alcohol use than all the others.

The analyses of psychological variables show differences in sense of coherence among the four groups of adolescents. Ordered from the highest to lowest score they are: competent, resilient, vulnerable, and maladaptative adolescents. In relation to emotional regulation, the competent adolescents have the highest score, followed by the resilient and vulnerable adolescents and, finally, the maladaptative adolescents. The resilient and competent adolescents present more curiosity and exploration and they see themselves as less obese than the maladaptative and vulnerable adolescents. In addition, there are differences among the four groups regarding satisfaction with body image. Ordered from highest to lowest they are: competent, resilient, vulnerable and maladaptative adolescents. Lastly, significant differences are found in parents' education, showing that the educational level of the competent and vulnerable adolescents' fathers is higher than that of the fathers of maladaptative adolescents. However, the educational level of the competent adolescents' mothers is higher than that of the mothers of maladaptative and vulnerable

TABLE 5 | Percentage of maladaptative, resilient, competent and vulnerable adolescents in relation to the sex (boys and girls), the type of educational center (public and private) and the habitat (urban and rural).

		Maladaptative (%)	Resilient (%)	Competent (%)	Vulnerable (%)
Sex	Boys	33.81	60.93	57.47	40.00
	Girls	66.19	39.07	42.53	60.00
Type of educational center	Public	66.59	61.26	62.34	67.14
	Private	33.41	38.74	37.66	32.86
Habitat	Urban	55.16	52.32	50.81	55.36
	Rural	44.84	47.68	49.19	44.64

adolescents. The resilient adolescents show an intermediate score between maladaptive and vulnerable adolescents for both the father and mother's education. There are significant differences in perceived family wealth, being higher in the resilient and competent adolescents than it is in the maladaptative ones (in this case the vulnerable adolescents are situated between the competent and maladaptive adolescents).

Research Question 2. The Study of the Resilient Adolescents

This second subsection focuses on those adolescents who, despite having low-quality parent-child relationships have a high global health score, that is to say, the resilient group (4.5% of the global sample and 13.4% of the participants classified as low-quality in parent-child relationship). This group of adolescents are compared with those which, having a low-quality parent-child relationship, have a low global health score, that is to say, the maladaptative group (18.9% of the global sample and 56.5% of the sample with low-quality parent-child relationships).

The results of the logistic regression analyses using the group of resilient adolescents as the reference value are shown below. Specifically, six models have been estimated, one for each group of independent variables (although sex and age have been included in all of them to prevent them to become confounding variables). Additionally, a global model is shown at the end, including only those variables that were found to be significant in previous models.

As can be seen in the first row of **Table 8**, although model 1 explained 10.8% of the total variability, being significant the variables sex and age (specifically, boys and younger adolescents have a higher probability of being resilient), using these demographic variables only the percentage of well-classified adolescents was 0%.

In model 2, concerning school context, the explained variance is 22.8 and 22.1% of the resilient adolescents are correctly classified. In this case, those adolescents with a higher perception of academic achievement, with an OR of 1.83 (95% *CI* = 1.44–2.33), and those with higher teacher support (*OR* = 1.19, 95% *CI* = 1.11–1.29), have a higher likelihood of being resilient.

In model 3, which includes the variables of peer context, the predictive capacity is 23.8%, with 19.2% of the adolescents in the resilient group being correctly classified. Significant variables in this model are models of behavior, satisfaction with friendships and being a victim of bullying. Adolescents who are more satisfied with their friendships are 1.5 times more likely to be resilient (95% *CI* = 1.26–1.79), whereas those that were victims of bullying more often have a lower likelihood of being resilient (*OR* = 0.53, 95% *CI* = 0.33-0.84). Likewise, those adolescents with a group of friends providing better models of behavior also show a higher likelihood of being resilient (*OR* = 1.07, 95% *CI* = 1.01–1.14).

Model 4 is devoted to variables related to lifestyles and shows a level of explained variance of 24.7%, with 25.6% of the resilient adolescents being correctly classified. Only two variables in this model are significant: breakfast frequency and moderate to vigorous physical activity. Specifically, those adolescents that engage in higher levels of moderate to vigorous physical activity increase their likelihood of being resilient in 1.37 times (95% *CI* = 1.22–1.54). Additionally, those adolescents who have breakfast more regularly are more likely to be resilient (*OR* = 1.24, 95% *CI* = 1.12–1.38).

Model 5 includes the psychological variables. Among the six specific models, this model shows the highest level of explained variance, which reaches 37% (30.4% of the adolescents in resilient group are correctly classified). The significant variables in this model are: sense of coherence, curiosity and exploration and satisfaction with body image. Sense of coherence stands out for its high OR, which is 3.18 (95% *CI* = 2.14–4.73), meaning that those adolescents with higher scores in this psychological construct have the highest likelihood of being resilient. Next, adolescents with a higher satisfaction with their body image are 1.83 times more likely to be resilient (*OR* = 1.83, 95% *CI* = 1.31–2.56). Lastly, those adolescents with higher scores in curiosity and exploration have a higher likelihood of being resilient (*OR* = 1.07, *CI* 95% = 1.03–1.11).

Model 6, referring to the socioeconomic variables, shows a lower predictive capacity than the previous models (13.1%), with only 3.5% of resilient adolescents being correctly classified. The only significant variable in this model is perceived family wealth, meaning that those who perceive a higher family wealth are 1.98 times more likely to be resilient (95% *CI* = 1.37–2.85).

Finally, in model 7 or the global model (the one which includes only the significant variables from previous models), the results show that the variables sex, age, models of behavior in the peer group and being a victim of bullying loose predictive capacity. Therefore, the global model includes the following nine variables: perceived academic achievement, perceived teacher support, satisfaction with friendships, breakfast frequency, moderate to vigorous physical activity, sense of coherence, curiosity and exploration, satisfaction with body image and perceived family wealth. This model stands out for its high predictive capacity, surpassing 50% of the explained variance (specifically, 51.8%). Additionally, there are a notably high proportion of correctly-classified resilient adolescents, specifically, 51.5%.

TABLE 6 | Descriptive statistics of the age, school context, peer context, lifestyle, psychological and socioeconomic variables between maladaptative, resilient, competent and vulnerable adolescents.

Variables		Maladaptative		Resilient		Competent		Vulnerable	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
	Age	14.92	1.07	14.52	1.10	14.28	1.10	14.84	1.10
School context	Perceived academic achievement	2.40	0.80	2.87	0.81	3.08	0.74	2.53	0.77
	Feelings toward school	2.37	0.88	2.67	0.86	3.01	0.88	2.60	0.87
	Perceived teacher support	3.19	0.88	3.73	0.90	4.09	0.84	3.60	0.88
Peer context	Perceived social support	5.18	1.63	5.81	1.44	6.11	1.41	5.70	1.46
	Models of behavior	2.91	0.43	3.08	0.42	3.16	0.42	2.99	0.42
	Satisfaction with friendships	7.78	1.85	8.87	1.28	9.05	1.35	8.27	1.77
	Having been bullied	1.32	0.78	1.09	0.42	1.14	0.53	1.23	0.73
	Having bullied others	1.32	0.74	1.24	0.58	1.14	0.51	1.19	0.59
Lifestyles	Breakfast frequency	5.05	2.36	6.16	1.68	6.44	1.47	5.77	2.02
	Fruit consumption	4.22	1.72	4.59	1.65	4.94	1.63	4.39	1.58
	Snack consumption	3.59	1.20	3.63	1.15	3.47	1.17	3.62	1.11
	MVPA	4.63	1.91	6.11	1.91	6.12	1.75	4.69	1.82
	VFA	4.51	1.67	5.38	1.52	5.42	1.37	4.50	1.67
	Dental hygiene	4.46	0.83	4.52	0.78	4.69	0.55	4.61	0.64
	Tobacco use	1.59	1.51	1.22	0.98	1.08	0.58	1.26	1.09
	Alcohol use	1.67	1.14	1.47	1.06	1.20	0.63	1.48	1.02
	Cannabis use	1.22	0.84	1.09	0.53	1.04	0.37	1.07	0.48
Psychological variables	Sense of coherence	3.72	0.78	4.72	0.81	5.41	0.87	4.32	0.75
	Emotional regulation	2.79	0.73	3.19	0.81	3.63	0.88	3.14	0.86
	Curiosity and exploration	2.85	0.81	3.39	0.84	3.48	0.91	3.03	0.84
	Perceived body image	2.55	0.91	2.98	0.67	3.01	0.62	2.71	0.82
	Satisfaction with body image	3.20	1.01	4.23	0.88	4.46	0.73	3.57	0.85
Socioeconomic variables	Father educational level	2.79	0.75	2.90	0.81	3.03	0.78	2.97	0.79
	Mother educational level	2.94	0.79	3.03	0.80	3.20	0.79	2.97	0.84
	Perceived family wealth	2.94	0.50	3.11	0.48	3.09	0.43	3.00	0.53

MVPA, Moderate to Vigorous Physical Activity; VPA, Vigorous Physical Activity.

The most influential independent variables, with ORs higher than 2, are perceived family wealth (*OR* = 2.83, 95% *CI* = 1.47–5.44) and sense of coherence (*OR* = 2.74, 95% *CI* = 1.84–4.10). Satisfaction with body image (*OR* = 1.84, 95% *CI* = 1.31–2.58) and perceived academic achievement (*OR* = 1.64, 95% *CI* = 1.13–2.38) also stand out. Lastly, more modest contributions were found for breakfast frequency (*OR* = 1.33, 95% *CI* = 1.13–1.56), satisfaction with friendships (*OR* = 1.31, 95% *CI* = 1.02–1.68), frequency of moderate to vigorous physical activity (*OR* = 1.24, 95% *CI* = 1.05–1.45), teacher support (*OR* = 1.19, 95% *CI* = 1.06–1.33) and curiosity and exploration (*OR* = 1.05, 95% *CI* = 1.01–1.10).

Research Question 3. The Study of Vulnerable Adolescents

This third section focuses on those adolescents who, despite having good-quality parent-child relationships show low global health scores, that is to say, the vulnerable group (3.9% of the global sample and 11.9% of the group of participants that showed high-quality in parent-child relationships). This group of adolescents are compared with those adolescents who, having a good-quality parent-child relationship, show high global health score, that is to say, the competent group (18.3% of the global sample and 56.1% of the group with good-quality parent-child relationships).

Results from the logistic regression analyses, taking the group of vulnerable adolescents as a reference value, are shown. As in the analyses of the resilient adolescents, six models have been estimated, one for each set of independent variables (including the variables sex and age in all of them, so that they do not become confounding variables). In addition, a global model is presented at the end in which only the significant variables from previous models are included.

As can be seen in the first row of **Table 9**, although model 1 overall explained 9.7% of total variability, with the variables sex and age being significant (specifically, girls and older adolescents

TABLE 7 | Mean comparisons test (ANOVA with Bonferroni correction for multiple comparisons and effect size) of age, school context, peer context, lifestyle, psychological and socioeconomic variables between maladaptative, resilient, competent, and vulnerable adolescents.

	Variables	Maladaptative/ Resilient	Maladaptative/ Competent	Maladaptative/ Vulnerable	Resilient/ Competent	Resilient/ Vulnerable	Competent/ Vulnerable
	Age	<0.001*	<0.001**	>0.999	0.003*	0.003*	<0.001**
School context	Perceived academic achievement	<0.001**	<0.001***	0.483	0.012*	<0.001*	<0.001**
	Feelings toward school	<0.001*	<0.001**	0.020*	<0.001*	>0.999	<0.001*
	Perceived teacher support	<0.001**	<0.001***	<0.001**	<0.001*	0.948	<0.001**
Peer context	Perceived social support	<0.001*	<0.001**	0.001*	0.118	>0.999	0.015*
	Models of behavior	<0.001*	<0.001**	0.304	0.123	0.324	<0.001*
	Satisfaction with friendships	<0.001**	<0.001**	0.003*	>0.999	0.005*	<0.001**
	Having been bullied	<0.001*	<0.001*	0.969	>0.999	0.329	0.685
	Having bullied others	>0.999	<0.001*	0.181	0.312	>0.999	>0.999
Lifestyles	Breakfast frequency	<0.001**	<0.001**	<0.001*	0.538	0.468	0.001*
	Fruit consumption	0.048	<0.001*	>0.999	0.083	>0.999	0.002*
	Snack consumption	>0.999	0.290	>0.999	0.571	>0.999	0.884
	MVPA	<0.001**	<0.001***	>0.999	>0.999	<0.001**	<0.001***
	VFA	<0.001**	<0.001**	>0.999	>0.999	<0.001**	<0.001**
	Dental hygiene	>0.999	<0.001*	0.115	0.025*	>0.999	>0.999
	Tobacco use	0.001*	<0.001*	0.008*	0.867	>0.999	0.466
	Alcohol use	0.071	<0.001**	0.138	0.005*	>0.999	0.007*
	Cannabis use	0.102	>0.999	0.053	>0.999	>0.999	>0.999
Psychological variables	Sense of coherence	<0.001***	<0.001***	<0.001**	<0.001***	<0.001**	<0.001***
	Emotional regulation	<0.001**	<0.001***	<0.001*	<0.001**	>0.999	<0.001**
	Curiosity and exploration	<0.001**	<0.001**	0.257	>0.999	0.016*	<0.001**
	Perceived body image	<0.001**	<0.001**	0.139	>0.999	0.011*	<0.001*
	Satisfaction with body image	<0.001***	<0.001***	<0.001*	0.017*	<0.001**	<0.001***
Socioeconomic variables	Father educational level	0.502	<0.001*	0.045*	0.301	>0.999	>0.999
	Mother educational level	0.827	<0.001*	>0.999	0.089	>0.999	0.009*
	Perceived family wealth	<0.001*	<0.001*	0.818	>0.999	0.223	0.184

MVPA, Moderate to Vigorous Physical Activity; VPA, Vigorous Physical Activity. Effect size interpretation: 0–0.19 = negligible (–), 0.20–0.49 = small (*), 0.50–0.79 = medium (**), 0.80 and above = high (***). The bold values indicates (small, medium, or high) effect size values.

have a higher probability of being vulnerable), the percentage of correctly classified adolescents using theses demographic variables only was 0%.

In model 2, regarding school context, the explained variance is 20.4% and the model correctly classifies 16.8% of the vulnerable adolescents. Specifically, those adolescents who perceive lower teacher support, with an OR of 0.87 (95% *CI* = 0.83–0.91), have less positive feelings toward school (*OR* = 0.77, 95% *CI* = 0.65–0.91) and worse academic achievement (*OR* = 0.60, 95% *CI* = 0.50–0.72) have a higher likelihood of being vulnerable.

Model 3, which includes the variables of peer context, shows a predictive capacity of 16.8%, with 10% of the vulnerable adolescents being correctly classified. Those adolescents who report lower perceived social support (*OR* = 0.97, 95% *CI* = 0.94–0.99) and less satisfaction with friendships (*OR* = 0.76, 95% *CI* = 0.69–0.83) are more likely to be vulnerable.

Model 4 is devoted to variables related to lifestyles and its explained variance level is 21%, with 17.5% of the vulnerable adolescents being correctly classified. In this model, alcohol use stands out, showing that those adolescents who show a higher frequency of alcohol use in the last 30 days are 1.45 times more likely to be vulnerable (95% *CI* = 1.20–1.76). Likewise, the adolescents who do less moderate to vigorous physical activity (*OR* = 0.74, 95% *CI* = 0.67–0.80) and less vigorous physical activity (*OR* = 0.89, 95% *CI* = 0.80–0.99) have a higher likelihood of being vulnerable.

The group of psychological variables, analyzed in model 5, has the highest level of explained variance among the six specific models. Specifically, the level of explained variance in model 5 is 44.7% and 52% of the vulnerable adolescents are correctly classified. As in the previous section regarding resilient adolescents, the significant variables in this model are: sense of

TABLE 8 | Logistic regression models on resilience by demographic, school context, peer context, lifestyle, psychological and socioeconomic variables.

		Model 1-OR (95%-CI) Demographic variables	Model 2-OR (95%-CI) School context	Model 3-OR (95%-CI) Peer context	Model 4-OR (95%-CI) Lifestyles	Model 5-OR (95%-CI) Psychological variables	Model 6-OR (95%-CI) Socioeconomic variables	Model 7-OR (95%-CI) Global
	Variables	***R*² = 0.108 (80.5/0.0%)**	***R*² = 0.228 (83.2/22.1%)**	***R*² = 0.238 (82.4/19.2%)**	***R*² = 0.247 (83.0/25.6%)**	***R*² = 0.370 (87.2/30.4%)**	***R*² = 0.131 (80.2/3.5%)**	***R*² = 0.518 (89.3/51.5%)**
Demographic variables	Sex (ref. girls)	3.23** (2.27–4.58)	3.57** (2.45–5.21)	3.60** (2.46–5.27)	2.21** (1.47–3.32)	1.35 (0.79–2.33)	3.23 (2.27–4.61)	1.35 (0.73–2.48)
	Age	0.70** (0.59–0.83)	0.73** (0.62–0.87)	0.72** (0.60–0.87)	0.74** (0.62–0.90)	0.72 (0.50–1.04)	0.69 (0.58–0.81)	0.88 (0.59–1.32)
	Type of educational center (ref. public)	1.40 (0.98–1.99)	NA	NA	NA	NA	NA	NA
	Habitat (ref. urban)	1.04 (0.74–1.47)	NA	NA	NA	NA	NA	NA
School context	Perceived academic achievement	NA	1.83** (1.44–2.33)	NA	NA	NA	NA	1.64** (1.13–2.38)
	Feelings toward school	NA	1.19 (0.96–1.50)	NA	NA	NA	NA	NA
	Perceived teacher support	NA	1.19** (1.11–1.29)	NA	NA	NA	NA	1.19** (1.06–1.33)
Peer context	Perceived social support	NA	NA	1.04 (1.00–1.08)	NA	NA	NA	NA
	Models of behavior	NA	NA	1.07* (1.01–1.14)	NA	NA	NA	1.04 (0.95–1.15)
	Satisfaction with friendships	NA	NA	1.50** (1.26–1.79)	NA	NA	NA	1.31* (1.02–1.68)
	Having been bullied	NA	NA	0.53** (0.33–0.84)	NA	NA	NA	0.82 (0.42–1.62)
	Having bullied others	NA	NA	0.92 (0.68–1.25)	NA	NA	NA	NA
Lifestyles	Breakfast frequency	NA	NA	NA	1.24** (1.12–1.38)	NA	NA	1.33** (1.13–1.56)
	Fruit consumption	NA	NA	NA	1.02 (0.91–1.15)	NA	NA	NA
	Snack consumption	NA	NA	NA	1.07 (0.90–1.26)	NA	NA	NA
	MVPA	NA	NA	NA	1.37** (1.22–1.54)	NA	NA	1.24** (1.05–1.45)
	VPA	NA	NA	NA	1.13 (0.97–1.30)	NA	NA	NA
	Dental hygiene	NA	NA	NA	1.17 (0.91–1.50)	NA	NA	NA
	Tobacco use	NA	NA	NA	0.88 (0.68–1.13)	NA	NA	NA
	Alcohol use	NA	NA	NA	0.92 (0.75–1.13)	NA	NA	NA
	Cannabis use	NA	NA	NA	0.89 (0.61–1.31)	NA	NA	NA
Psychological variables	Sense of coherence	NA	NA	NA	NA	3.18** (2.14–4.73)	NA	2.74** (1.84–4.10)
	Emotional regulation	NA	NA	NA	NA	1.01 (0.96–1.06)	NA	NA
	Curiosity and exploration	NA	NA	NA	NA	1.07** (1.03–1.11)	NA	1.05** (1.01–1.10)
	Perceived body image	NA	NA	NA	NA	1.17 (0.82–1.66)	NA	NA
	Satisfaction with body image	NA	NA	NA	NA	1.83** (1.31–2.56)	NA	1.84** (1.31–2.58)
Socio econom.	Father educational level	NA	NA	NA	NA	NA	1.08 (0.83–1.39)	NA
	Mother educational level	NA	NA	NA	NA	NA	1.06 (0.82–1.36)	NA
	Perceived family wealth	NA	NA	NA	NA	NA	1.98** (1.37–2.85)	2.83** (1.47–5.44)

*MVPA, Moderate to Vigorous Physical Activity; VPA, Vigorous Physical Activity; OR, Odds Ratio (95% CI = Confidence Interval at the 95% level); R^2 = Model explained variance (% correctly-classified total/% correctly-classified resilient group); NA, not applicable. *p < 0.05, **p < 0.01.*

TABLE 9 | Logistic regression models on vulnerability by school context, peer context, lifestyle, psychological and socioeconomic variables.

		Model 1-OR (95%-CI) Demographic variables	**Model 2-OR (95%-CI) School context**	**Model 3-OR (95%-CI) Peer context**	**Model 4-OR (95%-CI) Lifestyles**	**Model 5-OR (95%-CI) Psychological variables**	**Model 6-OR (95%-CI) Socioeconomic variables**	**Model 7-OR (95%-CI) Global**
	Variables	**$R^2 = 0.097$ (81.5/0.0%)**	**$R^2 = 0.204$ (82.3/16.8%)**	**$R^2 = 0.168$ (81.3/10.0%)**	**$R^2 = 0.210$ (82.8/17.5%)**	**$R^2 = 0.447$ (83.3/52.0%)**	**$R^2 = 0.112$ (82.4/2.7%)**	**$R^2 = 0.565$ (86.6/62.9%)**
Demographic variables	Sex (ref. girls)	0.46** (0.35–0.60)	0.35** (0.26–0.47)	0.41** (0.31–0.55)	0.62** (0.45–0.85)	0.59 (0.33–1.05)	0.52** (0.36–0.76)	0.76 (0.38–1.53)
	Age	1.60** (1.41–1.80)	1.53** (1.34–1.74)	1.57** (1.38–1.79)	1.42** (1.24–1.62)	1.36 (0.92–2.02)	1.62** (1.37–1.91)	1.31 (0.85–2.01)
	Type of educational center (ref. public)	0.77 (0.58–1.03)	NA	NA	NA	NA	NA	NA
	Habitat (ref. urban)	0.80 (0.61–1.06)	NA	NA	NA	NA	NA	NA
School context	Perceived academic achievement	NA	0.60** (0.50–0.72)	NA	NA	NA	NA	0.49** (0.32–0.76)
	Feelings toward school	NA	0.77** (0.65–0.91)	NA	NA	NA	NA	0.88 (0.62–1.27)
	Perceived teacher support	NA	0.87** (0.83–0.91)	NA	NA	NA	NA	0.85** (0.75–0.95)
Peer context	Perceived social support	NA	NA	0.97** (0.94–0.99)	NA	NA	NA	1.01 (0.91–1.13)
	Models of behavior	NA	NA	0.98 (0.94–1.02)	NA	NA	NA	NA
	Satisfaction with friendships	NA	NA	0.76** (0.69–0.83)	NA	NA	NA	0.92 (0.74–1.15)
	Having been bullied	NA	NA	1.12 (0.88–1.43)	NA	NA	NA	NA
	Having bullied others	NA	NA	1.10 (0.86–1.41)	NA	NA	NA	NA
Lifestyles	Breakfast frequency	NA	NA	NA	0.94 (0.87–1.02)	NA	NA	NA
	Fruit consumption	NA	NA	NA	0.99 (0.91–1.08)	NA	NA	NA
	Snack consumption	NA	NA	NA	1.05 (0.93–1.18)	NA	NA	NA
	MVPA	NA	NA	NA	0.74** (0.67–0.80)	NA	NA	0.70** (0.58–0.86)
	VPA	NA	NA	NA	0.89* (0.80–0.99)	NA	NA	0.84 (0.67–1.05)
	Dental hygiene	NA	NA	NA	0.70** (0.56–0.89)	NA	NA	1.54 (0.79–2.99)
	Tobacco use	NA	NA	NA	1.03 (0.84–1.26)	NA	NA	NA
	Alcohol use	NA	NA	NA	1.45** (1.20–1.76)	NA	NA	1.24 (0.84–1.82)
	Cannabis use	NA	NA	NA	1.19 (0.86–1.65)	NA	NA	NA
Psychological variables	Sense of coherence	NA	NA	NA	NA	0.27** (0.18–0.40)	NA	0.30** (0.19–0.45)
	Emotional regulation	NA	NA	NA	NA	0.99 (0.94–1.03)	NA	NA
	Curiosity and exploration	NA	NA	NA	NA	0.95** (0.92–0.98)	NA	0.96* (0.93–0.99)
	Perceived body image	NA	NA	NA	NA	0.76 (0.50–1.16)	NA	NA
	Satisfaction with body image	NA	NA	NA	NA	0.46** (0.35–0.69)	NA	0.50** (0.35–0.71)
Socioeconomic variables.	Father educational level	NA	NA	NA	NA	NA	1.19 (0.89–1.58)	NA
	Mother educational level	NA	NA	NA	NA	NA	0.66** (0.51–0.87)	0.83 (0.57–1.23)
	Perceived family wealth	NA	NA	NA	NA	NA	0.65 (0.41–1.02)	NA

*MVPA, Moderate to Vigorous Physical Activity; VPA, Vigorous Physical Activity; OR, Odds Ratio (95% CI, Confidence Interval at the 95% level); R^2 = Model explained variancel (% correctly-classified total/% correctly-classified vulnerable group); NA, not applicable. * $p < 0.05$, ** $p < 0.01$.*

coherence, curiosity and exploration and satisfaction with body image. The likelihood of being vulnerable is higher in those adolescents with a lower score in sense of coherence (*OR* = 0.27, 95% *CI* = 0.18–0.40), less satisfaction with their body image (*OR* = 0.46, 95% *CI* = 0.35–0.69) and a lower score in curiosity and exploration (*OR* = 0.95, 95% *CI* = 0.92–0.98).

Model 6, examining the socioeconomic variables, again shows a lower predictive capacity than previous models (11.2%), with only 2.7% of the vulnerable adolescents being correctly-classified. The mother's educational level is the only significant variable in this model, revealing that those adolescents whose mothers have a lower educational level exhibit a higher probability of being vulnerable (*OR* = 0.66, 95% *CI* = 0.51-0.87).

Lastly, in model 7 or the global model (in which only the significant variables from previous models have been included), the following six variables were significant: perceived academic achievement, perceived teacher support, moderate to vigorous physical activity, sense of coherence, curiosity and exploration and satisfaction with body image. The predictive capacity of this model is very high, with an explained variance level of 56.5%. This model was also able to correctly classify a high proportion of vulnerable adolescents, specifically 62.9%.

The independent variables which stand out in this model due to their ORs being closer to zero, and therefore their higher predictive capacity, are sense of coherence (*OR* = 0.30, 95% *CI* = 0.19-0.45), academic achievement (*OR* = 0.49, 95% *CI* = 0.32–0.76) and satisfaction with body image (*OR* = 0.50, 95% *CI* = 0.35–0.71). Moderate to vigorous physical activity (*OR* = 0.70, 95% *CI* = 0.58–0.86) and teacher support (*OR* = 0.85, 95% *CI* = 0.75–0.95) appear on an intermediate level. Lastly, the level of curiosity and exploration (*OR* = 0.96, 95% *CI* = 0.93–0.99) made the most modest contribution. Higher levels of the aforementioned variables are associated with a lower likelihood of belonging to the group of vulnerable adolescents.

DISCUSSION

The aim of this study was to characterize resilience and vulnerability in a large and representative sample of adolescents. This objective was first addressed separately on a number of potential levels of influence (demographic, school, peer, lifestyle, psychological, and socioeconomic variables) and later, in a more holistic approach, by integrating the factors in all the aforementioned levels.

A separate analysis of each of the two phenomena showed, first at all, that although there was a higher representation of boys and younger adolescents in the resilient group, and of girls and older adolescents in the vulnerable group, the variables sex and age were not sufficient to accurately predict adolescent adaptation. Previous research has found differences in wellbeing and adjustment between boys and girls, as well as according to age (Cavallo et al., 2006; Ramos et al., 2010), but at the same time there is notable diversity amongst adolescents of the same sex and age. This diversity tends to be related to the combination of life experiences and psychological characteristics of these adolescents. Hence the demographic variables (that were included in all the regression models) were insufficient to characterize such complex phenomena such as resilience and vulnerability and their significant effects disappeared when they were entered along with the rest of the variables in the final model. In fact, sex and age already lost their significant effects in previous models, specifically in those evaluating the contributions of psychological and socioeconomic variables. This is probably owing to that those models incorporated variables such as satisfaction with body image, which tends to be lower and more strongly associated with girls' wellbeing (Knauss et al., 2007; Mond et al., 2011), or family wealth, which tends to be assessed more negatively by older adolescents (Goodman et al., 2001). Therefore, it could be understood that these predictor variables (such as body image or family wealth) explain the predictive capacity of the variables sex and age on the phenomena resilience and vulnerability.

Beyond demographic variables, a look to the separate models for each set of predictors shows that a hierarchy based on the predictive capacity of each set of variables would be very similar for resilience and vulnerability: psychological variables in the first place, along with contextual and lifestyle variables, and more modest contributions of demographic and socioeconomic variables.

In addition, the final models for resilience and vulnerability also revealed a number of common factors for the explanation of both phenomena. In other words, these analyses also helped identify several factors that contributed significantly to explaining both resilience and vulnerability.

First at all, sense of coherence was one of the most important factors for both resilience and vulnerability. This construct, coming from the salutogenic model in the field of public health, has to do with a person's ability to interpret their social environments as predictable and ordered, their confidence that any life demand can be successfully dealt with as well as a motivational-emotional component that helps one to see difficult situations as challenges and facilitates an active engagement in problem-solving (Antonovsky, 1987). Therefore, the important contribution of sense of coherence to resilience and vulnerability should come as no surprise. On one hand, its links to some factors associated with successful adaptation in classic resilience studies, such as analytical skills, motivation to engage in the environment, self-efficacy and self-esteem (Masten, 2001; Hamill, 2003), are apparent in the prior description. In addition, research on sense of coherence indicates that its relationship with health and wellbeing is rooted in helping people mobilize other useful coping resources in stressful situations (Lindström and Eriksson, 2010), which has led to its inclusion in the health assets model as a *supra-order asset* for wellbeing (Morgan and Hernán, 2013). In this sense, one line of research that arises from the results obtained in the current study is the study on the processes that explain why a high sense of coherence would help resilient adolescents take full advantage of available resources, whereas low levels of the same would hamper the effective use of the apparently more abundant resources in the case of vulnerable adolescents.

Satisfaction with body image and perceived academic achievement also appeared as important explanatory variables in the analysis of both resilience and vulnerability. The significant contribution of satisfaction with body image is probably related to the importance of physical appearance for adolescents' positive self-perception. In this regard, numerous studies have found a significant relation between satisfaction with body image and self-esteem in adolescence (Tiggemann, 2005), this latter being a factor traditionally connected to successful adaptation (e.g., Dumont and Provost, 1999). Something similar can be said of the relationship between perceived academic achievement and self-efficacy (Danielsen et al., 2009), another fundamental protective factor in resilience research (Hamill, 2003). Additionally, previous research indicates that feeling competent in daily life is very important for the adaptation of individuals suffering adversity (Masten and Coatsworth, 1998). Therefore, it is likely that behind an adolescent who thinks that their teachers consider their academic achievement as good, there are various underlying beneficial elements for adaptation and wellbeing, such as experiences of competence in the school context, higher school connectedness or even a higher intellectual capacity (Masten et al., 1999; Blum, 2005).

In addition to perceived academic achievement, higher levels of teacher support increased the likelihood of showing resilience and diminished that of being part of the group of vulnerable adolescents. Studies about teachers' contribution to adolescent wellbeing also suggest that, regardless of the level of academic achievement, teacher support acts as an asset associated with wellbeing for all adolescents (e.g., García-Moya et al., 2015), which makes it fundamental to favor close and supportive teacher-student relationships.

Moderate to vigorous physical activity was also amongst the significant factors associated with resilience and vulnerability. Physical activity has been found to have protective effects in stressful situations (Gerber and Pühse, 2009; Silverman and Deuster, 2014), as well as it tends to reduce the likelihood of engaging in risk behaviors (Pate et al., 1996), therefore serving as a clear example of the importance of taking into account lifestyles' contributions to explaining resilience and vulnerability.

Finally, higher levels of curiosity and exploration increased the likelihood of being resilient and diminished that of being part of the vulnerable group. The curiosity and exploration construct reflects openness and interest in learning, good management of the uncertainty associated with new or unknown situations (Kashdan et al., 2009) and is associated with psychological and contextual variables significantly linked to adaptation and resilience. Specifically, high levels of curiosity and exploration are related to an active response in unfamiliar and challenging environments (Kashdan and Roberts, 2004) and have been linked to psychological variables such as intrinsic motivation and self-efficacy (Kashdan et al., 2004). Additionally, curiosity was also significantly associated with more positive social interactions (Kashdan and Roberts, 2004). Specifically, people with higher levels of curiosity and exploration generated more positive responses from strangers, who tended to be more responsive, participative, and interested in social exchanges with people with high curiosity. Despite this, the contribution of curiosity and exploration to the final model was relatively modest, probably due to its connections with other constructs, such as sense of coherence. The conceptual delimitation of curiosity and exploration is still under study (Kashdan et al., 2009), and with regards to sense of coherence one focus of analysis and debate is precisely its connection to other constructs in positive psychology (Lindström and Eriksson, 2010). Consequently, advancing in the conceptual delimitation of these constructs, identifying common elements and differences between them, is an important line of research (García-Moya and Morgan, 2016) that could contribute to a better understanding of resilience and vulnerability and, in general, of their role in promoting adolescent wellbeing and adjustment.

As explained in the previous lines, the vast majority of the examined factors operated by increasing the likelihood of good adaptation in resilient adolescents and diminishing it in vulnerable ones. Overall, this suggests more similarities than differences in the factors contributing to explaining resilience and vulnerability. These findings coincide with previous research pointing out that factors associated with resilience are not specific to this phenomenon, but that they are the manifestation of basic systems of human adaptation and, therefore, are influential in both adversity and non-adversity situations (Masten and Coatsworth, 1998; Masten, 2001). Additionally, some scholars have noted that protective factors identified in resilience and vulnerability studies frequently correspond to the positive pole of risk factors for maladaptation or, in other words, that in this type of research it is possible to identify factors in which one of their extremes facilitates successful adaptation while the opposite hampers it (Sameroff and Fiese, 2000; Fergus and Zimmerman, 2005; Luthar et al., 2015), which seems to coincide with findings in the present study.

Despite the predominant similarities described so far, results also revealed some differential aspects between the resilience and vulnerability phenomena. First, the psychological variables showed a larger explicative capacity in vulnerable adolescents than in resilient ones (R^2= 0.447 and 0.370, respectively), whereas factors related to school and peer contexts, especially the second, showed a stronger association with resilience than with vulnerability (R^2= 0.228 and 0.238 respectively in the models on resilience vs. 0.204 and 0.168 for vulnerability). Some research suggests that certain protective factors such as temperament (e.g., Werner and Smith, 1982) or intellectual capacity (Masten and Coatsworth, 1998), to name some classic examples, have a multiplier effect, i.e., they can contribute to a higher likelihood of encountering other positive events in life, giving rise to chain reactions favoring positive adaptation or, conversely, they can initiate cascading effects in which new risk factors are more probable. Applying a similar logic, it can be hypothesized that certain psychological variables, such as a low sense of coherence, a lower tendency toward curiosity and exploration, or a higher dissatisfaction with body image, could be preventing vulnerable adolescents from taking advantage of potential resources in extrafamily environments (school and peer contexts), whereas resilient adolescents, despite their more unfavorable family context (which was the indicator used for the definition of adversity in the current study), would be more likely to find and benefit from resources available in extrafamily environments thanks to their more positive profile in psychological variables.

Along these lines, prior research has documented the existence of compensatory effects from other contexts in only part of the adolescents exposed to low-quality family contexts (e.g., García-Moya et al., 2013b).

Second, three of the examined factors, specifically perceived family wealth, satisfaction with friendships and breakfast frequency, were only significant in the analysis of resilience. This means that these variables made a difference for adolescents exposed to adversity in the family context (resilient vs. maladaptative adolescents) but did not contribute to explain differences in adaptation between vulnerable and competent adolescents. Scientific literature has extensively documented that resilience has among its defining attributes an ability, despite adversity, to find and take advantage of any resources and opportunities in proximal environments.

In this sense, it is not surprising that being raised in a family environment with good socioeconomic resources opens a horizon of possibilities to resilient adolescents that they seem to know how to take advantage of. What is interesting in the findings of the present study is that although vulnerable and resilience adolescents reported similar levels of perceived family wealth, this factor made one of the most noticeable contributions in analyzing resilience (but not vulnerability). A reflection on the nature of the indicator used may help understand this finding. On the one hand, research suggests that perceived family wealth includes some of the elements which are common to objective indicators such as family affluence, and therefore, it can arguably be interpreted as indicative to some extent of the wider access to external resources and opportunities for development that families' socioeconomic level relates to Bornstein and Bradley (2003). However, research also indicates that subjective and objective measures are not assessing exactly the same content (Hartley et al., 2015; Elgar et al., 2016), since unlike objective indicators perceived family wealth may also incorporate a comparative assessment of the socioeconomic position of the adolescent's family in comparisons with that of others they related with (Moreno-Maldonado et al., under review). The levels of wealth perceived by resilient adolescents may therefore represent a relative socioeconomic advantage for these adolescents compared to their peers also exposed to adversity in family relationships (the maladaptative group).

Results on satisfaction with friendships can also be interpreted in a similar sense, i.e., that resilient individuals are able to take advantage of the potential resources they find. Peer support tends to be considered a protective factor in adversity situations (Olsson et al., 2003) and resilient adolescents in the present study probably illustrate very well the compensatory effects which are frequently mentioned in this field (e.g., Luthar et al., 2015): they belong to a group who, despite coming from families in which parent-child relationships are not good, is able to build positive relationships with their peer group and benefit from them (Lansford et al., 2003; Rubin et al., 2004). In a similar vein, Luthar et al. (2015) state that relationships with peers can become a "remedial" socializing context for children who grow up exposed to family adversity. In addition, positive peer relationships are indicative of good social competence, a fundamental skill in which resilient adolescents usually show positive results, which are comparable to those of competent adolescents and clearly more favorable than the social competence levels exhibited by maladaptative adolescents (Masten et al., 1999).

Finally, the fact that breakfast frequency was significant only in the analysis of resilience may be related to the fact that, as children gain more independence during adolescence, the importance of parental supervision in this behavior decreases while internalization of the habit and other personal characteristics, such as constancy and self-regulation, gain prominence (Kalavana et al., 2010). Given that breakfast frequency is also believed to act as a proxy for diverse socioeconomic and family aspects this is an issue which, in particular, would benefit from further research.

In any case, the comments that have been made throughout this discussion about a higher ability of resilient individuals to take advantage of potential resources in proximal contexts or the important role of psychological factors for explaining the resilience and vulnerability phenomena should not be interpreted as evidence that they are characteristics unrelated to the contextual experiences associated with resilience and vulnerability. As rightly pointed out by Luthar et al. (2015), contextual experiences indeed give shape, from the beginning and in a continuous transactional dynamic, to said skills or psychological resources.

This study has some limitations that should be taken into consideration in the interpretation of its findings. Firstly, its cross-sectional design means that the results must be interpreted on an associative level, not being possible to draw conclusions about the directionality of the relationships found. Secondly, adversity was defined using quality of parent-child relationships as a sole criterion. Although, as explained in the introduction, this is an well-informed decision, which draws on scientific literature that highlights the role of family as a basic system for human adaptation (Masten, 2001; Fergus and Zimmerman, 2005), previous research also shows the wide variety of life circumstances that can constitute adversity in childhood and adolescence (Luthar et al., 2000); consequently, it would be inappropriate to generalize these findings to other adverse circumstances. Finally, this study used a factorial health score as its measure of adaptation. Although, this measure is a sound and validated global health indicator (Ramos et al., 2010) whose characteristics fit well with key measurement issues in the empirical definition of positive adaptation (Luthar and Cushing, 1999), there is substantial evidence on the multi-dimensional nature of human adaptation, which makes individuals show dissimilar levels of adaptation in different areas (Luthar et al., 1993). Therefore, future research should complement the present study by conducting separated analyses of the contributions of the factors analyzed here to distinct areas of adaptation, mainly the following: academic, behavioral, social and emotional (Masten et al., 1999; Luthar et al., 2000).

Despite the aforementioned limitations, this study also has important strengths. In line with recommendations from some of the seminal reviews in this research field (Luthar et al., 2000, 2015; Masten, 2014), the elements of adversity and adaptation were clearly operationalized for the definition of resilience and vulnerability in the present study, which is fundamental for an adequate interpretation of its findings and its comparability with other studies. The criteria used for making the distinction and comparisons among the four adaptation groups (competent, vulnerable, resilient, and maladaptative) were also based in previous research (Tiêt and Huizinga, 2002). Additionally, this research adheres to the methodological rigor characteristics of the HBSC survey (Roberts et al., 2009), as well as it stands out for its large sample size, which allowed for a characterization of resilience and vulnerability phenomena in a representative and notably large sample. Although, the four groups may appear unbalanced in size, the representativeness of the initial sample is a guarantee that this is a relatively realistic reflection of the population. In addition, using effect size tests in all of the analyses minimizes the potential bias that such differences in the subgroups' size could case from a methodological point of view. The high predictive capacity of the models of resilience and vulnerability obtained, which reached levels of explained variance higher than 50%, is also outstanding. These values are considered high in the field of behavioral science (Cohen, 1988), being notably above the 10–20% that is usual for associations between protective factors and adaptation outcomes in resilience studies (Luthar et al., 2000). Finally, this study has three elements that are, to some extent, innovative. Firstly, factors traditionally receiving little attention as referred to in the introduction, such as lifestyles, satisfaction with body image, sense of coherence, curiosity and exploration and perceived family wealth, were analyzed in the present study of resilience and vulnerability. Secondly, this study included vulnerable adolescents, a population subgroup that had rarely been studied in previous research due to its limited sample size (Masten et al., 1999). Additionally, this work makes a valuable contribution regarding the prevalence of vulnerability and resilience in the general population. Given the difficulties associated with defining resilience and vulnerability and the limited methodological consensus with regards to the measures to use and how to apply them to a representative sample, it is understandable that prevalence studies are not available. In this regard, the present study found that vulnerable adolescents made up 3.9% of the global sample, representing 11.9% of the group that reported high-quality parent-child relationships. The resilience group represented 4.5% of the global sample, corresponding to 13.4% of the participants with low-quality parent-child relationships, which is in line with the findings of some longitudinal studies that have found a very low prevalence and stability in resilient coping (Cicchetti and Rogosch, 1997). Specifically, Bolger and Patterson (2003) found that between 6 and 21% of abused children were functioning competently during at least one of the temporal points examined in their longitudinal follow up from middle childhood to early adolescence, but less than 5% consistently maintained that competent functioning over time.

In addition to its strengths from a research perspective, which have just been highlighted, the fact that the present study provides valuable implications for the improvement of the methodological quality of interventions with resilient and vulnerable populations, which was one its guiding principles, should also be noted amongst its strong points.

Throughout these pages a number of important factors for adolescents' successful adaptation have been underlined. These include certain personal characteristics (such as sense of coherence, satisfaction with body image and curiosity and exploration), as well as some that characterize their lifestyles (regularity in healthy eating habits and physical activity) or that refer to the quality of their developmental contexts (such as satisfaction with peer relationships, academic achievement and teacher support). Therefore, all of these are elements to bear in mind in interventions aimed at promoting successful adaptation and wellbeing in adolescence (Olsson et al., 2003). Likewise, this study highlights the need to conduct further research devoted to developing reliable and valid indicators for the assessment of all these factors, both those that characterize the individual person and the ones that characterize their developmental contexts. These indicators will serve the double function of detecting subjects with different profiles of adaptation as well as of monitoring their evolution and evaluating the implemented interventions.

On a separate issue, it should be noted that some studies have advocated that interventions should be adjusted to the distinct developmental needs of adolescence (Kim et al., 2015). What the present research adds is that detecting different adaptation profiles would also serve to adjust interventions to every person's specific needs. On the one hand, some could argue that allocating powerful and costly resources to detect and intervene in vulnerable individuals, which represents 3.9% of adolescents, would not be an efficient strategy. However, it must be noted that this study used very demanding criteria to define the categories of vulnerability/resilience, and hence may have underestimated the prevalence figures. Additionally, it is well known from the accumulated evidence in previous research that life paths of vulnerable people will be full of difficulties in very different areas (this paper has provided some good examples of this). From an economic perspective, those adverse life paths will lead to a lot of public spending in the education, health, legal and judicial, and labor systems, amongst others (see Khan et al., 2015), if the direct and indirect costs to which these difficulties will give rise are taken into consideration; consequently, they should be detected as soon as possible. On the other hand, one should not forget that amongst the adaptation profiles considered in this paper, there were also a 18.88% of maladaptative adolescents with clear intervention needs, and in the remaining 77.3% of adolescents there will most likely always be areas of improvement and optimization in need of reinforcement. Similarly, it could also be thought that the resilient adolescents, for which our study shows a prevalence only slightly higher than 4%, would not need any intervention because they seem to resist adversity without help. It is true that these adolescents seem to have an admirable capacity to deal with adversity, but their resilience is not without limits. As can be seen when comparing them to the competent adolescents (please compare values of the resilient column with values of the competent column in **Tables 4**, **5**), resilient adolescents scored lower in an important number of variables. In other words, even though adolescents in the resilient group showed very high levels of adjustment despite coming from adverse family environments, their adjustment levels could still be higher if they were aided in taking full advantage of their skills and if interventions were implemented at the source of adversity. Needless to say, reducing adversity in their family environment should also be a top priority. Additionally, certain studies have already warned on the risks of underestimating resilient adolescents' needs for support, since some of these adolescents, despite being classified as resilient for showing excellent competence levels according to external and behavioral indicators, can nonetheless suffer from elevated levels of emotional distress (Luthar et al., 1993).

A final more general consideration should probably be added. In the dynamic relationship between research and intervention underlined in this paper, it should be emphasized that interventions should not work with models that explain development and change from a lineal or even interactive perspective, since empirical evidence shows data in favor of transactional models that involve much more complex multilevel dynamic systems (Sameroff, 2010). Therefore, even though all recent school intervention efforts aimed at strengthening life skills to optimize development and along the way prevent risk behaviors deserve our most sincere recognition and applause (Springer et al., 2004), the intervention that we defend here should go further. This guiding conceptual framework leads to the claim that intervention in adolescence should be preceded by an ambitious systematic and multi-sector intervention starting at the beginning of life. In this vein, as already noted by Luthar and Cicchetti (2000), interventions should take into consideration and simultaneously work on different levels of influence (individual, family and extrafamily) and should begin as early as possible. Community work with families and current steps toward promoting positive parenting very early in the baby's life are good points of reference in this direction (Rodrigo et al., 2012).

In summary, this study emphasizes the enormous potential of research on resilient and vulnerable individuals, both for creating scientific knowledge and for designing intervention guidelines. For a long time psychology overlooked both phenomena (vulnerability and resilience), due to the predominant scientific interest in central trends, i.e., toward what happened to the majority of people. Research was focused, on one hand, on those individuals that succumbed to adversity, and on the other, on those that showed strength as the result of having grown up surrounded by quality relationships. However, psychology must acknowledge the great deal that has been learnt since then by studying the limited percentage of people whose developmental trajectories apparently challenged the central-tendency hypotheses of that time: individuals who appeared to be strong and healthy despite adversity, as well as those who, despite growing up surrounded by strengths, seemed to be weak. Analysing the life trajectories of the first helps us to clarify what is desirable that all people have in their lives and the analysis of the life paths of the second, teaches us what is necessary to eradicate in all of them.

ETHICS STATEMENT

The study was approved by the ethics committee of Comité Ético de Experimentación de la Universidad de Sevilla. Parents of adolescents participating in the study received a letter with information about the study and informed consent model.

AUTHOR CONTRIBUTIONS

All authors conceived of the study, participated in its design and helped to draft the manuscript. Introduction was drafted by IG, methods and results were drafted by PR, FR and the discussion draft was done by IG, CM. All authors made suggestions and critical reviews to the initial draft and contributed to its improvement until reaching the final manuscript, which was read and approved by all authors.

ACKNOWLEDGMENTS

This study has been funded by the Ministerio de Sanidad, Servicios Sociales e Igualdad [Ministry of Health, Social Services and Equality] of Spain.

REFERENCES

Anteghini, M., Fonseca, H., Ireland, M., and Blum, R. W. (2001). Health risk behaviors and associated risk and protective factors among Brazilian Adolescents in Santos, Brazil. *J. Adolesc. Health.* 28, 295–302. doi: 10.1016/S1054-139X(00)00197-X

Antonovsky, A. (1987). *Unraveling the Mystery of Health*. San Francisco, CA: Jossey-Bass.

Blum, R. W. (2005). A case for school connectedness. *Adolesc. Learn.* 62, 16–20.

Bolger, K. E., and Patterson, C. J. (2003). "Sequelae of child maltreatment: Vulnerability and resilience," in *Resilience and Vulnerability: Adaptation in the Context of Childhood Adversity*, ed S. S. Luthar (New York, NY: Cambridge University Press), 156–181.

Bornstein, M. H., and Bradley, R. H. (2003).*Socioeconomic Status, Parenting, and Child Development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Buckner, J. C., Mezzacappa, E., and Beardslee, W. R. (2003). Characteristics of resilient youths living in poverty: the role of self-regulatory processes. *Dev. Psychopathol.* 15, 139–162. doi: 10.1017.S0954579403000087

Cantril, H. (1965). *The Pattern of Human Concerns*. New Brunswick, NJ: Rutgers University Press.

Cavallo, F., Zambon, A., Borraccino, A., Ravens-Sieberer, U., Torsheim, T., Lemma, P., et al. (2006). Girls growing through adolescence have a higher risk of poor health. *Qual. Life Res.* 15, 1577–1585. doi: 10.1007/s11136-006-0037-5

Cicchetti, D., and Rogosch, F. A. (1997).The role of self-organization in the promotion of resilience in maltreated children. *Dev. Psychopathol.* 9, 797–815.

Clarke-Stewart, A., and Dunn, J. (2006). *Families Count: Effects on Child and Adolescent Development*. Cambridge: Cambridge University Press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Science*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Conger, K. J., Rueter, M. A., and Conger, R. D. (2000). "The role of economic pressure in the lives of parents and their adolescents: The Family Stress Model," in *Negotiating Adolescence in Times of Social Change,* eds L. J.

Crockett, and R. K. Silbereisen (Cambridge: Cambridge University Press), 201–223.

Danielsen, A. G., Samdal, O., Hetland, J., and Wold, B. (2009). School-related social support and students' perceived life satisfaction. *J. Educ. Res.* 102, 303–320. doi: 10.3200/JOER.102.4.303-320

DeCoster, J., Iselin, A. M. R., and Gallucci, M. (2009). A conceptual and empirical examination of justifications for dichotomization. *Psychol. Methods* 14, 349–366. doi: 10.1037/a0016956

DuBois, D. L., Felner, R. D., Brand, S., Adan, A. M., and Evans, E. G. (1992).A prospective study of life stress, social support, and adaptation in early adolescence. *Child Dev.* 63, 542–557. doi: 10.1111/j.1467-8624.1992.tb01645.x

Dumont, M., and Provost, M. A. (1999).Resilience in adolescents: protective role of social support, coping strategies, self-esteem, and social activities on experience of stress and depression. *J. Youth Adolesc.* 28, 343–363. doi: 10.1023/A:1021637011732

Elgar, F. J., McKinnon, B., Torsheim, T., Schnohr, C. W., Mazur, J., Cavallo, F., et al. (2016). Patterns of socioeconomic inequality in adolescent health differ according to the measure of socioeconomic position. *Soc. Indic. Res.* 127, 1169–1180. doi: 10.1007/s11205-015-0994-6

Elliot, D. S. (1993). "Health-enhancing and health-compromising lifestyles," in *Promoting the Health of Adolescents: New Directions for the Twenty-first Century*, eds S. G. Millstein, A. C. Petersen and E. O. Nightingale (Oxford: Oxford University Press), 119–150.

Fergus, S., and Zimmerman, M. A. (2005). Adolescent resilience: a framework for understanding healthy development in the face of risk. *Annu. Rev. Public Healt.* 26, 399–419. doi: 10.1146/annurev.publhealth.26.021304.144357

Fergusson, D. M., and Linskey, M. T. (1996). Adolescent resiliency to family adversity. *J. Child Psychol. Psychiatry* 37, 281–292. doi: 10.1111/j.1469-7610.1996.tb01405.x

García-Moya, I., Brooks, F., Morgan, A., and Moreno, C. (2015). Subjective well-being in adolescence and teacher connectedness: a health asset analysis. *Health Educ. J.* 74, 641–654. doi: 10.1177/0017896914555039

García-Moya, I., Moreno, C., and Jiménez-Iglesias, A. (2013a). Building a composite factorial score for the assessment of quality of parent-child relationships in adolescence. *Eur. J. Dev. Psychol.* 10, 642–648. doi: 10.1080/17405629.2012.707781

García-Moya, I., Moreno, C., and Jiménez-Iglesias, A. (2013b). Understanding the joint effects of family and other developmental contexts on the sense of coherence (SOC): a person-focused analysis using the Classification Tree. *J. Adolesc.* 36, 913–923. doi: 10.1016/j.adolescence.2013.07.007

García-Moya, I., and Morgan, A. (2016). Salutogenesis' utility as a theory for guiding health promotion practice: the case of young people's well-being. *Health Promot Int.* doi: 10.1093/heapro/daw008. [Epub ahead of print].

Gerber, M., and Pühse, U. (2009). Review Article: do exercise and fitness protect against stress-induced health complaints? *Scand. J Public Health* 37, 801–819. doi: 10.1177/1403494809350522

Goodman, E., Adler, N. E., Kawachi, I., Frazier, A. L., Huang, B., and Colditz, G. A. (2001). Adolescents' perceptions of social status: development and evaluation of a new indicator. *Pediatrics* 108:e31. doi: 10.1542/peds.108.2.e31

Hamill, S. K. (2003). Resilience and self-efficacy: the importance of efficacy beliefs and coping mechanisms in resilient adolescents. *Colgate Univ. J. Sci.* 35, 115–146.

Hartley, J. E. K., Levin, K., and Currie, C. (2015). A new version of the HBSC Family Affluence Scale - FAS III: Scottish qualitative findings from the international FAS development study. *Child Indic. Res.* 9, 233–245. doi: 10.1007/s12187-015-9325-3

Hibell, B., Andersson, B., Ahlström, S., Balakireva, O., Bjarnason, T., Kokkevi, A., et al. (2000). *The 1999 ESPAD Report. Alcohol and Other Drug Use Among Students in 30 European Countries*. Stockholm: The Swedish Council for Information on Alcohol and Other Drugs.

Idler, E. L., and Benyamini, Y. (1997). Self-rated health and mortality: A review of 27 community studies. *J. Health Soc. Behav.* 38, 21–37. doi: 10.2307/2955359

Inchley, J., Currie, D., Young, T., Samdal, O., Torsheim, T., Augustson, L., et al. (2016). *Growing Up Unequal: Gender and Socioeconomic Differences in Young People's Health and Well-being, Vol. 7.* Health behaviour in school-aged children: International report from the 2013/2014 survey. Health policy for children and adolescents, World Health Organization Regional Office for Europe, Copenhagen.

Jain, S., Buka, S.L., Subramanian, S. V., and Molnar, B. E. (2012). Protective factors for youth exposed to violence: role of developmental assets in building emotional resilience. *Youth Viol. Juv. Justice* 10, 107–129. doi: 10.1177/1541204011424735

Jiménez-Iglesias, A., Moreno, C., Ramos, P., and Rivera, F. (2015). What family dimensions are important for health-related quality of life in adolescence? *J. Youth Stud.* 18, 53–67. doi: 10.1080/13676261.2014.933191

Kalavana, T. V., Maes, S., and De Gucht, V. (2010). Interpersonal and self-regulation determinants of healthy and unhealthy eating behavior in adolescents. *J. Health Psychol.* 15, 44–52. doi: 10.1177/1359105309345168

Karademas, E. C., Peppa, N., Fotiou, A., and Kokkevi, A. (2008). Family, school and health in children and adolescents: findings from the 2006 HBSC Study in Greece. *J. Health Psychol.* 13, 1012–1020. doi: 10.1177/1359105308097965

Kashdan, T. B., Gallagher, M. W., Silvia, P. J., Winterstein, B. P., Breen, W. E., Terhar, D., et al. (2009). The curiosity and exploration inventory-II. development, factor structure, and psychometrics. *J. Res. Pers.* 43, 987–998. doi: 10.1016/j.jrp.2009.04.011

Kashdan, T. B., and Roberts, J. E. (2004).Trait and state curiosity in the genesis of intimacy: differentiation from related constructs. *J. Soc. Clin. Psychol.* 23, 792–816. doi: 10.1521/jscp.23.6.792.54800

Kashdan, T. B., Rose, P., and Fincham, F. D. (2004). Curiosity and exploration: facilitating positive subjective experiences and personal growth opportunities. *J. Pers. Assess.* 82, 291–305. doi: 10.1207/s15327752jpa8203_05

Khan, L., Parsonage, M., and Stubbs, J. (2015). *Investing in Children's Mental Health. A Review of Evidence on the Costs and Benefits of Increased Service Provision*. London: Center for Mental Health.

Kim, B.K.E., Oesterle, S., Catalano, R.F., and Hawkins, J.D. (2015).Change in protective factors across adolescent development. *J. Appl. Dev. Psychol.* 40, 26–37. doi: 10.1016/j.appdev.2015.04.006

Klimidis, S., Minas, I. H., and Ata, A. W. (1992). The PBI-BC: A brief current form of the parental bonding instrument for adolescent research. *Compr Psychiatry.* 33, 374–377. doi: 10.1016/0010-440X(92)90058-X

Knauss, C., Paxton, S. J., and Alsaker, F. D. (2007). Relationships amongst body dissatisfaction, internalisation of the media body ideal and perceived pressure from media in adolescent girls and boys. *Body Image* 4, 353–360. doi: 10.1016/j.bodyim.2007.06.007

Lansford, J. E., Criss, M. M., Pettit, G. S., Dodge, K. A., and Bates, J. E. (2003). Friendship quality, peer group affiliation, and peer antisocial behavior as moderators of the link between negative parenting and adolescent externalizing behavior. *J. Res. Adolesc.* 13, 161–184. doi: 10.1111/1532-7795.1302002

Lengua, L. J. (2002). The contribution of emotionality and self-regulation to the understanding of children's response to multiple risk. *Child Dev.* 73, 144–161. doi: 10.1111/1467-8624.00397

Lindström, B., and Eriksson, M. (2010).*The Hitchhiker's Guide to Salutogenesis*. Helsinki: Folkhälsan Research Center Health Promotion Research.

Luthar, S. S., and Cicchetti, D. (2000).The construct of resilience: Implications for interventions and social policies. *Dev. Psychopathol.* 12, 857–885. doi: 10.1017/S0954579400004156

Luthar, S. S., Cicchetti, D., and Becker, B. (2000). The construct of resilience: a critical evaluation and guidelines for future work. *Child Dev.* 71, 543–562. doi: 10.1111/1467-624.00164

Luthar, S. S., Crossman, E. J., and Small, P. J. (2015). "Resilience and adversity," in *Handbook of Child Psychology and Developmental Science, Vol. III, 7th Edn.,* eds R. M. Lerner and M. E. Lamb (New York, NY: Wiley), 247–286.

Luthar, S. S., and Cushing, G. (1999). "Measurement issues in the empirical study of resilience," in *Resilience and Development. Positive Life Adaptation,*eds M. D. Glantz and J. L. Johnson (New York, NY: Kluwer Academic Publishers), 129–160.

Luthar, S. S., Doernberger, C. H., and Zigler, E. (1993). Resilience is not a unidimensional construct: insights from a prospective study of inner-city adolescents. *Dev. Psychopathol.* 5, 703–717.

MacDermott, S. T., Gullone, E., Allen, J. S., King, N. J., and Tonge, B. (2010). The Emotion Regulation Index for Children and Adolescents (ERICA). A Psychometric Investigation. *J. Psychopathol. Behav. Assess.* 32, 301–314. doi: 10.1007/s10862-009-9154-0

Marmot, M. (2010). "Fair society, healthy lifes,"in *The Marmot Review (Executive Summary). Strategic Review of Health Inequalities in England Post-2010*. (London: Institute of Health Equity). Available online at: http://www.instituteofhealthequity.org/projects/fair-society-healthy-lives-the-marmot-review

Masten, A. S. (2001). Ordinary magic. Resilience processes in development. *Am. Psychol.* 56, 227–238. doi: 10.1037//0003-066X.56.3.227

Masten, A. S. (2014). Global perspectives on resilience in children and youth. *Child Dev.* 85, 6–20. doi: 10.1111/cdev.12205

Masten, A. S., and Coatsworth, J. D. (1998).The development of competence in favorable and unfavorable environments. Lessons from research on successful children. *Am. Psychol.* 53, 205–220.

Masten, A. S., Hubbard, J. J., Gest, S. D., Tellegen, A., Garmezy, N., and Ramírez, M. (1999). Competence in the context of adversity: pathways to resilience and maladaptation from childhood to late adolescence. *Dev. Psychopathol.* 11, 143–169.

Masten, A. S., and Narayan, A. J. (2012).Child development in the context of disaster, war, and terrorism: pathways of risk and resilience. *Annu. Rev. Psychol.* 63, 227–257. doi: 10.1146/annurev-psych-120710-100356

Masten, A. S., and Reed, M. G. J. (2005). "Resilience in development,"in *Handbook of Positive Psychology,* eds C. R. Snyder and S. J. López (New York, NY: Oxford University Press), 74–88.

Masten, A.S., and Shaffer, A. (2006). "How families matter in child development: reflections from reasearch on risk and resilience," in *Families Count: Effects on Child and Adolescent Development,* eds A. Clarke-Stewart and J. Dunn (Cambridge: Cambridge University Press), 5–25.

McVie, S. (2014).The impact of bullying perpetration and victimization on later violence and psychological distress: a study of resilience among a Scottish youth cohort. *J. Sch Violence* 13, 39–58. doi: 10.1080/15388220.2013.841586

Mond, J., van den Berg, P., Boutelle, K., Hannan, P., and Neumark-Sztainer, D. (2011). Obesity, body dissatisfaction and emotional well-being in early and late adolescence: findings from the Project EAT study. *J. Adolesc. Health* 48, 373–378. doi: 10.1016/j.jadohealth.2010.07.022

Moreno, C., Ramos, P., Rivera, F., Jiménez-Iglesias, A., García-Moya, I., Sánchez-Queija, I., et al. (2016). *Informe Técnico de los Resultados Obtenidos por el Estudio Health Behaviour in School-aged Children (HBSC) 2014 en España*. Madrid: Ministerio de Sanidad, Servicios Sociales e Igualdad.

Morgan, A., Davies, M., and Ziglio, E. (2010). *Health Assets in a Global Context.* London: Springer.

Morgan, A., and Hernán, M. (2013). Promoting health and wellbeing through the asset model. *Rev. Esp. Sanid. Penit.* 15, 78–86. doi: 10.4321/S1575-06202013000300001

Olsson, C. A., Bond, L., Burns, J. M., Vella-Brodick, D. A., and Sawyer, S. M. (2003). Adolescence resilience: a concept analysis. *J. Adolesc.* 26, 1–11. doi: 10.1016/S0140-1971(02)00118-5

Olweus, D. (1996). *The Revised Olweus Bully/Victim Questionnaire*. Bergen: Mimeo, Research Center for Health Promotion (HEMILCenter), University of Bergen.

Orbach, I., and Mikulincer, M. (1998). Body investment scale: construction and validation of a body experience scale. *Psychol. Assess.* 10, 415–425.

Pate, R. R., Heath, G. W., Dowda, M., and Trost, S. G. (1996). Associations between physical activity and other health behaviors in a representative sample of US adolescents. *Am. J. Public Health* 86, 1577–1581. doi: 10.2105/AJPH.86.11.1577

Preacher, K. J., Rucker, D. D., MacCallum, R. C., and Nicewander, W. A. (2005). Use of the extreme groups approach: a critical reexamination and new recommendations. *Psychol. Methods* 10, 178–192. doi: 10.1037/1082-989X.10.2.178

Prochaska, J. J., Sallis, J. F., and Long, B. (2001). A physical activity screening measure for use with adolescents in primary care. *Arch. Pediatr. Adolesc. Med.* 155, 554–559. doi: 10.1001/archpedi.155.5.554

Ramos, P. (2010). *Lifestyles and Health in Adolescence.* Doctoral thesis, University of Seville. Available online at: http://www.injuve.es/observatorio/tesis-doctorales/estilos-de-vida-y-salud-en-la-adolescencia

Ramos, P., Moreno, C., Rivera, F., de Matos, M. G., and Morgan, A. (2012). Analysis of social inequalities in health through an integrated measure of perceived and experienced health in Spanish and Portuguese adolescents. *J. Health Psychol.* 17, 57–67. doi: 10.1177/1359105311406154

Ramos, P., Moreno, C., Rivera, F., and Pérez, P. J. (2010). Integrated analysis of the health and social inequalities of Spanish adolescents. *Int. J. Clin. Health Psychol.* 10, 477–498.

Ravens-Sieberer, U., Erhart, M., Torsheim, T., Hetland, J., Freeman, J., Danielson, M., et al. (2008). An international scoring system for selfreported health complaints in adolescents. *Eur. J. Public Health* 18, 294–299. doi: 10.1093/eurpub/ckn001

Ravens-Sieberer, U., Gosch, A., Abel, T., Auquier, P., Bellach, B. M., Bruil, J., et al. Group (2001). Quality of life in children and adolescents: a European public health perspective. *Soc. Prev. Med.* 46, 297–302. doi: 10.1007/BF01321080

Roberts, C., Freeman, J., Samdal, O., Schnohr, C., de Looze, M. E., NicGabhainn, S., et al. (2009). The Health Behaviour in School-aged Children (HBSC) study: methodological developments and current tensions. *Int. J. Public Health.* 54, 140–150. doi: 10.1007/s00038-009-5405-9

Rodrigo, M. J., Almeida, A., Spiel, C., and Koops, W. (2012). Introduction: evidence-based parent education programmes to promote positive parenting. *Eur. J. Dev. Psychol.* 9, 2–10. doi: 10.1080/17405629.2011.631282

Roosa, M. W. (2000). Some thought about resilience versus positive development, main effects versus interactions, and the value of resilience. *Child Dev.* 71, 567–569. doi: 10.1111/1467-8624.00166

Rubin, K. H., Dwyer, K. M., Booth-LaForce, C. L., Kim, A. H., Burgess, K. B., and Rose-Krasnor, L. (2004). Attachment, friendship, and psychosocial functioning in early adolescence. *J. Early Adolesc.* 24, 326–356. doi: 10.1177/0272431604268530

Rutter, M. (2006). "The promotion of resilience in the face of adversity," in *Families Count: Effects on Child and Adolescent Development,* eds A. Clarke-Stewart and J. Dunn (Cambridge: Cambridge University Press), 26–52.

Sameroff, A. J. (2010). A unified theory of development: a dialectic integration of nature and nurture. *Child Dev.* 81, 6–22. doi: 10.1111/j.1467-8624.2009.01378.x

Sameroff, A. J., and Fiese, B. H. (2000). "Transactional regulation: The developmental ecology of early intervention," in *Handbook of Early Childhood Intervention*, eds J. P. Shonkoff, and S. J. Meisels (New York, NY: Cambridge University Press), 135–159.

Sandín-Esteban, M.-P., and Sánchez-Martín, A. (2015). Resiliencia y éxito escolar en jóvenes inmigrantes. *Infanc. Aprendiz.* 38, 175–211. doi: 10.1080/02103702.2015.1009232

Silventoinen, K., Posthuma, D., Lahelma, E., Rose, R. J., and Kaprio, J. (2007). Genetic and environmental factors affecting self-rated health from age 16-25: a longitudinal study of Finnish twins. *Behav. Genet.* 37, 326–333. doi: 10.1007/s10519-006-9096-1

Silverman, M. N., and Deuster, P. A. (2014). Biological mechanisms underlying the role of physical fitness in health and resilience. *Interface Focus.* 4:20140040. doi: 10.1098/rsfs.2014.0040

Springer, J. F., Sale, E., Herman, J., Sambrano, S., Kasim, R., and Nistler, M. (2004).Characteristics of effective substance abuse prevention programs for high-risk youth. *J. Prim. Prev.* 25, 171–194. doi: 10.1023/B:JOPP.0000042388.63695.3f

Steinberg, L., and Silk, J. S. (2002). "Parenting adolescents," in *Handbook of Parenting*, *Vol. 1, Children and Parenting*, ed M. H. Bornstein (Mahwah, NJ: Lawrence Erlbaum Associates), 103–134.

Tiêt, Q. Q., and Huizinga, D. (2002).Dimensions of the construct of resilience and adaptation among inner-city youth. *J. Adolesc. Res.* 17, 260–276. doi: 10.1177/0743558402173003

Tiggemann, M. (2005). Body dissatisfaction and adolescent self-esteem: prospective findings. *Body Image* 2, 129–135. doi: 10.1016/j.bodyim.2005.03.006

Torsheim, T., Wold, B., and Samdal, O. (2000). The teacher and classmate support scale: Factor structure, test-retest reliability and validity in samples of 13- and 15-year old adolescents. *Sch. Psychol. Int.* 21, 195–212. doi: 10.1177/0143034300212006

Werner, E. E., and Smith, R. S. (1982).*Vulnerable But Invincible: A Study of Resilient Children*. New York, NY: McGraw-Hill.

WHO (1948). *WHO Constitution*. Geneva: World Health Organization.

WHO (2012). *Health 2020: The European Policy for Health and Wellbeing*. Copenhagen: WHO Regional Office for Europe.

WHO (2014). *Investing in Children: The European Child and Adolescent Strategy 2015-2020*. Copenhagen: WHO Regional Office for Europe.

Zimet, G. D., Dahlem, N. W., Zimet, S. G., and Farley, G. K. (1988). The multidimensional scale of perceived social support. *J. Pers. Assess.* 52, 30–41.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REVIEW
published: 19 February 2016
doi: 10.3389/fpsyg.2016.00206

Animal Models of Maladaptive Traits: Disorders in Sensorimotor Gating and Attentional Quantifiable Responses as Possible Endophenotypes

*Juan P. Vargas, Estrella Díaz, Manuel Portavella and Juan C. López**

Animal Behavior and Neuroscience Lab, Department of Experimental Psychology, Universidad de Sevilla, Seville, Spain

OPEN ACCESS

Edited by:
Pietro Cipresso,
IRCCS Istituto Auxologico Italiano,
Italy

Reviewed by:
Eddy A. Van Der Zee,
University of Groningen, Netherlands
Kenn Konstabel,
National Institute for Health
Development, Estonia

****Correspondence:***
Juan C. López
jclopez@us.es

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: *07 September 2015*
Accepted: *03 February 2016*
Published: *19 February 2016*

Citation:
Vargas JP, Díaz E, Portavella M and López JC (2016) Animal Models of Maladaptive Traits: Disorders in Sensorimotor Gating and Attentional Quantifiable Responses as Possible Endophenotypes. Front. Psychol. 7:206. doi: 10.3389/fpsyg.2016.00206

Traditional diagnostic scales are based on a number of symptoms to evaluate and classify mental diseases. In many cases, this process becomes subjective, since the patient must calibrate the magnitude of his/her symptoms and therefore the severity of his/her disorder. A completely different approach is based on the study of the more vulnerable traits of cognitive disorders. In this regard, animal models of mental illness could be a useful tool to characterize indicators of possible cognitive dysfunctions in humans. Specifically, several cognitive disorders such as schizophrenia involve a dysfunction in the mesocorticolimbic dopaminergic system during development. These variations in dopamine levels or dopamine receptor sensibility correlate with many behavioral disturbances. These behaviors may be included in a specific phenotype and may be analyzed under controlled conditions in the laboratory. The present study provides an introductory overview of different quantitative traits that could be used as a possible risk indicator for different mental disorders, helping to define a specific endophenotype. Specifically, we examine different experimental procedures to measure impaired response in attention linked to sensorimotor gating as a possible personality trait involved in maladaptive behaviors.

Keywords: dopamine, endophenotype, latent inhibition, mental disorder, prepulse inhibition

INTRODUCTION

The criteria used by current diagnostic scales are based on the analysis of external symptoms of the patient. Disorders such as attention deficit with hyperactivity or mental disorders such as schizophrenia are diagnosed based on symptoms that, in many cases, require the patient to evaluate their intensity. This situation creates a serious problem for the diagnosis, given the large amount of subjective information handled by the psychologist or the psychiatrist (Robbins et al., 2012).

The problem of subjectivity and comorbidity in diagnostic errors are, in part, a consequence of the absence of biological markers to facilitate proper classification of the disorder. With relative ease, the diagnostic manuals such as the DSM or ICD propose a continuous change in the criteria for inclusion or exclusion of a disorder due largely to the heterogeneity and complexity of symptoms that define that disorder. These are so complex that patients with different symptoms might have the same diagnosis, a fact that significantly increases the difficulty of providing proper treatment. This high comorbidity between various diseases indicates a clear deficiency in the classification system of mental disorders, preventing the identification of valid pathologies (Hyman, 2010). It is possible that the psychotherapeutic and pharmacological failures are largely due to this fact. Note for example that the therapeutic effectiveness of pharmacological treatments reaches approximately 50% (Wong et al., 2010).

Using a diverse group of pharmacological treatments to relieve disorders such as depression is also an indicator of the disparity of its diagnosis. For example, the use of inhibitors of serotonin reuptake is applied for a specific type of depressive symptoms, which differs from those used under MAO inhibitors or under tricyclics. The differential response of each patient to treatment indicates that disorders included in the same category should be treated with different principles. Alternatively, this phenomenon could be indirectly indicating that different types of disorders within a category may have a different biological basis.

An alternative to this traditional view is the characterisation of endophenotypes. An endophenotype is a quantitative measurable trait associated with a genetic predisposition (Gottesman and Shields, 1972, 1973). In contrast to the symptomatic view of psychopathology, the endophenotype analyses the characteristics that show possible brain vulnerability to suffer a specific type of disorder. The objective is the study and quantification of specific features that reflect a mental disorder associated with a biochemical sign (Hasler et al., 2006; Turetsky et al., 2007). Throughout its long history, the functional study of behavior in the laboratory has provided a number of indicators that could serve as markers for selective expression of the maladaptive behaviors. Applying this model to the field of psychopathology, mental disorders could be considered as extremes at one or both tails of these normal distributions (Miller and Rockstroh, 2013). From this point of view, psychopathology would view disorders as dimensional notions, and not as categories under a binary diagnosis (Hyman, 2010; Frances and Widiger, 2012; Morris and Cuthbert, 2012).

Here, we provide a set of measurable procedures sensitive enough to be used to identify possible endophenotypes developed from animal models. These endophenotypes are based on the correlation between brain processes and measurable responses of a subject that enable us to discriminate between different sets of symptoms, and facilitate new specific therapies. In addition, the evaluation of these traits could facilitate a more objective classification system of psychopathologies.

HOW DOES THE USE OF AN ANIMAL MODEL CONTRIBUTE TO PSYCHOPATHOLOGY CLASSIFICATION?

The recent developments in genetics and epigenetics allow us to better approach understanding behavior and facilitate the understanding of mental disorders. The fact that some behaviors have a Mendelian basis, suggests the possibility of finding simple mutations that affect behavior in a relatively specific manner. However, there are only a small group of features known as Mendelian traits (or traits 1:1) in relation to genotype. Mental disorders such as depression or schizophrenia are clearly polygenic, or may also be generated by various mutant alleles of the same gene and specific environmental conditions, making the analysis of their causes a complex procedure (Zahn-Waxler et al., 1988; Winokur and Kadrmas, 1989; Kidd, 1997; Moldin, 1997; Owen, 2000; Torrey and Yolken, 2000; Goldman, 2012). Moreover, these illnesses are the result of the interactions of both genetic and epigenetic factors. And although we now have suitable tools for genotype analysis, the fact that these etiological factors -genes and environment- interact to produce similar phenotypes, significantly increases the difficulty to precisely define the specific weight of each one in the generation of behavior (Plomin and Rende, 1991). Identifying what groups of genes may contribute to the expression of a disorder is a long process of molecular genetics. However, the identification of relating groups of genes with specific traits is currently a more achievable goal.

The use of animal models for the study of personality traits, vulnerability to certain disorders or substance abuse dependence is an interesting strategy for developing behavioral protocols in the laboratory. Although in some cases these models could show poor face and predictive validity, the construct validity associated with the etiology or mechanism of the underlying disorder is usually high (O'Donnell, 2011). For example, animal models of schizophrenia have been successful in evaluating risk factors (see **Table 1**). This fact is crucial in order to develop new pharmacological treatments or genetic therapies. However, the reduced face validity is often a problem when applying to human models.

The development of endophenotypes is one alternative to try to improve this model. Taking advantage of high construct validity, we can develop sensitive tests for quantifying specific traits. Measures such as latent inhibition (LI) or prepulse inhibition (PPI) are, among others, easily quantifiable under controlled conditions in the laboratory. In addition, we can use the advantage of these procedures in a similar way in both animals and humans, and the results are easily extrapolated from an animal model to a human model (Le Pen et al., 2011). While PPI is a very simple procedure seeking to analyze early attentional gating mechanisms, the LI is a learning process related to selective attention and habituation to irrelevant information (Lubow and Gewirtz, 1995; Swerdlow et al., 1996; Braff and Swerdlow, 1997). Animal models indicate that problems in the expression of PPI or LI correlate with cognitive deficits such as working memory or alternation behavior, locomotion activity such as hyperactivity induced by a dopamine receptor agonist, and some negative symptoms also described in pathologies such as schizophrenia (Flagstad et al., 2004; Le Pen et al., 2006; Moore et al., 2006; Hazane et al., 2009). For example, patients with schizophrenia show these symptoms associated with a dysfunctional prefrontal cortex (PfC; Manoach, 2003; Silver et al.,

TABLE 1 | Several animal models have studied schizophrenia.

		MAM	NVHL
Executive functions	Attentional processes	Flagstad et al., 2005; Featherstone et al., 2007	
	Working memory deficits	Flagstad et al., 2005; Hazane et al., 2009	Chambers et al., 1996; Lipska et al., 2002
	Perseveration	Moore et al., 2006; Hazane et al., 2009	Marquis et al., 2008
	Recognition deficits	Featherstone et al., 2007	Sams-Dodd et al., 1997; Bachevalier et al., 1999
Motivational behavior	Increased liability for addictive behaviors	Flagstad et al., 2005	Swerdlow et al., 2001; Brady et al., 2008
	Responses to stress	Le Pen et al., 2006; Hazane et al., 2009	Sams-Dodd et al., 1997
Activity	Hyperlocomotion	Le Pen et al., 2006; Moore et al., 2006; Penschuck et al., 2006; Hazane et al., 2009	Lipska et al., 1993; Wan et al., 1996
Information filtering mechanism	Sensorimotor gating deficits	Le Pen et al., 2006; Moore et al., 2006; Hazane et al., 2009	Swerdlow et al., 1995

Pharmacological models have used amphetamine, PCP or NMDA to simulate some of the symptoms. However, only two models have shown the illness as a developmental process. The neonatal ventral hippocampal lesion (NVHL) and MAM models showed marked maladaptive behavior when animals reached adulthood. Le Pen et al. (2011) and O'Donnell (2011) have described several behavioral procedures where we can find similar results with different techniques aimed at developing a dysfunctional PfC.

2003; Godsil et al., 2013), therefore, a behavioral test aimed to evaluate PfC function is a useful tool for an accurate differential diagnostic. The knowledge acquired in recent years on the use of a quantifiable measurement of these traits is boosting the development of unified models of diagnosis that include data from all levels, that is: genetics, biochemical, and behavioral levels.

But the question is, how can we contribute to this proposal? Consider, for instance, one of the most complex disorders, schizophrenia. Currently, schizophrenia is an umbrella term for a diverse group of disorders with possibly different etiologies. Focusing on PfC dysfunction, animal research has provided explanatory models to understand the possible development of this mental disturbance (O'Donnell, 2011; Godsil et al., 2013). Procedures aimed to alter gestation and fetal development such as the MAM model (Methylazoxymethanol), or techniques affecting the maturation process of PfC such as ventral hippocampus lesion in neonates, allow us to experimentally analyze this disorder (Waddington et al., 1999; Bramon et al., 2005; Chambers and Lipska, 2011). Both procedures show a clear PfC dysfunction (Tseng and O'Donnell, 2004, 2007). Cells unit recording studies indicate the possibility of a deficit in inhibitory GABAergic cells. This could be the cause of an excessive release of dopamine cells in the mesocortical system (Tseng and O'Donnell, 2004, 2007; O'Donnell, 2011; Godsil et al., 2013). This could be the reason that PPI or LI could be affected in these animal models and in schizophrenia. Both behavioral processes require an operative PfC for a normal expression. PPI and LI are very sensitive to disturbances in this structure. In this regard, a deficit in one or both processes could be a risk factor. On the whole, the characteristics of a dysfunctional PfC and the impairment in LI and PPI expression could be signs of a specific type of mental disorder, apart from the current model of mental illness where the disorder and its severity are expressed in terms of a scale filled out by the patient or a close family member.

DOPAMINERGIC SYSTEM AS A SIGN OF A POSSIBLE RISK FACTOR

The function of the dopamine neurotransmitter has attracted great interest because of its relationship with the processes of learning and with several mental disorders such as schizophrenia, depression, ADHT or addiction to a substance of abuse (Robbins, 1992; Feldman et al., 1997; Weiner, 2003; Grace and Sesack, 2010; Simpson et al., 2010; Wise, 2010; Milad and Rauch, 2012; Díaz et al., 2015). The distribution of dopaminergic neurons is abundant in the central nervous system. The midbrain neurons and their efferences to the ventral striatum and PfC play a special role in the learning process (Robbins and Everitt, 1996). Dopaminergic pathways of the ventral tegmental area (VTA) toward the nucleus accumbens (NAc) are closely linked to the motivational processes of learning (Berridge and Robinson, 1998; Berridge, 2007). Many stimulant drugs, such as cocaine or amphetamine, operate in this place, and their function significantly increases the release or reduces the reuptake of dopamine in the system.

Dopamine receptors belong to the G-protein coupled receptors family. All these receptors possess seven transmembrane domains and five subtypes of dopamine receptors according to their molecular characteristics. These have been grouped into two pharmacological families according to the effect produced by agonists and antagonists. D1 family includes the subtypes D1 and D5 receptors. Both stimulate adenylyl cyclase, producing cAMP. On the other hand, D2 receptor family includes the subtypes D2, D3, and D4. These receptors inhibit the formation of cAMP. The D1 receptor is the most abundant in the central nervous system (Missale et al., 1998). The greatest concentration of this receptor is found in the neostriatum, NAc, amygdala, and substantia nigra. However, its affinity for dopamine is relatively low. The D2 receptor is found in high concentrations in the neostriatum (GABAergic neurons) in the NAc and hippocampus, and with a moderate density in the substantia nigra, cerebral cortex, globus pallidus, thalamus, and hypothalamus. These data make D1 and D2 receptors specific targets for the study of cognitive, emotional, and motivational disorders. Electrophysiological studies have made important contributions concerning their functional activity in the mesolimbic system (O'Donnell and Grace, 1998; Moore et al., 1999; Grace, 2000; Floresco et al., 2001; Goto and Grace, 2008). These studies are of great relevance given the importance of these receptors in the processes of associative learning. Recent studies have shown that the D2 receptor is located in the projections of both the PfC and the amygdala in the form of autoreceptors (O'Donnell and Grace, 1995; Groenewegen et al., 1999; Goto and Grace, 2008). Specifically, D2 receptors are located in the presynaptic areas with the function of modulating the dopaminergic activity of the VTA over NAc through excitatory projections. That is why this receptor has been linked to the goal directed processes or controlled processes that require high attentional activity (O'Donnell and Grace, 1995; Goto and Grace, 2008). In contrast, the activity of D1 receptors in the mesolimbic system is different than the one described for D2. These are located in the post-synaptic cells of the NAc that receive glutamatergic afferences from the hippocampus and dopaminergic afferences from the VTA (O'Donnell and Grace, 1995; Groenewegen et al., 1999; Goto and Grace, 2008).

Disturbances in this system increase the risk of developing serious mental illness (O'Donnell, 2011; Godsil et al., 2013). Disorders such as schizophrenia have been linked directly to disturbances during brain development associated with the second trimester of pregnancy (Waddington et al., 1999; Bramon et al., 2005). Changes in the dopaminergic sensitivity and in the levels of dopamine or dopamine receptors volume could be the result of this process. Specifically, the family of D2 receptors seems to be more related to the disease process (Grace and Sesack, 2010; Simpson et al., 2010; Wise, 2010; Milad and Rauch, 2012), since the antagonists of these receptors such as haloperidol are effective in reducing symptoms (Lubow and Weiner, 2010). This is the reason why it was suggested a substantial increase of this type of receptors underlies this disorder as shown, for instance, in the studies of post-mortem tissue (Seeman and Nizkik, 1990).

Currently the drug treatment of disorders such as schizophrenia or ADHT act directly on the dopaminergic modulation in the brain. The changes that cause the blockade or stimulation of receptors of this neurotransmitter can be studied in the laboratory. Changes in the sensorimotor gating or selection of relevant stimulus of the environment can be considered as possible quantifiable traits directly related with the level of dopamine or dopaminergic receptors in the mesocortical and mesolimbic system. It is important to emphasize that injuries to PfC produce dopamine dysregulation and deficits in PPI and LI expression.

PPI OF STARTLE RESPONSE. A SENSORIMOTOR GATING MEASURE

The startle response to an intense stimulus is a reflex behavior that has been described in all mammals studied. This is a fast-twitch of the skeletal muscle that leads to processing environmental stimuli and guiding the attention of the subject to a possible threat. This type of response is interesting because it has been associated with specific genes that appear in schizophrenia and as a possible trait with endophenotypical characteristics. For example, Vaidyanathan et al. (2014) studied the startle blink reflex using a very large human sample. Analyzing the startle response, they found a heritable specific pattern of behavior in the sample. In addition, this trait was associated with candidate genes in the endophenotype of schizophrenia. However, although it is an automatic reaction, the outcome can be modulated by the previous presence of a stimulus of lower intensity, therefore PPI is defined as the attenuation of the startle response to an intense pulse when it is preceded by a lower-intensity prepulse stimulus. When the prepulse is perceived, the mechanism of startle is inhibited and the animal displays a lower response (Graham, 1975; Lüthy et al., 2003; Larrauri and Schmajuk, 2006).

The problems with sensorimotor gating have been linked with the levels of dopamine in the NAc. The NAc integrates information from different structures, and even though dopamine modulation in NAc is dependent on mesocortical and mesolimbic systems (Ellenbroek et al., 1996; Larrauri and Schmajuk, 2006), the selective modulation of PfC afferent transmission is especially relevant. PfC afferences could facilitate behaviors oriented to specific goals, and a dopamine deficit could be involved in the incapacity to control the behavior (Goto and Grace, 2008).

It should be noted that the dopaminergic innervation of the PfC increases progressively through adolescence until adulthood. In this period, we can find modifications in density, shape and organization of the circuits (Kalsbeek et al., 1988; Benes et al., 2000; Seamans and Yang, 2004; Segalowitz and Davies, 2004; Manitt et al., 2011; Naneix et al., 2012). A mature circuit allows the dopaminergic neurons to fit their responses in an adaptive way, modulating their response in correlation with environmental changes (Spear, 2000; Tseng and O'Donnell, 2004, 2007; O'Donnell, 2011; Cass et al., 2013; Godsil et al., 2013). Currently, it is estimated that delays or alterations in the maturation process of the PfC dopamine system could be the cause of a large number of mental disorders (O'Donnell, 2011; Godsil et al., 2013). Specifically, a poor inhibitory capacity of the PfC over the NAc may be the major etiological factor in severe disorders such as schizophrenia. In fact, a deficit in the response to the pulse has been observed in different types of cognitive disorders, and it is specifically relevant in patients with schizophrenia (Braff et al., 1992, 2001a,b). Thus, a reduced PPI could be used as a trait for attentional deficit, besides being included as a schizotypy personality trait or a possible endophenotype of schizophrenia (Cadenhead et al., 2000; Braff, 2010; O'Donnell, 2011).

However, this trait is not specific for patients with schizophrenia but indicates a trait of vulnerability, and it is very clear in patients with schizophrenia. In this regard, the PPI deficit could be a necessary condition as a risk factor of schizophrenia, but it could not be sufficient by itself. The PPI deficit might be found in several disorders, and a pathological process such as schizophrenia needs other indicators.

PPI, DOPAMINE AND IMPULSIVITY: A TRAIT, A NEUROTRANSMITTER AND A QUANTIFIABLE MEASURE NOT ASSOCIATED EXCLUSIVELY WITH SCHIZOPHRENIA

PPI is an easy system to measure in animals including humans. It has been used in animal models of schizophrenia, even though there are several studies where this procedure has been correlated with impulsivity traits. López et al. (2015) analyzed the PPI in rats classified as impulsive by an autoshaping procedure. Animals designated as sign trackers showed approach behavior to a conditional stimulus before delivery of unconditional stimulus. Specifically, for sign tracker animals (STa) the conditional stimulus could be a surrogate of the unconditional stimulus (Flagel et al., 2007; Robinson and Flagel, 2009). These kind of animals showed high levels of dopamine in NAc, but only in the presence of a conditional stimulus (Flagel et al., 2011). These data were consistent with the results of López et al. (2015) using a PPI procedure. In fact, the STa showed a lower PPI response to stimuli of low intensity. This reduced inhibitory ability of the STa showed a difference in the behavioral pattern in normal animals. Furthermore, these data may indicate that ST subjects may be more vulnerable to cognitive disorders in which dopamine is involved.

An important question about the vulnerability of STa to an impulsive behavior comes from specific activity of D2 subtype dopamine receptor. This receptor is located presynaptically on PfC terminals, and has been related with a selective modulation of the NAc to facilitate goal-directed behaviors (Goto and Grace, 2008). In addition, several psychopathologies associated with PfC have shown a deficit between this structure and the projections to NAc (O'Donnell, 2011). López et al. (2015) found a possible vulnerability from STa, since these animals showed a large sensibility of D2 receptor to the administration of an agonist such as quinpirole. This drug affected only STa performance, indicating that this type of trait differs from that observed in schizophrenia. It would be appropriate at this stage to point out the difference between an animal model of impulsivity and an animal model of schizophrenia regarding a dysfunction in PfC. These models have developed several protocols to evaluate attentional processes, and LI is a perfect candidate to discriminate between impulsivity and schizophrenia, because it allows for evaluating attention and executive functions, both specific to PfC function. Impulsive models of animals have found differences in incentive salience of the conditional stimulus, but not in attentional problems (Berridge and Robinson, 1998; Berridge, 2007) such as in schizophrenia models.

LI, DOPAMINE AND ATTENTIONAL DEFICITS

LI is a learning process observed when the acquisition of a conditional response to a conditioned stimulus paired with a reinforcer is retarded if the same stimulus has previously been pre-exposed in the absence of the reinforcer. LI pharmacology has been associated almost exclusively with the use of an animal model of schizophrenia, and is therefore largely consistent with the pharmacology of schizophrenia (Lubow and Kaplan, 2010; Lubow and Weiner, 2010; Díaz et al., 2015). Specifically, because some of the symptoms of schizophrenia are characterized by an inability to filter, or ignore irrelevant or unimportant stimuli, an anomalous LI was proposed as a tool for the study of possible deficits of attention (Lubow and Weiner, 2010).

Again, dopaminergic activity of the NAc is the essential neural substrate for its expression. Animal models have shown that the primary role of the NAc is to restrict the expression of LI under certain conditions, and thus ensure that the LI is flexible and sensitive to environmental demands. It is important to highlight that, in the absence of modulator mechanisms responsible for restricting the expression of LI to the specific conditions, the effects of an irrelevant stimulus would be extremely robust and maladaptive. In this regard, LI might reflect the psychological processes that are impaired in schizophrenia, since most of the patients showed a reduced expression of this phenomenon. The identification of brain regions whose damage leads to disrupt the LI, joined with the studies of different parameters of expression in animal models, can provide important information on the dysfunctional brain circuits in schizophrenia. In previous decades it was suggested that some kind of hyperactivity of the dopaminergic systems represent a primary biochemical alteration in schizophrenia, which apparently constituted at least a plausible justification for biochemical alteration in this disorder (Iversen, 1976).

To gain insight into quantifiable attentional processes in LI, Díaz et al. (2014) analyzed the effect of various types of pre-exposure to a stimulus. The results indicated that there is a transfer from the ventral to the dorsal striatum in the processing of environmental information. In addition, the dorsomedial striatum is key to encode stimuli when these become irrelevant due to the lack of consequences after their presentation. A deficit in PfC could be the cause of a loss of transfer from ventral to dorsal striatum. Currently there are some laboratories working on this possibility. The inability to modulate dopamine in NAc does not allow for attentional disengagement, showing a persistent state of continuous attention.

The inability of encoding irrelevant information is one of the clearest deficits observed in patients with schizophrenia. Many modern learning theories assume that the amount of attention to a signal depends on how well the signal predicts the significant event of the past. Schizophrenia is associated with attention deficit and recent theories of psychosis have argued that positive symptoms such as delusions and hallucinations are related to a lack of selective attention. Patients with schizophrenia, who had severe positive symptoms, showed a clear difficulty in discriminating between predictive and non-predictive cues when compared to healthy adults. In addition, the rate of learning about non-predictive signals correlated with more severe positive symptoms in schizophrenia. These results suggest that the positive symptoms of schizophrenia were associated with increased attention, both to signals that are likely to be predictive and to those that are not predictive for causal learning. This selective attention deficit was the result of learning irrelevant causal associations (Morris et al., 2013). In this regard, the development of specific protocols to differentiate the expression of LI could be used as a possible risk factor in the population.

However, the complexity of this disorder suggests the possibility of different etiological factors may underlie the disease. At present there are many contradictory results regarding whether IL is affected in schizophrenia. Lubow and Kaplan (2010) addressed this issue in a recent review. They emphasize the difference between positive and negative symptoms in relation to the expression of IL. For instance, patients with high levels of negative symptoms and low of positive showed a potentiated LI. This data is relevant, because they could be observing different symptoms of the illness or different illness.

CONCLUDING REMARKS

The mesocortical input of dopamine and the PfC play a critical role in normal cognitive processes and in several neuropsychiatric diseases. This dopamine input regulates aspects of working memory, planning and attention, among others. Similarly, some disturbances may be the basis for a variety of positive and negative symptoms, and therefore of many of the cognitive deficits associated with mental illness. Despite intensive research, we still have a lack of understanding of the basic principles of dopamine activity in the PfC and all the mesolimbic system. In recent years, there has been considerable effort to understand the cellular mechanisms of modulation of dopamine neurons in the PfC and its relationship with behavior. However, the results of these efforts have often led to contradictions and disputes (Nieoullon, 2002). Given the complexity of the function of the mesolimbic and the dopaminergic systems, the development of new tools will be necessary to facilitate discrimination of diagnostics and to provide a more objective assessment of the current classification systems. Namely, we suggest a shift or reconsideration in diagnostic scales adding other indicators. Clinical psychology has many tools to evaluate PfC dysfunction (for a review see Gruszka et al., 2010). We propose that PPI and LI could help to develop a new classification system, where we could distinguish between a psychotic illness such as schizophrenia by a dysfunction in PfC dopamine from other types of schizophrenia included in current scales. As we indicated above, current classification systems could be considering a diverse group of disorders under the same term of schizophrenia illness, and the different combination of positive and negative symptoms could indicate the severity of the disorder. The in depth analysis of these mechanisms, combined with genetic factors, is a new view that could facilitate the development of diagnostic categories in a more specific way and, therefore, a new therapeutic perspective in the future.

AUTHOR CONTRIBUTIONS

All authors contributed similarly in the theoretical development of the manuscript.

FUNDING

This research was supported by Ministerio de Economía y Competitividad (PSI2012- 32445 grant).

REFERENCES

Bachevalier, J., Alvarado, M. C., and Malkova, L. (1999). Memory and socio-emotional behavior in monkeys after hippocampal damage incurred in infancy or in adulthood. *Biol. Psychiatry* 46, 329–339. doi: 10.1016/S0006-3223(99)00123-7

Benes, F. M., Taylor, J. B., and Cunningham, M. C. (2000). Convergence and plasticity of monoaminergic systems in the medial prefrontal cortex during the postnatal period: implications for the development of psychopathology. *Cereb. Cortex* 10, 1014–1027. doi: 10.1093/cercor/10.10.1014

Berridge, K. (2007). The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology* 191, 391–431. doi: 10.1007/s00213-006-0578-x

Berridge, K. C., and Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Res. Rev.* 28, 309–369. doi: 10.1016/S0165-0173(98)00019-8

Brady, A. M., McCallum, S. E., Glick, S. D., and O'Donnell, P. (2008). Enhanced methamphetamine self-administration in a neurodevelopmental rat model of schizophrenia. *Psychopharmacology* 200, 205–215. doi: 10.1007/s00213-008-1195-7

Braff, D., Geyer, M., Light, G., Sprock, J., Perry, W., Cadenhead, K., et al. (2001a). Impact of prepulse characteristics on the detection of sensorimotor gating deficits in schizophrenia. *Schizophr. Res.* 49, 171–178. doi: 10.1016/S0920-9964(00)00139-0

Braff, D. L., Geyer, M. A., and Swerdlow, N. R. (2001b). Human studies of prepulse inhibition of startle: normal subjects, patient groups, and pharmacological studies. *Psychopharmacology* 156, 234–258. doi: 10.1007/s002130100810

Braff, D. L. (2010). Prepulse inhibition of the startle reflex: a window on the brain in schizophrenia. *Curr. Top Behav. Neurosci.* 4, 349–371. doi: 10.1007/7854_2010_61

Braff, D. L., Grillon, C., and Geyer, M. A. (1992). Gating and habituation of the startle reflex in schizophrenic patients. *Arch. Gen. Psychiatry* 49, 206–215. doi: 10.1001/archpsyc.1992.01820030038005

Braff, D. L., and Swerdlow, N. R. (1997). Neuroanatomy of schizophrenia. *Schizophr. Bull.* 23, 509–512. doi: 10.1093/schbul/23.3.509

Bramon, E., Walshe, M., McDonald, C., Martín, B., Toulopoulou, T., Wickham, H., et al. (2005). Dermatoglyphics and schizophrenia: a meta-analysis and investigation of the impact of obstetric complications upon a-b ridge count. *Schizophr. Res.* 75, 399–404. doi: 10.1016/j.schres.2004.08.022

Cadenhead, K. S., Light, G. A., Geyer, M. A., and Braff, D. L. (2000). Sensory gating deficits assessed by the P50 event-related potential in subjects with schizotypal personality disorder. *Am. J. Psychiatry* 157, 55–59. doi: 10.1176/ajp.157.1.55

Cass, D. K., Thomases, D. R., Caballero, A., and Tseng, K. Y. (2013). Developmental disruption of gamma-aminobutyric acid function in the medial prefrontal cortex by noncontingent cocaine exposure during early adolescence. *Biol. Psychiatry* 74, 490–501. doi: 10.1016/j.biopsych.2013.02.021

Chambers, R. A., and Lipska, B. K. (2011). "A method to the madness: producing the neonatal ventral hippocampal lesion rat model of schizophrenia," in *Animal Models of Schizophrenia and Related Disorders*, ed. P. O´Donnell (New York: Humana Press), 1–24.

Chambers, R. A., Moore, J., McEvoy, J. P., and Levin, E. D. (1996). Cognitive effects of neonatal hippocampal lesions in a rat model of schizophrenia. *Neuropsychopharmacology* 15, 587–594. doi: 10.1016/S0893-133X(96)00132-7

Díaz, E., Medellín, J., Sánchez, N., Vargas, J. P., and López, J. C. (2015). Involvement of D1 and D2 dopamine receptor in the recovery processes of stimuli in latent inhibition. *Psychopharmacology* 232, 4337–4346. doi: 10.1007/s00213-015-4063-2

Díaz, E., Vargas, J. P., Quintero, E., de la Casa, L. G., O'Donnell, P., and López, J. C. (2014). Differential implication of dorsolateral and dorsomedial striatum in encoding and recovery processes of latent inhibition. *Neurobiol. Learn. Mem.* 111, 19–25. doi: 10.1016/j.nlm.2014.02.008

Ellenbroek, B. A., Budde, S., and Cools, A. R. (1996). Prepulse inhibition and latent inhibition: the role of dopamine in the medial prefrontal cortex. *Neuroscience* 2, 535–542. doi: 10.1016/0306-4522(96)00307-7

Featherstone, R. E., Rizos, Z., Nobrega, J. N., Kapur, S., and Fletcher, P. J. (2007). Gestational methylazoxymethanol acetate treatment impairs select cognitive functions: parallels to schizophrenia. *Neuropsychopharmacology* 32, 483–492. doi: 10.1038/sj.npp.1301223

Feldman, R. S., Meyer, J. S., and Quenzer, L. F. (1997). *Principles of Neuropsychopharmacology*. Sunderland, MA: Sinauer Associates Inc.

Flagel, S. B., Clark, J. J., Robinson, T. E., Mayo, L., Czuj, A., Willuhn, I., et al. (2011). A selective role for dopamine in stimulus-reward learning. *Nature* 469, 53–57. doi: 10.1038/nature09588

Flagel, S. B., Watson, S. J., Robinson, T. E., and Akil, H. (2007). Individual differences in the propensity to approach signals vs goals promote different adaptations in the dopamine system of rats. *Psychopharmacology* 191, 599–607. doi: 10.1007/s00213-006-0535-8

Flagstad, P., Glenthoj, B. Y., and Didriksen, M. (2005). Cognitive deficits caused by late gestational disruption of neurogenesis in rats: a preclinical model of schizophrenia. *Neuropsychopharmacology* 30, 250–260. doi: 10.1038/sj.npp.1300625

Flagstad, P., Mork, A., Glenthoj, B. Y., Van Beek, J., Michael-Titus, A. T., and Didriksen, M. (2004). Disruption of neurogenesis on gestational day 17 in the rat causes behavioral changes relevant to positive and negative schizophrenia symptoms and alters amphetamine-induced dopamine release in nucleus accumbens. *Neuropsychopharmacology* 29, 2052–2064. doi: 10.1038/sj.npp.1300516

Floresco, S. B., Todd, C. L., and Grace, A. A. (2001). Glutamatergic afferents from the hippocampus to the nucleus accumbens regulate activity of ventral tegmental area dopamine neurons. *J. Neurosci.* 21, 4915–4922.

Frances, A. J., and Widiger, T. (2012). Psychiatric diagnoses: lessons learned from the DSM-IV past and cautions for the DSM-5 future. *Annu. Rev. Clin. Psychol.* 8, 109–130. doi: 10.1146/annurev-clinpsy-032511-143102

Godsil, B. P., Kissc, J. P., Spedding, M., and Jay, T. M. (2013). The hippocampal–prefrontal pathway: the weak link in psychiatric disorders? *Eur. Neuropsychopharmacol.* 23, 1165–1181. doi: 10.1016/j.euroneuro.2012.10.018

Goldman, D. (2012). *Our Genes, Our Choices: How Genotype and Gene Interactions Affect Behavior*. Waltham, MA: Academic Press.

Goto, Y., and Grace, A. (2008). Limbic and cortical information processing in the nucleus accumbens. *Trends Neurosci.* 31, 552–558. doi: 10.1016/j.tins.2008.08.002

Gottesman, I. I., and Shields, J. (1972). *Schizophrenia and Genetics: A Twin Study Vantage Point*. New York: Academic Press.

Gottesman, I. I., and Shields, J. (1973). Genetic theorizing and schizophrenia. *Br. J. Psychiatry* 122, 15–30. doi: 10.1192/bjp.122.1.15

Grace, A. A. (2000). Gating of information flow within the limbic system and the pathophysiology of schizophrenia. *Brain Res.* 31, 331–342. doi: 10.1016/S0165-0173(99)00049-1

Grace, A. A., and Sesack, S. (2010). The cortico-basal ganglia reward network: microcircuitry. *Neuropsychopharmacology* 3, 4–26. doi: 10.1038/npp.2009.93

Graham, F. K. (1975). The more or less startling effects of weak prestimulation. *Psychophysiology* 12, 238–248. doi: 10.1111/j.1469-8986.1975.tb01284.x

Groenewegen, H. G., Wright, C. I., Beijer, V. J., and Voorn, P. (1999). Convergence and segregation of ventral striatal inputs and outputs. *Ann. N. Y. Acad. Sci.* 877, 49–64. doi: 10.1111/j.1749-6632.1999.tb09260.x

Gruszka, A., Hampshire, A., and Owen, A. M. (2010). "Learned irrelevance revisited: pathology-based individual differences, normal variation and neural correlates," in *Handbook of Individual Differences in Cognition, Attention, Memory, and Executive Control*, eds A. Gruszka, G. Matthews, and B. Szymura (New York, NY: Springer), 127–143.

Hasler, G., Drevets, W. C., Gould, T. D., Gottesman, I. I., and Janji, H. K. (2006). Toward constructing an endophenotype strategy for bipolar disorders. *Biol. Psychiatry* 60, 93–105. doi: 10.1016/j.biopsych.2005.11.006

Hazane, F., Krebs, M. O., Jay, T. M., and Le Pen, G. (2009). Behavioral perturbations after prenatal neurogenesis disturbance in female rat. *Neurotox. Res.* 15, 311–320. doi: 10.1007/s12640-009-9035-z

Hyman, S. E. (2010). Diagnosis of mental disorders: the problem of reification. *Annu. Rev. Clin. Psychol.* 6, 155–179. doi: 10.1146/annurev.clinpsy.3.022806.091532

Iversen, L. (1976). Dopamine in the brain and its possible role in madness. *Trends Biochem. Sci.* 1, 121–123. doi: 10.1016/0968-0004(76)90027-X

Kalsbeek, A., Voorn, P., Buijs, R. M., Pool, C. W., and Uylings, H. B. (1988). Development of the dopaminergic innervation in the prefrontal cortex of the rat. *J. Comp. Neurol.* 269, 58–72. doi: 10.1002/cne.902690105

Kidd, K. K. (1997). Can we find genes for schizophrenia? *Am. J. Med. Genet.* 74, 104–111. doi: 10.1002/(SICI)1096-8628(19970221)74:1<104::AID-AJMG21>3.0.CO;2-U

Larrauri, J., and Schmajuk, N. (2006). "Prepulse inhibition mechanisms and cognitive processes: a review and model," in *Neurotransmitter Interactions and Cognitive Function*, ed. E. D. Levin (Basel: Birkhäuser Verlag), 245–278.

Le Pen, G., Bellon, A., Krebs, M. O., and Jay, T. M. (2011). "Gestational MAM (Methylazoxymethanol) administration: a promising animal model for psychosis onset," in *Animal Models of Schizophrenia and Related Disorders*, ed. P. O´Donnell (New York: Humana Press), 25–77.

Le Pen, G., Gourevitch, R., Hazane, F., Hoareau, C., Jay, T. M., and Krebs, M. O. (2006). Peripubertal maturation after developmental disturbance: a model for psychosis onset in the rat. *Neuroscience* 143, 395–405. doi: 10.1016/j.neuroscience.2006.08.004

Lipska, B. K., Aultman, J. M., Verma, A., Weinberger, D. R., and Moghaddam, B. (2002). Neonatal damage of the ventral hippocampus impairs working memory in the rat. *Neuropsychopharmacology* 27, 47–54. doi: 10.1016/S0893-133X(02)00282-8

Lipska, B. K., Jaskiw, G. E., and Weinberger, D. R. (1993). Postpubertal emergence of hyperresponsiveness to stress and to amphetamine after neonatal excitotoxic hippocampal damage: a potential animal model of schizophrenia. *Neuropsychopharmacology* 90, 67–75. doi: 10.1038/npp.1993.44

López, J. C., Karlsson, R. M., and O'Donnell, P. (2015). Dopamine D2 modulation of sign and goal tracking in rats. *Neuropsychopharmacology* 40, 2096–2102. doi: 10.1038/npp.2015.68

Lubow, R. E., and Gewirtz, J. C. (1995). Latent inhibition in humans: data, theory, and implications for schizophrenia. *Psychol. Bull.* 117, 87–103. doi: 10.1037/0033-2909.117.1.87

Lubow, R. E., and Kaplan, O. (2010). "Psychopathology and individual differences in latent inhibition: schizophrenia and schizotypality," in *Handbook of Individual Differences in Cognition. Attention, Memory, and Executive Control*, eds A. Gruszka, G. Matthews, and B. Szymura (New York, NY: Springer), 181–193.

Lubow, R. E., and Weiner, I. (2010). "Issues in latent inhibition research and theory," in *Latent Inhibition Cognition, Neuroscience and Applications to Schizophrenia*, eds R. E. Lubow and I. Weiner (Cambridge: Cambridge University Press), 531–557.

Lüthy, M., Blumenthal, T., Langewitz, W., Kiss, A., Keller, U., and Schächinger, H. (2003). Prepulse inhibition of the human startle eye blink response by visual food cues. *Appetite* 41, 191–195. doi: 10.1016/S0195-6663(03)00080-1

Manitt, C., Mimee, A., Eng, C., Pokinko, M., Stroh, T., Cooper, H. M., et al. (2011). The netrin receptor DCC is required in the pubertal organization of mesocortical dopamine circuitry. *J. Neurosci.* 31, 8381–8394. doi: 10.1523/JNEUROSCI.0606-11.2011

Manoach, D. S. (2003). Prefrontal cortex dysfunction during working memory performance in schizophrenia: reconciling discrepant findings. *Schizophr. Res.* 60, 285–298. doi: 10.1016/S0920-9964(02)00294-3

Marquis, J. P., Goulet, S., and Dore, F. Y. (2008). Neonatal ventral hippocampus lesions disrupt extra-dimensional shift and alter dendritic spine density in the medial prefrontal cortex of juvenile rats. *Neurobiol. Learn. Mem.* 90, 339–346. doi: 10.1016/j.nlm.2008.04.005

Milad, M. R., and Rauch, S. L. (2012). Obsessive-compulsive disorder: beyond segregated cortico-striatal pathways. *Trends Cogn. Sci.* 16, 43–51. doi: 10.1016/j.tics.2011.11.003

Miller, G. A., and Rockstroh, B. (2013). Endophenotypes in psychopathology research: where do we stand? *Annu. Rev. Clin. Psychol.* 9, 177–213. doi: 10.1146/annurev-clinpsy-050212-185540

Missale, C., Russel, N. S., Robinson, M., and Caron, M. (1998). Dopamine receptors: from structure to function. *Physiol. Rev.* 78, 189–225.

Moldin, S. O. (1997). The maddening hunt for madness genes. *Nat. Genet.* 17, 127–129. doi: 10.1038/ng1097-127

Moore, H., Jentsch, J. D., Ghajarnia, M., Geyer, M. A., and Grace, A. A. (2006). A neurobehavioral systems analysis of adult rats exposed to methylazoxymethanol acetate on E17: implications for the neuropathology of schizophrenia. *Biol. Psychiatry* 60, 253–264. doi: 10.1016/j.biopsych.2006.01.003

Moore, H., West, A. R., and Grace, A. A. (1999). The regulation of forebrain dopamine transmission: relevance to the pathophysiology and psychopathology of schizophrenia. *Biol. Psychiatry* 46, 40–55. doi: 10.1016/S0006-3223(99)00078-5

Morris, R., Griffiths, O., Le Pelle, E., and Weickert, T. (2013). Attention to irrelevant cues is related to positive symptoms in schizophrenia. *Schizophr. Bull.* 39, 575–582. doi: 10.1093/schbul/sbr192

Morris, S. E., and Cuthbert, B. N. (2012). Research domain criteria: cognitive systems, neural circuits, and dimensions of behavior. *Dialogues Clin. Neurosci.* 14, 29–37.

Naneix, F., Marchand, A. R., Di Scala, G., Pape, J.-R., and Coutureau, E. (2012). Parallel maturation of goal-directed behavior and dopaminergic systems during adolescence. *J. Neurosci.* 32, 16223–16232. doi: 10.1523/JNEUROSCI.3080-12.2012

Nieoullon, A. (2002). Dopamine and the regulation of cognition and attention. *Progr. Neurobiol.* 67, 53–83. doi: 10.1016/S0301-0082(02)00011-4

O'Donnell, P. (2011). Adolescent onset of cortical disinhibition in schizophrenia: insights from animal models. *Schizophr. Bull.* 37, 484–492. doi: 10.1093/schbul/sbr028

O'Donnell, P., and Grace, A. A. (1995). Synaptic interactions among excitatory afferents to nucleus accumbens neurons: hippocampal gating of prefrontal cortical input. *J. Neurosci.* 15, 3622–3639.

O'Donnell, P., and Grace, A. A. (1998). Dysfunctions in multiple interrelated systems as the neurobiological bases of schizophrenic symptom clusters. *Schizophr. Bull.* 24, 267–283. doi: 10.1093/oxfordjournals.schbul.a033325

Owen, M. J. (2000). Molecular genetic studies of schizophrenia. *Brain Res. Rev.* 31, 179–186. doi: 10.1016/S0165-0173(99)00035-1

Penschuck, S., Flagstad, P., Didriksen, M., Leist, M., and Michael-Titus, A. T. (2006). Decrease in parvalbumin-expressing neurons in the hippocampus and increased phencyclidine- induced locomotor activity in the rat methylazoxymethanol (MAM) model of schizophrenia. *Eur. J. Neurosci.* 23, 279–284. doi: 10.1111/j.1460-9568.2005.04536.x

Plomin, R., and Rende, R. (1991). Human behavioral genetics. *Annu. Rev. Psychol.* 42, 161–190. doi: 10.1146/annurev.ps.42.020191.001113

Robbins, T. W. (1992). Milestone in dopamine research. *Semin. Neurosci.* 4, 93–97. doi: 10.1016/1044-5765(92)90007-O

Robbins, T. W., and Everitt, B. J. (1996). Neurobehavioural mechanisms of reward and motivation. *Curr. Opin. Neurobiol.* 6, 228–236. doi: 10.1016/S0959-4388(96)80077-8

Robbins, T. W., Gillan, C. M., Smith, D. G., Wit, S., and Ersche, K. D. (2012). Neurocognitive endophenotypes of impulsivity and compulsivity: towards dimensional psychiatry. *Trends Cogn. Sci.* 16, 81–91. doi: 10.1016/j.tics.2011.11.009

Robinson, T. E., and Flagel, S. B. (2009). Dissociating the predictive and incentive motivational properties of reward-related cues through the study of individual differences. *Biol. Psychiatry* 65, 869–873. doi: 10.1016/j.biopsych.2008.09.006

Sams-Dodd, F., Lipska, B. K., and Weinberger, D. R. (1997). Neonatal lesions of the rat ventral hippocampus result in hyperlocomotion and deficits in social behavior in adulthood. *Psychopharmacology* 132, 303–310. doi: 10.1007/s002130050349

Seamans, J. K., and Yang, C. R. (2004). The principal features and mechanisms of dopamine modulation in the prefrontal cortex. *Prog. Neurobiol.* 74, 1–58. doi: 10.1016/j.pneurobio.2004.10.002

Seeman, P., and Nizkik, H. B. (1990). Dopamine receptors and transporters in Parkinson's disease and schizophrenia. *FASEB J.* 4, 2737–2744.

Segalowitz, S. J., and Davies, P. L. (2004). Charting the maturation of the frontal lobe: an electrophysiological strategy. *Brain Cogn.* 55, 116–133. doi: 10.1016/S0278-2626(03)00283-5

Silver, H., Feldman, P., Bilker, W., and Gur, R. C. (2003). Working memory deficit as a core neuropsychological dysfunction in schizophrenia. *Am. J. Psychiatry* 160, 1809–1816. doi: 10.1176/appi.ajp.160.10.1809

Simpson, E. H., Kellendonk, C., and Kandel, E. (2010). A possible role for the striatum in the pathogenesis of the cognitive symptoms of schizophrenia. *Neuron* 65, 585–596. doi: 10.1016/j.neuron.2010.02.014

Spear, L. P. (2000). The adolescent brain and age-related behavioral manifestations. *Neurosci.Biobehav. Rev.* 24, 417–463.

Swerdlow, N. R., Braff, D. L., Hartston, H., Perry, W., and Geyer, M. A. (1996). Latent inhibition in schizophrenia. *Schizophr. Res.* 20, 91–103. doi: 10.1016/0920-9964(95)00097-6

Swerdlow, N. R., Halim, N., Hanlon, F. M., Platten, A., and Auerbach, P. P. (2001). Lesion size and amphetamine hyperlocomotion after neonatal ventral hippocampal lesions: more is less. *Brain Res. Bull.* 55, 71–77. doi: 10.1016/S0361-9230(01)00492-0

Swerdlow, N. R., Lipska, B. K., Weinberger, D. R., Braff, D. L., Jas-kiw, G. E., and Geyer, M. A. (1995). Increased sensitivity to the sensorimotor gating-disruptive effects of apomorphine after lesions of medial prefrontal cortex or ventral hippocampus in adult rats. *Psychopharmacology* 122, 27–34. doi: 10.1007/BF02246438

Torrey, E. F., and Yolken, R. H. (2000). Familial and genetic mechanisms in schizophrenia. *Brain Res. Rev.* 31, 113–117. doi: 10.1016/S0165-0173(99)00028-4

Tseng, K. Y., and O'Donnell, P. (2004). Dopamine-glutamate interactions controlling prefrontal cortical pyramidal cell excitability involve multiple signalling mechanisms. *J. Neurosci.* 24, 5131–5139. doi: 10.1523/JNEUROSCI.1021-04.2004

Tseng, K.-Y., and O'Donnell, P. (2007). Dopamine modulation of prefrontal cortical interneurons changes during adolescence. *Cereb. Cortex* 17, 1235–1240. doi: 10.1093/cercor/bhl034

Turetsky, B. I., Calkins, M. E., Light, G. A., Olincy, A., Radant, A. D., and Swerlow, N. R. (2007). Neurophysiological endophenotypes of schizophrenia: the viability of selected candidate measures. *Schizophr. Bull.* 33, 69–94. doi: 10.1093/schbul/sbl060

Vaidyanathan, U., Malone, S. M., Miller, M. B., Mcgue, M., and Iacono, W. G. (2014). Heritability and molecular genetic basis of acoustic startle eye blink and affectively modulated startle response: a genome-wide association study. *Psychophysiology* 51, 1285–1299. doi: 10.1111/psyp.12348

Waddington, J. L., Lane, A., Larkin, C., and O'Callaghan, E. (1999). The neurodevelopmental basis of schizophrenia: clinical clues from cerebro-craniofacial dysmorphogenesis, and the roots of a lifetime trajectory of disease. *Biol. Psychiatry* 46, 31–39. doi: 10.1016/S0006-3223(99)00055-4

Wan, R. Q., Giovanni, A., Kafka, S. H., and Corbett, R. (1996). Neonatal hippocampal lesions induced hyperresponsiveness to amphetamine: behavioral and in vivo microdialysis studies. *Behav. Brain Res.* 78, 211–223. doi: 10.1016/0166-4328(95)00251-0

Weiner, I. (2003). The "two-headed" latent inhibition model of schizophrenia: modeling positive and negative symptoms and their treatment. *Psychopharmacology* 169, 257–297. doi: 10.1007/s00213-002-1313-x

Winokur, G., and Kadrmas, A. (1989). A polyepisodic course in bipolar illness: possible clinical relationships. *Comparat. Psychiatry* 30, 121–127. doi: 10.1016/0010-440X(89)90063-1

Wise, R. A. (2010). Roles for nigrostriatal -not just mesocorticolimbic-dopamine in reward and addiction. *Trends Neurosci.* 32, 517–524. doi: 10.1016/j.tins.2009.06.004

Wong, E. H., Yocca, F., Smith, M. A., and Lee, C. M. (2010). Challenges and opportunities for drug discovery in psychiatric disorders: the drug hunters' perspective. *Int. J. Neuropsychopharmacol.* 13, 1269–1284. doi: 10.1017/S1461145710000866

Zahn-Waxler, C., Mayfíeld, A., Radke-Yarrow, M., McKnew, D. H., Cytryn, L., and Davenport, Y. B. (1988). A follow-up investigation of offspríng of parents with bipolar disorder. *Am. J. Psychiatry* 145, 506–509. doi: 10.1176/ajp.145.4.506

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 25 October 2016
doi: 10.3389/fpsyg.2016.01659

Reading Ability Development from Kindergarten to Junior Secondary: Latent Transition Analyses with Growth Mixture Modeling

Yuan Liu[1, 2], *Hongyun Liu*[3, 4]* *and Kit-tai Hau*[5]

[1] Faculty of Psychology, Southwest University, Chongqing, China, [2] Key Laboratory of Cognition and Personality, Southwest University, Ministry of Education, Chongqing, China, [3] Beijing Key Laboratory of Applied Experimental Psychology, Beijing Normal University, Beijing, China, [4] School of Psychology, Beijing Normal University, Beijing, China, [5] Department of Psychology, Chinese University of Hong Kong, Hong Kong, Hong Kong

The present study examined the reading ability development of children in the large scale Early Childhood Longitudinal Study (Kindergarten Class of 1998-99 data; Tourangeau et al., 2009) under the dynamic systems. To depict children's growth pattern, we extended the measurement part of latent transition analysis to the growth mixture model and found that the new model fitted the data well. Results also revealed that most of the children stayed in the same ability group with few cross-level changes in their classes. After adding the environmental factors as predictors, analyses showed that children receiving higher teachers' ratings, with higher socioeconomic status, and of above average poverty status, would have higher probability to transit into the higher ability group.

Keywords: reading development, latent transition analysis, growth mixture model, dynamical systems, social rating

OPEN ACCESS

Edited by:
Salvador Chacón-Moscoso,
University of Seville, Spain

Reviewed by:
Ratna Nandakumar,
University of Delaware, USA
Jocelyn Holden Bolin,
Ball State University, USA

****Correspondence:***
Hongyun Liu
hyliu@bnu.edu.cn

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *09 April 2016*
Accepted: *11 October 2016*
Published: *25 October 2016*

Citation:
Liu Y, Liu H and Hau K-t (2016) Reading Ability Development from Kindergarten to Junior Secondary: Latent Transition Analyses with Growth Mixture Modeling. Front. Psychol. 7:1659. doi: 10.3389/fpsyg.2016.01659

INTRODUCTION

Reading is an important activity composing of various sub-skills which grow at different speed. In reality, students are nurtured in a dynamic system where they are not only self-organizing, but also interacting and being substantially affected by the psychosocial environment (Votruba-Drzal et al., 2008; Ding et al., 2013; Iruka et al., 2014). In such a system, one under-researched area is the effect of young students' social environment at home and at school on their learning to read behavior. The purpose of the present study, therefore, was to explain the pattern of reading development and to depict the relations between the developmental pattern and children's behavior as perceived by their parents and teachers. We applied and explored with the application of the latest appropriate statistical method—the latent transition analysis with growth mixture model on a large scale longitudinal survey (Early Childhood Longitudinal Study, Kindergarten Class of 1998-99, ECLS-K, Tourangeau et al., 2009).

READING DEVELOPMENT: NON-CONTINUOUS PATTERN AND GROUPING

Reading can be seen as a way of meaning extraction which requires the working of different sub-skills on the text (Stahl, 1997; Clay, 2001; Rodgers, 2004). Recent research has highlighted the need to look more closely at the different skills. Word reading, therefore, might have to be separated from reading comprehension because the former includes some of the basic phonological abilities, letter knowledge, and short-term memory (Muter et al., 2004; Kendeou et al., 2009), whereas the latter may need inference, monitoring, and knowledge of the story structure (Vellutino et al., 2007; Kendeou et al., 2009).

The mastery process of the language is, however, quite different for different subskills, such as for the constrained and unconstrained skills (e.g., Paris, 2005, 2009; Paris et al., 2005). Children's reading ability grows irregularly with spurts and stops (de Weerth et al., 1999). For example, with substantial individual differences, children's language competence may grow extremely rapidly before Spring Grade 1 but may decline thereafter (Palardy, 2010; Kieffer, 2012). Verhoeven et al. (2011) showed that the different patterns of the reading development were distinct from those around Grade 2.

Since the reading development pattern may differ from phase to phase, researchers are very interested in tracing and examining the growth trajectories. Paris (2005) suggested that when calibrating the unconstrained skills to the constrained skills, reading development follows a non-continuous growing pattern. This may not be easily detected when a simple linear growth modeling is used. Thus, for example, Quinn et al. (2015) have to use a two-part model to depict separately the developmental trajectories of the vocabulary knowledge and the reading comprehension through Grade 1 to Grade 4. Their bivariate model showed that vocabulary knowledge acted as a causal indicator of the subsequent reading comprehension growth. In summary, if researchers intend to depict the full picture of the reading developmental trajectory, a continuous growth model may not be suitable. Students stay at different "stages" with adaption to the new context using different reading skills.

A more sophisticated issue is that not all students share the same growing pattern across stages (Kaplan, 2002; Pianta et al., 2008). Empirically, these differential patterns in growth can be analyzed by (i) differentiating children into language ability groups and (ii) tracing their changes in groups as they progress in schools. For example, while most students develop rapidly before Spring Grade 1 and then slow down afterwards, some children may have a consistently slow growth rate (Kaplan, 2005; Kapland, 2008; Palardy, 2010).

The variation in growth rate is more likely to occur in the lower grades—as early as first grade (Ferrer et al., 2015), or around age of eight (Stanovich, 1986). Studies also showed that the dyslexic reader would probably grow at a slow pace that hardly enables the children to catch up with other typical readers (Grimm et al., 2010; Ferrer et al., 2015). The grouping phenomenon among slow developers is potentially harmful to them, since this low-ability-group students may have lower self-efficacy or motivation to learn let alone their ability shortage. Thus, it is important to find the conducive factors to facilitate these low ability students to "transit" into the higher competence group.

To solve the above challenging questions, we need a combined model to depict the various developing patterns with spurs and spots. Furthermore, as students' growth is determined by their current pre-exiting ability as well as by other influential factors in the environment, a dynamic systems model was adopted to analyze the interplay of these factors.

READING DEVELOPMENT IN DYNAMIC SYSTEMS

To depict and explore the reading development, two issues should be noticed. Firstly, the sub-skills are correlated among each other. For example, Verhoeven et al. (2011) showed that the vocabulary at the beginning phase could predict word decoding and reading comprehension at the early stages of development. From Grade 2 onwards, word decoding competence in turn predicted later vocabulary development. Reading ability develops under the effects of the formal skills (Oakhill and Cain, 2012). Secondly, children live in a complicated environment where many of the external factors may influence the reading development. Thus, a *dynamic systems* view should be introduced when describing such a development.

The dynamic systems theory originates from natural science studies (for a review, see van Geert, 2003). According to this perspective, individual development is a consequence of the dynamic interactions within an individual and between an individual and the environment. In the last two decades, the dynamic systems view has been intensively discussed and widely applied, especially in language development research (Robinson and Mervis, 1999; van Geert and Steenbeek, 2005; Hollenstein, 2011; van Geert, 2011).

According to the dynamic systems, reading development can be described in terms of the *change*, *interactions,* and *conjoint analysis* of the individual and environment systems (Clay, 1977, 2001). For example, Clay (2001) believed that individuals would be able to construct and self-organize with their potential ability. They will push through the boundaries and improve their knowledge with their skills already mastered. So, proficient readers are able to mobilize the processing systems to fit the challenges of different texts by using environmental cues such as visual and motor stimulants. Kainz and Vernon-Feagans (2007) showed that the acquisition of reading ability was not isolated from the outside world. Kainz and Veron-Feagans worked with their colleague and developed a system of the dynamic circles involving the individuals, families, classrooms and school systems. This would be helpful to children's reading development and possibly help their transitions into higher ability groups (Kainz and Vernon-Feagans, 2007; Vernon-Feagans et al., 2008).

Among various factors in the social environment, teachers and parents' perception and attitude on students' study behaviors play important roles. These factors and their interplay vary from one individual to another and crucially affect students' academic outcomes. Ladd et al. (1999) *Child* × *Environment* model provides further explanation on how the quality of children's relationships can directly and indirectly influence school achievement from a dynamic system perspective. In the model, they show that children's initial behavior or the background factors influence their relationships with peers and teachers. Peer and teacher relationships in the school environment enhance or sometimes adversely affect student's achievement. For example, it is likely the students from lower socioeconomic backgrounds would be benefitted more by teachers who employed a more interpersonal approach of instruction, such as incorporating mixed group work, using peer tutoring, and solving problems with partners (Jung, 2014). Other studies have also consistently shown that high quality teacher-child relationship is conducive to high achievement (Davis, 2003; Pianta and Stuhlman, 2004; Hughes and Kwok, 2007; O'Connor and McCartney, 2007; Hughes et al., 2008). This relationship is also influenced by children's social behavior, such as their classroom engagement, which in turn affects children's achievement and academic outcomes (Cohen, 1997; Hughes and Kwok, 2007; O'Connor and McCartney, 2007).

From a dynamic systems perspective, teachers and parents could offer help to speed up children's transition into higher ability groups (Cho et al., 2013; Eyden et al., 2014). For example, teachers and parents' perceptions of students' ability and effort are closely related to children's academic achievement (Rytkönen et al., 2007; Natale et al., 2009; Longobardi et al., 2011). Particularly, since highly motivated children are perceived as talented and effortful (Upadyaya et al., 2012), parents and teachers' positive perceptions on children would be conducive to children's development. Upadyaya and Eccles (2015) showed that teachers' perceptions on ability and effort could predict the subsequent reading ability in a longitudinal study. It is thus quite important how teachers and parents perceive and show to the students their positive evaluation. This is because at the early elementary school years, children often assimilate teachers' perceptions in formulation the judgment of their own ability (Rosenholtz and Simpson, 1984; Tiedemann, 2000). From another perspective, children's educational aspiration partially reflected their parents and teachers' expectation on them as well, thus highlighting the importance of setting an appropriate but sufficiently high educational aspiration (Kuklinski and Weinstein, 2001; Herbert and Stipek, 2005).

THE PRESENT STUDY

Two important issues would be addressed in the present study. Firstly, we were interested in the transition showing students' potency to develop their abilities. There are patterns shared by children in the same group in that they improved in their mastery of different reading skills, and thus grew together from one stage (lower ability groups) to the next (higher ability groups). Secondly and more importantly, we are interested in those environmental variables, especially the parents and teachers' perception on the children, that might facilitate such a transition.

Driven by the research questions, we had several research questions to examine under the dynamic systems theory. First, according to the integrated view of dynamic systems theory, a self-organizing process reflected an auto-regression development. We would examine whether and how extensive the subsequent ability status was determined by the previous status. Second, we would examine how much individual differences existed in students' growth trajectories. Finally, the contribution of parents' and teachers' perception on students' growth would be examined.

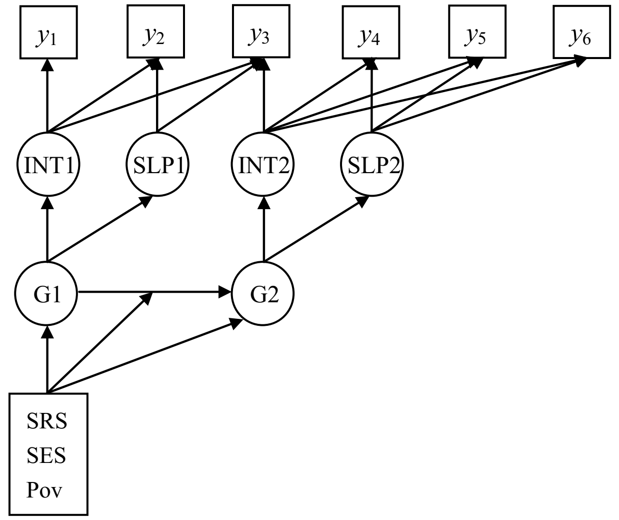


FIGURE 1 | Latent transition analysis with growth mixture model (LTA-GMM) on reading development. The dynamic systems of reading development show the relations among the reading ability, initial state, slopes of two stages, and latent groups with social rating (SRS), SES, and poverty as the environmental factors. $y_1 - y_6$, reading ability indicators from Wave 1 to Wave 6; INT1, Stage 1 Intercept; SLP1, Stage 1 Slope; INT2, Stage 2 Intercept; SLP2, Stage 2 Slope; G1, Stage 1 Grouping (two groups); G2, Stage 2 Grouping (two groups).

METHODS

Participants

We used the publicly available data in the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K) (Tourangeau et al., 2009)[1] to examine our research questions. This data set was developed under the National Center for Education Statistics (NCES). We chose the ECLS-K because it focused on children's early school experiences from kindergarten to Grade 8, and the longitudinal data displayed students' long-term trajectory development. Furthermore, ECLS-K adopted a multi-source, multi-method approach, which included interviews with parents, data from principals and teachers, information from student records, and direct assessment on children (including reading, mathematics and science cognitive items). The study was in alignment with the dynamic systems theory, in which various environmental variables were considered.

In total, seven waves of measures of reading assessment were available in the data set (C1R4RSCL–C7R4RSCL). As the data at Fall Grade 1 (C3R4RSCL) contained only 30% of the total sample, without jeopardizing the generalization of our conclusion, it was not included in our study. The remaining data points were from Fall Kindergarten, Spring Kindergarten, Spring Grade 1, Spring Grade 3, Spring Grade 5 and Spring Grade 8 ($y_1 - y_6$ in **Figure 1**).

[1] Retrieved from http://nces.ed.gov/ecls/kindergarten.asp.

Together with the parents' and teacher's questionnaires, 7803 children's questionnaires were available in our analyses.

There were 456 individuals with missing covariate values, and totally 1033 individuals with missing values on one or more of the covariates or indicators. For the missing rate of each variable, other than the slightly higher rate at y_1 (7.0%), all other ranged from 0.5 to 4.8% only, with an overall average missing rate of 2.6%. Generally, the missing pattern of the present dataset could be treated as missing at random, so that the multiple imputation method by Mplus 7.0 (Muthén and Muthén, 2012) could be appropriately used. We generated 10 datasets, and the sample size 7803 was applied to the analyses with either the null model or with covariates being included. Basic information among the variables is shown in **Table 1**.

Measures

Reading Ability

The reading items were drawn from assessments used in other large-scale studies of similar-aged youth, including the National Assessment of Educational Progress (NAEP), the National Education Longitudinal Study of 1988 (NELS:88), the Education Longitudinal Study of 2002 (ELS:2002), the Texas Assessment of Knowledge and Skills (TAKS), and previous rounds of the ECLS-K. The reading items in ECLS-K were repeatedly measured with ten levels of the reading ability (see **Figure 2**). Each new wave was recalibrated to the former one and tests at each wave included some identical items so that the instruments at different waves could be linked on the same IRT scales (represented on the same unit of measurement). Specifically, in the collection of the Grade 8 data which was used in the present analysis, all the proficiency scores for the former levels were re-estimated to be pooled with the latest wave (see Tourangeau et al., 2009 for details).

Social Rating

The social rating was the evaluation of the children's behavior by parents and teachers. The items were obtained from the Social Rating Scale (SRS) Approaches to Learning scales of the ECLS-K Parent and Teacher questionnaires. The SRS survey items comprised of parents' and teachers' ratings on how frequent and whether students had those study-related behaviors or not. The scale contained items such as intrinsic motivation, persistence/attention, and study habits. These ratings by teachers and parents, rather not self-reported by children, reflected students' social behaviors as perceived by the others, thus shows the interaction between students and their guardians.

A four-point scale was used, with "1 = never" and "4 = very often." Parents' SRS was collected annually except in the third, fifth, and eighth grades, while teachers' SRS was not collected at the eighth grade. In the study, the SRS in Fall Kindergarten was used to predict the transition of latent class. All these items were used as continuous variables in the present analyses (see Tourangeau et al., 2009).

Background Information

While many studies had investigated the relationships among Socio-economic status (SES), poverty, race, minority and achievement, which were generally used as the background variables (e.g., Hattie, 2009; OECD, 2013). Specifically, SES referred to students' relative position in the social hierarchy, directly reflected the resources at home, and was often used as an important controlling variable. Both SES and poverty status measured important characteristics of the background family information and were thus chosen in our analysis (see Tourangeau et al., 2009).

Analyses Procedure

Model Definition

The latent transition analysis (LTA) (Prochaska and Velicer, 1997) was used to analyze the longitudinal transitions. The auto-regression part of the LTA model described appropriately the self-organizing process under the dynamic systems. LTA also allowed us to add environmental covariates to moderate the auto-regression process. With the LTA model, the measurement part could be further replaced according to different contexts and situations.

TABLE 1 | Correlations and descriptive statistics of variables used in the analyses.

	1	2	3	4	5	6	7	8	9	10
1. Parent Rating	–									
2. Teacher Rating	0.23	–								
3. SES	0.18	0.19	–							
4. Poverty	0.12	0.15	0.49	–						
5. y_1	0.21	0.39	0.43	0.28	–					
6. y_2	0.22	0.39	0.39	0.27	0.80	–				
7. y_3	0.22	0.39	0.39	0.29	0.69	0.78	–			
8. y_4	0.23	0.38	0.45	0.33	0.61	0.67	0.76	–		
9. y_5	0.23	0.36	0.46	0.32	0.59	0.63	0.72	0.85	–	
10. y_6	0.21	0.34	0.48	0.31	0.53	0.55	0.61	0.75	0.79	–
M	3.13	3.06	0.11	1.84	−1.26	−0.68	0.16	0.82	1.08	1.34
SD	0.22	0.42	0.63	0.14	0.26	0.24	0.19	0.09	0.08	0.15

$y_1 - y_6$ = reading ability indicators from Wave 1 to Wave 6.

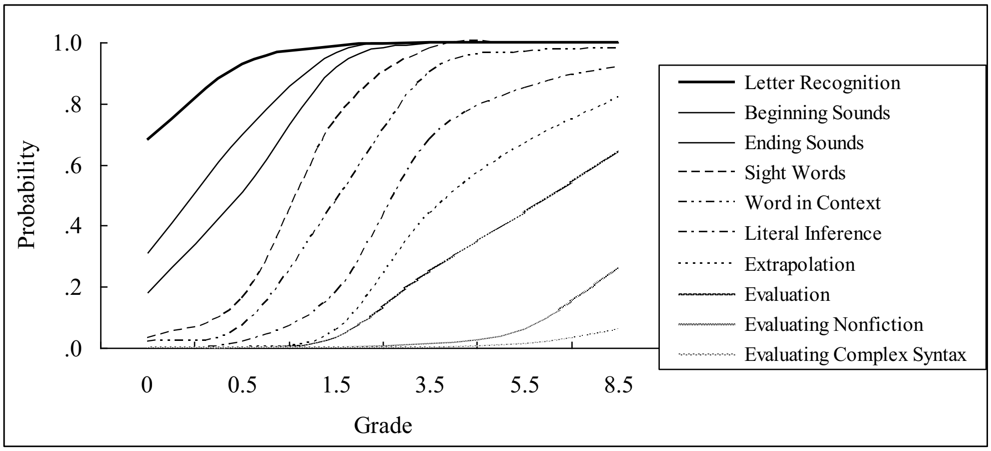


FIGURE 2 | Probability of mastery of different proficiency levels at different grades. The ten levels of reading proficiencies were: (1) Letter Knowledge, identifying upper- and lower-case letters of the alphabet; (2) Beginning Sounds, associating letters with sounds at the beginning of words; (3) Ending Sounds, associating letters with sounds at the end of words; (4) Sight Words, recognizing common "sight" words; (5) Words in Context, reading words in context; (6) Literal Inference, making inferences using cues that were directly stated with key words in text; (7) Extrapolation, identifying clues used to make inferences; (8) Evaluation, demonstrating understanding of author's craft and making connections between a problem in the narrative and similar life problems; (9) Evaluating Nonfiction, comprehension of biographical and expository text; and (10) Evaluating Complex Syntax, evaluating complex syntax and understanding high-level vocabulary (Tourangeau et al., 2009).

As an extension, we took advantage of the growth mixture model (GMM) to replace the measurement part of the original LTA (see Muthén et al., 2012). The GMM model could detect the growth of the reading skills by allowing individual differences in growth rate within each group, in contrast to the more stringent requirement with little individual differences allowed at each point of time.

According to earlier studies (Votruba-Drzal et al., 2008; Kieffer, 2012), the Spring Grade 1 (y_3) was chosen to be the cut point of two stages. Thus, ($y_1 - y_3$) were the indicators of Stage 1 (kindergarten stage) with latent growth factors INT 1 and SLP 1 (intercept/initial state and slope) classified into latent groups (G1); whereas ($y_3 - y_6$) were the indicators of Stage 2 (primary to junior high school) with latent growth factors INT2 and SLP2 classified into groups (G2, see **Figure 1**).

Implementing the 3-Step Analysis

Specifically, in testing the effects due to environmental facilitating factors, covariates have to be introduced into the LTA. When adding these covariates, it is necessary to find appropriate ways to control for the characteristics that predict the membership in the different latent classes. Therefore, a *3-step Maximum Likelihood Method* (referred to the *3-step* approach in subsequent discussion) was used (see Collins and Lanza, 2010; Vermunt, 2010; Asparouhov and Muthén, 2014, see also Liu and Liu, 2015 for details).

In the first step in the 3-step LTA, GMM was used to get the classification of latent class for each stage, using the indicators at their respective stage only. For example, when estimating GMM at Stage 1, y_1 to y_3 were used as indicators, with y_4 to y_6 and the covariates serving as auxiliary variables; the proportions of each latent class were recorded. Similarly, GMM was conducted at Stage 2. In the second step, using the classification outcomes and the proportions given by Mplus, the classification error was computed for each latent class. With the odds ratio computed by the second step as the starting value of each latent class, LTA (G2 was regressed on G1) with the covariates (G1, G2, and the transition from G1 to G2, respectively, regressed on covariates) was applied (for detailed syntax, see Asparouhov and Muthén, 2014).

Model Selection Indices

The selection of the number of the latent classes has been a topic of much discussion (e.g., Nylund et al., 2007; Tofighi and Enders, 2008; Peugh and Fan, 2012). Most studies suggested that the BIC (Bayesian information criterion) value should be the best choice because it was a sample based index which also penalized sophisticated model. Tofighi and Enders (2008) in their simulation study showed that a sample size adjusted BIC (aBIC) was an even better index, and thus was used in our study. A smaller BIC/aBIC value indicated better model fit for nesting models. Besides, the entropy value was to measure how well a mixture model separated the classes. An entropy value close to 1 indicated good classification certainty. Asparouhov and Muthén (2014) suggested that an entropy level of 0.6 or higher might provide sufficient good classification for the 3-step method.

RESULTS

Selection of the Proper Model

As LTA was used in combination with GMM, the original GMM analyses were examined first. The piecewise GMM ($y_1 - y_3$ as Piece 1 and $y_3 - y_6$ as Piece 2) null model was chosen. We conducted the exploration analyses from 2 to 4 classes (see **Table 2**). The model fit indices, $-2LL$, BIC, and aBIC,

TABLE 2 | Model comparison and selection.

	BIC	aBIC	−2*LL*	Entropy
GMM_2c	17231	17170	17060	0.914
GMM_3c	16966	16893	16760	0.902
GMM_4c	17302	17216	17060	0.957
LTA-GMM_2c	11269	11171	10991	0.734
LTA-GMM_3c	12098	11958	11704	0.897
LTA-GMM_2c (3-step)	8099	8094	8082	0.915

2c, 2 classes; 3c, 3 classes; 4c, 4 classes; 3-step, 3 step method.

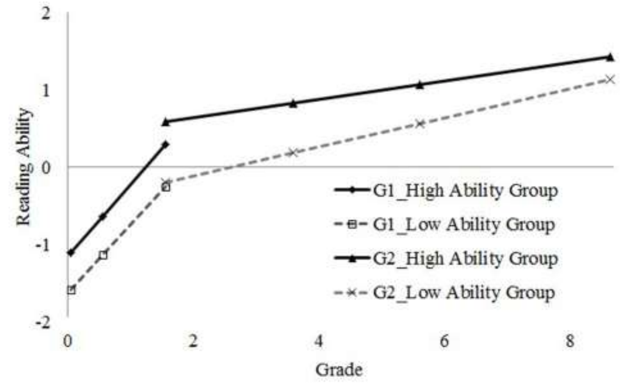


FIGURE 3 | Reading growth trajectories. G1, Grouping at Stage 1; G2, Grouping at Stage 2.

TABLE 3 | Class counts, proportions and conditional transition probability for the final model solution.

	Stage 1	Stage 2	Class count	Class proportion	Transition probability
Combination 1	High	High	7093	0.909	1.000
Combination 2	High	Low	0	0.000	0.000
Combination 3	Low	High	336	0.043	0.474
Combination 4	Low	Low	374	0.048	0.526

High, High Ability Group; Low, Low Ability Group.

TABLE 4 | Parameter estimates of the growth factors for the final model solution.

	High ability group				Low ability group			
	Est.	*SE*	*t*	*p*	Est.	*SE*	*t*	*p*
STAGE 1								
Means								
INT 1	−1.10	0.02	−72.62	<0.001	−1.58	0.03	−59.18	<0.001
SLP 1	0.93	0.01	198.43	<0.001	0.89	0.02	60.21	<0.001
VARIANCES								
INT 1	0.18	0.01	34.52	<0.001	0.16	0.01	18.69	<0.001
SLP 1	0.04	0.00	11.68	<0.001	0.09	0.01	15.12	<0.001
COVARIANCE								
INT 1 with SLP 1	−0.05	0.00	−21.32	<0.001	−0.05	0.00	−21.32	<0.001
STAGE 2								
Means								
INT 2	0.41	0.00	215.30	<0.001	−0.48	0.02	−30.13	<0.001
SLP 2	0.12	0.00	463.48	<0.001	0.19	0.00	139.32	<0.001
VARIANCES								
INT 2	0.01	0.00	10.94	<0.001	0.01	0.00	3.93	<0.001
SLP 2	0.00	0.00	36.12	<0.001	0.00	0.00	12.13	<0.001
COVARIANCE								
INT 2 with SLP 2	0.01	0.00	55.28	<0.001	0.01	0.00	55.28	<0.001

consistently supported a 3-class model. Then we checked the class proportion to ensure the empirical significance. For the 2-class model, the proportion was 0.95 and 0.05 for each class; for the 3-class model, the proportion was 0.93, 0.05, and 0.02; for the 4-class model, two groups contained 0 individuals. It was evident that the third group in a 3-class model was so tiny (less than 5%) and would not contribute substantially and empirically to the model, so the 2-class model was retained.

We then conducted the GMM-LTA null model, using two stages of growth but without any covariate. Results showed that the 2-class model was the best according to the selection criteria (BIC and aBIC), with slightly worse but acceptable entropy value (see **Table 2**).

Finally, we conducted the 3-step GMM-LTA. BIC was 8099 with an entropy value of 0.92. The information criteria and entropy value indicated that the 3-step model was the best. The final model consisted of two groups at two stages, respectively (**Figure 3**).

Grouping Membership

The classification results are shown in **Table 3**, and the parameter estimates for the growth factors are shown in **Table 4**. At Stage 1, most of the students were classified into the high ability group (90.9%, with initial ability of −1.10). The other 9.1% were in the low ability group with a lower initial status (−1.56). The growing rate (slope) of the high ability group (0.93) was slightly faster than that of the low ability group (0.89), but with quite similar pattern seen from the trajectory in **Figure 3**. For Stage 2, from Spring Grade 1 to Grade 8, children in different classes had different growing rates. There were 95.2% in the high ability group with an initial ability of 0.41 and a growth rate of 0.12, while 4.8% in the low ability group had an initial ability of −0.48 and a growth rate of 0.19.

After grouping, there were two groups in each stage; so four possible combinations of sub-groups were formed (**Table 3**). Combination 1 (90.9%), which contains individuals classified in the high ability groups at both Stages 1 and 2, had the largest proportion. Combination 4 referred to individuals classified as low ability at both stages contained 4.8% of the population. This showed that most students' growth was stable (totally 95.7% of the population). There were about 4.3% of students being classified as Combination 3, who moved from the low ability group to the high ability group across time. No individual was in Combination 2, indicating that there was no reversed pattern (changed from high ability group to low ability group).

A transition probability showed that, once classified into the high group, students would have a 100% probability staying in the high ability group thereafter. In contrast, children starting in the low group would likely be in the low ability group at Stage 2 but had a considerably high probability to transit into the high ability group at Stage 2.

Effect of the Environmental Factors

We set the significant level at p <.001 for this study with a large sample size. Results (see **Table 5**) showed that the covariates could predict the Stage 1's classification. Other than the background variables, both parents and teachers' higher ratings were associated with children's higher reading ability (with the lower ability group as the reference) at Stage 1. The Stage 2's classification could be predicted positively only by the parents' rating and SES level, with higher parents' rating and SES related to better children's performance (i.e., classified in the higher ability group). In contrast, higher teachers' rating was related to lower students' performance (being classified in the lower ability group; β = −6.02, odds ratio = 0.00).

Interactive effects with grouping transition were examined. It was found that when teachers' ratings (β = 5.66, odds ratio = 288) were more positive, then the children had a higher chance to transit from the lower to the higher ability group. Specifically, when the teachers' ratings were one unit higher, the low ability children at Stage 1 would have 288 times higher probability in transiting to the high ability group at Stage 2. However, the effects due to parents' ratings (β = −2.77, odds ratio = 0.06) and SES (β = −4.14, odds ratio = 0.02) were negligible.

TABLE 5 | Dynamic systems model involving environmental factors.

	β	*SE*	*t*	*p*	Odds ratio
STAGE 1 GROUPING ON[a]					
Parent Rating	0.65	0.06	10.82	<0.001	1.91
Teacher Rating	1.95	0.04	47.69	<0.001	7.05
SES	1.65	0.07	25.43	<0.001	5.22
Poverty	0.59	0.07	8.77	<0.001	1.81
STAGE 2 GROUPING ON[a]					
Parent Rating	2.58	0.16	15.80	<0.001	13.24
Teacher Rating	−6.02	1.27	−4.75	<0.001	0.00
SES	3.66	0.88	4.14	<0.001	38.79
Poverty	−8.57	3.41	−2.52	0.012	0.00
TRANSITION (COMBINATION 3) ON[b]					
Parent Rating	−2.77	0.18	−15.12	<0.001	0.06
Teacher Rating	5.66	1.27	4.45	<0.001	287.75
SES	−4.14	0.89	−4.67	<0.001	0.02
Poverty	8.48	3.41	2.49	0.013	4812.21

[a] *Classification was regressed on the covariates.*

[b] *Stage 2 High ability group (cf. low ability group) was regressed on the covariates in Stage 1 low ability group.*

DISCUSSION

Developing Patterns

The present study showed the advanced 3-step GMM-LTA model well described the complex longitudinal ECLS-K database set in the dynamic systems model. The developmental trend showed a fast grow from kindergarten to Spring Grade 1 and then a slowing down to a plateau on time beyond. A closer examination of the reading ability scores (**Figure 2**) showed that the formal five levels of reading proficiency were more related to Paris's constrained skills which were close perfection after Spring Grade 1. After this time spot, students continuously learned unconstrained skills. From **Table 4**, statistical evidence showed that the variances were much smaller at Stage 2 than those at Stage 1, especially for their growth rates which had little variance at Stage 2. This indicates the non-normal distribution across the development from kindergarten to Grade 8. It is necessary, therefore, to analyze the reading skills separately at different stage, where sub-skills developed with quite different speeds and patterns.

The grouping results were consistent with the literature (Grimm et al., 2010; Ferrer et al., 2015) in that two groups with different ability levels could be differentiated. The classification indicates that most of the students were classified in high ability group, either at Stage 1 or Stage 2. We can thus treat the high ability group as the reference "normal" developing pattern, since it contained more than 90% of the population. So, students classified in the lower ability group were those likely to have reading problems. According to Ferrer et al. (2015), the grouping differentiation could emerge as early as Grade 1; our study indicates that the grouping may emerge even earlier. However, students still had a considerable chance to transit into the higher ability group through the self-organizing progress (conditional probability was 0.47). Educators should pay more attention to children's early reading problems as early as possible before they develop into more serious language learning problems.

Environmental Facilitators

We found that all the factors being examined had substantial effects on the grouping at Stage 1. Contradictory results were found, however, in the prediction of Stage 2 grouping/transition. The results showed that, parents' rating and SES positively predicted Stage 2 grouping, whereas they negatively predicted the transition. Vice versa, teachers' rating negatively predicted classification, but positively predicted the transition. These contradictions may reflect problems in the long-term prediction efficiency. When we took the transition prediction terms out of our model, all predictions on Stage 2 grouping showed negative estimates (ranged from −0.48 to −0.10), but with quite small or non-significant effects (odds ratios ranged from 0.70 to 0.91). So the social environmental variables collected at Wave 1 may have less predictive power to the subsequent ability, especially for a long-term growth (8.5 years). This is somehow similar to the previous study (Upadyaya and Eccles, 2015) which showed that teachers' perception of the effort of students could predict the subsequent reading ability with a small interval (1 year) only. Further investigations on the prediction power in long term studies would be useful.

As for the transition, the results showed that teachers' ratings had larger effects in predicting the transition probability than that of the parents'. This reveals that the teachers' ratings are probably more accurate as compared to those of the parents', which might be explained by the Child × Environment model (Ladd et al., 1999; Pianta and Stuhlman, 2004). To illustrate, the teacher-student relationship is a mediator influenced by the effect of school behavior and other background or cognitive variables on children's achievement. With the accurate perceptions, teachers may adopt more efficient approaches on students' learning. Teachers' interaction with students is thus playing as a *proximal factor* influencing the achievement influencing academic achievement more directly, while school entries (family variables) are *distal factors*. On the other hand, longitudinal studies show that teachers' perception of the students (either ability or effort) can predict subsequent children's self-concept (Natale et al., 2009); teachers are significant socialization agents whose perception greatly impact children's self-concept formation (Madon et al., 2001), and thus have a great impact on students ability. To summarize, we are alerted again of the important role of the teacher-student relationships, since students spend more time in school with their teachers when they progress in schools. In contrast, their after-class activities with parents may reduce so that the parents' evaluations become less accurate and predictive of children's reading performance.

As for background variables, SES is a potentially useful predictor of children's reading performance, particularly on grouping but not on transition. Meta-analysis (e.g., Hattie, 2009) showed that SES has a moderate impact ($d = 0.57$) on academic achievement. In the present research, we took SES as one of the important home background variables, used it as a controlling covariate, and showed that it had influence on grouping. It is logical, therefore, to pay greater attention to the reading development of students from lower SES background (e.g., Ladd et al., 1999; Jung, 2014).

LIMITATIONS AND FUTURE DIRECTIONS

One possible limitation is that we used a two-stage model to analyze the data. This was mainly decided from the general trajectory of the reading growth of the data and findings from earlier studies (Kaplan, 2005; Kapland, 2008; Palardy, 2010; Kieffer, 2012). However, the problem is that the interval of the stage (especially at Stage 2) is quite large with the time points of data collection being several years apart. There is a possibility, therefore, that students grow in discernible stages crossing a long period of time. If the intervals of the data collection had been much smaller, we would have been more confident to use the growth modeling within each stage. An alternative is to use the non-linear model to build the GMM (e.g. Grimm et al., 2010). But it requires demanding measures. Future studies could further explore the possible trajectories of reading development, identify the proper cutoff for each stage, and describe the most suitable trend within each stage.

We also notice that long-term effects and growth patterns are less well predicted by the social environmental covariates. These covariates may include the home and teachers' social environmental factors which generally have smaller effects than those of more direct variables such as teaching and school (for meta-analysis, see Hattie, 2009). One possible direction of the future study is, therefore, to focus on the short-term prediction of a set of more comprehensive social environmental factors from schools (teachers, peers, etc.) and families (parents, etc.). Another possibility is to treat the covariate as a time-varying variable in multilevel structure (Vermunt et al., 1999; Bartolucci et al., 2011). That means, in our analyses, the social rating recorded at Kindergarten, Spring Grade 1 Spring and Fall Grade 5 can all be treated as multiple indicators affecting the transition at different time points. Especially under the condition with a large interval of measuring time, time-varying measures would then produce more accurate prediction.

CONCLUSIONS

In summary, the study contributes in showing that: (i) the LTA-GMM fitted the data well; (ii) most of the children stayed in the same ability group with practically few cross-level class changes in the transition; (iii) children receiving higher teachers' ratings and with higher SES, and of above average poverty status, would have higher chance to transit into the higher ability level group. The findings supported the importance of the moderating effects of these social environmental facilitators on the patterns of children's reading development.

AUTHOR CONTRIBUTIONS

YL contributes the most to the article. HL is the corresponding author who organizes and helps conducting the analysis. KH helps a lot providing useful suggestions on modeling and revising the article.

FUNDING

The manuscript is supported by the National Natural Science Foundation of China (No. 31571152).

ACKNOWLEDGMENTS

We acknowledge the financial support from the Collaborative Innovation Center of Assessment toward Basic Education Quality at Beijing Normal University.

REFERENCES

Asparouhov, T., and Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using mplus. *Struct. Equat. Model.* 21, 329–341. doi: 10.1080/10705511.2014.915181

Bartolucci, F., Pennoni, F., and Vittadini, G. (2011). Assessment of school performance through a multilevel latent Markov Rasch model. *J. Educ. Behav. Stat.* 36, 491–522. doi: 10.3102/1076998610381396

Cho, S.-J., Cohen, A. S., and Bottge, B. (2013). Detecting intervention effects using a multilevel latent transition analysis with a mixture IRT model. *Psychometrika* 78, 576–600. doi: 10.1007/s11336-012-9314-0

Clay, M. M. (1977). *Reading: The patterning of Complex Behaviour*. Auckland: Heinemann Educational Books.

Clay, M. M. (2001). *Change Over Time in Children's Literacy Development*. Auckland: Heinemann Educational Books.

Cohen, E. G. (1997). "Understanding status problems: sources and consequences," in *Working for Equity in Heterogeneous Classrooms: Sociological Theory in Practice,* eds E. G. Cohen and R. A. Lotan (New York, NY: Teachers College Press), 61–76.

Collins, L. M., and Lanza, S. T. (2010). *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. Hoboken, NJ: John Wiley & Sons.

Davis, H. A. (2003). Conceptualizing the role and influence of student-teacher relationships on children's social and cognitive development. *Educ. Psychol.* 38, 207–234. doi: 10.1207/S15326985EP3804_2

de Weerth, C., van Geert, P., and Hoijtink, H. (1999). Intraindividual variability in infant behavior. *Dev. Psychol.* 35, 1102–1112. doi: 10.1037/0012-1649.35.4.1102

Ding, C., Richardson, L., and Schnell, T. (2013). A developmental perspective on word literacy from kindergarten through the second grade. *J. Educ. Res.* 106, 132–145. doi: 10.1080/00220671.2012.667009

Eyden, J., Robinson, E. J., and Einav, S. (2014). Children's trust in unexpected oral versus printed suggestions: limitations of the power of print. *Br. J. Dev. Psychol.* 32, 430–439. doi: 10.1111/bjdp.12054

Ferrer, E., Shaywitz, B. A., Holahan, J. M., Marchione, K. E., Michaels, R., and Shaywitz, S. E. (2015). Achievement gap in reading is present as early as first grade and persists through adolescence. *J. Pediatr.* 167, 1121–1125.e2. doi: 10.1016/j.jpeds.2015.07.045

Grimm, K. J., Ram, N., and Estabrook, R. (2010). Nonlinear structured growth mixture models in mplus and openmx. *Multivariate Behav. Res.* 45, 887–909. doi: 10.1080/00273171.2010.531230

Hattie, J. A. C. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. London: Routledge.

Herbert, J., and Stipek, D. (2005). The emergence of gender differences in children's perceptions of their academic competence. *J. Appl. Dev. Psychol.* 26, 276–295. doi: 10.1016/j.appdev.2005.02.007

Hollenstein, T. (2011). Twenty years of dynamic systems approaches to development: significant contributions, challenges, and future directions. *Child Dev. Perspect.* 5, 256–259. doi: 10.1111/j.1750-8606.2011.00210.x

Hughes, J. N., and Kwok, O.-M. (2007). Influence of student-teacher and parent-teacher relationships on lower achieving readers' engagement and achievement in the primary grades. *J. Educ. Psychol.* 99, 39–51. doi: 10.1037/0022-0663.99.1.39

Hughes, J. N., Luo, W., Kwok, O.-M., and Loyd, L. K. (2008). Teacher-student support, effortful engagement, and achievement: A 3-year longitudinal study. *J. Educ. Psychol.* 100, 1–14. doi: 10.1037/0022-0663.100.1.1

Iruka, I. U., Gardner-Neblett, N., Matthews, J., and Winn, D.-M. C. (2014). Preschool to kindergarten transition patterns for African American boys. *Early Child. Res. Q.* 29, 106–117. doi: 10.1016/j.ecresq.2013.11.004

Jung, E. (2014). Examining differences in kindergarteners' mathematics learning: a closer look at instruction, socioeconomic status, and race. *J. Educ. Res.* 107, 429–439. doi: 10.1080/00220671.2013.833074

Kainz, K., and Vernon-Feagans, L. (2007). The ecology of early reading development for children in poverty. *Elem. Sch. J.* 107, 407–427. doi: 10.1086/518621

Kaplan, D. (2002). Methodological advances in the analysis of individual growth with relevance to education policy. *Peabody J. Educ.* 77, 189–215. doi: 10.1207/S15327930PJE7704_9

Kaplan, D. (2005). Finite mixture dynamic regression modeling of panel data with implications for dynamic response analysis. *J. Educ. Behav. Stat.* 30, 169–187. doi: 10.3102/10769986030002169

Kapland, D. (2008). An overview of Markov chain methods for the study of stage-sequential developmental processes. *Dev. Psychol.* 44, 457–467. doi: 10.1037/0012-1649.44.2.457

Kendeou, P., Van den Broek, P., White, M. J., and Lynch, J. S. (2009). Predicting reading comprehension in early elementary school: the independent contributions of oral language and decoding skills. *J. Educ. Psychol.* 101, 765. doi: 10.1037/a0015956

Kieffer, M. J. (2012). Before and after third grade: Longitudinal evidence for the shifting role of socioeconomic status in reading growth. *Read. Writ.* 25, 1725–1746. doi: 10.1007/s11145-011-9339-2

Kuklinski, M. R., and Weinstein, R. S. (2001). Classroom and developmental differences in a path model of teacher expectancy effects. *Child Dev.* 72, 1554–1578. doi: 10.1111/1467-8624.00365

Ladd, G. W., Birch, S. H., and Buhs, E. S. (1999). Children's social and scholastic lives in kindergarten: related spheres of influence? *Child Dev.* 70, 1373–1400. doi: 10.1111/1467-8624.00101

Liu, Y., and Liu, H. (2015). "Piecewise growth mixture modeling: Language development of young children," in *Paper presented at 6th World Conference on Learning, Teaching and Educational Leadership* (Paris).

Longobardi, E., Rossi-Arnaud, C., and Spataro, P. (2011). A longitudinal examination of early communicative development: evidence from a parent-report questionnaire. *Br. J. Dev. Psychol.* 29, 572–592. doi: 10.1348/026151010X523473

Madon, S., Smith, A., Jussim, L., Russell, D. W., Eccles, J., Palumbo, P., et al. (2001). Am i as you see me or do you see me as i am? Self-fulfilling prophecies and self-verification. *Pers. Soc. Psychol. Bull.* 27, 1214–1224. doi: 10.1177/0146167201279013

Muter, V., Hulme, C., Snowling, M. J., and Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: evidence from a longitudinal study. *Dev. Psychol.* 40:665. doi: 10.1037/0012-1649.40.5.665

Muthén, B., Muthén, L., and Asparouhov, T. (2012). *Latent Variable Modeling Using Mplus*. Beijing: National Survey Reserach Center at Renmin University of China.

Muthén, L., and Muthén, B. (2012). *Mplus (Version 7.0)*. Los Angeles, CA: Muthén & Muthén.

Natale, K., Viljaranta, J., Lerkkanen, M. K., Poikkeus, A. M., and Nurmi, J. E. (2009). Cross-lagged associations between kindergarten teachers' causal attributions and children's task motivation and performance in reading. *Educ. Psychol.* 29, 603–619. doi: 10.1080/01443410903165912

Nylund, K. L., Asparouhov, T., and Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Struct. Equat. Model.* 14, 535–569. doi: 10.1080/10705510701575396

O'Connor, E., and McCartney, K. (2007). Examining teacher–child relationships and achievement as part of an ecological model of development. *Am. Educ. Res. J.* 44, 340–369. doi: 10.3102/0002831207302172

Oakhill, J. V., and Cain, K. (2012). The precursors of reading ability in young readers: evidence from a four-year longitudinal study. *Sci. Stud. Read.* 16, 91–121. doi: 10.1080/10888438.2010.529219

OECD (2013). *Pisa 2012 Results: Ready to Learn – Students' Engagement, Drive and Self-Beliefs, Vol. 3*, PISA: OECD Publishing. Available online at: http://dx.doi.org/10.1787/9789264201170-en

Palardy, G. J. (2010). The multilevel crossed random effects growth model for estimating teacher and school effects: issues and extensions. *Educ. Psychol. Meas.* 70, 401–419. doi: 10.1177/0013164409355693

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Read. Res. Q.* 40, 184–202. doi: 10.1598/RRQ.40.2.3

Paris, S. G. (2009). "Constrained reading skills—So what?" in *Paper Presented at the 58th Tearbook of the National Reading Conference*, Oak Creek, WI.

Paris, S. G., Carpenter, R. D., Paris, A. H., and Hamilton, E. E. (2005). "Spurious and genuine correlates of children's reading comprehension," in *Children's Reading Comprehension and Assessment*, eds S. G. Paris and S. A. Stahl (Mahwah, NJ: Lawrence Erlbaum Associates), 131–160.

Peugh, J., and Fan, X. (2012). How well does growth mixture modeling identify heterogeneous growth trajectories? A simulation study examining gmm's performance characteristics. *Struct. Equat. Model.* 19, 204–226. doi: 10.1080/10705511.2012.659618

Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R., and Morrison, F. J. (2008). Classroom effects on children's achievement trajectories in elementary school. *Am. Educ. Res. J.* 45, 365–397. doi: 10.3102/0002831207308230

Pianta, R. C., and Stuhlman, M. W. (2004). Teacher-child relationships and children's success in the first years of school. *Sch. Psych. Rev.* 33, 444–458.

Prochaska, J. O., and Velicer, W. F. (1997). The transtheoretical model of health behavior change. *Am. J. Health Promot.* 12, 38–48. doi: 10.4278/0890-1171-12.1.38

Quinn, J. M., Wagner, R. K., Petscher, Y., and Lopez, D. (2015). Developmental relations between vocabulary knowledge and reading comprehension: a latent change score modeling study. *Child Dev.* 86, 159–175. doi: 10.1111/cdev.12292

Robinson, B. F., and Mervis, C. B. (1999). Comparing productive vocabulary measures from the CDI and a systematic diary study. *J. Child Lang.* 26, 177–185. doi: 10.1017/S0305000998003663

Rodgers, E. M. (2004). Interactions that scaffold reading performance. *J. Lit. Res.* 36, 501–532. doi: 10.1207/s15548430jlr3604_4

Rosenholtz, S. J., and Simpson, C. (1984). The formation of ability conceptions: developmental trend or social construction? *Rev. Educ. Res.* 54, 31–63. doi: 10.3102/00346543054001031

Rytkönen, K., Aunola, K., and Nurmi, J. E. (2007). Do parents' causal attributions predict the accuracy and bias in their children's self-concept of maths ability? A longitudinal study. *Educ. Psychol.* 27, 771–788. doi: 10.1080/01443410701309316

Stahl, S. A. (1997). "Instructional models in reading: an introduction," in *Instructional Models in Reading*, eds S. A. Stahl and D. A. Hayes (Mahwah, NJ: Lawrence Erlbaum Associates), 1–29.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Read. Res. Q.* 21, 360–407. doi: 10.1598/RRQ.21.4.1

Tiedemann, J. (2000). Parents' gender stereotypes and teachers' beliefs as predictors of children's concept of their mathematical ability in elementary school. *J. Educ. Psychol.* 92:144. doi: 10.1037/0022-0663.92.1.144

Tofighi, D., and Enders, C. K. (2008). *Identifying the Correct Number of Classes in Growth Mixture Models, Vol. 13* (Charlotte, NC: Information Age Publishing Inc.).

Tourangeau, K., Nord, C., Lê, T., Sorongon, A., Najarian, M., and Hausken, E. (2009). *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), Combined User's Manual for the ECLS-K Eighth-Grade and K-8 Full Sample Data Files and Electronic Codebooks (NCES 2009-004).* Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Upadyaya, K., and Eccles, J. (2015). Do teachers' perceptions of children's math and reading related ability and effort predict children's self-concept of ability in math and reading? *Educ. Psychol.* 35, 110–127. doi: 10.1080/01443410.2014.915927

Upadyaya, K., Viljaranta, J., Lerkkanen, M.-K., Poikkeus, A.-M., and Nurmi, J.-E. (2012). Cross-lagged relations between kindergarten teachers' causal attributions, and children's interest value and performance in mathematics. *Soc. Psychol. Educ.* 15, 181–206. doi: 10.1007/s11218-011-9171-1

van Geert, P. (2003). "Dynamic systems approaches and modeling of developmental processes," in *Handbook of Developmental Psychology*, eds J. Valsiner and K. J. Connolly (Trowbridge: Sage), 640–672.

van Geert, P. (2011). The contribution of complex dynamic systems to development. *Child Dev. Perspect.* 5, 273–278. doi: 10.1111/j.1750-8606.2011.00197.x

van Geert, P., and Steenbeek, H. (2005). Explaining after by before: basic aspects of a dynamic systems approach to the study of development. *Dev. Rev.* 25, 408–442. doi: 10.1016/j.dr.2005.10.003

Vellutino, F. R., Tunmer, W. E., Jaccard, J. J., and Chen, R. (2007). Components of reading ability: multivariate evidence for a convergent skills model of reading development. *Sci. Stud. Read.* 11, 3–32. doi: 10.1080/10888430709336632

Verhoeven, L., van Leeuwe, J., and Vermeer, A. (2011). Vocabulary growth and reading development across the elementary school years. *Sci. Stud. Read.* 15, 8–25. doi: 10.1080/10888438.2011.536125

Vermunt, J. K. (2010). Latent class modeling with covariates: two improved three-step approaches. *Polit. Anal.* 18, 450–469. doi: 10.1093/pan/mpq025

Vermunt, J. K., Langeheine, R., and Bockenholt, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *J. Educ. Behav. Stat.* 24, 179–207. doi: 10.3102/10769986024002179

Vernon-Feagans, L., Odom, E., Pancsofar, N., and Kainz, K. (2008). "Comments on Farkas and Hibel: a transactional/ecological model of readiness and inequality," in *Disparities in School Readiness,* eds A. Booth and A. C. Crouter (New York, NY: Lawrence Erlbaum Associates), 61–78.

Votruba-Drzal, E., Li-Grining, C. P., and Maldonado-Carreño, C. (2008). A developmental perspective on full- versus part-day kindergarten and children's academic trajectories through fifth grade. *Child Dev.* 79, 957–978. doi: 10.1111/j.1467-8624.2008.01170.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 21 July 2016
doi: 10.3389/fpsyg.2016.01092

Does Exercise Improve Cognitive Performance? A Conservative Message from Lord's Paradox

Sicong Liu, Jean-Charles Lebeau and Gershon Tenenbaum*

Department of Educational Psychology and Learning System, Florida State University, Tallahassee, FL, USA

Although extant meta-analyses support the notion that exercise results in cognitive performance enhancement, methodology shortcomings are noted among primary evidence. The present study examined relevant randomized controlled trials (RCTs) published in the past 20 years (1996–2015) for methodological concerns arise from Lord's paradox. Our analysis revealed that RCTs supporting the positive effect of exercise on cognition are likely to include Type I Error(s). This result can be attributed to the use of gain score analysis on pretest-posttest data as well as the presence of control group superiority over the exercise group on baseline cognitive measures. To improve accuracy of causal inferences in this area, analysis of covariance on pretest-posttest data is recommended under the assumption of group equivalence. Important experimental procedures are discussed to maintain group equivalence.

Keywords: exercise intervention, cognition, gain score analysis, ANCOVA, experimental group equivalence, false positive error, review

Edited by:
Jason C. Immekus,
University of Louisville, USA

Reviewed by:
Evgueni Borokhovski,
Concordia University, Canada
Daniel Saverio John Costa,
University of Sydney, Australia

****Correspondence:***
Sicong Liu
64zone@gmail.com

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *26 October 2015*
Accepted: *06 July 2016*
Published: *21 July 2016*

Citation:
Liu S, Lebeau J-C and Tenenbaum G (2016) Does Exercise Improve Cognitive Performance? A Conservative Message from Lord's Paradox. Front. Psychol. 7:1092. doi: 10.3389/fpsyg.2016.01092

INTRODUCTION

Does exercise enhance cognitive functioning in human beings? Meta-analyses have provided support for the beneficial effect of exercise on cognitive performance with effect sizes (g) ranging from 0.097 for acute exercise (Chang et al., 2012) to 0.158 for chronic exercise (Smith et al., 2010). Additionally, some authors have reported on several underlying mechanisms by considering evidence from behavioral and psychophysiological studies (for a review, see Hillman et al., 2008). These arguments seem to offer convincing evidence that exercise results in cognitive performance enhancement. The present study takes a critical perspective on this conclusion by assessing methodological characteristics of relevant evidence.

The most relevant evidence comes from exercise-cognition randomized controlled trials (RCT). First, these RCTs are considered clinical trials. According to World Health Organization (2015, para. 3) and the International Committee of Medical Journal Editors (Laine et al., 2007, p. 275), a clinical trial "is any research study that prospectively assigns human participants or groups of humans to one or more health-related interventions to evaluate the effects on health outcomes." Second, RCT is generally regarded as the best design for testing causal relationship because it makes group equivalence likely on all covariates (Freedman et al., 2007; Torgerson, 2009).

Several Exercise-cognition RCTs' findings support the causal relationship between exercise and cognition. For example, Chang et al. (2012) reported a larger effect size from RCTs ($d = 0.19$) compared to those from either quasi-experimental or observational designs ($d = -0.02$ and $d = -0.14$, respectively). These results have led some authors to conclude that exercise benefits cognition in a population ranging from children to older adults. Although such message is exciting, as Rubin (1974) cautioned, the relevance of evidence to answering research questions is not solely determined by the choice of research design but many other factors. Guided by this message, we examined exercise-cognition RCTs published in the past 20 years for potential methodological shortcomings.

Why are Errors Possible

When analyzing pretest-posttest data from RCTs, researchers typically apply two group-comparison strategies to draw causal inferences: analysis of covariance and gain score analysis (Vickers and Altman, 2001; Van Breukelen, 2006). *Analysis of Covariance* (ANCOVA)[1] refers to the approach where posttest scores are compared between groups, adjusting for baseline scores (as covariates in the linear model). Assuming baseline group equivalence, *Analysis of Partial Variance* is a parallel of this strategy (Cohen et al., 2013). The alternative approach, *Gain Score Analysis* (GSA), considers the gain score (i.e., posttest minus pretest) as the criterion for group comparison. Forms of GSA include repeated-measures analysis of variance (RM ANOVA), gain score *t*-test, and ANOVA of gain score, among others. Researchers' choice between ANCOVA and GSA often leads to disparate conclusions, an inconsistency historically termed "Lord's Paradox" (Lord, 1967).

Lord's paradox generated a lasting research effort and a consensus was reached among methodologists. The consensus is that, as long as baseline group equivalence is likely by randomization (such as in a RCT design), investigators should choose ANCOVA in drawing causal conclusions, because ANCOVA has a higher testing power and unbiased effect estimate compared to GSA (Cronbach and Furby, 1970; Huck and McLean, 1975; Holland and Rubin, 1983; Miller and Chapman, 2001; Senn, 2006; Van Breukelen, 2006). However, when baseline group equivalence is unlikely (such as in a quasi-experimental design), none of the statistical procedures enables to "control for" such a flaw, and thus no causal inferences should be attempted (Campbell and Stanley, 1963; Lord, 1967; Cronbach and Furby, 1970; Meehl, 1970; Senn, 2006; Van Breukelen, 2006). To reiterate previous points with an analogy, perfect dishes ("causal inferences") come from fresh raw food ("baseline group equivalence") and skillful cooking ("ANCOVA"), whereas no perfect dishes can be made from non-fresh food ("baseline group non-equivalence") irrespective of how skillful the cook is.

Given Lord's paradox conclusion, strong evidence for causal inferences can be obtained only if (a) baseline group equivalence is likely, and (b) pretest-posttest data are analyzed using ANCOVA. In practice, researchers never know with certainty that a given RCT has baseline group equivalence, but they can ascertain baseline group non-equivalence when group baseline measures show statistical differences. Assuming that baseline group equivalence is achieved by identifying no baseline group differences on any baseline measures (which is a likely portrait of a given RCT, at least on baseline measures statistically tested), researchers should choose ANCOVA over GSA when comparing groups.

One advantage of ANCOVA over GSA is an increased power. Originally, ANCOVA was not developed to "control" for anything but to enhance the testing power of independent variables (Miller and Chapman, 2001). For instance, assuming identical within-group variance between pretest and posttest, Van Breukelen (2006) quantified that ANCOVA requires only 75% of the sample size of ANOVA of gain score (i.e., one form of GSA) to detect the same effect when the pretest-posttest correlation is 0.50. The other advantage of ANCOVA over GSA has to do with effect estimate accuracy. Specifically, ANCOVA produces the unbiased effect estimate, whereas GSA can generate under- or over- estimated effect size depending on the situation of baseline group imbalance (Vickers and Altman, 2001).

Baseline group imbalance is the descriptive difference between groups on baseline measures. If an exercise-cognition RCT has only two groups (i.e., one control and one exercise group), the control group and the exercise group have an equal chance to perform better than the other descriptively on a cognitive task at baseline. The interpretation of "better" is task specific. For instance, a shorter reaction time (RT) is better in simple reaction time tasks (e.g., Stroop Color), whereas a larger value is better in time-limited memory tasks (e.g., Digit Symbol). If the control group has baseline superiority (*control-BS*) by having, for instance, a shorter RT than that of the exercise group on the Stroop Color task, the adoption of GSA will lead to an over-estimate of exercise's benefits on cognition. Conversely, baseline exercise group superiority (*exercise-BS*) will generate an under-estimated effect with the GSA method (Vickers and Altman, 2001).

Baseline measures are usually negatively correlated with gain scores (Cronbach and Furby, 1970; Knapp and Schafer, 2009), a phenomenon known as "regression to the mean" (Galton, 1886; Bland and Altman, 1994). In such instances, the bias due to GSA's failure to account for baseline group imbalance can be larger. As a consequence, the Type I error (i.e., false positive) from control-BS and Type II error (i.e., false negative) from exercise-BS are likely to happen when using GSA. For example, Bland and Altman (2011) reported that comparing a baseline with a follow-up separately in each group by using *t*-test (i.e., one form of GSA) could raise the actual alpha level to be as high as 0.50 when comparing two groups and 0.75 when comparing three groups, depending on the power of a specific test. To make things worse, Bland and Altman's results were based on one outcome measure. When an exercise-cognition RCT assesses the effect of exercise on multiple cognitive measures (which is often the case), the practice of having a presumable false positive threshold (e.g., $\alpha = 0.05$) could turn meaningless.

How to Test for Possible Errors

Rather than assessing the effect of exercise on cognition by considering potential moderators, a procedure common to meta-analytic studies, the focus of the present study was to determine whether exercise-cognition RCTs published in the past 20 years (1996–2015) involve false positives or false negatives due to GSA application in pretest-posttest data analysis. We provided

[1] In this paper, the key distinction between ANCOVA and GSA is how researchers use the baseline measure. Although researchers can choose variables (e.g., age) as covariates in testing group difference on gain scores, these analyses are not what we mean by ANCOVA here.

a simple test to achieve this goal. Because group assignment was random, one would expect an equal chance for control-BS and exercise-BS on a certain cognitive measure. In other words, across all RCTs in our review, we expect half RCTs to show control-BS and the other half to have exercise-BS. In terms of a probability distribution, if we assume that X represents the number of RCTs showing control-BS, we would expect the probability of observing X, P (X), to follow a binomial distribution:

$$P(X) \sim \text{Binomial}(n, k)$$

where n represents the total number of RCTs examined and k symbolizes the expected probability ($k = 0.5$) of getting control-BS in a given exercise-cognition RCT[2] . Similarly, if researchers select randomly between GSA and ANCOVA, we should expect the group comparison strategy to follow the same binomial distribution with the only difference being that X is representing the number of RCTs employing GSA.

In order to detect possible false positive and/or negative errors among exercise-cognition RCTs using GSA, we must check for independence between baseline group imbalance (i.e., control-BS vs. exercise-BS) an statistical significance test result (i.e., significant vs. non-significant). If baseline group imbalance were independent to statistical significance test result, we would expect X, representing the number of RCTs using GSA that showed control-BS, to continue following the binomial distribution when conditioned on statistical test result. Assuming that Y stands for the statistical test result that has two possible outcomes (i.e., significant or non-significant), we will have the following conditional binomial distribution:

$$\mathrm{P}(X|Y) \sim \text{Binomial}(n|Y,\ k)$$

where n is the total number of RCTs using GSA method and k still takes the value of 0.5.

To summarize, we had three hypotheses in the present study. First, we hypothesized that, among all the RCTs, half of them should demonstrate control-BS and the other half should show exercise-BS due to randomization. Second, we hypothesized that researchers, as a group, selected between GSA and ANCOVA without preference, and therefore half of the RCTs should employ GSA and the other half should use ANCOVA as a group-comparison strategy. Lastly, we hypothesized that, when GSA-RCTs are counted separately based on whether they are positive (i.e., include at least one significant finding) or negative (i.e., include no significant findings), more control-BS (than exercise-BS) GSA-RCTs should be found in positive GSA-RCTs, whereas more exercise-BS (than control-BS) GSA-RCTs should be found in negative GSA-RCTs.

METHODS

Literature Search and Inclusion Criteria

The second author (J.-C. L.) conducted a literature search in April and May 2015 using SPORTDiscus, Web of Science, and Google Scholar databases. The search strategy utilized the following key words within full documents: (*exercise* OR *physical activity*) AND (*cognition* OR *cognitive performance*) AND *randomized controlled trial*. A manual search of reference list from key studies (e.g., meta-analysis) was also performed. The first author (S. L.) screened studies by title and abstract, then by full documentation. Trial authors were contacted when required information was missing. In total, 38 RCTs were considered for coding. However, five articles were excluded because they were missing information and corresponding authors were unable to respond to our request by July 1, 2015. The final set of studies consisted of 33 exercise-cognition RCTs.

The following inclusion criteria were applied to the exercise-cognition RCTs: (a) studies were published between January1996 and May 2015, (b) randomization is evident at the individual level, (c) the design included pre- and post-intervention measures on cognitive tasks such as perception, intelligence, academic achievement, memory, executive function, and cognitive impairment, (d) exercise intervention focused on aerobic, resistance training, or a combination of both, (e) studies included a passive control (e.g., waiting list), an active control (that can have a cognitive, physical, or social focus), or a combination of both (see Scherder et al., 2005), and (f) group differences were tested on cognitive measures. If multiple exercise intensities were used within an RCT, we regarded the group receiving the highest intensity as the exercise group and compared it to the control group. For example, if an RCT has two exercise groups (e.g., participants exercising at 60 and 70% of their $\mathrm{VO}_{2\max}$) and a reading control group, the group exercising at 70% $\mathrm{VO}_{2\max}$ was selected as the treatment group and was compared to the control group. In addition, if the two exercise groups differed in exercise modality (i.e., aerobic training and resistance training), we compared each of these exercise groups to the control group, respectively, and the results were coded under a given RCT. Furthermore, if multiple interventions were included and at least one of the groups received an intervention focusing on elements other than exercise (e.g., cognitive training), only the exercise group was considered as a treatment group and was compared to the control group. Finally, if multiple follow-up measurements were available after the intervention period, we chose the immediate post-intervention measurement as the post-test measure. Details of the literature search and study selection were shown in a flowchart (**Figure 1**).

Coding and Reliability

The first two authors discussed and settled coding variables to be included in the coding sheet. One author (S. L.) independently coded all the studies. The coded variables focused on the information relevant to the focus of the study, which is to check potential Type I and Type II errors in exercise-cognition RCTs. Therefore, for every cognitive task, we coded the targeted cognitive process (e.g., executive functioning), baseline group imbalance (control-BS vs. exercise-BS), and statistical test result (significant vs. non-significant). Other key methodological information were also coded including (a) group-comparison strategy in pretest-posttest data analysis (ANCOVA vs. GSA), (b) the form of control (passive vs. active), (c) the presence or

[2] We chose k instead of p to avoid confusion later when reporting the probability of our hypothesis testing.

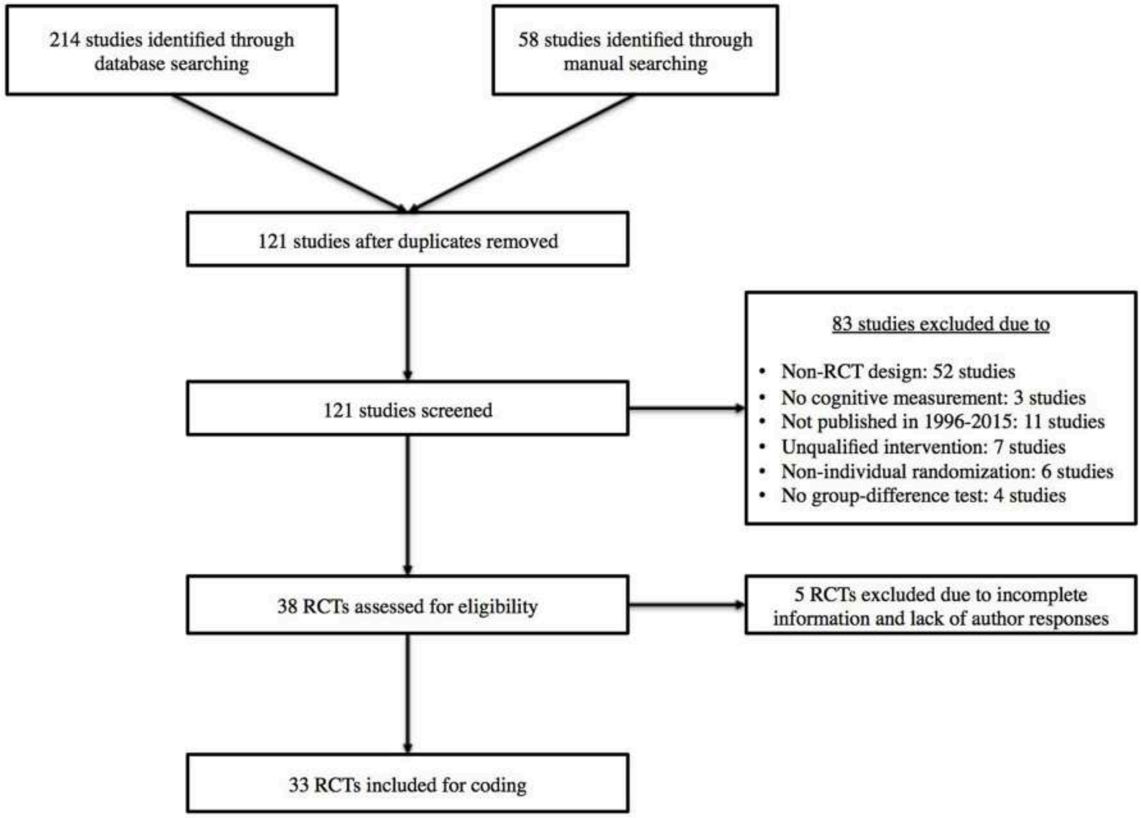

214 studies identified through database searching
58 studies identified through manual searching
121 studies after duplicates removed
121 studies screened
83 studies excluded due to
• Non-RCT design: 52 studies
• No cognitive measurement: 3 studies
• Not published in 1996-2015: 11 studies
• Unqualified intervention: 7 studies
• Non-individual randomization: 6 studies
• No group-difference test: 4 studies
38 RCTs assessed for eligibility
5 RCTs excluded due to incomplete information and lack of author responses
33 RCTs included for coding

FIGURE 1 | Flowchart of study selection.

absence of randomization procedure, (d) testing baseline group equivalence on cognitive measure(s), (e) the use of blinding procedures (i.e., single-, double-, or triple-blind), (f) explicit inclusion of intention-to-treat (ITT) analysis, (g) presence of *a priori* power analysis, (h) total participant number and number of groups (enabling participant number per group to be calculated), and (i) the presence or absence of pre-registering the trial. **Table 1** displays the coded information for each study included.

Eleven articles (33.3% of total) were randomly selected and separately coded to produce inter-coder reliability. A research assistant blinded to the study purposes completed the coding. Inter-rater reliability was calculated using Cohen's *Kappa* coefficient for each coding variable (**Table 2**). Following Landis and Koch's (1977) recommendations, we considered *Kappa* values between 0.61 and 0.80 as substantial and above 0.80 as very good. All the coded variables in the present study showed very good reliability. Coding discrepancies were resolved by re-visiting studies and discussion.

RCT Count and Statistical Analysis

We categorized and counted all the RCTs regarding their group-comparison strategy and baseline group imbalance. For group-comparison strategy, we categorized a given RCT into GSA-RCT if it used *gain scores* as the criterion in comparing groups. We classified an RCT as ANCOVA-RCT if the outcome variable was the post-test score while controlling for baseline score as covariate, or if analysis of partial variance was used.

Although we coded baseline group imbalance for every cognitive task within an RCT, we later counted the number of RCT regarding their baseline group imbalance favorableness (control-BS vs. exercise-BS). This ensured an equal weight for every RCT given their varying number of cognitive measures. For example, one RCT reported 42 cognitive measures but several RCTs reported only one cognitive measure. In this case, the 42-task RCT would be over-weighted if the count were made at the task level. We applied the "dominance rule" in judging whether a given RCT favors control-BS or exercise-BS. For example, if an RCT used four cognitive measures, we coded it as favoring control-BS if three of the four measures had better performing control group at baseline. Due to within-study measurement dependence, multiple cognitive measures tended to show homogeneous results with respect to baseline group imbalance. Among 33 RCTs, we applied the dominance rule to 14 RCTs. Two RCTs showed equal number of cognitive measures between control-BS and exercise-BS, and thus were dropped from the final count on baseline group imbalance.

We also made "conditional count" among GSA-RCTs. First, all the RCTs were screened for GSA employment. Then, GSA-RCTs were categorized as either positive (i.e., having at least one significant finding) or negative (i.e., having no significant

TABLE 1 | Study coding sequenced by group comparison strategy and study positivity.

Authors and Year	Grp. (T/C)	Sig.	Anal.	Control	Random	Test Base.	Blind	ITT	Power	*N* (Grp. #)	Prereg.
Williamson et al., 2009	C/C	N	ANCOVA	A-Cog.	N	N	Single	N	Y	102(2)	Y
Scherder et al., 2005	E/E	Y	ANCOVA	Both	N	Y	Single	N	N	43(3)	N
Lautenschlager et al., 2008	E/E	Y	ANCOVA	A-Cog.	Y	Y	Single	Y	Y	170(2)	Y
Liu-Ambrose et al., 2010	C/C	Y	ANCOVA	A-Phy.	Y	N	Single	Y	Y	155(3)	Y
Davis et al., 2011	E/E	Y	ANCOVA	P	N	N	Single	Y	Y	171(2)	Y
Nagamatsu et al., 2012	E/E	Y	ANCOVA	A-Phy.	N	N	Single	N	N	86(3)	Y
Okumiya et al., 1996	E/E	N	GSA	P	N	Y	Single	N	N	42(2)	N
Lemmink and Visscher, 2005	E/E	N	GSA	A-Cog.	N	N	N	N	N	16(2)	N
Foley et al., 2008	E/E	N	GSA	A-Phy.	N	Y	N	Y	N	20(2)	N
Krogh et al., 2009	E/E	N	GSA	A-Phy.	Y	N	Single	Y	N	165(3)	Y
Kimura et al., 2010	E/E	N	GSA	A-Cog.	N	Y	Single	N	N	171(2)	N
Varela et al., 2012	C/C	N	GSA	A-Mix	N	N	Single	Y	N	68(3)	N
Ruscheweyh et al., 2011	C/C	N	GSA	P	N	N	Single	N	N	62(3)	N
Linde and Alfermann, 2014	E/E	N	GSA	P	Y	Y	Single	Y	N	70(4)	N
Ruiz et al., 2015	E/E	N	GSA	A-Mix	N	Y	Single	Y	N	40(2)	N
Williams and Lord, 1997	E/E	Y	GSA	P	N	Y	N	N	N	187(2)	N
Emery et al., 1998	C/C	Y	GSA	P	Y	N	N	N	N	79(2)	N
Erickson et al., 2011	E/E	Y	GSA	A-Phy.	N	N	Single	N	N	120(2)	N
Bakken et al., 2001	C/C	Y	GSA	P	N	N	N	N	N	15(2)	N
Kramer et al., 2001	C/C	Y	GSA	A-Phy.	N	N	N	N	N	124(2)	N
Fabre et al., 2002	C/C	Y	GSA	A-Soc.	N	Y	N	N	N	32(4)	N
Netz et al., 2007	C/C	Y	GSA	A-Cog.	N	Y	Single	N	N	59(3)	N
Busse et al., 2008	C/C	Y	GSA	P	N	N	N	N	N	31(2)	N
Chang and Etnier, 2009	C/C	Y	GSA	A-Cog.	N	N	N	N	N	41(2)	N
Barella et al., 2010	E/C	Y	GSA	A-Soc.	N	N	N	N	N	40(2)	N
Muscari et al., 2010	C/C	Y	GSA	A-Cog.	N	Y	Single	Y	Y	120(2)	N
Ellemberg and St-Louis-Deschênes, 2010	N/N	Y	GSA	A-Cog.	N	N	N	N	N	72(2)	N
Kamijo et al., 2011	C/C	Y	GSA	P	N	N	N	N	N	43(2)	N
Chang et al., 2011	C/C	Y	GSA	A-Cog.	N	Y	N	N	Y	42(2)	N
Hopkins et al., 2012	C/C	Y	GSA	P	N	N	N	N	N	75(4)	N
Maki et al., 2012	E/E	Y	GSA	A-Cog.	N	Y	N	Y	N	150(2)	N
Liu-Ambrose et al., 2012	C/C	Y	GSA	A-Phy.	Y	N	Single	Y	Y	155(3)	Y
Hillman et al., 2014	N/C	Y	GSA	P	Y	N	Single	Y	Y	221(2)	Y

Year, Year of publication; Grp, (T/C), Baseline group imbalance (total count/conditional count); Sig., Study positivity (at least one significant test result identified by corresponding RCT); Anal., Group comparison strategy in pretest-posttest data analysis; Control, Form of control group; Random, Described random allocation procedures; Test Base, Tested baseline group equivalence on cognitive measures; Blind, Blinding procedures reported; ITT, Explicitly mentioned following intention-to-treat principle; Power, Performed a priori power analysis; N (Grp.), Total sample size (number of groups); Prereg., Pre-registered the trial. Liu-Ambrose et al. (2012) reported data dependence with Liu-Ambrose et al. (2010); E, Exercise-BS; C, Control-BS; Y, Yes; N, No; GSA, Gain score analysis; ANCOVA, Analysis of covariance; A-Cog., Active control with a cognitive focus; A-Phy., Active control with a physical focus; A-Soc., Active control with a social focus; A-Mix, Active control with more than one focus (e.g., cognitive and social); P, Passive control, Both, A control group consisting both actively and passively controlled participants; Single, Single blinding procedure (i.e., cognitive task assessors).

findings). The "conditional count" process was very similar to the previous count except that a RCT's baseline group imbalance was decided only on those cognitive measures fitting the positive/negative category. Specifically, if a GSA-RCT had at least one significant result (i.e., positive study), its baseline group imbalance was determined on all significant cognitive measures. If a GSA-RCT had no significant results (i.e., negative study), all its cognitive measures were included to determine its baseline group imbalance. These decisions were made for two reasons. First, some positive RCTs employed only one cognitive task (which reached statistical significance). Second, we could bias the negative RCT count regarding baseline group imbalance if we retained the non-significant measures from positive RCTs and recycled them in the negative RCT count.

During the "conditional count," we applied the dominance rule to only one GSA-RCT because it included one cognitive measure supporting control-BS and one cognitive measure with description-wise equal baseline between the control and exercise group; and thus it was counted as control-BS. In addition, one positive GSA-RCT reported a control-BS on one cognitive measure and exercise-BS on the other cognitive measure. This RCT was subsequently classified as neutral and was dropped from the final conditional count. We used the R version 3.2.0 (R Core Team, 2015) to estimate the probability of obtaining those counts based on continuity-corrected binomial distributions. Whereas the first two hypotheses had two-sided tests, the third hypothesis had one-sided test. The alpha level was set at 0.05.

RESULTS

Table 3 summarizes results pertaining to the first two hypotheses. The first hypothesis assumed that the occurrence of control-BS and exercise-BS are equally likely. Among all the RCTs ($n = 31$), we observed that 16 RCTs resulted in a control-BS and 15 RCTs in an exercise-BS (two RCTs were dropped in the count because they showed no clear favorableness between control-BS and exercise-BS). The probability of detecting this result met our expectation, $\hat{k} = 0.52$, $p = 0.99$, with a 95% CI of (0.33, 0.69). The second hypothesis assumed that the incidence of GSA and ANCOVA as a group comparison strategy are equal among RCTs. The count revealed 27 GSA-RCTs and 6 ANCOVA-RCTs. The test of such occurrence reached significance, $\hat{k} = 0.82$, $p < 0.001$, with a 95% CI of (0.64, 0.92). Therefore, we rejected the second hypothesis and concluded that researchers predominantly used GSA over ANCOVA in analyzing pretest-posttest data.

Table 4 displays results for the third hypothesis, which tested independence between baseline group imbalance and statistical significance test result among GSA-RCTs. Among positive GSA-RCTs ($n = 17$), 14 resulted in a control-BS and three in exercise-BS. This pattern reached significant level, $\hat{k} = 0.82$, $p = 0.006$, with a 95% CI of (0.60, 1.00). Among the negative GSA-RCTs ($n = 9$), two studies had a control-BS and seven had exercise-BS. This observation was not significant, $\hat{k} = 0.22$, $p = 0.09$, with a 95% CI of (0.00, 0.55). Thus, baseline group imbalance was related to statistical test in that more control-BS GSA-RCTs (which had over-estimated effect sizes) than exercise-BS GSA-RCTs resulted in significant results.

TABLE 2 | Kappa coefficients for coding variables.

Coding Variable	Kappa
Cognitive task	1.00
Baseline group imbalance (Control vs. Exercise)	0.92
Group difference results (significant vs. non-significant)	1.00
Group comparison strategy (GSA vs. ANCOVA)	0.85
Form of control	1.00
Description of randomization	1.00
Baseline group equivalence test on cognitive measures	1.00
Description of blinding	0.80
Intention-to-treat principle (ITT)	1.00
A priori power analysis	1.00
Total participant number and number of groups	1.00
Trial pre-registration	1.00

TABLE 3 | The probability of observed RCT counts regarding baseline group imbalance and group comparison strategy.

	Group ($N = 31$)		Strategy ($N = 33$)	
	Control	**Exercise**	**GSA**	**ANCOVA**
RCT Count	16	15	27	6
$\hat{k}$ (95% C.I.)	0.52 (0.33, 0.69)		0.82 (0.64, 0.92)	
p	0.99		<0.001	

Group, Baseline group imbalance; Control, Control-BS; Exercise, Exercise-BS; Strategy, Group-comparison strategy used in pretest-posttest data analysis; GSA, Gain score analysis; ANCOVA, Analysis of covariance.

TABLE 4 | The probability of observed conditional count on GSA-RCTs regarding baseline group imbalance.

	Positive ($n = 17$)		Negative ($n = 9$)	
	Control	**Exercise**	**Control**	**Exercise**
RCT Count	14	3	2	7
$\hat{k}$ (95% C.I.)	0.82 (0.60, 1.00)		0.22 (0.00,0.55)	
p	0.006		0.09	

Positive, GSA-RCTs identifying at least one significant finding; Negative, GSA-RCTs identifying no significant findings; Control, Control-BS; Exercise = Exercise-BS.

DISCUSSION

The objective of the present study was to determine whether exercise-cognition RCTs published in the past 20 years (1996–2015) include false positives or false negatives due to the ignorance of Lord's paradox (i.e., performing GSA in analyzing pretest-posttest data). Overall, several findings emerged from this study. First, baseline group superiority was found to be randomly determined among all the RCTs, with an equal probability of control-BS and exercise-BS. Second, GSA was the more popular group comparison strategy (27 RCTs) compared to ANCOVA (6 RCTs). Lastly, evidence suggested that positive GSA-RCTs were likely to include false positive errors because 82% (14 out of 17 studies) of them tested on over-estimated effect sizes. However, no clear evidence supported false negative errors among negative GSA-RCTs although a descriptive consistency was revealed.

Given findings that GSA is prevalent and misleading, it is necessary to re-emphasize the adoption of ANCOVA in pretest-posttest data analysis. The employment of ANCOVA could eliminate the biased effect estimate due to baseline group imbalance and increase testing power, thus reducing inferential errors. However, choosing ANCOVA as group comparison strategy is only half the story because ANCOVA enhances causal inferences only when group equivalence is likely. The other half, baseline group equivalence, depends on multiple factors during the experimental process. Some important factors are discussed next.

Randomization Procedures

One factor influencing group equivalence is randomization procedure. According to Schulz (1996), randomization consists of two stages: generation of unpredictable assignment sequence and concealment of that sequence until group allocation occurs. The first stage is related to the reliability of the randomizing tool (e.g., computer algorithm), and is often mistakenly identified as randomization itself. Consequently, sequence-concealment often receives insufficient attention, which introduces bias that emerges from the predictability of participant allocation. Ideally, the information on participant allocation should be revealed "as late as possible." As an example, Newell (1992) reported an anecdotal story of a surgeon who tosses a sterilized coin after a patient's abdomen was opened to decide which "treatment" he should perform. Although a little extreme, it highlights the importance of concealing participants' allocation information from experimenters. **Table 1** shows that only 7 out of 33 RCTs described randomization tools and even fewer RCTs described sequence-concealment procedures. In a couple of occasions, the randomization was done with imbalanced assignment ratio (e.g., 2:1 in assigning participants to exercise and control group, respectively) and no justifications were offered. Therefore, it is encouraged to report the randomization tool and to describe procedures for concealing the randomization sequence. In cases of imbalanced group assignment ratios, justifications are required.

Baseline Check

Prior to intervention, researchers must examine group equivalence on baseline measures. To foster such an examination, the CONSORT (Consolidated Standards of Reporting Trials) statement (Schulz et al., 2010) suggests reporting baseline data of demographic and clinical characteristics for each group. Concerning the CONSORT statement and the difficulty in conducting double-blind trials in exercise-cognition area, we recommend researchers to examine baseline group equivalence using both significance tests and subjective judgments. Baseline significance tests can alert researchers to factors interfering with randomization (e.g., no double-blinding); even when no significant group differences are identified at baseline, researchers must still review descriptive group imbalance on its size and prognostic strength (Altman, 1985). If meaningful group differences are found on any of the baseline measures (regardless of test significance), researchers could take different approaches in solving the problem, depending on how many baseline measures showed group differences. For instance, researchers can block participants when only few baseline measures (i.e., one or two) showed group differences in baseline check, or can re-randomize participants when more baseline variables exhibited group differences (Rubin, 2008).

Single-Blinding and Differential Expectation

Blinding procedure also affects group equivalence. When participants were assigned to either exercise or control group, it was challenging (if not impossible) to blind them to their respective interventions. In the present review, 18 out of the 33 RCTs reported blinding procedures and all of them were "single-blinded" (i.e., cognitive task assessors were blinded to participants' group assignment). No RCTs reported blinding participants to their group assignments. This raises the concern that participants may show differential expectations due to open group assignment. Such a possibility is consistent with the idea of "unmatched task" for the control group in the literature dealing with the effect of exercise on cognition (Brisswalter et al., 2002). The concern of differential expectation can also be evidenced by the diversity of control conditions in **Table 1**. This diversity reveals little agreement among researchers in speculating an active control for exercise intervention. To help select and/or design a good control, we recommend an empirical solution. That is, researchers should measure differential expectation. Although, preliminary effort has been made to survey differential group expectations prior to intervention (e.g., Stothart et al., 2014), we echoed Boot et al. (2013) in suggesting future research to consider testing differential expectation either during or after the intervention period. The optimal active control of exercise intervention must equate expectations on all these periods.

Intention-to-Treat Principle

Intention-to-Treat (ITT) is a widely accepted principle in analyzing clinical trials. ITT prevents group non-equivalence due to participant dropout (e.g., differential attrition) by including all the randomized participants in data analysis based on their intended treatment assignment (Gillings and Koch, 1991). The ideal situation for ITT would be having complete data for all the randomized participants (Hollis and Campbell, 1999). However, attrition is typically inevitable for clinical trials. In order to include participants with incomplete data into the analysis, missing values need to be handled. Some missing value imputation methods are available. For example, methods based on multiple imputation or maximum likelihood are generally recommended, but special considerations must be given to specific situations (Enders, 2010). However, no statistical methods can perfectly fix experimental flaws. When applying ITT, it is necessary to develop protocols (e.g., excluding likely exercise-intolerant participants before randomization) to ensure that participant adherence rate is roughly 80% or higher (Gillings and Koch, 1991; Montori and Guyatt, 2001). Regardless of adherence rate for a given RCT, a sensitivity test should always be performed to compare the ITT analysis results (as primary outcome) with the complete-case analysis results (Gillings and Koch, 1991). Compatible result of the sensitivity test precludes the concern of differential attrition, whereas incompatibility suggests this threat to internal validity. In short, future investigations are advised to include protocols that maximize adherence rate, to follow ITT principle, and to perform sensitivity analysis. Two other important elements of clinical trials are discussed next, although they do not affect group equivalence directly.

Power

Despite that no clear evidence of false negative errors was observed in the present study, it was still important to make sure that each RCT has sufficient power so that false negative errors could be minimized. Among all the RCTs included, only eight of 33 RCTs reported performing an *a priori* power analysis. Depending on the inputted parameters, the sample sizes varied among these RCTs. However, the average group size among the RCTs with *a priori* power analysis was about 65 participants, whereas the average group size for those not performing an a priori power analysis was about 32 participants[3]. It seems that a substantial proportion of exercise-cognition RCTs was underpowered, and thus could lead to false negative errors. It might be argued that 23 out of 33 included RCTs had at least one significant result, and thus false negative errors should not be a concern. However, 23 out of 33 RCTs having at least one positive result is not an evidence of sufficient power. First, we showed that false positive errors are likely to be included in those 17 positive GSA-RCTs, and by extension in the 23 positive RCTs. Second, as highlighted by Rubin (1974), a poorly implemented experiment can maintain many errors and ultimately be irrelevant to testing the research question. An experiment should follow optimal procedures (including *a priori* power analysis) for its conclusions to appropriately address research questions.

Researcher Degrees of Freedom and Trial Pre-registration

Although researchers are following the best paradigm including fixed set of practices, they still make decisions on quite some circumstances. These decision-calling circumstances are regarded as the *researcher degrees of freedom* (Simmons et al., 2011). It includes, among others, types of measure used in data collection, group-comparison strategies employed for data analysis, and type of data reported. When considering the researcher degrees of freedom with publication bias, an increased likelihood of Type I error would follow. For example, Gelman and Loken (2013) argued that data analysis strategies could be unwittingly conditioned on data patterns, which allow for false positive findings. To restrict researcher degrees of freedom by increasing clinical trial transparency, the International Committee of Medical Journal Editors (ICMJE) declared a trial's pre-registration as a condition for publishing in its 11 member journals in 2004 (De Angelis et al., 2004). ICMJE only recognizes registries meeting several criteria, including being free to public access, electronically searchable, open to all registrants, run by not-for-profit organization, as well as able to ensure validity of registration data by offering a mechanism. For example, www.clinicaltrials.gov maintained by the U.S. National Institute of Health is a qualified registry, even though many other registries have become available since 2004 (Humphreys et al., 2013) maintained by the U.S. National Institute of Health is a qualified registry, even though many other registries have become available since 2004 (Humphreys et al., 2013). It is by revealing critical trial information before participant enrollment that trial pre-registration combats researcher degrees of freedom. By pre-registering trials, researchers can still make changes afterwards as long as they offer good justifications. Although pre-registration has been the rule in clinical trial publication for almost 10 years (Laine et al., 2007), it is not true among exercise-cognition RCTs because only 8 out of 27 studies published in 2005 and later had trial pre-registration (**Table 1**). Therefore, we recommend future exercise-cognition RCTs to follow ICMJE's guidelines and make trial pre-registrations before enrolling participants.

Limitations

Several limitations in the present study are worth pointing out. First, we only focused on group comparison strategies in analyzing pretest-posttest data in exercise-cognition RCTs because it generates good evidence to evaluate the claim that exercise benefits cognition, and it is a design shared by all the exercise-cognition RCTs. Second, although ANCOVA should be used in analyzing pretest-posttest data in RCTs given group equivalence, it should be noted that ANCOVA was developed under several statistical assumptions, among which the assumption of homogeneity of regression slopes should receive particular attention (Miller and Chapman, 2001). However, these assumptions should not be used as an excuse to choose GSA against ANCOVA because GSA shares the same set of assumptions and because of ANCOVA's robustness and flexibility under assumption violation (Huck and McLean, 1975). Lastly, the counting process may have introduced bias in our conclusions, especially for the conditional count. We made the counts at trial level rather than at task level, and thus applied the "dominance rule" in order to maintain equal weight among exercise-cognition RCTs. Even though a better approach may be possible, evidence supported our decision. For example, we applied the "dominance rule" only to a minority of collected RCTs and the marginal count met the exact expectation from a probability point of view. Among the 33 RCTs, only two RCTs switched the group regarding baseline superiority between the marginal count and the conditional count.

CONCLUSION

Although exercise-cognition RCTs showed randomness of baseline group imbalance, RCTs adopting GSA as group comparison strategy were likely to have false positive errors and thus weakened the overall exercise-benefit-cognition claim. Future research will benefit from employing ANCOVA in analyzing pretest-posttest data while maintaining baseline group equivalence. Several suggestions have been offered to maintain baseline group equivalence in future research. It is likely that the results of current study are not limited to the effect of exercise on cognition and could potentially be extended to RCTs in other domains.

AUTHOR CONTRIBUTIONS

Conceived and designed the study: SL, JL. Searched publications: JL. Screened publications, coded data, and analyzed results: SL. Calculated inter-rater reliability: JL. Contributed to the writing of this manuscript: SL, JL, GT.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Yu-Kai Chang and Dr. Walter R. Boot for their review of the initial draft of this paper.

[3] This information was calculated based on the "*N* (Grp.)" column of **Table 1**.

REFERENCES

Altman, D. G. (1985). Comparability of randomised groups. *Statistician* 34, 125–136. doi: 10.2307/2987510

*Bakken, R. C., Carey, J. R., Di Fabio, R. P., Erlandson, T. J., Hake, J. L., and Intihar, T. W. (2001). Effect of aerobic exercise on tracking performance in elderly people: a pilot study. *Phys. Ther.* 81, 1870–1879.

*Barella, L. A., Etnier, J. L., and Chang, Y. K. (2010). The immediate and delayed effects of an acute bout of exercise on cognitive performance of healthy older adults. *J. Aging Phys. Act.* 18, 87–98.

Bland, J. M., and Altman, D. G. (1994). Regression towards the mean. *Br. Med. J.* 308:1499. doi: 10.1136/bmj.308.6942.1499

Bland, J. M., and Altman, D. G. (2011). Comparisons against baseline within randomised groups are often used and can be highly misleading. *Trials* 12, 1–7. doi: 10.1186/1745-6215-12-264

Boot, W. R., Simons, D. J., Stothart, C., and Stutts, C. (2013). The pervasive problem with placebos in psychology: why active control groups are not sufficient to rule out placebo effects. *Perspect. Psychol. Sci.* 8, 445–454. doi: 10.1177/1745691613491271

Brisswalter, J., Collardeau, M., and René, A. (2002). Effects of acute physical exercise characteristics on cognitive performance. *Sports Med.* 32, 555–566. doi: 10.2165/00007256-200232090-00002

*Busse, A. L., Magaldi, R. M., Coelho, V. A., Melo, A. C., Betoni, R. A., and Santarem, J. M. (2008). Effects of resistance training exercise on cognitive performance in elderly individuals with memory impairment: results of a controlled trial. *Einstein* 6, 402–407. doi: 10.1590/S1679-45082013000200003

Campbell, D. T., and Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research*. Chicago, IL: Rand McNally.

*Chang, Y. K., and Etnier, J. L. (2009). Effects of an acute bout of localized resistance exercise on cognitive performance in middle-aged adults: a randomized controlled trial study. *Psychol. Sport Exerc.* 10, 19–24. doi: 10.1016/j.psychsport.2008.05.004

Chang, Y. K., Labban, J. D., Gapin, J. I., and Etnier, J. L. (2012). The effects of acute exercise on cognitive performance: a meta-analysis. *Brain Res.* 1453, 87–101. doi: 10.1016/j.brainres.2012.02.068

*Chang, Y. K., Tsai, C. L., Hung, T. M., So, E. C., Chen, F. T., and Etnier, J. L. (2011). Effects of acute exercise on executive function: a study with a Tower of London task. *J. Sport Exerc. Psychol.* 33, 847–865.

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2013). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Mahwah, NJ: Routledge.

Cronbach, L. J., and Furby, L. (1970). How we should measure "change": or should we? *Psychol. Bull.* 74, 68–80. doi: 10.1037/h0029382

*Davis, C. L., Tomporowski, P. D., McDowell, J. E., Austin, B. P., Miller, P. H., Yanasak, N. E., et al. (2011). Exercise improves executive function and achievement and alters brain activation in overweight children: a randomized, controlled trial. *Health Psychol.* 30:91. doi: 10.1037/a0021766

De Angelis, C., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., et al. (2004). Clinical trial registration: a statement from the international committee of medical journal editors. *N. Engl. J. Med.* 351, 1250–1251. doi: 10.1056/NEJMe048225

*Ellemberg, D., and St-Louis-Deschênes, M. (2010). The effect of acute physical exercise on cognitive function during development. *Psychol. Sport Exerc.* 11, 122–126. doi: 10.1016/j.psychsport.2009.09.006

*Emery, C. F., Schein, R. L., Hauck, E. R., and MacIntyre, N. R. (1998). Psychological and cognitive outcomes of a randomized trial of exercise among patients with chronic obstructive pulmonary disease. *Health Psychol.* 17, 232–240. doi: 10.1037/0278-6133.17.3.232

Enders, C. (2010). *Applied Missing Data Analysis.* New York, NY: The Guilford Press.

*Erickson, K. I., Voss, M. W., Prakash, R. S., Basak, C., Szabo, A., Chaddock, L., et al. (2011). Exercise training increases size of hippocampus and improves memory. *Proc. Natl. Acad. Sci. U.S.A.* 108, 3017–3022. doi: 10.1073/pnas.1015950108

*Fabre, C., Chamari, K., Mucci, P., Massé-Biron, J., and Préfaut, C. (2002). Improvement of cognitive function by mental and/or individualized aerobic training in healthy elderly subjects. *Int. J. Sports Med.* 23, 415–421. doi: 10.1055/s-2002-33735

*Foley, L. S., Prapavessis, H., Osuch, E. A., De Pace, J. A., Murphy, B. A., and Podolinsky, N. J. (2008). An examination of potential mechanisms for exercise as a treatment for depression: a pilot study. *Ment. Health Phys. Act.* 1, 69–73. doi: 10.1016/j.mhpa.2008.07.001

Freedman, D., Pisani, R., and Purves, R. (2007). *Statistics, 4th Edn.* New York, NY: W.W. Norton and Company.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *J. Anthropol. Inst*. 15, 246–263. doi: 10.2307/2841583

Gelman, A., and Loken, E. (2013). *The Garden of Forking Paths: Why Multiple Comparisons Can be a Problem, Even When There is No "Fishing Expedition" or "p-Hacking" and the Research Hypothesis was Posited Ahead of Time.* Technical Report, Department of Statistics, Columbia University. Available online at: www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf (August 30, 2015).

Gillings, D., and Koch, G. (1991). The application of the principle of intention–to–treat to the analysis of clinical trials. *Drug Infect. J.* 25, 411–424. doi: 10.1177/009286159102500311

Hillman, C. H., Erickson, K. I., and Kramer, A. F. (2008). Be smart, exercise your heart: exercise effects on brain and cognition. *Nat. Rev. Neurosci.* 9, 58–65. doi: 10.1038/nrn2298

*Hillman, C. H., Pontifex, M. B., Castelli, D. M., Khan, N. A., Raine, L. B., Scudder, M. R., et al. (2014). Effects of the FITKids randomized controlled trial on executive control and brain function. *Pediatrics* 134, e1063–e1071. doi: 10.1542/peds.2013-3219

Holland, P. W., and Rubin, D. B. (1983). "On Lord's paradox," in *Principals of Modern Psychological Measurement,* eds H. Wainer and S. Messick (Hillsdale, NJ: Erlbaum), 3–25.

Hollis, S., and Campbell, F. (1999). What is meant by intention to treat analysis? Survey of published randomised controlled trials. *Br. Med. J.* 319, 670–674. doi: 10.1136/bmj.319.7211.670

*Hopkins, M. E., Davis, F. C., VanTieghem, M. R., Whalen, P. J., and Bucci, D. J. (2012). Differential effects of acute and regular physical exercise on cognition and affect. *Neuroscience* 215, 59–68. doi: 10.1016/j.neuroscience.2012.04.056

Huck, S. W., and McLean, R. A. (1975). Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: a potentially confusing task. *Psychol. Bull.* 82, 511–518. doi: 10.1037/h0076767

Humphreys, M., de la Sierra, R. S., and Van der Windt, P. (2013). Fishing, commitment, and communication: a proposal for comprehensive nonbinding research registration. *Polit. Anal.* 21, 1–20. doi: 10.1093/pan/mps021

*Kamijo, K., Pontifex, M. B., O'Leary, K. C., Scudder, M. R., Wu, C. T., Castelli, D. M., et al. (2011). The effects of an afterschool physical activity program on working memory in preadolescent children. *Dev. Sci.* 14, 1046–1058. doi: 10.1111/j.1467-7687.2011.01054.x

*Kimura, K., Obuchi, S., Arai, T., Nagasawa, H., Shiba, Y., Watanabe, S., et al. (2010). The influence of short-term strength training on health-related quality of life and executive cognitive function. *J. Physiol. Anthropol.* 29, 95–101. doi: 10.2114/jpa2.29.95

Knapp, T. R., and Schafer, W. D. (2009). From gain score t to ANCOVA F (and vice versa). *Pract. Assess. Res. Eval.* 14, 1–7.

*Kramer, A. F., Hahn, S., McAuley, E., Cohen, N. J., Banich, M. T., Harrison, C., et al. (2001). "Exercise, aging and cognition: Healthy body, healthy mind," in *Human Factors Interventions for the Health Care of Older Adults,* eds A. D. Fisk and W. Rogers (Hillsdale, NJ: Erlbaum), 91–120.

*Krogh, J., Saltin, B., Gluud, C., and Nordentoft, M. (2009). The DEMO trial: a randomized, parallel-group, observer-blinded clinical trial of strength versus aerobic versus relaxation training for patients with mild to moderate depression. *J. Clin. Psychiatry* 70, 790–800. doi: 10.4088/JCP.08m04241

Laine, C., Horton, R., DeAngelis, C. D., Drazen, J. M., Frizelle, F. A., Godlee, F., et al. (2007). Clinical trial registration: looking back and moving ahead. *N. Engl. J. Med.* 356, 2734–2736. doi: 10.1056/NEJMe078110

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310

*Lautenschlager, N. T., Cox, K. L., Flicker, L., Foster, J. K., van Bockxmeer, F. M., Xiao, J., et al. (2008). Effect of physical activity on cognitive function in older adults at risk for Alzheimer disease: a randomized trial. *JAMA* 300, 1027–1037. doi: 10.1001/jama.300.9.1027

*Lemmink, K. A., and Visscher, C. (2005). Effect of intermittent exercise on multiple-choice reaction times of soccer players. *Percept. Mot. Skills* 100, 85–95. doi: 10.2466/pms.100.1.85-95

*Linde, K., and Alfermann, D. (2014). Single versus combined cognitive and physical activity effects on fluid cognitive abilities of healthy older adults: a 4-month randomized controlled trial with follow-up. *J. Aging Phys. Act.* 22, 302–313. doi: 10.1123/JAPA.2012-0149

*Liu-Ambrose, T., Nagamatsu, L. S., Graf, P., Beattie, B. L., Ashe, M. C., and Handy, T. C. (2010). Resistance training and executive functions: a 12-month randomized controlled trial. *Arch. Intern. Med.* 170, 170–178. doi: 10.1001/archinternmed.2009.494

*Liu-Ambrose, T., Nagamatsu, L. S., Voss, M. W., Khan, K. M., and Handy, T. C. (2012). Resistance training and functional plasticity of the aging brain: a 12-month randomized controlled trial. *Neurobiol. Aging* 33, 1690–1698. doi: 10.1016/j.neurobiolaging.2011.05.010

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychol. Bull.* 68, 304–305. doi: 10.1037/h0025105

*Maki, Y., Ura, C., Yamaguchi, T., Murai, T., Isahai, M., Kaiho, A., et al. (2012). Effects of intervention using a community-based walking program for prevention of mental decline: a randomized controlled trial. *J. Am. Geriatr. Soc.* 60, 505–510. doi: 10.1111/j.1532-5415.2011.03838.x

Meehl, P. E. (1970). "Nuisance variables and the ex post facto design," in *Minnesota Studies in the Philosophy of Science, Vol. IV, Analyses of Theories and Methods of Physics and Psychology,* eds M. Radner and S. Winokur (Minneapolis, MN: University of Minnesota Press), 373–402.

Miller, G. A., and Chapman, J. P. (2001). Misunderstanding analysis of covariance. *J. Abnorm. Psychol.* 110, 40–48. doi: 10.1037/0021-843X.110.1.40

Montori, V. M., and Guyatt, G. H. (2001). Intention-to-treat principle. *Can. Med. Assoc. J.* 165, 1339–1341.

*Muscari, A., Giannoni, C., Pierpaoli, L., Berzigotti, A., Maietta, P., Foschi, E., et al. (2010). Chronic endurance exercise training prevents aging—related cognitive decline in healthy older adults: a randomized controlled trial. *Int. J. Geriatr. Psychiatry* 25, 1055–1064. doi: 10.1002/gps.2462

*Nagamatsu, L. S., Handy, T. C., Hsu, C. L., Voss, M., and Liu-Ambrose, T. (2012). Resistance training promotes cognitive and functional brain plasticity in seniors with probable mild cognitive impairment. *Arch. Intern. Med.* 172, 666–668. doi: 10.1001/archinternmed.2012.379

*Netz, Y., Tomer, R., Axelrad, S., Argov, E., and Inbar, O. (2007). The effect of a single aerobic training session on cognitive flexibility in late middle-aged adults. *Int. J. Sports Med.* 28, 82–87. doi: 10.1055/s-2006-924027

Newell, D. J. (1992). Intention-to-treat analysis: implications for quantitative and qualitative research. *Int. J. Epidemiol.* 21, 837–841. doi: 10.1093/ije/21.5.837

*Okumiya, K., Matsubayashi, K., Wada, T., Kimura, S., and Ozawa, T. (1996). Effects of exercise on neurobehavioral function in community-dwelling older people more than 75 years of age. *J. Am. Geriatr. Soc.* 44, 569–572. doi: 10.1111/j.1532-5415.1996.tb01444.x

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org/

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66, 688–701. doi: 10.1037/h0037350

Rubin, D. B. (2008). Comment: the design and analysis of gold standard randomized experiments. *J. Am. Stat. Assoc.* 103, 1350–1353. doi: 10.1198/016214508000001011

*Ruiz, J. R., Gil-Bea, F., Bustamante-Ara, N., Rodríguez-Romo, G., Fiuza-Luces, C., Serra-Rexach, J. A., et al. (2015). Resistance training does not have an effect on cognition or related serum biomarkers in nonagenarians: a randomized controlled trial. *Int. J. Sports Med.* 36, 54–60. doi: 10.1055/s-0034-1375693

*Ruscheweyh, R., Willemer, C., Krüger, K., Duning, T., Warnecke, T., Sommer, J., et al. (2011). Physical activity and memory functions: an interventional study. *Neurobiol. Aging* 32, 1304–1319. doi: 10.1016/j.neurobiolaging.2009.08.001

*Scherder, E. J., Van Paasschen, J., Deijen, J. B., Van Der Knokke, S., Orlebeke, J. F. K., Burgers, I., et al. (2005). Physical activity and executive functions in the elderly with mild cognitive impairment. *Aging Ment. Health* 9, 272–280. doi: 10.1080/13607860500089930

Schulz, K. F. (1996). Randomised trials, human nature, and reporting guidelines. *Lancet* 348, 596–598. doi: 10.1016/S0140-6736(96)01201-9

Schulz, K. F., Altman, D. G., and Moher, D. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC Med.* 8, 1–9. doi: 10.1016/j.ijsu.2010.09.006

Senn, S. (2006). Change from baseline and analysis of covariance revisited. *Stat. Med.* 25, 4334–4344. doi: 10.1002/sim.2682

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22, 1359–1366. doi: 10.1177/0956797611417632

Smith, P. J., Blumenthal, J. A., Hoffman, B. M., Cooper, H., Strauman, T. A., Welsh-Bohmer, K., et al. (2010). Aerobic exercise and neurocognitive performance: a meta-analytic review of randomized controlled trials. *Psychosom. Med.* 72, 239–252. doi: 10.1097/PSY.0b013e3181d14633

Stothart, C. R., Simons, D. J., Boot, W. R., and Kramer, A. F. (2014). Is the effect of aerobic exercise on cognition a placebo effect?. *PLoS ONE* 9:e109557. doi: 10.1371/journal.pone.0109557

Torgerson, C. J. (2009). Randomised controlled trials in education research: a case study of an individually randomised pragmatic trial. *Education* 3–13, 37, 313–321. doi: 10.1080/03004270903099918

Van Breukelen, G. J. (2006). ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *J. Clin. Epidemiol.* 59, 920–925. doi: 10.1016/j.jclinepi.2006.02.007

*Varela, S., Ayán, C., Cancela, J. M., and Martín, V. (2012). Effects of two different intensities of aerobic exercise on elderly people with mild cognitive impairment: a randomized pilot study. *Clin. Rehabil.* 26, 442–450. doi: 10.1177/0269215511425835

Vickers, A. J., and Altman, D. G. (2001). Analysing controlled trials with baseline and follow up measurements. *Br. Med. J.* 323, 1123–1124. doi: 10.1136/bmj.323.7321.1123

*Williams, P., and Lord, S. R. (1997). Effects of group exercise on cognitive functioning and mood in older women. *Aust. N. Z. J. Public Health* 21, 45–52. doi: 10.1111/j.1467-842X.1997.tb01653.x

*Williamson, J. D., Espeland, M., Kritchevsky, S. B., Newman, A. B., King, A. C., Pahor, M., et al. (2009). Changes in cognitive function in a randomized trial of physical activity: results of the lifestyle interventions and independence for elders pilot study. *J. Gerontol. Series A Biol. Sci. Med. Sci.* 64A, 688–694. doi: 10.1093/gerona/glp014

World Health Organization (2015). *WHO.INT*. *International Clinical Trials Registry Platform*. Available online at: http://www.who.int/ictrp/en

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*References marked with an asterisk indicate studies included in **Table 1**.

ORIGINAL RESEARCH
published: 10 May 2016
doi: 10.3389/fpsyg.2016.00664

Incremental Validity and Informant Effect from a Multi-Method Perspective: Assessing Relations between Parental Acceptance and Children's Behavioral Problems

Eva Izquierdo-Sotorrío[1], Francisco P. Holgado-Tello[1,2] and Miguel Á. Carrasco[1]*

[1] Department of Personality, Assessment and Psychological Treatments, Faculty of Psychology, National University of Distance Education, Madrid, Spain, [2] Department of Behavioral Science Methodology, Faculty of Psychology, National University of Distance Education, Madrid, Spain

This study examines the relationships between perceived parental acceptance and children's behavioral problems (externalizing and internalizing) from a multi-informant perspective. Using mothers, fathers, and children as sources of information, we explore the informant effect and incremental validity. The sample was composed of 681 participants (227 children, 227 fathers, and 227 mothers). Children's (40% boys) ages ranged from 9 to 17 years ($M = 12.52$, $SD = 1.81$). Parents and children completed both the Parental Acceptance Rejection/Control Questionnaire (PARQ/Control) and the check list of the Achenbach System of Empirically Based Assessment (ASEBA). Statistical analyses were based on the correlated uniqueness multitrait-multimethod matrix (model MTMM) by structural equations and different hierarchical regression analyses. Results showed a significant informant effect and a different incremental validity related to which combination of sources was considered. A multi-informant perspective rather than a single one increased the predictive value. Our results suggest that mother–father or child–father combinations seem to be the best way to optimize the multi-informant method in order to predict children's behavioral problems based on perceived parental acceptance.

Keywords: incremental validity, multiple informants, parental acceptance-rejection, behavioral problems, children, hierarchical regression, structural equations models, informant effect

OPEN ACCESS

Edited by:
Jason C. Immekus,
University of Louisville, USA

Reviewed by:
Christian Wandeler,
California State University, Fresno, USA
Lourdes Ezpeleta,
Universitat Autònoma de Barcelona, Spain

****Correspondence:***
Eva Izquierdo-Sotorrío
eva.izq@cop.es

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *13 November 2015*
Accepted: *21 April 2016*
Published: *10 May 2016*

Citation:
Izquierdo-Sotorrío E, Holgado-Tello FP and Carrasco MÁ (2016) Incremental Validity and Informant Effect from a Multi-Method Perspective: Assessing Relations between Parental Acceptance and Children's Behavioral Problems. Front. Psychol. 7:664. doi: 10.3389/fpsyg.2016.00664

INTRODUCTION

The progress of psychology is inextricably linked to the development of new and more refined methods and strategies for measuring psychological concepts, models, and intervention programs (Eid and Diener, 2006). A multi-informant approach offers insights into scientific phenomena and can contribute to confirming psychological theories in a way that a single-informant approach cannot. Due to the complexity of constructs evaluated and developmental factors that take place in children's psychological adjustment, their assessment is mainly multimodal (e.g., rating scales, interviews, and observations), multi-informant (e.g., child, parents, teachers, and mates), and/or multi-trait (Eyde et al., 1993; Ollendick and Hersen, 1993; Mash and Terdal, 1997; Duhig et al., 2000; Meyer et al., 2001; Johnston and Murray, 2003; Achenbach, 2006;

Hunsley and Mash, 2007). Specifically for informant assessment, the most reliable source of information on a target's psychological characteristics is not to be found in his or her self-ratings, nor it is guaranteed by single informant ratings; rather, it is found in the combination of the judgments from the community of the target's knowledgeable informants. According to this, the multi-informant assessment is mostly accepted by the psychological assessment community as an adequate and useful procedure, since rarely is a unique measure sufficient for providing all the required information needed to form an accurate judgment (Meyer and Archer, 2001; Garb, 2003; De Los Reyes and Kazdin, 2004; Carrasco et al., 2008; Hughes and Gullone, 2010). However, informant effects represent bias that can derive from the use of the same source of information in the assessment of different traits, the knowledge of informants, the observability of assessed traits, the judgment of informants, or the social desirability, among other factors (Cheng and Furnham, 2004; Neyer, 2006). For these reasons, determining the extent to which an informant effect is affecting the assessment of constructs and its relations is an important goal in determining the real construct validity. Individual reports often yield inconsistent data and discrepancies that can create considerable uncertainties in designing interventions and drawing conclusions from research (Klein, 1991; Epkins, 1993; Jané et al., 2000; De Los Reyes and Kazdin, 2004, 2005, 2006; Achenbach, 2006; Goodman et al., 2010; De Los Reyes et al., 2015). For instance, associations between constructs tend to be largest: (*a*) when a single informant is used, because of shared method variance (Neyer, 2006); (*b*), when the assessment of interventions has a large effect on parent reports vs. observed child behaviors of children's externalizing problems (Tarver et al., 2014); or (*c*) when family members experience their interaction differently and therefore have dissimilar views on parenting and parent child relations (e.g., Lanz et al., 2001; Hoeve et al., 2009). A key reason for these uncertainties originates from the near-exclusive focus on mental health research as applied to whether informant discrepancies reflect measurement error or reporting biases (e.g., Richters, 1992; De Los Reyes, 2011). Consequently, what remains unclear is whether a multi-informant approach to assessment validly captures contextual variations displayed in children's behavioral problems or whether it instead reflects different perceptions or beliefs about what a symptom is, and, finally, which informants ought to be included in assessments of children and adolescents.

Regarding this last point, another important issue from a multi-informant approach is the differential contribution of a particular source of information in relation to others. That is, the incremental validity or degree to which adding a new informant to the assessment consistently increases the predictive power and decision making (Garb, 2003; Hunsley, 2003; Hunsley and Mash, 2005). Unfortunately, the incremental validity inherent in using and combining multiple assessment methods has not undergone wide empirical testing in the literature on either adult or child assessment (Mash and Terdal, 1997; Hunsley, 2002). Thus, strong psychometric properties of the individual measures are necessary but do not provide sufficient conditions to ensure the incremental validity of incorporating these measures into the assessment process. Furthermore, not only is the research that deals directly with incremental validity in child assessment relatively small, the incremental validity of mothers' vs. fathers' reports has seldom been tested (Johnston and Murray, 2003).

With regard to cross-informant use, some studies support the incremental value of adults' over children's information when externalizing problems are measured (Loeber et al., 1991; Carrasco et al., 2008). However, the use of adults' information in children's assessment does not always augment the value of using only one source of information (Biederman et al., 1990). On the other hand, for older children, when assessing internalizing problems or covert behaviors, there is some evidence for the incremental value of youth self-reports over parents reports (Langhinrichsen et al., 1990; Cantwell et al., 1997; Johnston and Murray, 2003).

One of the most consistent observations in the field of child assessment is the correspondence levels between informants' reports, which range from low to moderate in magnitude (Achenbach et al., 1987; Duhig et al., 2000; Achenbach, 2011; Markon et al., 2011; De Los Reyes et al., 2015). The evidence usually shows that pairs of informants who observed children in the same context (e.g., pairs of parents or pairs of teachers) tend to show greater levels of correspondence than pairs of informants who observed children in different contexts (e.g., parent and teacher). Accordingly, some studies have found that the cross-informant agreement was moderate to high between mother and father, and moderate to low between father–child and mother–child pairs (Grigorenko et al., 2010; Weitkamp et al., 2013). Correspondence between mothers and children tend to be higher than correspondence between fathers and children (Grigorenko et al., 2010) and mother–child reports tend to find a greater endorsement than father–child reports (Lapouse and Monk, 1958; Achenbach et al., 1987; Stanger and Lewis, 1993; De Los Reyes et al., 2015). Also, the confluence of informants' reports about children's externalizing problems (e.g., aggression and hyperactivity concerns) tends to be higher than that concerning internalizing problems (e.g., anxiety and depression). In this regard, maternal and paternal reports show moderate correspondence when rating internalizing behavior problems in children and a larger correspondence in ratings of externalizing behavior problems in children (Achenbach et al., 1987; Duhig et al., 2000; Grigorenko et al., 2010). This evidence may reflect the greater correspondence between reports of directly observable behaviors than internalized behaviors. There is also evidence supporting claims that the degree of acquaintance between parents and children is a factor that leads to different parental ratings (Hughes and Gullone, 2010). The variability of correspondence found between the different pairs of informants is probably reflective of both the potential informant effect and the differential contribution of each source of information to the assessment's target. Furthermore, we would like to remark that the variation of the responses will be due to real differences from individual subjects, and the variation of the subjects on the variable won't be a continuous uniform distribution, but its favorable or unfavorable position on the studied object will be according to their perception (Likert, 1932).

This study tries to explore from a multi-informant approach the relations between parental acceptance and children's internalizing and externalizing problems. Perceived parental acceptance is one of the main factors involved in children's psychological adjustment, as is shown from the interpersonal acceptance-rejection theory (IPARTheory; Rohner, 1986; Rohner et al., 2012). Parental rejection (the opposite of parental acceptance) implies the absence or a significant withdrawal of parental warmth, affection, care, comfort, concern, nurturance, support, or love, and the presence of a variety of physically and psychologically hurtful behaviors and effects (Rohner and Khaleque, 2005; Rohner et al., 2012). Meta-analysis studies on this subject have found that rejection has consistently negative effects on the psychological adjustment and behavioral functioning of both children and adults worldwide (Khaleque and Rohner, 2002; Rohner and Khaleque, 2005; Rohner et al., 2012). The same body of research also shows that children who perceive their parents as being rejecting tend to experience distress, and in turn develop a specific cluster of internalizing (i.e., emotional instability, depression) and externalizing (i.e., aggression, delinquency) problems (McLeod et al., 2007; Hoeve et al., 2009; Rohner and Khaleque, 2010; Khaleque and Rohner, 2012; Khaleque, 2015; Ramírez-Lucas et al., 2015). However, no studies from this perspective have been conducted, to our knowledge, that explore either the informant effect or the incremental validity of parents' and children's perceived parental acceptance on externalizing and internalizing behavioral problems. Accordingly, no specific results are expected and no particular hypotheses are going to be tested. The first aim of this study is to test for evidence of informant effects related to the links between parental acceptance and children's behavioral problems as measured by children, fathers, and mothers through a round-robin design, in which all informants rate all targets. The second aim is to explore the incremental validity of the informants. Specifically, we deal with two questions: (1) Are there significant informant effects predicting children's behavioral problems based on perceived parental acceptance? (2) What is the incremental validity of the children's perceived parental acceptance over the parent's perceived parental acceptance in predicting the children's behavioral problems?

MATERIALS AND METHODS

Sample

The sample was composed of 681 participants (227 children, 227 fathers, and 227 mothers). Children's (40% boys; $n = 90$) ages ranged from 9 to 17 years ($M = 12.52$, $SD = 1.81$): 37% ($n = 61$) were between 9 and11 years, 47% ($n = 107$) were between 12 and 13 years, 20% ($n = 46$) were between 14 and 15, and 6% ($n = 13$) were between 16 and 17 years.

All of the children attended school, the majority lived in two-parent households (91%), and the mean number of siblings was three. Of the parents, 88% of fathers and 70% of mothers were employed. Occupational titles for mothers and fathers (respectively) were: major professionals (17 and 17%), lesser professionals (40 and 33%), semi-skilled workers (18 and 26%), and unskilled workers (25 and 24%). The mothers' and fathers' education levels were: university studies (40 and 35%), high school studies (40 and 57%), and primary studies (20 and 8%).

This sample is part of a larger sample of a general study about parental acceptance and children's psychological adjustment in the Spanish population. Children were selected according to mother–father–child matched participation. This sample represents 22% of the total sample ($N = 1036$). The total sample was randomly selected from public schools and publically funded private schools in different cities and communities of Spain. The participation rate of the total families was 91.5%.

No significant differences were found between participant and non-participant families in the demographic variables (i.e., child's sex, age, and socioeconomic level).

Measures

All measures were filled in by children, mothers, and fathers using the appropriate versions of the instruments described below.

Parental Acceptance

Four versions of the *Parental Acceptance-Rejection/Control Questionnaire* were used to report on perceived parental acceptance, two for children (mother and father versions, one to report about each parent) and two for parents (one version for mothers and another version for fathers). Children filled in both mother and father versions (*Parental Acceptance-Rejection/Control Questionnaire,* Child PARQ/Control: *mother-short version for children* and Child PARQ/Control: *father -short version for children).* Mothers filled in mother versions and fathers filled in father versions (*Parental Acceptance-Rejection/Control Questionnaire*, PARQ/Control: Mother- *short version for parents* and, PARQ/Control: father- *short version for parents;* Rohner, 1990; Rohner and Khaleque, 2005; Spanish adaptation by Del Barrio et al., 2014). The short versions of the PARQ/Control for children and for parents consist of 29-item. The PARQ/Control for children is a self-reporting questionnaires with four scales measuring warmth/affection [e.g., "My mother (father) says nice things about me"], hostility/aggression [e.g., "My mother (father) gets angry at me easily"], indifference/neglect [e.g., "My mother (father) pays no attention to me"], and undifferentiated rejection [e.g., "My mother (father) does not really love me"], plus a parental control (permissive-strictness) scale built into it. The PARQ/Control for mothers and fathers are self-reports with the same scales as the version for children; the difference with the children version is that items ask about the mother or father her/himself (e.g., "I get angry at my son easily"). The mother and father versions of the PARQ/Control (short forms) are identical, with the exception of the title changing according to which parent is being assessed. In all versions items are scored on a 4-point Likert-type scale ranging from 4 (*almost always true*) through 1 (*almost never true*). The sum of the first four scales (24 items) constitutes a measure of overall perceived maternal and paternal acceptance/rejection (with the entire warmth scale reverse scored). A greater score indicates a perception of greater parental rejection. Evidence regarding the validity and reliability of the PARQ/Control has been very well supported (Khaleque and Rohner, 2002; Rohner and Khaleque, 2005). Coefficient alphas for the total score in this sample are 0.88 for fathers and

0.97 for mothers in the children versions; and 0.88 for fathers and 0.88 for mothers in the parent version.

Children's Behavioral Problems

Two versions from the Achenbach System Evidence Based Assessment (Achenbach and Rescorla, 2007) were used to report on the children's behavioral problems: one for children (YSR) and one for parents (CBCL). Fathers and mothers inform separately about the children's behavioral problems on the CBCL version. The *Youth Self-Report* (YSR) is composed of two parts, the first assessing various psychosocial skills and competences, and the second consisting of a check-list of 112 items assessing a large number of behavioral problems, which are aggregated into two broad dimensions: internalizing (anxiety/depression, withdrawal, somatic complaints) and externalizing (breaking rules, aggressive behavior) problems. The items are scored on a 3-point Likert-type scale with anchors of 0 (*not true*), 1 (*somewhat or sometimes true*), and 2 (*very true or often true*). The *Children's Behavioral Check List* (CBCL) is similar to YSR, with the exception of having one item more (113 "Other problems"). For the purpose of this study, we only use the check-lists and the two broad dimensions: externalizing and internalizing behavioral problems.

For this sample, the Cronbach's alphas were 0.75 for the internalizing scale, and 0.73 for the externalizing scale on the YSR version; 0.79 and 0.78 for the internalizing scale, and 0.80 and 0.77 for the externalizing scale on the father-CBCL and mother-CBCL, respectively.

Procedure

Once the cluster sample of schools was selected, an authorization from the school board and an informed consent form from each child's responsible guardian were collected. Participation was voluntary. The instruments were administered collectively to each school class group in their own classrooms by research personnel trained for this task.

To explore the potential informant effect, we started with the correlated uniqueness model MTMM (Multitrait-multimethod Matrix; Byrne, 1998). According to the correlated uniqueness model, if the different sources are adding systematic variability to the model, we should find significant correlations between errors of the dependent variables reported by the same informant. At the same time, no matter what the global fit of the model is, a significant increase in the model fit should be noted. Second, we used a different hierarchical regression analysis to determine the magnitude of the incremental validity.

Data was analyzed using LISREL 8.9 and SPSS version 20.0 for Windows (SPSS WIN).

Design and Variables

A round-robin design was employed, in which fathers, mothers, and children separately completed all the instruments used. The independent variables were parental acceptance levels as perceived by children, mothers, and fathers. The dependent variables were children's externalizing and internalizing problems, reported separately by fathers, mothers, and children.

Results

In **Table 1** is included the correlation matrix among the variables used. According to the Multitrait-Multimethod matrix logit, if any informant effect exists the Monosource-Multitrait correlation should be higher than the Multisource-Multitrait one. If we focus on the dependent variables, we observe that the correlation between the internalizing and the externalizing problems informed by children ($r_{int-ext}$) is 0.54 (monosource-multitrait). This value is higher than other multisource-multitrait correlations such as $r_{\text{int-pext}} = 0.15$; $r_{\text{int-mext}} = 0.19$; $r_{\text{pint-pext}} = 0.12$; or $r_{\text{mint-ext}} = 0.02$. These results should take us to think about a possible informant effect. The same pattern is found in other variables. Thus, the correlation intra-informant for the same two variables is higher than the correlation inter-informants.

In order to obtain more evidences about the informant effect, we tested two models. In the first one (model 1), all the PARQ measures (PARQP, PARQM, MPARQ, and PPARQ) were predictors of all the criterion variables (INT, EXT, MINT, MEXT, PINT, and PEXT; **Figure 1**). The second model (model 2), was essentially the same, but included the correlations between the errors of each criterion variable reported by each informant (children, mothers, and fathers; **Figure 2**). We established that if we observed significant correlations between these errors in the second model, and the fit was improved, then it could be reasonable to think about an informant effect.

The fit indexes obtained for the first model were: $\chi^2 = 482.66$, df = 21; $p = 0.00$; CFI = 0.57; RMSEA = 0.30; GFI = 0.35; AGFI = 0.35; GFI = 0.75; RMR = 0.15. For model 2, we obtained: $\chi^2 = 236.01$, df = 18; $p = 0.00$; CFI = 0.79; RMSEA = 0.22; GFI = 0.86; AGFI = 0.56; RMR = 0.12.

Logically, in terms of fit indexes, both models are not necessarily accepted because we are not looking for a predictive model to explain the relationship between the variables. According to our premise, we should test whether the errors of the various criterion measures from the same informant are correlated. In this sense, model 2 improves the fit of the model 1 ($\Delta\chi^2 = 246.65$; Δdf = 3), and the correlations between the errors of the criterion variables reported by the same source of information are significant [$e_{\text{int\_ext}} = 0.43$, *Critical Ratio* (CR) = 9.74; $e_{\text{mint\_mext}} = 0.28$, CR = 5.31; $e_{\text{pint\_pext}} = 0.43$, CR = 10.69].

These results show a significant effect of the informant. As we can see in **Figure 2**, children and fathers are the informants that add more variability to the model; that is, the covariance of errors between children's internalizing and externalizing problems are higher when they are reported by fathers and by children than when they are reported by mothers. In order to quantify the magnitude of the contributions of the various informants, and their incremental validity, we conducted six hierarchical regression analyses.

The results from the hierarchical regression analyses are shown in **Table 2**. The contribution of perceived parental acceptance on behavioral problems is organized by the three informants (mothers, fathers, and children) and by the children's externalizing and internalizing problems.

TABLE 1 | Correlation matrix.

	PARQP	PARQM	EXT	INT	MPARQ	MEXT	MINT	PPARQ	PEXT	PINT
PARQP	–									
PARQM	0.56**	–								
EXT	0.40**	0.41**	–							
INT	0.23**	0.17*	0.54**	–						
MPARQ	0.30**	0.39**	0.23**	0.08	–					
MEXT	0.30**	0.23**	0.34**	0.19**	0.34**	–				
MINT	0.10	0.09	0.02	0.17*	0.19**	0.36**	–			
PPARQ	0.38**	0.24**	0.23**	0.14*	0.38**	0.27**	0.17*	–		
PEXT	0.30**	0.20**	0.29**	0.15*	0.34**	0.75**	0.27**	0.27**	–	
PINT	0.20**	0.14*	0.12	0.13*	0.19**	0.39**	0.48**	0.18**	0.58**	–
Mean	35.48	33.00	13.49	17.33	36.30	5.12	6.73	36.91	4.69	5.45
SD	8.71	8.96	9.70	10.38	4.79	4.85	8.09	6.18	4.54	5.07

*Ext. Prob., Externalizing problems; Int. Prob., Internalizing problems; Pac, paternal acceptance; Mac, maternal acceptance. *p < 0.05, **p < 0.01.*

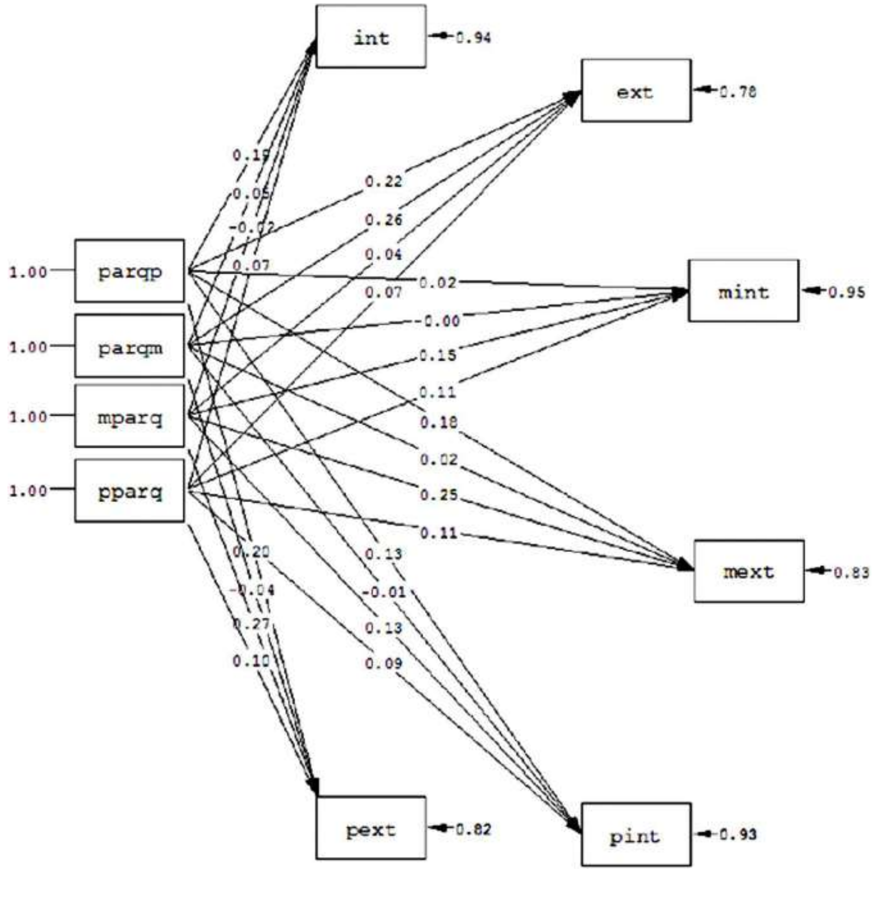


FIGURE 1 | Parental acceptance predicting children's behavioral problems from a multi-informant method with uncorrelated errors (Model 1). Parqp, paternal acceptance reported by children; parqm, maternal acceptance reported by children; mparq, maternal acceptance reported by mothers; pparq, paternal acceptance reported by fathers; int, internalizing problems reported by children; ext, externalizing problems reported by children; mint, internalizing problems reported by mothers; mext, externalizing problems reported by mothers; pext, externalizing problems reported by fathers; pint, internalizing problems reported by fathers.

When the informant referencing the child's behavioral problems is the mother, the maternal acceptance reported by mothers shows the largest increment of R^2, especially for externalizing problems. However, paternal acceptance reported by fathers made a significant contribution to externalizing problems (not internalizing), and maternal acceptance reported by mothers made a significant contribution to both internalizing and externalizing behavioral problems. Parental acceptance

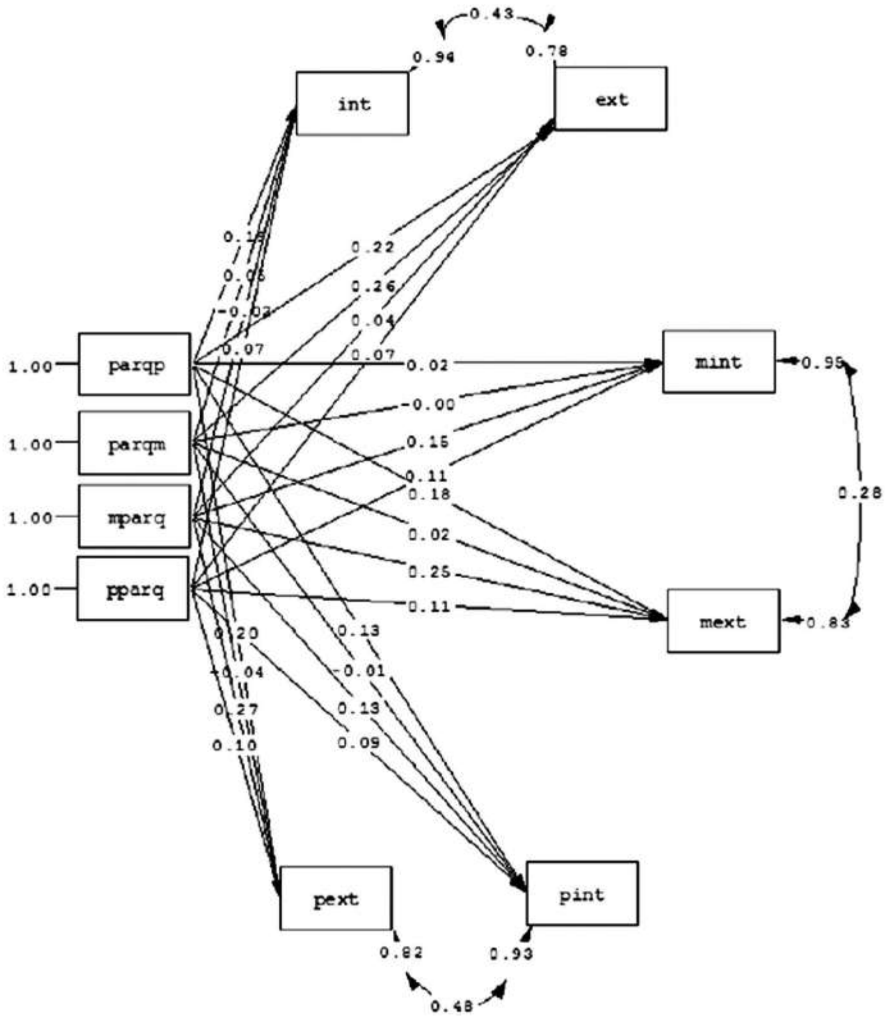


FIGURE 2 | Parental acceptance predicting children's behavioral problems from a multi-informant method with correlated errors (Model 2). Parqp, paternal acceptance reported by children; parqm, maternal acceptance reported by children; mparq, maternal acceptance reported by mothers; pparq, paternal acceptance reported by fathers; int, internalizing problems reported by children; ext, externalizing problems reported by children; mint, internalizing problems reported by mothers; mext, externalizing problems reported by mothers; pext, externalizing problems reported by fathers; pint, internalizing problems reported by fathers.

(maternal or paternal) perceived by children does not make any significant contribution to behavioral problems. Parental acceptance reported by fathers and maternal acceptance reported by mothers considered together become to explain 19% of the variance on externalizing problems.

When the informant referencing the child's behavioral problems is the father, the same pattern was found, with the exception of the instance of externalizing problems seen in step 4, wherein children make a significant contribution. Parental acceptance reported by fathers, mothers, and children considered together become to explain the 40% of the variance on externalizing problems.

Finally, when the informant referencing the child's behavioral problems is the child, the largest increase occurs in step 4, when children report on parental acceptance. Nevertheless, both paternal and maternal acceptances were significant predictors of externalizing problems (not internalizing problems), while only paternal acceptance was significant for internalizing problems. The increase in the variance explained by the parental acceptance perceived by children is 13% for externalizing problems and 4% for internalizing. Parental acceptance reported by fathers and children (the significant sources of information) considered together become to explain the 11% of the variance on externalizing problems and 14% on internalizing problems.

DISCUSSION

Method effects and incremental validity are two important issues for construct validity. The analysis of empirical similarities and differences between self and others as informants contribute to the knowledge of consistency of measures, its reliability

TABLE 2 | Hierarchical regression analyses predicting children's behavioral problems by multi-informants.

	Mother informant				Father informant				Child informant			
	Externalizing problems		Internalizing problems		Externalizing problems		Internalizing problems		Externalizing problems		Internalizing problems	
	β	**$R^2/\Delta R^2$**	**β**	**$R^2/\Delta R^2$**	**β**	**$R^2/\Delta R^2$**	**β**	**$R^2/\Delta R^2$**	**β**	**$R^2/\Delta R^2$**	**β**	**$R^2/\Delta R^2$**
Step 1		$R^2 = 0.02$		$R^2 = 0.00$		$R^2 = 0.02$		$R^2 = 0.01$		$R^2 = 0.08$		$R^2 = 0.03$
Age	−0.04	$\Delta R^2 = 0.02$	−0.07	$\Delta R^2 = 0.00$	−0.05	$\Delta R^2 = 0.02$	−0.01	$\Delta R^2 = 0.01$	−0.13*	$\Delta R^2 = 0.08$**	0.05	$\Delta R^2 = 0.03$*
Sex	0.13		0.05		0.13*		0.11		0.24**		0.16*	
Step 2		$R^2 = 0.06$		$R^2 = 0.02$		$R^2 = 0.08$		$R^2 = 0.03$		$R^2 = 0.10$		$R^2 = 0.05$
Pac by Father	0.20**	$\Delta R^2 = 0.04$**	0.13	$\Delta R^2 = 0.01$	0.24**	$\Delta R^2 = 0.06$**	0.13	$\Delta R^2 = 0.01$	0.17**	$\Delta R^2 = 0.02$**	0.13*	$\Delta R^2 = 0.02$*
Step 3		$R^2 = 0.13$		$R^2 = 0.04$		$R^2 = 0.15$		$R^2 = 0.05$		$R^2 = 0.12$		$R^2 = 0.05$
Mac by Mother	0.29**	$\Delta R^2 = 0.07$**	0.15*	$\Delta R^2 = 0.02$*	0.29**	$\Delta R^2 = 0.07$**	0.18*	$\Delta R^2 = 0.02$*	0.13	$\Delta R^2 = 0.02$*	0.00	$\Delta R^2 = 0.00$
Step 4		$R^2 = 0.15$		$R^2 = 0.04$		$R^2 = 0.17$		$R^2 = 0.07$		$R^2 = 0.01$		$R^2 = 0.09$
Pac by child	0.14	$\Delta R^2 = 0.02$	0.02	$\Delta R^2 = 0.00$	0.19*	$\Delta R^2 = 0.02$*	0.14	$\Delta R^2 = 0.01$	0.17*	$\Delta R^2 = 0.13$**	0.16*	$\Delta R^2 = 0.04$**
Mac by child	0.03		0.00		0.04		0.00		0.27**		0.08	

*Pac, paternal acceptance; Mac, maternal acceptance. *$p < 0.05$, **$p < 0.01$.*

and accuracy, and its validity in terms of behavior prediction (Kenny, 1994; Neyer, 2006). This study dealt with two questions: (1) Are there significant informant effects predicting children's behavioral problems from perceived parental acceptance? (2) What is the incremental validity of children's perceived parental acceptance over parents' perceived parental acceptance in predicting children's behavioral problems?

In relation to the first question, our findings confirm a significant informant effect, which shows that the predictive values are different from one informant to the others when predicting behavioral problems in children based on perceived parental acceptance. Consequently, the magnitude of relations in terms of behavior prediction between parental acceptance and children's externalizing and internalizing problems depends on the source of information used (i.e., children, mothers, or fathers). When the informant speaking on the child's behavioral problems is the mother, maternal acceptance perceived by mothers and paternal acceptance perceived by fathers are the best predictors of children's externalizing problems, while the best predictor for internalizing problems is only the maternal acceptance informed by mothers. The information provided by children about parental acceptance does not make any contribution to the behavioral problems reported upon by mothers. Likewise, the same pattern emerges when the informant about the child's behavioral problems is the father, except that children make a significant contribution to informing on externalizing problems (not internalizing). However, when children act as informants on their own behavioral problems, the pattern found is completely different; maternal acceptance as assessed by mothers does not make any contribution to the children's behavioral problems. Only paternal acceptance reported by fathers or children predicts the externalizing and internalizing problems; additionally, maternal acceptance reported by children predicts internalizing (not externalizing) problems.

The significant predictive value of perceived parental acceptance and children's psychological adjustment is very well supported in family research (Khaleque and Rohner, 2012; Rohner et al., 2012), but no studies have been conducted to explore the informant effect of parental acceptance on children's behavioral problems. Our results support this significant relation regardless of the source of information. Furthermore, our findings are consistent with previous studies that have found an informant effect reflected on the low or moderate confluence between children and parents on the information given by each of them (Achenbach et al., 1987; Rescorla et al., 2013; De Los Reyes et al., 2015). There are numerous prospective reasons for these results, such as the potential biased perception of informants (i.e., parents tending to perceive and inform about less or more problems than children), the information that informants use to rate the scales (i.e., family and school), conceptions of what constitutes abnormal behavior (Richters, 1992), the informants' own emotional state (Chilcoat and Breslau, 1997; Najman et al., 2000; Berg-Nielsen et al., 2003), the closeness of parent–child relationships (Hughes and Gullone, 2010), or the observability of behaviors (De Los Reyes and Kazdin, 2005).

According to previous studies (Stanger et al., 1992; Duhig et al., 2000), our results support the different predictive utility that a multiaxial assessment approach may have in children's outcomes, specifically in predicting the children's externalizing and internalizing behavioral problems from the parental acceptance construct. In this regard, when parents report about the children's behavioral problems, both fathers (paternal acceptance) and mothers (maternal acceptance) tend to be the best informants to predict externalizing problems, while mothers (maternal acceptance) excel at predicting internalizing ones. However, when children report about their own behavioral problems, children (paternal acceptance to externalizing and internalizing problems, and maternal acceptance to internalizing ones) and fathers (paternal acceptance) tend to be the best informants to predict all kinds of children's behavioral problems.

Research does not yet allow us to make a conclusion about to what extent maternal or paternal acceptance will make a higher or lower contribution to children's psychological problems. Some studies suggest that maternal parenting is more strongly associated with children's emotional and behavioral problems than paternal parenting (Rosnati et al., 2007; Meunier et al., 2012), while other studies find that the opposite is true (Flouri and Buchanan, 2002; Khaleque and Rohner, 2011). Probably on the basis of this contribution differences could be the externalized–internalized nature of behavioral problems, as well as the informant effect. Accordingly, the greater contribution of maternal acceptance to the children's problems could be explained by the closeness of the mother–child relationship and by the fact that mothers tend to have more knowledge about the children's behavioral problems (mainly about the internalizing ones), possibly because mothers generally spend more time with their children than fathers (Renk et al., 2003; De Los Reyes and Kazdin, 2005), or because mothers could be perceived by their offspring to have higher interpersonal power and prestige than fathers (Carrasco et al., 2014). Paternal acceptance may become more relevant to externalizing problems than internalizing because of the nature of father–child relationships, which tend to be more focused on leisure activities (Torres et al., 2014) and goal-oriented behaviors (Leaper et al., 1998; Tenenbaum and Leaper, 2003). The informant effect that our study shows is consistent with the studies that found a higher contribution of paternal acceptance vs. maternal acceptance when the informants are children (Flouri and Buchanan, 2002; Bosco et al., 2003; Khaleque and Rohner, 2011) or teachers (Mattanah, 2001). Maternal parenting tends to be a stronger predictor of children's behavioral problems when parents are the source of information (Gryczkowski et al., 2010), but this is not always confirmed (Hoeve et al., 2009).

Regarding the second question concerning how incremental validity was also affected by the source of information on the children's behavioral problems, our results suggest that there are differential contributions of one source of information over the others and a subsequent incremental validity related to which combination of sources is considered. More specifically, when the informant about the child's behavioral problems is the mother, both father's and mother's information about parental acceptance increases the predictive validity for externalizing problems, but only the mother's information does this (maternal acceptance) for internalizing. However, when the informant about the child's behavioral problems is the father, then mothers, fathers, and children increase the predictive validity for externalizing problems. Nevertheless, only the mother's information about maternal acceptance has significant predictive value on internalizing problems. Finally, when the informant about the child's behavioral problems is the child, then mothers, children, and fathers increase the predictive validity for externalizing problems, but only fathers (not mothers) and children do this for internalizing problems. It is important to highlight that mothers have the higher incremental validity when parents (mothers or fathers) inform about children's problems, but that children make the larger contribution to incremental validity when they self-report about their own behavioral problems. These results support the children's ability to be introspective and to assess their own thoughts and feelings even better than adults (Bidaut-Russell et al., 1995; Johnston and Murray, 2003). These results are also consistent with the studies that support the incremental value of adult informants compared with the child's reports on externalizing problems (Loeber et al., 1990, 1991).

Furthermore, our results support that single informants (parents or children) produced significantly stronger effects than multiple informants (parents and children). That is, when the same informant provides information about parental acceptance (predictor) and the children's outcomes (dependent variable), this single informant tends to reach the higher incremental validity. It is probably due to shared method variance (Campbell and Fiske, 1959). This effect may be particularly prominent when children are the source of information. Although asking children to report on parenting and their own behavioral problems can lead to inflated effect size estimates, children could provide the best information about themselves and the perceived parent–child relationships. The higher incremental validity of mothers on children's internalizing problems is consistent with the higher predictive value of maternal acceptance on internalizing behaviors, as previously discussed.

When fathers are the source of information, the rest of the informants (children and mothers) add significant incremental validity. This could be because fathers sometimes have less knowledge of children's day-to-day lives, meaning that more information is needed from mothers and children to predict children's behavioral problems. However, when children are the source of information, the incremental validity is mainly added by fathers. This may be because of overlapped information from mothers and children, as these would share more information about the emotional lives of the children. It is consistent with the higher agreement between mothers and children than between fathers and children (Schneewind and Ruppert, 2013; Leung and Shek, 2014). The closer relationship of mother and child can account for a higher concurrence on the information provided by these informants, and therefore, the parent with a closer relationship will give much redundant information when added to the one given by the child. In cultures like that of Spain, where gender and parental roles are still quite differentiated, it is common for mothers to spend more time than fathers with the children, which could be a reason why the mother does not add significant information when the child is used as the primary informant. Similarly, when the mother is the primary informant, the child does not add additional significant information.

Considering all these results as a whole, it can be concluded that the child is the best source of information about parental acceptance when we are trying to predict the children's behavioral problems (both externalizing and internalizing) reported by the own child. However, when the behavioral problems are informed by the parents, the parental acceptance information provided by them will be the data with better predictive value for children's externalizing problems. This changes when we deal with children's internalizing problems that are reported by the parents, in which case the mother's information will be the most predictive one.

A few limitations should be considered for future lines of research. First, this study focused on the general population instead of a clinical sample, meaning that generalization of the current findings to clinical populations should be made with caution, and future research should consider how these two samples may differ both quantitatively and qualitatively. Second, the lack of analysis by sex and age as moderators may be particularly relevant (Crick and Grotpeter, 1995; Johnston and Murray, 2003; Hughes and Gullone, 2010) in terms of informant effect and incremental validity. Studies about sex and age differences in the perception of parental acceptance and the expression of internalizing or externalizing problems symptoms may lead to variations in informant agreement and in relationships between parental acceptance and children's symptoms. Third, the parent's social desirability could minimize their reports about any adverse parenting experiences (i.e., rejection) affecting the level of parent–child agreement. Four, different methods of evaluation such as observations, rating scales, and self-reports should be explored in addition to the informant method. Future studies conducted from a developmental and gender perspective with a multi-measure perspective and using clinical samples are advised in order to bring more light to the informant effect and incremental validity.

Despite the above limitations, the findings of the present study have important practical implications. Considering previous analysis, a multi-informant perspective rather than a single should be considered in order to increase the predictive value and the incremental validity when we try to predict children's internalizing and externalizing problems. Our results suggest that mother–father or child–father informant pairs seem to be the way to optimize the combinations of sources of information in order to predict children's behavioral problems from parental acceptance. Nevertheless, a child may give enough information to make future decisions, and if we have to add only one informant to the assessment, this should be the father. There is a clear need for more research from a multi-method perspective in the child assessment field, rather than having blind faith in a "more are better" approach to getting informants (Johnston and Murray, 2003), which will lead to an optimization of empirically based children's assessment (Carrasco et al., 2008).

AUTHOR CONTRIBUTIONS

The tasks of each individual author are described in the folloing lines. EI-S: Bibliographic review, preparation of data matrices, drafting the theoretical contents, drafting the discusion, writing and preparing manuscript for sending. FH-T: Collection of data, statistical analysis, drafting the methodology and results. MC: Collection of data, statistical analysis, drafting the methodology and results, theoretical contents review, team coordination.

ACKNOWLEDGMENTS

This research is included in the Project PSI2011-28925 and it is supported by grants from the Spanish Government, *Ministerio de Ciencia e Innovación.*

REFERENCES

Achenbach, T. M. (2006). As others see us: clinical and research implications of cross-informant correlations for psychopathology. *Curr. Dir. Psychol. Sci.* 15, 94–98. doi: 10.1111/j.0963-7214.2006.00414.x

Achenbach, T. M. (2011). Commentary: definitely more than measurement error: but how should we understand and deal with informant discrepancies? *J. Clin. Child Adolesc. Psychol.* 40, 80–86. doi: 10.1080/15374416.2011.533416

Achenbach, T. M., McConaughy, S. H., and Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. *Psychol. Bull.* 101, 213–232. doi: 10.1037/0033-2909.101.2.213

Achenbach, T. M., and Rescorla, L. A. (2007). *Multicultural Supplement to the Manual for the ASEBA: School-Age Forms & Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, and Families.

Berg-Nielsen, T. S., Vika, A., and Dahl, A. (2003). When adolescents disagree with their mothers: CBCL-YSR discrepancies related to maternal depression and adolescent self-esteem. *Child Care Health Dev.* 29, 207–213. doi: 10.1046/j.1365-2214.2003.00332.x

Bidaut-Russell, M., Reich, W., Cottler, L. B., Robins, L. N., Compton, W. M., and Mattison, R. E. (1995). The Diagnostic Interview Schedule for Children (PC-DISC v. 3.0): parents and adolescents suggest reasons for expecting discrepant answers. *J. Abnorm. Child Psychol.* 23, 641–659. doi: 10.1007/BF01447667

Biederman, J., Keenan, K., and Faraone, S. V. (1990). Parent-based diagnosis of attention deficit disorder predicts a diagnosis based on teacher report. *J. Am. Acad. Child Adolesc. Psychiatry* 29, 698–701. doi: 10.1097/00004583-199009000-00004

Bosco, G. L., Renk, K., Dinger, T. M., Epstein, M. K., and Phares, V. (2003). The connections between adolescents' perceptions of parents, parental psychological symptoms, and adolescent functioning. *J. Appl. Dev. Psychol.* 24, 179–200. doi: 10.1016/S0193-3973(03)00044-3

Byrne, B. (1998). *Structural Equation Modelling with LISREL, PRELIS y SIMPLIS: Basic Concepts, Applications and Programming*. London: LEA.

Campbell, D. T., and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* 56, 81–105. doi: 10.1037/h0046016

Cantwell, D. P., Lewinsohn, P. M., Rohde, P., and Seeley, J. R. (1997). Correspondence between adolescent report and parent report of psychiatric diagnostic data. *J. Am. Acad. Child Adolesc. Psychiatry* 36, 610–619. doi: 10.1097/00004583-199705000-00011

Carrasco, M. A., Holgado, F. P., Del Barrio, M. V., and Barbero, M. I (2008). Validez incremental: un estudio aplicado con diversas fuentes informantes y medidas [Incremental validity: an applied study using different informants and measurements]. *Acción Psicol.* 5, 65–76.

Carrasco, M. Á., Holgado, F. P., and Delgado, B. (2014). Cuestionario interpersonal de poder y prestigio parental (3PQ): dimensionalidad y propiedades psicométricas en niños españoles [The perceived interpersonal parental power and prestige questionnaire (3PQ): dimensionality and psychometric properties in Spain]. *Acción Psicol.* 11, 47–60. doi: 10.5944/ap.11.2.14174

Cheng, H., and Furnham, A. (2004). Perceived parental rearing style, self-esteem and self-criticism as predictors of happiness. *J. Happiness Stud.* 5, 1–21. doi: 10.1023/B:JOHS.0000021704.35267.05

Chilcoat, H. D., and Breslau, N. (1997). Does psychiatric history bias mothers' reports? An application of a new analytic approach. *J. Am. Acad. Child Adolesc. Psychiatry* 36, 971–979.

Crick, N. R., and Grotpeter, J. K. (1995). Relational aggression, gender, and social-psychological adjustment. *Child Dev.* 66, 710–722. doi: 10.2307/1131945

De Los Reyes, A. (2011). Introduction to the special section: more than measurement error: discovering meaning behind informant discrepancies in clinical assessments of children and adolescents. *J. Clin. Child Adolesc. Psychol.* 40, 1–9. doi: 10.1080/15374416.2011.533405

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A. G., Burgers, D. E., et al. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychol. Bull.* 144, 858–900. doi: 10.1037/a0038498

De Los Reyes, A., and Kazdin, A. E. (2004). Measuring informant discrepancies in clinical child research. *Psychol. Assess.* 16, 330–334. doi: 10.1037/1040-3590.16.3.330

De Los Reyes, A., and Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: a critical review, theoretical framework, and recommendations for further study. *Psychol. Bull.* 131, 483–509. doi: 10.1037/0033-2909.131.4.483

De Los Reyes, A., and Kazdin, A. E. (2006). Informant discrepancies in assessing child dysfunction relate to dysfunction within mother-child interactions. *J. Child Fam. Stud.* 15, 645–663. doi: 10.1007/s10826-006-9031-3

Del Barrio, V. D., Ramírez-Uclés, I., Romero, C., and Carrasco, M. A. (2014). Adaptación del child-PARQ/Control: versiones para el padre y la madre en población infantil y adolescente española [Adaptation of the child-PARQ/Control mother and father versions in Spanish child and adolescent population]. *Acción Psicol.* 11, 27–46. doi: 10.5944/ap.11.2.14173

Duhig, A. M., Renk, K., Epstein, M. K., and Phares, V. (2000). Interparental agreement on internalizing, externalizing, and total behavior problems: a meta-analysis. *Clin. Psychol. Sci. Pract.* 7, 435–453. doi: 10.1093/clipsy.7.4.435

Eid, M. E., and Diener, E. E. (eds) (2006). *Handbook of Multimethod Measurement in Psychology*. Washington, DC: American Psychological Association.

Epkins, C. (1993). A preliminary comparison of teacher ratings and child self-report of depression, anxiety, and aggression in inpatient and elementary school samples. *J. Abnorm. Child Psychol.* 21, 649–661. doi: 10.1007/BF00916448

Eyde, L. D., Robertson, G. J., Krug, S. E., Moreland, K. L., Robertson, A. G., Shewan, C. M., et al. (1993). *Responsible Test Use: Case Studies for Assessing Human Behavior*. Washington, DC: American Psychological Association.

Flouri, E., and Buchanan, A. (2002). What predicts good relationships with parents in adolescents and partners in adult life: findings from the 1958 British Birth Cohort. *J. Fam. Psychol.* 16, 196–198. doi: 10.1037/0893-3200.16.2.186

Garb, H. N. (2003). Incremental validity and the assessment of psychopathology in adults. *Psychol. Assess.* 15, 508–520. doi: 10.1037/1040-3590.15.4.508

Goodman, K. L., De Los Reyes, A., and Bradshaw, C. P. (2010). Understanding and using informants' reporting discrepancies of youth victimization: a conceptual model and recommendations for research. *Clin. Child Fam. Psychol. Rev.* 13, 366–383. doi: 10.1007/s10567-010-0076-x

Grigorenko, E. L., Geiser, C., Slobodskaya, H. R., and Francis, D. J. (2010). Cross-informant symptoms from CBCL, TRF, and YSR: trait and method variance in a normative sample of Russian youths. *Psychol. Assess.* 22, 893–911. doi: 10.1037/a0020703

Gryczkowski, M. R., Jordan, S. S., and Mercer, S. H. (2010). Differential relations between mothers' and fathers' parenting practices and child externalizing behavior. *J. Child Fam. Stud.* 19, 539–546. doi: 10.1007/s10826-009-9326-2

Hoeve, M., Dubas, J. S., Eichelsheim, V. I., Laan, P. H. V. D., Smeenk, W., and Gerris, J. R. M. (2009). The relationship between parenting and delinquency: a meta-analysis. *J. Abnorm. Child Psychol.* 37, 749–775. doi: 10.1007/s10802-009-9310-8

Hughes, E. K., and Gullone, E. (2010). Discrepancies between adolescent, mother, and father reports of adolescent internalizing symptom levels and their association with parent symptoms. *J. Clin. Psychol.* 66, 978–995. doi: 10.1002/jclp.20695

Hunsley, J. (2002). Psychological testing and psychological assessment: a closer examination. *Am. Psychol.* 57, 139–140. doi: 10.1037/0003-066X.57.2.139

Hunsley, J. (2003). Introduction to the special section on incremental validity and utility in clinical assessment. *Psychol. Assess.* 15, 443–445. doi: 10.1037/1040-3590.15.4.443

Hunsley, J., and Mash, E. J. (2005). Introduction to the special section on developing guidelines for the evidence based assessment (EBA) of adult disorders. *Psychol. Assess.* 17, 251–255. doi: 10.1037/1040-3590.17.3.251

Hunsley, J., and Mash, E. J. (2007). Evidence-based assessment. *Annu. Rev. Clin. Psychol.* 3, 29–51. doi: 10.1146/annurev.clinpsy.3.022806.091419

Jané, M., Araneda, N., Valero, S., and Doménech, E. (2000). Evaluación de la sintomatología depresiva del preescolar: correspondencia entre los informes de padres y de maestros. [Assessment of preschool symptomatology depression: correspondence between reports by parents and teachers]. *Psicothema* 12, 212–215.

Johnston, C. H., and Murray, C. (2003). Incremental validity in the psychological assessment of children and adolescents. *Psychol. Assess.* 15, 496–507. doi: 10.1037/1040-3590.15.4.496

Kenny, D. A. (1994). *Interpersonal Perception: A Social Relations Analysis*. New York, NY: Guilford Press.

Khaleque, A. (2015). Perceived parental neglect, and children's psychological maladjustment, and negative personality dispositions: a meta-analysis of multi-cultural studies. *J. Child Fam. Stud.* 24, 1419–1428. doi: 10.1007/s10826-014-9948-x

Khaleque, A., and Rohner, R. P. (2002). Perceived parental acceptance rejection and psychological adjustment: a meta-analysis of cross-cultural and intracultural studies. *J. Marriage Fam.* 64, 54–64. doi: 10.1111/j.1741-3737.2002.00054.x

Khaleque, A., and Rohner, R. P. (2011). Transnational relations between perceived parental acceptance and personality dispositions of children and adults: a meta-analytic review. *Pers. Soc. Psychol. Rev.* 16, 103–115. doi: 10.1177/1088868311418986

Khaleque, A., and Rohner, R. P. (2012). Pancultural associations between perceived parental acceptance and psychological adjustment of children and adults: a meta-analytic review of worldwide research. *J. Cross Cult. Psychol.* 43, 784–800. doi: 10.1177/0022022111406120

Klein, R. G. (1991). Parent-child agreement in clinical assessment of anxiety and other psychopathology: a review. *J. Anxiety Disord.* 5, 187–198. doi: 10.1016/0887-6185(91)90028-R

Langhinrichsen, J., Lichtenstein, E., Seely, J. R., Hops, H., Ary, D. V., Tildesley, E., et al. (1990). Parent-adolescent congruence for adolescent substance abuse. *J. Youth Adolesc.* 19, 623–635. doi: 10.1007/BF01537181

Lanz, M., Scabini, E., Vermulst, A. A., and Gerris, J. R. M. (2001). Congruence on child rearing in families with early adolescent and middle adolescent children. *Int. J. Behav. Dev.* 25, 133–139. doi: 10.1080/01650250042000104

Lapouse, R., and Monk, M. A. (1958). An epidemiologic study of behavior characteristics in children. *Am. J. Public Health Nations Health* 48, 1134–1144. doi: 10.2105/AJPH.48.9.1134

Leaper, C., Anderson, K. J., and Sanders, P. (1998). Moderators of gender effects on parents' talk to their children. *Dev. Psychol.* 34, 3–27. doi: 10.1037/0012-1649.34.1.3

Leung, J. T., and Shek, D. T. (2014). Parent-adolescent discrepancies in perceived parenting characteristics and adolescent developmental outcomes in poor chinese families. *J. Child Fam. Stud.* 23, 200–213. doi: 10.1007/s10826-013-9775-5

Likert, R. S. (1932). Technique for the measurement of attitudes. *Arch. Psychol.* 140, 44–53.

Loeber, R., Green, S. M., and Lahey, B. B. (1990). Mental health professionals' perception of the utility of children, mothers, and teachers as informants on childhood psychopathology. *J. Clin. Child Psychol.* 19, 136–143.

Loeber, R., Green, S. M., Lahey, B. B., and Stouthamer-Loeber, M. (1991). Differences and similarities between children, mothers and teachers as informants on disruptive behaviour disorders. *J. Abnorm. Child Psychol.* 19, 75–95. doi: 10.1007/BF00910566

Markon, K. E., Chmielewski, M., and Miller, C. J. (2011). The reliability and validity of discrete and continuous measures of psychopathology: a quantitative review. *Psychol. Bull.* 137, 856–879. doi: 10.1037/a0023678

Mash, E. J., and Terdal, L. G. (1997). *Assessment of Childhood Disorders*, 3rd Edn. New York, NY: Guilford Press.

Mattanah, J. F. (2001). Parental psychological autonomy and children's academic competence and behavioral adjustment in late childhood: more than just limit-setting and warmth. *Merrill Parker Q.* 47, 355–376. doi: 10.1353/mpq.2001.0017

McLeod, B. D., Weisz, J. R., and Wood, J. J. (2007). Examining the association between parenting and childhood depression: a meta-analysis. *Clin. Psychol. Rev.* 27, 986–1003. doi: 10.1016/j.cpr.2006.09.002

Meunier, J. C. H., Bisceglia, R., and Jenkis, J. M. (2012). Differential parenting and children's behavioral problems: curvilinear associations and mother–father combined effects. *Dev. Psychol.* 48, 987–1002. doi: 10.1037/a0026321

Meyer, G. J., and Archer, R. P. (2001). The hard science of Rorschach research: what do we know and where do we go? *Psychol. Assess.* 13, 486–502. doi: 10.1037/1040-3590.13.4.486

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment: a review of evidence and issues. *Am. Psychol.* 56, 128–165. doi: 10.1037/0003-066X.56.2.128

Najman, J. M., Williams, G. M., Nikles, J., Spence, S., Bor, W., O'Callaghan, M., et al. (2000). Mothers' mental illness and child behavior problems: cause-effect association or observation bias? *J. Am. Acad. Child Adolesc. Psychiatry* 39, 592–602. doi: 10.1097/00004583-200005000-00013

Neyer, F. J. (2006). "Informant assessment," in *Handbook of Multimethod Measurement in Psychology*, eds M. E. Eid and E. E. Diener (Washington, DC: American Psychological Association), 43–59.

Ollendick, T. H., and Hersen, M. E. (1993). *Handbook of Child and Adolescent Assessment*. Boston, MA: Allyn & Bacon.

Ramírez-Lucas, A., Ferrando, M., and Sainz, A. (2015). Influyen los estilos parentales y la inteligencia emocional de los padres en el desarrollo emocional de sus hijos escolarizados en 2° ciclo de educación infantil? [Do parental styles and parents' emotional intelligence influence their children's emotional development in kindergarten school?]. *Acción Psicol.* 12, 65–78. doi: 10.5944/ap.12.1.14314

Renk, K., Roberts, R., Roddenberry, A., Luick, M., Hillhouse, S., Meehan, C., et al. (2003). Mothers, fathers, gender role, and time parents spend with their children. *Sex Roles* 48, 305–315. doi: 10.1023/A:1022934412910

Rescorla, L. A., Ginzburg, S., Achenbach, T. M., Ivanova, M. Y., Almqvist, F., Begovac, I., et al. (2013). Cross-informant agreement between parent-reported and adolescent self-reported problems in 25 societies. *J. Clin. Child Adolesc. Psychol.* 42, 262–273. doi: 10.1080/15374416.2012.717870

Richters, J. E. (1992). Depressed mothers as informants about their children: a critical review of the evidence for distortion. *Psychol. Bull.* 112, 485–499. doi: 10.1037/0033-2909.112.3.485

Rohner, R. P. (1986). *The Warmth Dimension: Foundations of Parental Acceptance-Rejection Theory*. Beverly Hills, CA: Sage Publications, Inc.

Rohner, R. P. (1990). *Handbook for the Study of Parental Acceptance and Rejection*. Storrs, CT: University of Connecticut.

Rohner, R. P., and Khaleque, A. (eds) (2005). *Handbook for the Study of Parental Acceptance and Rejection*, 4th Edn. Storrs, CT: Rohner Research Publications, 187–226.

Rohner, R. P., and Khaleque, A. (2010). Testing central postulates of parental acceptance-rejection theory (PARTheory): a meta-analysis of cross-cultural studies. *J. Fam. Theory Rev.* 3, 73–87. doi: 10.1111/j.1756-2589.2010.00040.x

Rohner, R. P., Khaleque, A., and Cournoyer, D. E. (2012). "Parental acceptance-rejection theory, methods, and implications," in *Handbook for the Study of Parental Acceptance and Rejection*, 4th Edn, eds R. P. Rohner and A. Khaleque (Storrs, CT: Rohner Research Publications), 1–35.

Rosnati, R., Iafrate, R., and Scabini, E. (2007). Parent–adolescent communication in foster, inter-country adoptive, and biological Italian families: gender and generational differences. *Int. J. Psychol.* 42, 36–45. doi: 10.1080/00207590500412128

Schneewind, K. A., and Ruppert, S. (2013). *Personality and Family Development: An Intergenerational Longitudinal Comparison*. New York, NY: Psychology Press.

Stanger, C., and Lewis, M. (1993). Agreement among parents, teachers, and children on internalizing and externalizing behavior problems. *J. Clin. Child Psychol.* 22, 107–116. doi: 10.1207/s15374424jccp2201_11

Stanger, C., McConaughy, S. H., and Achenbach, T. M. (1992). Three-year course of behavioral/emotional problems in a national sample of 4- to 16-year-olds: II. Predictors of syndromes. *J. Am. Acad. Child Adolesc. Psychiatry* 31, 941–950. doi: 10.1097/00004583-199209000-00024

Tarver, J., Daley, D., Lockwood, J., and Sayal, K. (2014). Are self-directed parenting interventions sufficient for externalising behaviour problems in childhood? A systematic review and meta-analysis. *Eur. Child Adolesc. Psychiatry* 23, 1123–1137. doi: 10.1007/s00787-014-0556-5

Tenenbaum, H. R., and Leaper, C. (2003). Parent-child conversations about science: the socialization of gender inequities? *Dev. Psychol.* 39, 34–47. doi: 10.1037/0012-1649.39.1.34

Torres, N., Veríssimo, M., Monteiro, L., Ribeiro, O., and Santos, A. J. (2014). Domains of father involvement, social competence and problem behavior in preschool children. *J. Fam. Stud.* 20, 188–203. doi: 10.5172/jfs.2014.20.3.188

Weitkamp, K., Daniels, J., Rosenthal, S., Romer, G., and Wiegand-Grefe, S. (2013). Health-related quality of life: cross-informant agreement of father, mother, and self-report for children and adolescents in outpatient psychotherapy treatment. *Child Adolesc. Ment. Health* 18, 88–94. doi: 10.1111/j.1475-3588.2012.00656.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on
You Tube G+ f