The Spanish Journal of Psychology (2023), **26**, e13, 1–11.
© The Author(s), 2023. Published by Cambridge University Press on behalf of Universidad Complutense de Madrid and Colegio Oficial de la Psicología de Madrid
doi:10.1017/SJP.2023.14

How do you Behave as a Psychometrician? Research Conduct in the Context of Psychometric Research

Pablo Ezequiel Flores-Kanter[1,2] **and Mariano Mosquera**[1]

[1] *Universidad Católica de Córdoba (Argentina)*
[2] *Consejo Nacional de Investigaciones Científicas y Técnicas (Argentina)*

Abstract. The identification of fraudulent and questionable research conduct is not something new. However, in the last 12 years the aim has been to identify specific problems and concrete solutions applicable to each area of knowledge. For example, previous work has focused on questionable and responsible research conducts associated with clinical assessment, measurement practices in psychology and related sciences; or applicable to specific areas of study, such as suicidology. One area of study that merits further study of questionable and responsible research behaviors is psychometrics. Focusing on psychometric research is important and necessary, as without adequate evidence of construct validity the overall validity of the research is at least debatable. Our interest here is to (a) identifying questionable research conduct specifically linked to psychometric studies; and (b) promoting greater awareness and widespread application of responsible research conduct in psychometrics research. We believe that the identification and recognition of these conducts is important and will help us to improve our daily work as psychometricians.

Received 24 October 2022; Revised 30 March 2023; Accepted 19 April 2023

Keywords: HARKing, measures, open science framework (OSF), p-hacking

Critical analysis of the general methods, procedures and forms of doing science, as well as the identification of fraudulent and questionable conduct in research, is not something new (Barber, 1976). However, in the last 12 years, the detailed review of scientific work has been extended to other fields of knowledge and other research designs and, particularly in psychology and other related sciences, it has been resumed with great force (Chin et al., 2023). These ideas have been discussed from the perspective of the so-called *replicability crisis* (Nosek et al., 2022; Simmons et al., 2011). Revisions and proposals of viable explanations and solutions are still now being produced; yet the scientific literature, and in particular the meta-scientific studies (i.e., studies on the way science is done), have been giving concrete recommendations regard questionable and responsible research conduct.

One of the pending objectives is to adjust these suggestions to the specific contexts of each discipline and subdiscipline within psychological science (Chin et al., 2023; Kirtley et al., 2022; Tackett et al., 2017). A correct identification of questionable research conduct and dissemination of responsible research conduct adapted to each specific area of research is essential in order to achieve a more generalized knowledge and adherence within the scientific community (Bosma & Granger, 2022; Waldman & Lilienfeld, 2016; Steneck, 2006). Some of these proposals have recently been published and there appeared examples in the field of psychological-clinical evaluation (Tackett et al., 2019) and in the context of more general measurement practices (Flake & Fried, 2020; Lilienfeld & Strother, 2020). Although these relevant antecedents have undoubtedly helped to identify questionable research conduct and promote responsible research practice in the field of psychological evaluation or measurement use in general, here we want to focus on a set of questionable research conduct in psychometrics[1] that we believe merit further

Correspondence concerning this article should be addressed to Pablo Ezequiel Flores-Kanter. Universidad Católica de Córdoba. Centro de Bioética. Córdoba (Argentina).

E-mail: ezequielfk@gmail.com; pablo.floreskanter@conicet.gov.ar

Funding Statement: This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Conflicts of Interest: None.

Authorship credit: Pablo Ezequiel Flores-Kanter: Conceptualization, writing-original draft preparation, writing-reviewing and editing, and supervision. Mariano Mosquera: Writing-reviewing and editing, and supervision.

Data Sharing: Not applicable.

[1] Psychometric studies refer to research that aims to provide evidence of the construct validity and reliability of a measurement (see Clark & Watson, 2019; and Flake et al., 2017).

How to cite this article:
Flores-Kanter, P. E., & Mosquera, M. (2023). How do you behave as a psychometrician? Research conduct in the context of psychometric research. *The Spanish Journal of Psychology*, *26*. e13. Doi:10.1017/SJP.2023.14

attention: *Fit-hacking*, *model-HARKing* and *emphasizing new -measurements or estimation- models.* We believe that the focus on these questionable research conducts in psychometric studies is relevant because the more general validity of our research results depends on the validity of the interpretations we can make of our measurements (Flake et al., 2022; Lilienfeld & Strother, 2020). In this sense, the focus on measurement is fundamental since it is at the base of scientific progress and of (valid) interpretation of research results (Clark & Watson, 2019; Flake et al., 2017; Flake & Fried, 2020). All this has consequences for the applied field: If the evidence behind the theories is not based on properly validated measurements[2], those theories are not correctly translated into practical applications (Bosma & Granger, 2022; Lewis, 2021). In addition to focusing on questionable research conduct in psychometrics, the aim of this paper is to provide resources that enable psychometric researchers to protect themselves against questionable research conduct, with a focus on practices related to transparency and the open science framework. All in all, we believe that in the context of psychometric studies it is important to continue to identify questionable research conduct that is specific to this type of research, and that it is also necessary to adapt the recommendations on responsible research conduct in this area.

Summarizing, the purpose of this paper is twofold: (a) To identify questionable research conduct specifically linked to psychometric studies; and (b) to promote greater awareness and widespread application of responsible research conduct in psychometrics research. To this end, we will first develop the more general concepts of research conduct and associated variables. We will then focus specifically on the identification of questionable research conduct in psychometrics, differentiating between questionable conduct linked to practices and questionable conduct linked to reporting. Finally, we will address the topic of transparency practices and the use of the open science framework as inherent actions of responsible research conduct, focusing here also on their applicability and relevance to psychometrics.

Research Conduct in Psychometrics

Behavior in psychometrics research can be analyzed based on more general models of research conduct. Setneck (2006) proposes one such model, pointing out that research conduct can be understood as a continuum: From the ideal conduct, Responsible Research Conduct (RRC), to the worst conduct, which is characterized by Practices of Fabrication, Falsification and Plagiarism (FFP). Questionable Research Conduct (QRC) falls in the middle point of this continuum. It is also important to differentiate research practices from reporting practices (Manapat et al., 2022; Munafò et al., 2017)[3]. Research and report practices have a direct impact on the reliability of science, that is, on the replicability, robustness, and reproducibility of scientific findings[4]. While FFP describes unequivocal, easily documented actions deserving severe sanctions (Steneck, 2006), QRC tend to be more difficult to define (i.e. they are not unequivocal), they occur more frequently (Munafò et al., 2017) and are more difficult to identify as bad practices by the researchers and institutions involved (i.e., researchers and institutions disagree on whether these practices are actually harmful or engage in QRC without being aware of their deleterious effects). QRC is a general term referring to the misuse or non-optimal use of methodological-statistical procedures, from which non-robust (e.g., overfitting), invalid, and biased results are more likely obtained (Antonakis, 2017; Munafò et al., 2017; Nelson et al., 2018; Waldman & Lilienfeld, 2016). Most of the literature about QRC also highlights the problem of flexibility or researcher degrees of freedom. In the context of QRC, the researcher has a "garden of forking paths" (Gelman & Loken, 2014) to make decisions about the method to be applied in each step, which can be exploited in favor of the researcher (intentionally or not) to achieve the desired results. So, QRC "often involve hidden research decisions" (Chin et al., 2023), and these practices distort the accuracy of research reports when are not reported transparently. Moreover, "Such practices produce biases because undisclosed flexibility … allows researchers to selectively under- or over-fit models and exploit noise in a way that goes uncorrected…" (Chin et al., 2023, p. 3). In these sense, QRC can impact on the reliability and validity of scientific research in a large and widespread manner (Chin et al., 2023), and this impact can be even greater than that of the FFP

[2] Measurements with well-established (i.e., robust, and solid), up-to-date, and relevant evidence of construct validity in the population(s) of interest.

[3] In difference to the terminology generally used in these cases, here we speak of Questionable Research Conduct instead of Questionable Research Practice. The reason for this is that we consider it important to differentiate between research practices and reporting practices within research conduct. Therefore, we use the term Questionable Research Conduct here as the more general term that includes both research and reporting practices.

[4] Following the definitions of Nosek et al. (2022), the term *replication* refers to the study of the reliability or consistency of a previous scientific finding, using data different from the original research; the term *robustness* refers to the reliability or consistency of a previous scientific finding, using the same data and a different analysis strategy than the original study; and the term *reproducibility* refers to the reliability or consistency of a previous scientific finding, using the same data and the same analysis strategy of the original study.

https://doi.org/10.1017/SJP.2023.14 Published online by Cambridge University Press

(Munafò et al., 2017). This is why international literature has focused primarily on identifying questionable research conduct and promoting responsible research conduct that serves as a preventive vaccine to avoid the adverse effects of QRC (Chin et al., 2023).

Examples of very widespread QRC, which have been the focus of study internationally, are the following:

1. Using multiple comparisons from, for example, exclusion of atypical cases, inclusion of covariates, incorporation of more cases to the sample, with the aim of finding statistically significant results (i.e., p-hacking; Nelson et al., 2018).
2. Generating hypotheses and theoretical explanations from the results obtained and presenting these explanations and theories as if they had been proposed prior to data collection. In other words, presenting exploratory research as confirmatory (i.e.: HARKing; Munafò et al., 2017).
3. Emphasizing new, and statistically significant, results (i.e., selectively reporting positive results), not mentioning the results that have not reached statistical significance (i.e., negative results) (Antonakis, 2017).

In the following, we will focus on QRC specific to psychometric studies and offer recommendations for promoting RRC in this area. Based on the taxonomy previously presented, we will separate QRC linked to practices from those linked to reporting.

Questionable Research Conduct in Psychometrics (QRC$_{\Psi metrics}$[5])

In psychometrics we also have at our disposal a "garden of forking paths" which can be exploited in favor to achieve the desired results. Let us look at some of these "tricks" in more detail.

QRC: Practice-Related Research (QRC-P$_{\Psi metrics}$)

We consider it important to highlight the following QRC-.P$_{\psi metrics}$, which we refer as follows:

Fit-hacking: Using different types of strategies in the model estimation-specification with the aim of finding an acceptable or optimal fit[6]. In this context, the publication of over-adjusted models (i.e., overfitting) is a very common practice. One example is the specification, within confirmatory factor analyses, of measurement models[7] that incorporate error covariances or correlation between errors, as many as necessary to exceed the cut-off points for model fit (Flores-Kanter et al., 2021). Another example is the application of unjustified-inappropriate complex models, such as bifactor confirmatory models, which facilitate obtaining acceptable and optimal fit indicators (Flores-Kanter et al., 2022; Haywood et al., 2021; Reise et al., 2016). The Positive and Negative Affective Affective Schedule (PANAS) studies serve as a good example to visualise both examples of QRC-$_{\psi metrics}$. In the case of incorporating error covariances, most studies on the PANAS have implemented this strategy to reach the minimum cut-off points to consider the fit of the measurement model acceptable (e.g., CFI > .90) (Flores-Kanter et al., 2022). In the case of bifactor models, mathematically equivalent to specifying covariances between all pairs of errors (Matsunaga, 2008), researchers have found in it a model that hardly presents indicators of poor fit (Flores-Kanter et al., 2018). Other good examples of the misuse of bifactor models can be found in the case of psychometric analysis of measures of depression (Heinrich et al., 2018) and psychopathology (Bonifay et al., 2016). These fit-hacking related practices undermine the external validity and reliability of the findings and are very similar to the behavior described as p-hacking in other disciplinary contexts.

Does the aforementioned mean that the practice of specifying covariances between pairs of uniqueness or bifactor models is, per se, bad practice? We state categorically that, per se, these models do not refer to bad practice. Indeed, there are concrete situations where the use of correlations between errors as well as the specification of traditional bifactor models can be justified and recommended (see, for example, Eid et al., 2017). Conversely, the downside of such practices lies in how these models are used, interpreted, and reported, not in the models themselves (Box, 1979; McElreath, 2020). In the case at hand, what is questionable is the indiscriminate and unjustified use of covariances between pairs of errors or bifactor models with the sole objective of reaching the cut-off points typically established for the model fit indicators. Added to this is the tendency to generate a discourse, after the results are known, that persuades the reader that the model is theoretically valid and procedurally sound. We will return to this point later when discussing Model-HARKing.

[5] We use the subscript $_{\psi metrics}$ to specify that we are referring to such practices within psychometric studies.

[6] There are different fit metrics that are used in the context of psychometric studies to assess the adjustment of a given model. For example, in the context of confirmatory factor analysis, traditional cutoffs used in empirical studies are SRMR $\leq$ .08, RMSEA $\leq$ .06, and CFI $\geq$ .96 (see McNeish & Wolf, 2021 for further discussion on this topic).

[7] The term measurement model is used within the context of Structural equation modelling (SEM). SEM is a technique that can handle latent (LV) and manifest (MV) variables (i.e., items or indicators). In theoretical terms, SEM consists of two different models: A measurement model (i.e., outer models) that account for the relationship between the MVs and the LVs; and a structural model (i.e., inner model) that account for the relationships between the LVs (Sarnacchiaro & Boccia, 2018).

https://doi.org/10.1017/SJP.2023.14 Published online by Cambridge University Press

Emphasizing new -measurements or estimation- models: There is a clear tendency to massively apply a new estimation method or measurement model without a clear justification for the choice and without a critical use of it. This has been seen in the case of bifactor models mentioned previously (Bonifay et al., 2016; Flores-Kanter et al., 2018), but a similar trend can also be verified in the case of psychometric network models (Burger et al., 2022). The emphasis on publishing a new measurement model or a new estimation method at the expense of generating a critical view of the measure in question prevents the proper advancement of psychometrics (Flake & Fried, 2020), consequently generating unnecessary noise in widely applied studies (Lewis, 2021). Previous studies have also warned about the tendency to create new scales without considering their overlap with existing scales, and without assessing their incremental value in relation to previous measures of the same or similar constructs (Rosenbusch et al., 2020). All of the above can be linked primarily to the research bias that Antonakis (2017) defined as *neophilia desease* (i.e., a tendency to show novel or spectacular results which are likely to be wrong); but it is also associated with *theorrea disease*, in the sense that many psychometric researchers do not engage critically with the theoretical aspects of the measures they assess, focusing almost exclusively on a single source of evidence of construct validity, usually the structural or external source (Flake & Fried, 2020; Lilienfeld & Strother, 2020). Here again, is a QRC-P$_{\Psi metrics}$ applying novel psychometric/measurement models or proposing a new measure? Of course not; what is questionable lies in a- the uncritical and unjustified application of models, selecting them not on the basis of critical and theoretically relevant reasoning, but on the basis of their novelty and associated publication advantage; and b- the generation of new measure in a superficial manner, i.e. without adequately considering their overlap with previous measures and their incremental value.

There are many other practices in psychometrics that can be included in the category we have called here QRC-P$_{\Psi metrics}$. Here we have made an arbitrary selection of two sets of practices, *fit-hacking* and *emphasizing new -measurements or estimation- models*, which we consider to be widespread and to which we believe more attention should be given. However, we encourage the reader to delve deeper into the extensive literature on other forms of questionable practices in psychometrics. We mention below some examples of these QRC-P$_{\Psi metrics}$ that have been identified in previous contributions[8]:

1. Inferring the measurements that are derived from an instrument, and basing the choice of the measures, solely on the basis of the name given to the instrument (Lilienfeld & Strother, 2020).
2. Misapplication of internal consistency indicators and the exclusive use of Cronbach's alpha coefficient (Cho, 2021).
3. The elimination of items in order to achieve acceptable internal consistency (Ulrich & Miller, 2018).
4. Inappropriate use of factor estimation methods and procedures associated with exploratory factor analysis (Ferrando et al., 2022; Lloret-Segura et al., 2014).
5. The debatable use of sum scores (Widaman & Revelle, 2023) and item parcel approach (Matsunaga, 2008).
6. Using fit indexes arbitrarily and with different cutoffs to support or reject the fit of a model (McNeish & Wolf, 2021).
7. Exclusive consideration of the structural or external phase as final evidence of construct validity (Flake et al., 2017; Lilienfeld & Strother, 2020).

In sum, in psychometrics we have a great deal of methodological flexibility available that can be exploited to our advantage to achieve the desired result; but not only that, we also have ways to convince editors, reviewers and readers of the relevance of our (questionable) approach, the novelty, relevance and necessity of our (forced) findings (i.e., theoretical and practical implication). In the following, we will make references to these QRC linked to the report.

QRC: Report-Related Research (QRC-R$_{\Psi metrics}$)

QRC-R$_{\Psi metrics}$ refer to non-transparent/non-accessible report of the steps and procedures applied in the investigation, as well as data and other research materials. Although this last case does not necessarily imply that QRC$_{\Psi metrics}$ have been presented referring to the misuse or non-optimal use of methodological-statistical procedures, the lack of transparency, and the impediment of access to the steps, procedures and research materials, makes it difficult to evaluate/review the entirely research.

The second group of QRC-R$_{\Psi metrics}$ has recently been identified in psychological assessment in general. Authors such as Flake and Fried (2020) have indicated certain uses and behaviors as questionable practices in measurement and have especially highlighted the need to promote a more open and transparent reporting methodology in the area. Here, we are interested in emphasizing another particular QRC-R$_{\Psi metrics}$, which we have named *model-HARKing*. Originally, the acronym HARKing is used to refer to the behavior of "hypothesizing after the results are known" (Munafò et al., 2017). In psychometrics it is possible to identify

[8] We are grateful to the reviewer "Esther Maassen" who suggested incorporating many of these previous contributions.

similar conducts, which are mainly evidenced in the way the report is presented. In there, the overlapping of exploratory and confirmatory objectives and/or analyses is common and widespread. The biggest problem is that, as happens in other fields of knowledge (Fife & Rodgers, 2022), an approach that is entirely exploratory is presented as confirmatory. Thus, with the term model-HARKing we try to draw attention to those forms of reporting that, after knowing which model presents a better fit (generally achieved from the behaviors we have mentioned as fit-hacking), aim to assemble the whole document in coherence with this result; not making visible the fact that this model did not emerge from a confirmatory approach but rather an exploratory one (e.g., from the modification indices[9] resulting from the specification of a given measurement model). This has led, for example, to innumerable factorial solutions being proposed for the same measure, all of which find a "reasonable" explanation within a given body of theory (see the examples presented in Flores-Kanter et al., 2021, and Fried et al., 2022). The latter is also associated with theorrea desease (Antonakis, 2017), in that psychometric researchers seem more concerned with showing a line of argumentation consistent with the best fitting model finding, rather than valuing an indicator of poor fit as an opportunity to critically reflect on the theoretical aspects of the measurements they assess, and as an opportunity to focusing on all sources of construct validity (i.e., substantive, structural and external), and not exclusively on the structural (in the majority of cases, on factor analysis) or external phase.

Let us now consider what vaccines are available to prevent the emergence of these diseases in psychometric research.

Responsible Research Conduct in Psychometrics ($RRC_{\Psi metrics}$)

As psychometric researchers we must do the best we can, trying to apply the best practices suggested and enabled at the time. However, what is recommended or conceived as good practice at one time may no longer be recommendable later; and no matter how well-intentioned we may be, we will always be susceptible to mistakes. Moreover, there will surely always be alternative ways of psychometrically modelling our problem (i.e., methodological flexibility; see Manapat et al., 2022). As we tried to express in the previous paragraphs, the criticism is not of statistical models per se, but of their unjustified and inappropriate use, and the way in which the procedure followed is detailed and the research report is written up. At this point, there are two components that we consider key and mutually related to the achievement of $RRC_{\Psi metrics}$: achieving greater transparency and aligning with the principles of open science.

Be Transparent

Psychometric studies should be explicit, clear, and systematic in the procedures and steps followed throughout the research process. Transparency of information, thought in terms of cooperation, contributes to strengthening the idea of peer control in the scientific community, not only as a way to legitimize research results but also to produce scientific advances. Promoting transparent is an essential step, as it facilitates the detection of errors, makes it possible to make pertinent corrections and enables the reproducibility of scientific findings.

In the context of qualitative-applied research in psychology, considerations regarding explicitness about assumptions and justifications of methodological choices, and recommendations on transparency have been addressed, developed and refined for decades. Yet quantitative science interest on these factors is only recent (Lewis, 2021). Tuval-Mashiach (2017) has proposed a model of transparency for qualitative research which summarizes these qualitative-applied science contributions. According to this model, authors should be able to clearly express, in general terms, the "what, how, and why" of their research procedures. We will now extend this model to psychometric research.

There are three basic questions that must be answered to achieve an adequate level of transparency in psychometric research reports. We speak of a report in a broad sense, including here not only the paper but all that annexed material (e.g., supplementary material in an external open access repository) that develops each of these questions with adequate detail:

"What I did": The procedure, method or approach used must be named with correct language (i.e., using the statistical-methodological term that is widely used and supported by the scientific community of reference). This may seem trivial at first glance, but the ambiguous use of certain terms in psychometrics undermines the replicability, robustness, and reproducibility of psychometrics findings (see for example Cho, 2021; on the denomination of reliability coefficients). Researchers should also follow international guidelines and standards on the reports of each specific design. Although the American Psychological Association has not yet presented a standard format for all types of psychometric studies, there is currently a guide for reporting studies that use Structural Equation Models (Appelbaum et al., 2018, p. 18, Table 7). This guide can be extended to, and serve as a model for, the reporting of

[9] Modification indices refer to a statistic usually obtained in the context of confirmatory factor analysis, which shows what re-specifications can be made in order to improve the model fit.

https://doi.org/10.1017/SJP.2023.14 Published online by Cambridge University Press

other types of psychometric studies (e.g., exploratory factor analysis). Using the guide for reporting studies that use Structural Equation Models (Appelbaum et al., 2018) is strongly recommended as it will enable that the information and details necessary to understand the investigation (the "What I did") should be present in the report.

"How I did it": the necessary degree of transparency is achieved when an external or independent researcher can repeat -and clearly understand- the steps and procedures described in a report. The correct reporting of "How I did it" can be achieved by making good use of open science tools. These open science tools will be described in detail later. We will simply say here that using open science tools is a good way to answer the "How I did it", given that integrates a wide variety of resources for open and reproducible science, giving the options to express the research workflow-procedure in a transparent, clear, and complete way.

"Why I did it": the researcher must be able to explain why a given method was chosen and justify such choice by comparing alternative methods. This is extremely important in the field of psychometrics where many times the decision about the applied methods (e.g., factor estimation methods; rotation methods; coefficients considered) depends on the software usually used by, or available to, the researcher (Lloret-Segura et al., 2014). Given the manifest tendency to use without reason certain procedures repeated in the literature simply because they appear cited by a respected authority in the field (McNeish & Wolf, 2021), it is relevant to reflect upon the motivation to choose a given method. A clear practice of this is in the interpretation made (i.e., cut-off points considered) of commonly applied fit indicators such as the comparative fit index (CFI) and the root mean square error of approximation (RMSEA) (McNeish & Wolf, 2021).

This model of transparency is not only applicable to the method procedures but must be transversal to the other facets of the research (or empirical cycle, see Tijdink et al., 2021), which corresponds to the sections of the report or paper (e.g., introduction and methods). For example, authors should report in a transparent manner the procedure used to research the background on the subject. In psychometric studies, this is fundamental to understand the state of the art of the proposed measurement models, as well as the applied psychometric procedures. In this sense, we recommend following the guidance offered in "Conducting a Meta-Analysis in the Age of Open Science" paper, particularly with regard to documenting the procedures of research and revision of antecedents in a transparent manner, and in an open science framework (Moreau & Gamble, 2020).

Be Open Science

In the context of analyzing and responding to QRC, the international literature has called for greater use of so-called open science practices (OSPs; Munafò et al., 2017; Nelson et al., 2018). These include a diverse set of practices, including sharing of data and code, preregistration and preprints, among others. As stated in the case of transparency practices, it is important to note that OSPs are not a guarantee, on their own, of validity and robustness in the reported procedure and findings. Instead, the value in OSPs lies in the fact that they facilitate scrutiny and evaluation of the entire research process, also making it easier to detect and correct honest errors in research (Chin et al., 2023). External repositories for OSPs are a great tool for psychometric research and its use should be widely encouraged. There are highly valuable technological resources, among which we strongly recommend the Open Science Framework (OSF),[10] since it is free and open source. In addition, it integrates a wide variety of resources for open and reproducible science, giving all the options to express the research workflow in a transparent, clear and complete way. Among OSF resources, we suggest:

Pre-registration of projects/research plans: With this, the hypotheses and planning of analytical methods and procedures can be shared in advance. Objectives, hypotheses and procedures planned in advance (i.e., prior to the execution of the investigation) are clearly differentiated in a more transparent way from objectives, procedures and hypotheses derived from the course of investigation itself (i.e., in the execution of the investigation or after it). An example of this is the difference between confirmatory objectives or analyses and exploratory ones, which is closely linked to the model-HARKing behavior mentioned above. The preregistration of projects/research plans is, therefore, a good antidote to fit-hacking and model-HARKing behaviors, given that psychometric researchers can use this resource to clearly delimit, prior to the actual execution of the research, fundamental aspects such as: the measurement and structural models to be estimated; the estimation method; the fit indicators to be considered and the cut-off points considered. This prior delimitation, and restriction-transparenting of the researcher degrees of freedom, denotes a clear baseline that allow then evaluate the parts of the report that correspond to a confirmatory approach from those that are exploratory in nature. Of course, if a psychometric researcher has published a pre-registration prior to conducting his research, it will be more difficult to publish a report in which the inherent methodological flexibility will be exploited in favor of the researcher to achieve the

[10] https://osf.io.

https://doi.org/10.1017/SJP.2023.14 Published online by Cambridge University Press

desired results (i.e., fit-hacking), and the report will be presented in a hidden research decisions manner (i.e., model-HARKing).

Open Database and Open Code: The platform allows uploading both the data (i.e., raw data and processing data) as well as the code-syntax that was used to carry out the analyses in the given software[11]. Publishing the code-syntax is useful for promoting reproducibility, as it allows the same analytical steps to be followed, but also promotes quality control through increased opportunities to find bugs in the code (Laurinavichyute et al., 2022). All of this is especially important in psychometrics, given the existence of so many analytical options and software availability. We also suggest taking into account the guidance for a correct presentation (that promotes reproducibility and replicability) of the software's syntax and information (Buchanan et al., 2021; Epskamp, 2019), and follow the TIER protocol[12]. Also, the publication of the database should respect the conditions known by the acronym FAIR (Buchanan et al., 2021; Levenstein & Lyle, 2018). FAIR refers to conditions of findability (with adequate metadata), accessibility, interoperability (adaptation to systems) and reuse (open licenses). It is important to mention that the open data movement acknowledges that there are always restrictions that may be valid (such as personal data, for example), and it is important to declare the restrictions applied to open data (Meyer, 2018).

Preprint research report: Preprints are open access documents (i.e., research reports) published on a specific server, made available to receive comments from peers in a given discipline (Moshontz et al., 2021). Preprints may refer to a preliminary version of a document, for example, being in a state prior to a peer review process; or it can be an accepted version of a paper to be published in a scientific journal. In the latter case, a version not edited by the journal in which the document has been accepted is uploaded, letting the reader know about the differences between this preprint version and the version formally published in the scientific journal. The psychology-specific preprint server hosted by the OSF is called PsyArxiv[13]. A preprint in PsyArxiv may be integrated as part of a larger project in OSF, meaning that it can have all associated data, protocols, and other study materials published along with it. The use of this resource in psychometrics, as in all disciplines, broadens the scope of access of many publications in scientific journals, reducing restrictions.

Conclusions

As scientists in the psychological field, we are witnessing a present time full of positive changes. Every day the movement that seeks to promote the credibility and replicability of psychological research upon the basis of transparency is becoming stronger (Mellor et al., 2018). The evidence suggests that the majority of researchers agree with the principles of transparency and open science in research. However, it has also been shown that the concrete application of these principles and practices is not homogeneous in all scientific disciplines or in all subdisciplines of a given knowledge area. While we have presented relevant background information that certainly helps to promote responsible conduct in research in the field of psychological assessment in general, these analyses and recommendations have not yet focused specifically on the psychometric studies proper. This paper has begun to fill this gap.

Given all the above, we believe it is important to give some recommendations considering the different levels involved, which have an influence on questionable research practices (Tijdink et al., 2021). At the individual level, it is important for psychometric researchers to be aware of these questionable research practices and to be able to identify the biases associated with these trends (Antonakis, 2017). At a more general level, it is essential that journals should begin, as a first step, to adapt their editorial processes to models of responsible conduct in research regarding transparency and open science. It is also important that research ethics committees, or centers in charge of the ethical evaluation of psychometric projects should incorporate, promote and adhere to these practices.

The present article only offers an initial approach to the problem of questionable research practices in psychometrics. It is necessary to carry out meta-scientific investigations and systematics reviews that help to investigate the frequency of QRC$_{\Psi metrics}$, the factors associated with these, and the uses and factors that contribute to RRC$_{\Psi metrics}$ (Chin et al., 2023). These future investigations should provide the context in which research work is carried out in order to verify variabilities depending on the countries involved. For example, we think that QRC$_{\Psi metrics}$ are especially widespread in our South-Central American region, and that the use of RRC$_{\Psi metrics}$ is not yet widespread in these countries. However, future studies should provide empirical data in this regard. Also tutorials showing the steps to generate a transparent report in psychometrics should be published (Luong & Flake, 2022; PsyTeachR Team et al.,

[11] One of the reviewers suggested indicating, and we agreed with him, the need to provide open access also to the total output file produced by software programmes such as M-plus or the Lavaan package in R. As the reviewer says, "the total output file is very informative for other researchers in order to replicate the reported results".

[12] It is available at https://www.projecttier.org/tier-protocol/.

[13] https://psyarxiv.com/.

2022) to facilitate the use of external open access repositories (e.g., OSF), as well as to promote good practices associated with information access (e.g., recommendations for providing open access to databases; how to present the information and organize it in the external repository; how to organize an open access code). Lastly, reporting standards for psychometrics studies should be promoted. These should consider not only the points to be presented in the paper, but also the aspects that should be developed in supplementary materials. In addition, this guideline or set of guidelines should cover the entire spectrum of approaches in the psychometric field, from more exploratory or less restrictive approaches to confirmatory or more restrictive approaches (e.g., Exploratory Factor Analysis -EFA-, Ferrando et al., 2022; Exploratory Structural Equation Modeling -ESEM-, Marsh et al., 2014; Confirmatory Factor Analysis -CFA-; see also Morin et al., 2020); as well as new approaches in psychometrics (e.g., Exploratory Graph Analysis -EGA-, Golino & Epskamp, 2017) and other type of psychometrics approaches such as item response theory (IRT; Raykov et al., 2017). Also, the specific objectives of the psychometric study (e.g., construction, adaptation, validation; Ferrando et al., 2022) and the type of test applied (e.g., experimental manipulations; Chester & Lasko, 2021) should be considered.

To summarize, we believe that the final objective pursued by the contributions about the replicability crisis in, particularly, the behavioral sciences should not be underestimated; the same is also true about the meta-scientific studies that identify questionable research conduct as well as insights developed to promote responsible conduct in research. The objective is "to construct reliable and valid knowledge about how the mind works, and how the mind influences our behavior and vice versa" (Lewis, 2021, p. 10). Bearing this general objective in mind, and following Lewis (2021), it is necessary not only to promote good methodological practices and transparency in individual studies, but also to promote greater heterogeneity and integration between diverse methodologies, different populations, and research groups (see also, Wagenmakers et al., 2022). Only then shall behavioral sciences overcome this moment of crisis and achieve more valid and reliable knowledge. It is also important to note that while transparency and open science practices can enhance the evaluation of research validity and robustness, it is crucial to supplement them with critical appraisal that can distinguish between strong (i.e., robust-valid) and weak research practices (Chin et al., 2023). As pointed out by Antonakis (2017), a useful science is one that, in addition to accounting for the rigor (i.e., robustness, accuracy, and reliability of the research) can respond to the following generic questions: (a) So what? Reports if the theoretical or empirical contribution adds up to cumulative research efforts; and (b) Will it make a difference? Refers the extent to which the finding can inform basic or applied research, so that we can better understand the phenomenon and/or inform policy or practice. This is in line with the conclusion of Rohrer et al. (2022, p. 11), which we agree with and consider relevant for psychometric studies: "Our vision is one in which psychological research is inherently transparent and collaborative, collectively striving toward greater robustness and culmination of knowledge."

Finally, we would like to emphasize that this manuscript has not been written with the aim of blaming anyone in particular but, on the contrary, in the spirit of constructive criticism in a field in which we, as psychometric researchers and authors of this proposal, are no strangers. Indeed, we recognize that we ourselves identify with many of the biases outlined above, we have conducted some of these QRC$_{\Psi metrics}$, and we have only recently begun to incorporate tools consistent with the OSPs. As Rohrer et al. (2022, p. 10) say "This unfortunate situation can occur without any ill intention on the part of researchers, and we do not mean to imply that researchers who use these models are bad at their job or (even worse) do not care about the truthfulness of their claims—they are simply implementing practices that they have been taught and that often result in interesting sounding empirical claims." We believe that the identification and recognition of these conducts is important and will help us to improve our daily work as psychometricians. All in all, we hope that this work will help generate greater awareness of QRC$_{\Psi metrics}$ and adherence to RRC$_{\Psi metrics}$.

References

Antonakis, J. (2017). On doing better science: From thrill of discovery to policy implications. *The Leadership Quarterly*, *28* (1), 5–21. https://doi.org/10.1016/j.leaqua.2017.01.006

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Barber, T. X. (1976). *Pitfalls in human research: Ten pivotal points*. Pergamon Press.

Bonifay, W., Lane, S. P., & Reise, S. P. (2016). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychological Science*, *5*(1), 184–186. https://doi.org/10.1177/2167702616657069

Bosma, C. M., & Granger, A. M. (2022). Sharing is caring: Ethical implications of transparent research in psychology. *American Psychologist*, *77*(4), 565–575. https://doi.org/10.1037/amp0001002

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. L. Wilkinson, *Robustness in statistics* (pp. 201–236). https://doi.org/10.1016/b978-0-12-438150-6.50018-2

Buchanan, E. M., Crain, S. E., Cunningham, A. L., Johnson, H. R., Stash, H., Papadatou-Pastou, M., Isager, P. M., Carlsson, R., & Aczel, B. (2021). Getting started creating data dictionaries: How to create a shareable data set. *Advances in Methods and Practices in Psychological Science*, *4*(1). https://doi.org/10.1177/2515245920928007

Burger, J., Isvoranu, A.-M., Lunansky, G., Haslbeck, J. M. B., Epskamp, S., Hoekstra, R. H. A., Fried, E. I., Borsboom, D., & Blanken, T. F. (2022). Reporting standards for psychological network analyses in cross-sectional data. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000471

Chester, D. S., & Lasko, E. N. (2021). Construct validation of experimental manipulations in social psychology: Current practices and recommendations for the future. *Perspectives on Psychological Science*, *16*(2), 377–395. https://doi.org/10.1177/1745691620950684

Chin, J. M., Pickett, J. T., Vazire, S., & Holcombe, A. O. (2023). Questionable research practices and open science in quantitative criminology. *Journal of Quantitative Criminology*, *39*, 21–51. https://doi.org/10.1007/s10940-021-09525-6

Cho, E. (2021). Neither Cronbach's Alpha nor McDonald's Omega: A commentary on Sijtsma and Pfadt. *Psychometrika*, *86*(4), 877–886. https://doi.org/10.1007/s11336-021-09801-1

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods*, *22*(3), 541–562. https://doi.org/10.1037/met0000083

Epskamp, S. (2019). Reproducibility and replicability in a fast-paced methodological world. *Advances in Methods and Practices in Psychological Science*, *2*(2), 145–155. https://doi.org/10.1177/2515245919847421

Ferrando, P. J., Lorenzo-Seva, U., Hernández-Dorado, A., & Muñiz, J. (2022). Decálogo para el análisis factorial de los ítems de un test [Decalogue for the factor analysis of test items]. *Psicothema*, *34*, 7–17. https://doi.org/10.7334/psicothema2021.456

Fife, D. A., & Rodgers, J. L. (2022). Understanding the exploratory/confirmatory data analysis continuum: Moving beyond the "replication crisis". *American Psychologist*, *77*(3), 453–466. https://doi.org/10.1037/amp0000886

Flake, J. K., Davidson, I. J., Wong, O., & Pek, J. (2022). *Construct validity and the validity of replication studies: A systematic review*. PsyArXiv. https://doi.org/10.31234/osf.io/369qj

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Flores-Kanter, P. E., Dominguez-Lara, S., Trógolo, M. A., & Medrano, L. A. (2018). Best practices in the use of bifactor models: Conceptual grounds, fit indices and complementary indicators. *Revista Evaluar*, *18*(3). 44–48. https://doi.org/10.35670/1667-4545.v18.n3.22221

Flores-Kanter, P. E., Garrido, L. E., Moretti, L. S., & Medrano, L. A. (2021). A modern network approach to revisiting the Positive and Negative Affective Schedule (PANAS) construct validity. *Journal of Clinical Psychology*, *77*(10), 2370–2404. https://doi.org/10.1002/jclp.23191

Flores-Kanter, P. E., Toro, R., & Alvarado, J. M. (2022). Internal Structure of Beck Hopelessness scale: An analysis of method effects using the CT-C(M–1) model. *Journal of Personality Assessment*, *104*, 408–416. https://doi.org/10.1080/00223891.2021.1942021

Fried, E. I., Flake, J. K., & Robinaugh, D. J. (2022). Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology*, *1*(6), 358–368. https://doi.org/10.1038/s44159-022-00050-2

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460. https://doi.org/10.1511/2014.111.460

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PLOS ONE*, *12*(6), Article e0174035. https://doi.org/10.1371/journal.pone.0174035

Haywood, D., Baughman, F. D., Mullan, B. A., & Heslop, K. R. (2021). Going "up" to move forward: S-1 Bifactor models and the study of neurocognitive abilities in psychopathology. *International Journal of Environmental Research and Public Health*, *18*(14), 7413. https://doi.org/10.3390/ijerph18147413

Heinrich, M., Zagorscak, P., Eid, M., & Knaevelsrud, C. (2018). Giving g a meaning: An application of the bifactor-(S-1) approach to realize a more symptom-oriented modeling of the Beck Depression Inventory–II. *Assessment*, *27*(7), 1429–1447. https://doi.org/10.1177/1073191118803738

Kirtley, O. J., Janssens, J. J., & Kaurin, A. (2022). Open science in suicide research is open for business. *Crisis*, *43*(5), 355–360. https://doi.org/10.1027/0227-5910/a000859

Laurinavichyute, A., Yadav, H., & Vasishth, S. (2022). Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy. *Journal of Memory and Language*, *125*, Article 104332. https://doi.org/10.1016/j.jml.2022.104332

Levenstein, M. C., & Lyle, J. A. (2018). Data: Sharing is caring. *Advances in Methods and Practices in Psychological Science*, *1*(1), 95–103. https://doi.org/10.1177/2515245918758319

Lewis, N. A., Jr. (2021). What counts as good science? How the battle for methodological legitimacy affects public psychology. *American Psychologist*, *76*(8), 1323–1333. https://doi.org/10.1037/amp0000870

Lilienfeld, S. O., & Strother, A. N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology/Psychologie Canadienne*, *61*(4), 281–288. https://doi.org/10.1037/cap0000236

https://doi.org/10.1017/SJP.2023.14 Published online by Cambridge University Press

Lloret-Segura, S., Ferreres-Traver, A., Hernández-Baeza, A., & Tomás-Marco, I. (2014). El análisis factorial exploratorio de los ítems: Una guía práctica, revisada y actualizada [Exploratory item factor analysis: A practical guide revised and up-dated]. *Anales de Psicología*, *30*(3), 1151–1169. https://doi.org/10.6018/analesps.30.3.199361

Luong, R., & Flake, J. K. (2022). Measurement invariance testing using confirmatory factor analysis and alignment optimization: A tutorial for transparent analysis planning and reporting. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000441

Manapat, P. D., Anderson, S. F., & Edwards, M. C. (2022). A revised and expanded taxonomy for understanding heterogeneity in research and reporting practices. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000488

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*(1), 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

Matsunaga, M. (2008). Item parceling in structural equation modeling: A primer. *Communication Methods and Measures*, *2* (4), 260–293. https://doi.org/10.1080/19312450802458935

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

McNeish, D., & Wolf, M. G. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods.* Advance online publication. https://doi.org/10.1037/met0000425

Mellor, D. T., Vazire, S., & Lindsay, D. S. (2018). *Transparent science: A more credible, reproducible, and publishable way to do science.* PsyArXiv. https://doi.org/10.31234/osf.io/7wkdn

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science*, *1*(1), 131–144. https://doi.org/10.1177/2515245917747656

Moreau, D., & Gamble, B. (2020). Conducting a meta-analysis in the age of open science: Tools, tips, and practical recommendations. *Psychological Methods*, *27*(3), 426–432. https://doi.org/10.1037/met0000351

Morin, A. J. S., Myers, N. D., & Lee, S. (2020). Modern factor analytic techniques. In G. Tenenbaum & R. C. Eklund (Eds.), *Handbook of sport psychology*, 1044–1073. Portico. https://doi.org/10.1002/9781119568124.ch51

Moshontz, H., Binion, G., Walton, H., Brown, B. T., & Syed, M. (2021). A guide to posting and managing preprints. *Advances in Methods and Practices in Psychological Science*, *4*(2). https://doi.org/10.1177/25152459211019948

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), Article 0021. https://doi.org/10.1038/s41562-016-0021

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*(1), 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*(1), 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

PsyTeachR Team. (2022). Open-source tutorials benefit the field. *Nature Reviews Psychology*, *1*, 312–313. https://doi.org/10.1038/s44159-022-00058-8

Raykov, T., Dimitrov, D. M., Marcoulides, G. A., & Harrison, M. (2017). On the connections between item response theory and classical test theory: A note on true score evaluation for polytomous items via item response modeling. *Educational and Psychological Measurement*, *79*(6), 1198–1209. https://doi.org/10.1177/0013164417745949

Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? Application of iteratively reweighted least squares to the Rosenberg Self-Esteem scale. *Multivariate Behavioral Research*, *51*, 818–838. https://doi.org/10.1080/00273171.2016.1243461

Rohrer, J. M., Hünermund, P., Arslan, R. C., & Elson, M. (2022). That's a lot to process! Pitfalls of popular path models. *Advances in Methods and Practices in Psychological Science*, *5*(2). https://doi.org/10.1177/25152459221095827

Rosenbusch, H., Wanders, F., & Pit, I. L. (2020). The Semantic Scale Network: An online tool to detect semantic overlap of psychological scales and prevent scale redundancies. *Psychological Methods*, *25*(3), 380–392. https://doi.org/10.1037/met0000244

Sarnacchiaro, P., & Boccia, F. (2018). Some remarks on measurement models in the structural equation model: An application for socially responsible food consumption. *Journal of Applied Statistics*, *45*(7), 1193–1208. https://doi.org/10.1080/02664763.2017.1363162

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Steneck, N. H. (2006). Fostering integrity in research: Definitions, current knowledge, and future directions. *Science and Engineering Ethics*, *12*(1), 53–74.

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., Oltmanns, T. F., & Shrout, P. E. (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science*, *12*(5), 742–756. https://doi.org/10.1177/1745691617690042

Tackett, J. L., Brandes, C. M., & Reardon, K. W. (2019). Leveraging the Open Science Framework in clinical psychological assessment research. *Psychological Assessment*, *31*(12), 1386–1394. https://doi.org/10.1037/pas0000583

Tijdink, J. K., Horbach, S. P. J. M., Nuijten, M. B., & O'Neill, G. (2021). Towards a Research Agenda for Promoting Responsible Research Practices. *Journal of Empirical Research on Human Research Ethics*, *16*(4), 450–460. https://doi.org/10.1177/15562646211018916

Tuval-Mashiach, R. (2017). Raising the curtain: The importance of transparency in qualitative research.

https://doi.org/10.1017/SJP.2023.14 Published online by Cambridge University Press

Qualitative Psychology, *4*(2), 126–138. https://doi.org/10.1037/qup0000062

Ulrich, R., & Miller, J. (2018). Some properties of *p*-curves, with an application to gradual publication bias. *Psychological Methods*, *23*(3), 546–560. https://doi.org/10.1037/met0000125

Wagenmakers, E.-J., Sarafoglou, A., & Aczel, B. (2022). One statistical analysis must not rule them all. *Nature*, *605*(7910), 423–425. https://doi.org/10.1038/d41586-022-01332-8

Waldman, I. D., & Lilienfeld, S. O. (2016). Thinking about data, research methods, and statistical analyses: Commentary on Sijtsma's (2014) "Playing with data." *Psychometrika*, *81*(1), 16–26. https://doi.org/10.1007/s11336-015-9447-z

Widaman, K. F., & Revelle, W. (2023). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods*, *55*, 788–806. https://doi.org/10.3758/s13428-022-01849-w

https://doi.org/10.1017/SJP.2023.14 Published online by Cambridge University Press