Check for updates

Empirical Article

Advances in Methods and Practices in Psychological Science
January-March 2025, Vol. 8, No. 1, pp. 1–16
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/25152459241306432
www.psychologicalscience.org/AMPPS

VALID: A Checklist-Based Approach for Improving Validity in Psychological Research

Susanne Kerschbaumer[1], Martin Voracek[1], Balazs Aczél[2], Samantha F. Anderson[3], Brandon M. Booth[4], Erin M. Buchanan[5], Rickard Carlsson[6], Daniel W. Heck[7], Anu Pauliina Hiekkaranta[8], Rink Hoekstra[9], Julian D. Karch[10], Ginette Lafit[11], Zhicheng Lin[12,13], Siwei Liu[14], David P. MacKinnon[3], Emma L. McGorray[15], David Moreau[16], Marietta Papadatou-Pastou[17], Helena Paterson[18], Robert A. Perera[19], Daniel J. Schad[20], David K. Sewell[21], Moin Syed[22], Louis Tay[23], Jorge N. Tendeiro[24], Michael D. Toland[25], Wolf Vanpaemel[11], Don van Ravenzwaaij[26], Lieke Voncken[27], and Ulrich S. Tran[1]

[1]Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna, Vienna, Austria; [2]Institute of Psychology, ELTE Eotvos Lorand University, Budapest, Hungary; [3]Department of Psychology, Arizona State University, Tempe, Arizona; [4]Department of Computer Science, University of Memphis, Memphis, Tennessee; [5]Analytics, Harrisburg University of Science and Technology, Harrisburg, Pennsylvania; [6]Department of Psychology, Linnaeus University, Kalmar, Sweden; [7]Department of Psychology, University of Marburg, Marburg, Germany; [8]Center for Contextual Psychiatry, Department of Neurosciences, KU Leuven, Leuven, Belgium; [9]Department of Educational Sciences, University of Groningen, Groningen, Netherlands; [10]Methodology and Statistics Department, Institute of Psychology, Leiden University, Leiden, Netherlands; [11]Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium; [12]Department of Psychology, University of Science and Technology of China, Hefei, China; [13]School of Humanities and Social Science, The Chinese University of Hong Kong, Shenzhen, China; [14]Department of Human Ecology, University of California Davis, Davis, California; [15]Department of Psychology, Northwestern University, Evanston, Illinois; [16]School of Psychology and Centre for Brain Research, The University of Auckland, Auckland, New Zealand; [17]Department of Primary Education, National and Kapodistrian University of Athens, Athens, Greece; [18]School of Psychology and Neuroscience, University of Glasgow, Glasgow, Scotland; [19]Department of Biostatistics, School of Medicine, Virginia Commonwealth University, Richmond, Virginia; [20]Institute for Mind, Brain and Behavior, HMU Health and Medical University, Potsdam, Germany; [21]School of Psychology, The University of Queensland, St Lucia, Australia; [22]Department of Psychology, University of Minnesota, Minneapolis, Minnesota; [23]Department of Psychological Sciences, Purdue University, West Lafayette, Indiana; [24]Graduate School of Advanced Science and Engineering, Hiroshima University, Higashihiroshima, Japan; [25]Herb Innovation Center, University of Toledo, Toledo, Ohio; [26]Department of Psychometrics and Statistics, University of Groningen, Groningen, Netherlands; and [27]Department of Methodology and Statistics, Tilburg University, Tilburg, Netherlands

Abstract

In response to the replication and confidence crisis across various empirical disciplines, ensuring the validity of research has gained attention. High validity is crucial for obtaining replicable and robust study outcomes when both exploring new questions and replicating previous findings. In this study, we aimed to address this issue by developing a comprehensive

Corresponding Author:
Susanne Kerschbaumer, Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna, Vienna, A-1010, Austria
Email: susanne.kerschbaumer@univie.ac.at

Creative Commons NonCommercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/), which permits noncommercial use, reproduction, and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

checklist to assist researchers in enhancing and monitoring the validity of their research. After systematically analyzing previous findings on validity, a comprehensive list of potential checklist items was compiled. Over the course of three rounds, more than 30 interdisciplinary and psychological-science experts participated in a Delphi study. Experts rated items on their importance and were given the opportunity to propose novel items as well as improve existing ones. This process resulted in a final set of 91 items, organized according to common stages of a research project. The VALID checklist is accessible online (https://www.validchecklist.com/) and provides researchers with an adaptable, versatile tool to monitor and improve the validity of their research and to suit their specific needs. By focusing on adaptiveness during its development, VALID encompasses 331 unique checklist versions, making it a one-stop solution suitable for a wide range of projects, designs, and requirements.

Keywords
validity, tailored online checklist, open practices, method reform, study quality appraisal, Delphi study, open data, open materials, preregistered

Received 2/16/24; Revision accepted 11/19/24

The ongoing replication and confidence crisis in multiple empirical disciplines has renewed the attention on research validity, whether in testing new research questions or replicating prior findings. For psychological research to have a positive impact on society, its findings must be valid first, and therefore, the validity of research is of utmost importance (Kenny, 2019). The term "validity" has been in use for decades (Newton & Shaw, 2013), and validity considerations date back as far as the discipline itself (Slaney, 2017). Over the course of more than 20 years, Campbell and colleagues first proposed the terms "internal validity" and "external validity" (Campbell, 1957; Campbell & Stanley, 1963), which later were expanded by the concepts "construct validity" (originally devised by Cronbach & Meehl, 1955) and "statistical-conclusion validity" (Cook & Campbell, 1979; Shadish et al., 2002). Whereas internal validity focuses on the identification of the independent variable as cause of the change in the dependent variable, external validity is concerned with the generalizability of study results (Campbell, 1957; Campbell & Stanley, 1963). Construct validity refers to the appropriate operationalization of the theoretical concept, and statistical-conclusion validity focuses on the conclusions drawn from the data analysis (Cook & Campbell, 1979).

Questionable research practices, such as *p*-hacking, selective reporting (Nelson et al., 2018), and HARKing (Kerr, 1998), pose a threat to the validity and integrity of psychological research. Nelson et al. (2018) illustrated that those practices are rarely employed intentionally or with ill meaning; rather, they are unknowingly used by well-meaning researchers. Therefore, it is important to support researchers in avoiding those practices and ensuring the validity of their research. Although scholarly debates surrounding the replication crisis have emphasized the importance of valid and robust research, recommendations on reaching this goal are scattered throughout the methods literature. One of the goals of our account was thus to bring together those recommendations.

Unfortunately, there is a lack of consensus and guidance on how to ensure the validity of an empirical research project throughout its different stages. Other aspects of research have been addressed in initiatives such as the PRISMA guidelines for systematic reviews (Page et al., 2021), which are accompanied by a checklist to facilitate the implementation by researchers. Evaluations of systematic reviews published before and after the original release of the PRISMA guidelines (Moher et al., 2009) show an increase in quality of reporting, completeness of reporting, and methodological quality (Panić et al., 2013; Tunis et al., 2013), suggesting the usefulness and success of checklists for enhancing research quality.

Although various authors have examined one or multiple types of validity and shared recommendations on how to improve them (e.g., for a review of checklists on external validity, see Dyrvig et al., 2014; for a review of validity guidelines applicable to in vivo animal studies, see Henderson et al., 2013), no overarching tool built to complement the research process has yet been designed.

One published, thematically adjacent initiative focused specifically on helping reviewers assess the validity of submissions (Seaboat; Schiavone et al., 2023). Although this tool centers around four types of validity (internal validity, external validity, construct validity, statistical-conclusion validity) and can be used for more than only reviewing submissions (e.g., reviewing one's own work), it is not explicitly built to support researchers during their projects. To address this gap, we aimed to develop a checklist structured by the stages of the research process (planning, data collection, data analysis, reporting)

that provides targeted support during the whole research process. A particular focus was set on the planning of research because this has not yet been addressed in any other tool that we are aware of.

The starting point of the item-development process was Campbell, Cook, and Stanley's previously introduced framework for research validity. This four-validity framework has recently been used to assess and discuss the state of psychological science (e.g., Fabrigar et al., 2020; Kenny, 2019; Schiavone et al., 2023). It was thus used in the current study to build on previous findings and benefit from most researchers' familiarity with these terms (Schiavone et al., 2023).

In the current work, we present an online tool that automatically generates an adaptive checklist tailored to a specific study design of interest. The outcome, the VALID checklist, is designed with the researcher and the research process in mind and should serve as an easy-to-use roadmap guiding researchers in the enhancement of the validity of their projects.

Disclosure

This study was preregistered on December 15, 2022, on OSF (https://doi.org/10.17605/OSF.IO/GX2K6), and all Delphi study surveys, the full results of all Delphi study rounds, and further information on the development of the VALID checklist are provided on OSF (https://doi.org/10.17605/OSF.IO/GWAVU). The source code for the VALID checklist tool is openly accessible on GitHub (https://github.com/susik98/VALID-Checklist). An initial literature review was started before the preregistration of the study, but the systematic review of the literature to generate the checklist items (Stage 1; see below), which is detailed in the preregistration, started only after preregistration. The Delphi study was preregistered before its start. One aspect of the Delphi study's rating procedure was, however, changed during the study because of participants' feedback. This change is explained in detail in the Method section. We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. The study involved only informed researchers as voluntary, expert-status participants; hence, ethical approval was not sought for.

Our study comprised three stages; the Delphi study was the second and main stage. Figure 1 provides an overview of the development process of our tool and key information on the individual stages of this process. The article is organized according to these stages (instead of using general method and results sections) because this allows us to present the methods and results within the context of each stage, thereby providing a clearer narrative of our research process.

Stage 1: Literature Review

A preliminary list of items was created by means of a literature review in December 2022. Google Scholar served as a starting point for the search of works on the validity of empirical research. For each of the used keywords/combinations (see Table 1 and Appendix A in the Supplemental Material available online), the first five pages of results (50 results) were screened for eligibility. With the development of the checklist in mind, we used keywords focused on the improvement and/or lack of

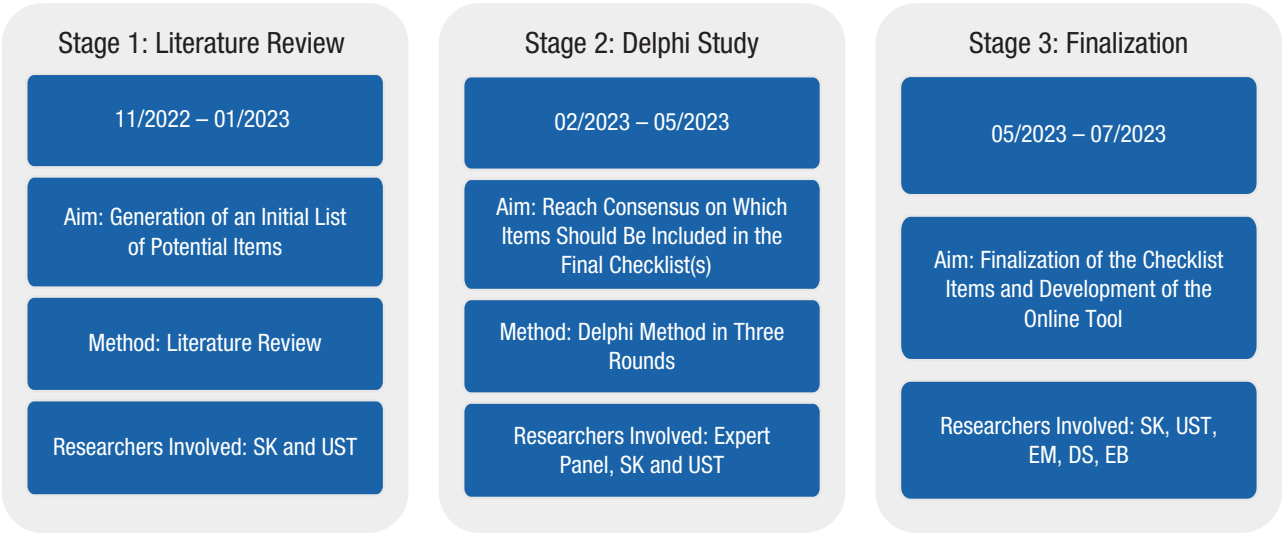


Fig. 1. Overview of the development process of the tool. SK = S. Kerschbaumer; UST = U. S. Tran; EM = E. McGorray; DS = D. Sewell; EMB = E. M. Buchanan.

Table 1. Literature Search Keywords and Combinations

Keywords	
German	Validität erhöhen, Validität Psychologie Experiment
English	Validity experiments, lack of validity psychology, lack of validity psychology experiment, improving validity experiment, improving validity research, validity experiments, validity psychological research, improving validity psychology, validity checklist experiment, external validity checklist, validity checklist, validity experiment psychology, validity guideline, validity study guidelines

validity, general guidance in research validity, and research validity itself. Afterward, backward and forward citation searches (partially via Connected Papers; https://www.connectedpapers.com) were used to find additional works that might be of relevance. Works recommended by colleagues that had not been already identified through this literature search were also included in the literature corpus.

The literature review focused on English- and German-language works even though we were open to include relevant French literature as well. Articles were deemed eligible if they either discussed issues related to any type of validity or provided guidance on the improvement of research validity. Neither issues nor guidance had to be the main part of the work; any mention of either qualified the article for inclusion. In addition, there were no publication-date limitations to the search or inclusion of articles. After completion of Stage 1, during the beginning of Stage 2 (see Stage 2: Delphi Study section), we were made aware of the validity tool by Schiavone et al. (2023). Because the manuscript was kindly provided to us by the authors before its dissemination as a preprint, we were able to screen the literature reviewed in their study as well and include any works that were of relevance for the current tool and that had not already been included before.

The literature included for this study was published between 1998 and 2023 and contained works authored by a total of 105 researchers from diverse fields. Articles were published in 28 different journals (27 in English, one in German), but the literature corpus also included four monographs (Hartig et al., 2020; Krantz & Dalal, 2000; Krosnick, 2018; Kuhlmann, 2007). The largest number of articles from a single journal were from the integrative psychological journal *Perspectives on Psychological Science* (six articles). However, relevant literature was also from fields other than psychology, for example, from pharmacology and pharmacy (*American Journal of Health-System Pharmacy*), computer and information science (*European Journal of Information Systems*), medicine (*PLOS Medicine*), and philosophy of science (*Philosophy of the Social Sciences*). For a PRISMA flowchart (Page et al., 2021) of the literature review, see Figure 2;

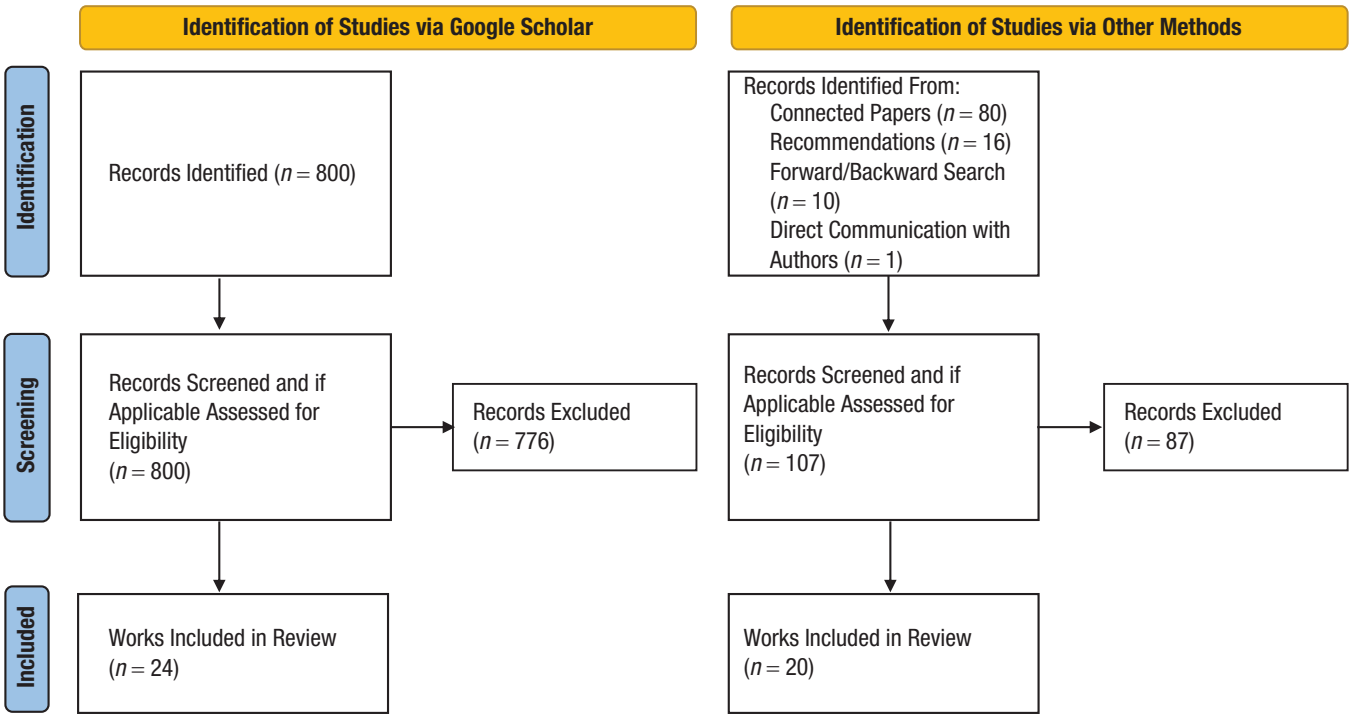


Fig. 2. PRISMA flowchart of literature review.

for the full list of included sources, see Appendix B in the Supplemental Material.

From the compiled literature, 94 items were extracted by S. Kerschbaumer, revised for clarity by U. S. Tran, and accordingly adjusted based on his feedback. For the items, either explicit recommendations by the article's authors were adopted or criticism of existing practices was reformulated as items. The preliminary list of 94 items (see Appendix C in the Supplemental Material) was then sorted by research-project stage (planning, data collection, data analysis, reporting) and grouped into smaller categories, such as "power" or "theoretical foundation." In addition, some items were marked as being not generally applicable but only in certain cases, for example, because of resource limitations or design specifications. This was specified in the item texts by modifiers such as "If: Use of groups/conditions" or modifiers akin to "If possible."

Stage 2: Delphi Study

Methodological considerations

Delphi studies typically consist of several rounds during which the experts' opinions are anonymously collected with the goal of reaching consensus on the specific question asked (Niederberger & Renn, 2023). The anonymity of experts during the rounds allows for the expression of opinions uninfluenced by other participants. In addition, because of the option of using online tools, the Delphi method allows for the inclusion of international experts and collaboration of geographically dispersed participants (Donohoe & Needham, 2009).

Although there is limited empirical evidence regarding best practices for this technique, which raises some concerns (for a review of 80 Delphi studies in the field of health care, see Belton et al., 2019; Boulkedid et al., 2011; Humphrey-Murto & de Wit, 2019), there is potential for a thoughtful and effective application of this method.

One concern is the lack of consensus on the definition of consensus itself and the absence of a priori decisions on its understanding in each specific case (Diamond et al., 2014). There are also no clear recommendations regarding the choice of participants, their number, and the number of rounds (Belton et al., 2019; Humphrey-Murto & de Wit, 2019). Belton et al. (2019) suggested using common sense to define inclusion criteria and search strategies for inviting experts.

Despite several aspects of the Delphi method yet to be critically and systematically examined, it still provides a useful tool for measuring experts' opinions on topics, especially when combined with preregistration of its implementation and transparent reporting of all choices made and results found (Belton et al., 2019; Hohmann et al., 2018; Humphrey-Murto & de Wit, 2019; Niederberger & Renn, 2023; Taylor, 2019).

Participants

We compiled a list of experts for the Delphi study by extracting all authors of articles published in the years 2021 and 2022 in the two flagship journals for research methods *Psychological Methods* (Volumes 26 and 27) and *Advances in Methods and Practices in Psychological Science* (Volumes 4 and 5). These two journals both focus on methodological issues currently relevant to the field of psychology, and we thus assumed authors to be highly familiar with the topic of validity in psychological research. We aimed to assemble a panel of 20 experts (which is the common upper bound of the minimum range recommended by Belton et al., 2019, and reported for published Delphi studies by Taylor, 2019) and expected a response rate of approximately 5%. Therefore, we planned to contact around 400 experts. The goal was to allow for a broad range of opinions, backgrounds, and academic ages among the experts. Consequently, we aimed for a pragmatically large sample and sought to also include authors not currently working in research institutes and universities (e.g., practitioners in hospitals, statistical-software developers). Whereas most participants were affiliated with universities, colleges, or research institutes, some individuals indicated working in hospitals and commercial institutions. For details on the initial search for experts, see Figure 3 and Appendix D in the Supplemental Material.

After removing duplicates and authors with missing contact information (see Fig. 3), all 358 remaining authors who had published in the selected journals during the 2-year period were invited to participate in a Delphi Study concerning the validity of psychological research via email at the beginning of February 2023. Exceptions were made in the case of three large studies (> 20 authors), for which only the first/corresponding author was invited to not bias the sample, and articles published in special sections, which were not included because of the conceivable specificity of these contents. The opportunity for coauthorship was offered to all experts as an acknowledgment of their contribution. Appendix D in the Supplemental Material contains the email text(s) sent out to all experts and a detailed report on the number of emails (not) successfully delivered.

During the sign-up period for the Delphi study (February 8–26, 2023), 58 experts signed up to take part in the project, of which about 70% were male. For self-reported key demographic and professional data on these signed-up experts, see Table 2 and Figure 4. All signed-up experts were invited to participate in each of the rounds because the invitation did not depend on participation in all previous rounds.

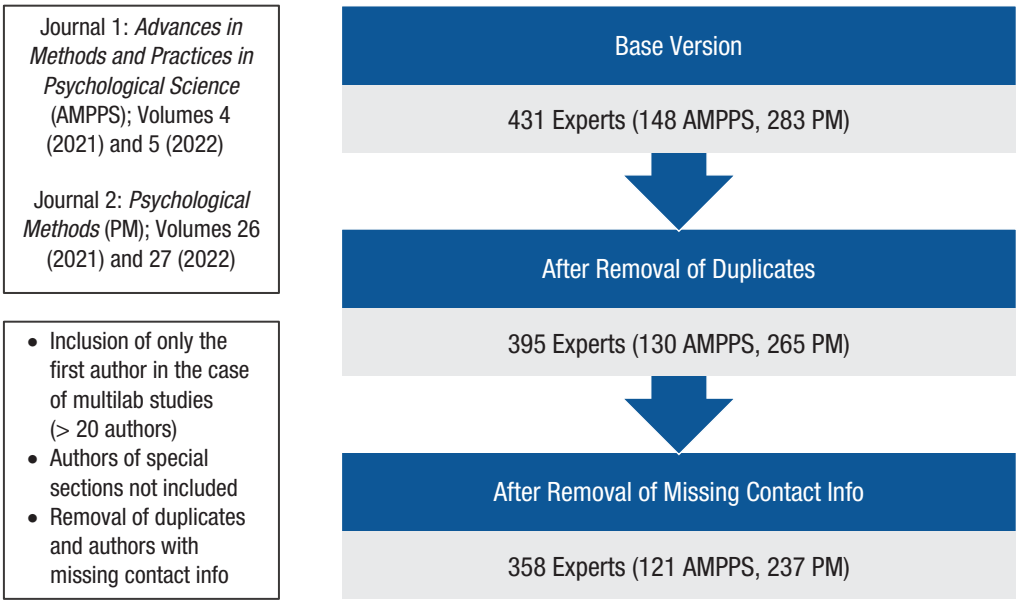


Fig. 3. Expert search strategy and expert list.

Rating procedure

For the rating of the items, the online survey tool SoSciSurvey with a server at the University of Vienna (https://sosci.univie.ac.at) was used. The surveys of each round of the Delphi study were accessible through a password-protected link sent out via email to the experts for a duration of approximately 2 weeks. The surveys contained a short introduction to the project and the rating procedure, the used definition of validity, a consent form, and the items to be rated. During the second round, additional information on the rating procedure was provided in response to open questions after the first round.

To decide on the inclusion of an item in the final checklist, the experts were asked to rate the importance of the inclusion of each individual item on a 5-point Likert-type scale (1 = *strongly disagree*, 5 = *strongly agree*; the additional response option of "unclear" was included in case the experts did not feel able to rate the item). In addition, open-ended questions offered the experts a chance to provide feedback on the items' contents or propose new ones that were missing. For an item to be included in the checklist, at least 70% of the experts needed to rate the item with 4 or 5 (this criterion was applied recently also by Sharma et al., 2021). If an item did not reach this cutoff in its first round, it was modified based on the experts' feedback and rerated in the next round. Items that did not receive any feedback and did not reach the minimum agreement to be included but had an agreement rate of at least 40% were rerated in their original form in the next round. To keep the survey at a manageable length, which we deemed important to minimize dropouts, items that did not reach a minimum agreement of 40% ratings of 4 or 5 were excluded. This exclusion criterion was not preregistered.

After each round, the ratings for the individual items and any comments were compiled and analyzed. The items were then grouped into three categories: red (excluded), yellow (neither excluded nor included, to be rerated), and green (included). If the green items received minor feedback, they were adjusted without rerating. If the feedback led to a substantial change in the item, it was rerated in its new form. The degree of change was determined by S. Kerschbaumer and U. S. Tran, and agreement was reached in each case. Based on feedback received during the second round and to avoid a high number of yellow items, the rating procedure was adapted during the last round such that experts had to vote directly in favor of or against the inclusion of each remaining item (i.e., forced choice).

Delphi study results

The following section provides a detailed picture of the Delphi study's results sorted by the Delphi study's rounds. Figure 5 contains an overview of the study's outcome in the form of a flowchart.

Round 1 of the Delphi study. In the first round (February 27 to March 14, 2023), 38 experts rated a total of 92[1] items, of which (a) 45 were directly accepted into the checklist, (b) 40 were to be rerated during the next round,

Table 2. Academic Research Experience, Self-Reported Main Scopus Subject Areas, and Geographical Distribution of Participating Experts

Characteristics	*n*	%
Research experience		
< 5 years	1	2
5–10 years	21	36
11–20 years	25	43
> 20 years	11	19
Scopus subject areas		
Psychology	39	67
Multidisciplinary	9	16
Social sciences	6	10
Computer science	2	3
Health professions	1	2
Other	1	2
Countries		
United States	22	40
Netherlands	9	16
Germany	6	11
Belgium	3	5
Scotland	2	4
Norway	2	4
Canada	2	4
Sweden	2	4
China	1	2
Greece	1	2
Australia	1	2
New Zealand	1	2
Northern Ireland	1	2
Hungary	1	2
Japan	1	2

and (c) seven were rejected (Fig. 6). In addition, 10 new items were proposed by the experts to be rated in the second round. The analysis of 146 detailed open-ended comments provided by the experts showed that the initial list was perceived as rather focused on null hypothesis significance testing (NHST; e.g., "I believe the use of NHST, or alternatives, should be the choice of the researcher") and that the rating of items applicable in only certain contexts (e.g., designs using groups and/or conditions) was not entirely clear to some experts.

The first issue, the focus on NHST, was partially solved by including new items on Bayesian analysis that were proposed during this first round. In addition, we explicitly asked the experts in the invitation email and the introduction section of the second-round survey to propose new items to make the checklist more well rounded regarding different approaches to significance testing and data analysis. To solve the second issue (applicability of items in certain contexts only), additional information on the rating of context-specific items was provided at the beginning of the second survey.

Out of the 92 items, 28 items (approximately 30%) received three or more ratings of "unclear." S. Kerschbaumer and U. S. Tran discussed each individual case with a focus on making the item wording clearer and more concise. Because of a high number of unclear ratings and/or comments by the experts expressing that the item is not suitable for the checklist, five of the yellow items were additionally excluded following their review by S. Kerschbaumer and U. S. Tran. Three already accepted items were minorly adapted based on the experts' feedback, and 15 previously accepted items underwent substantial changes and were thus rerated in their new form in the next round. For the full results of this and all following rounds, see the OSF project webpage (https://doi.org/10.17605/OSF.IO/GWAVU).

Round 2 of the Delphi study. The second rating round was open from March 27 to April 14, 2023, during which 33 experts rated 60 items. This led to the direct acceptance of 37 items, and three items were rejected. Twenty items were categorized as yellow and thus improved based on the experts' comments before moving on to the third round. Following the call for new items, 11 new items on Bayesian statistics were proposed, and three items relevant for users of linear mixed-effects models were proposed. In addition, one new item concerning the connection between theory and measurement was introduced.

The analysis of the second round's 79 open-ended comments ultimately led to the change of the rating system in the third round because one expert mentioned difficulties in rating the items on a scale of this granularity. Apart from this, the comments included the suggestion of a more consistent wording of the items (e.g., change of "Consideration of a wide range of similar concepts in discriminant validity analysis" to "Considering a wide range of similar concepts in discriminant validity analysis" to be in line with other items such as "Pretesting/piloting of experimental manipulations"). The number of unclear items decreased notably, and only eight out of 60 items (approximately 13%) received an unclear rating. Again, because of a high number of unclear ratings and/or comments by the experts expressing that the item is not suitable for the checklist, five of the yellow items were additionally excluded following their review by S. Kerschbaumer and U. S. Tran. Out of the accepted items, 12 were slightly adapted based on the experts' feedback without being rerated, and two items were significantly changed and thus rerated in the third round.

Round 3 of the Delphi study. During the last round (April 24 to May 9, 2023), which consisted of 32 items, 25 items were accepted into the checklist, and seven items were excluded. In this round, 34 experts participated in the rating process. During the analysis of the 70 open-ended

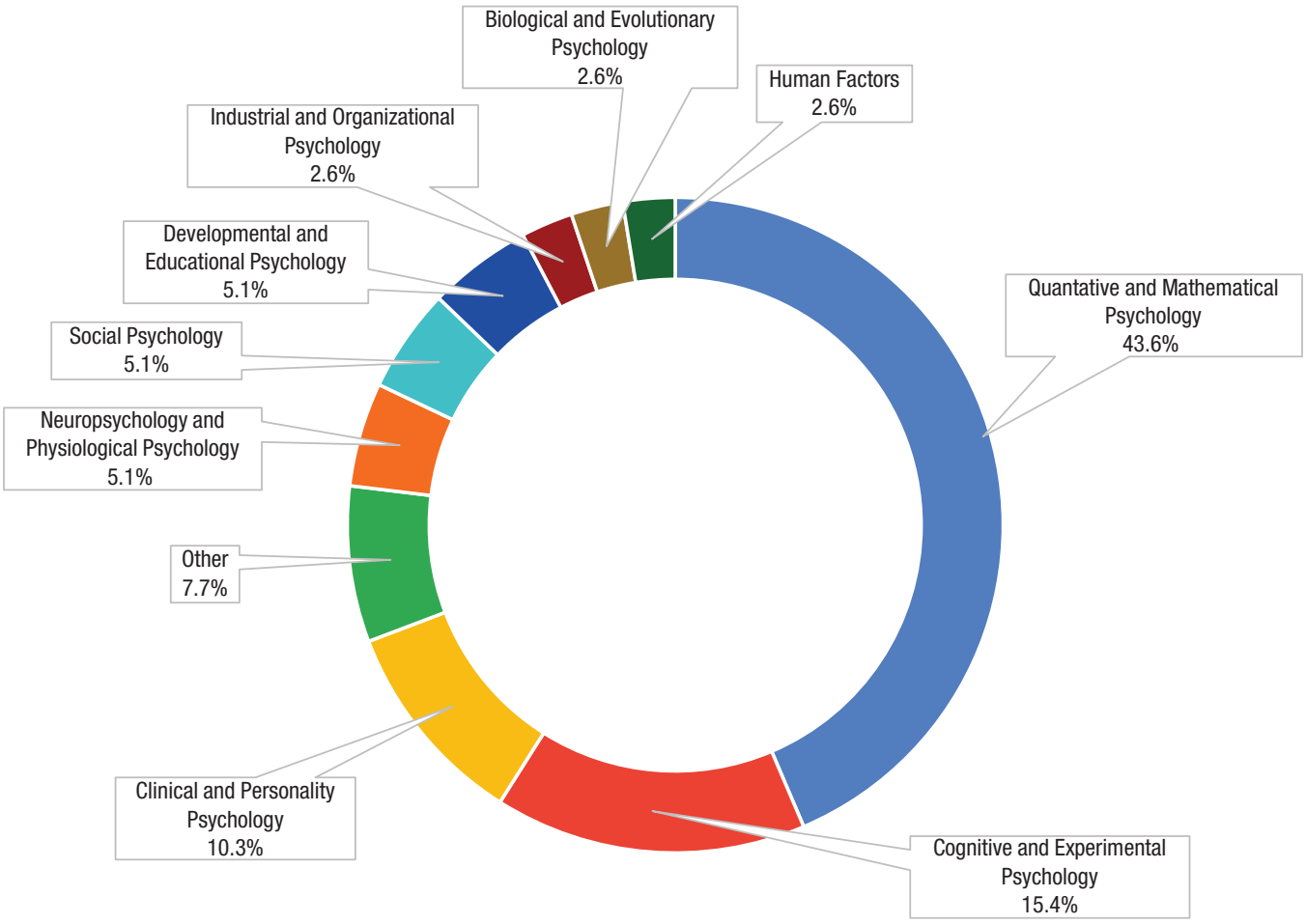


Fig. 4. Self-reported subfields of experts in psychology.

comments, we recognized that some of the experts rated items as unclear if they did not feel capable to rate it, for example, because of a lack of familiarity with the specific analysis technique (e.g., "While I have work[e]d with Bayesian analyses, not being an expert on the topic I responded with 'unclear' when I wasn't entirely familiar with the procedure being described"). Unclear ratings affected mostly the items on Bayesian analysis. Because the inclusion of an item was determined by the percentage of ratings of 4 or 5 (or "yes" in the last round), a high number of unclear ratings affected its acceptance. Yet the comments indicated that some or all the unclear ratings could be understood as "I do not wish to vote" rather than "should not be included." In an open discussion, S. Kerschbaumer and U. S. Tran thus decided to include two items on Bayesian analysis that were marginally under the limit of 70% agreement (66.67% and 69.70%). Furthermore, one new item proposed during the third round ("Following established reporting guidelines (e.g., Bayesian Analysis Reporting Guidelines (Kruschke, 2021) or the Discussion points for Bayesian inference (Aczél et al., 2020)") was directly added to the checklist without the rating of the experts because both S. Kerschbaumer and U. S. Tran agreed on its importance.

After the three rounds, the final list contained a total of 91 items for the stages of planning (39 items), data collection (10 items), data analysis (22 items), and data reporting (20 items).

Stage 3: Grouping of Items Into Categories

After completion of the rating procedure, the items were grouped by S. Kerschbaumer into categories to allow for the adoption of the VALID checklist to individual projects. This grouping was discussed with U. S. Tran and finalized after reaching a consensus on every item through thorough discussion. In addition to the grouping of the items, two native English speakers recruited from the expert panel (D. Sewell, E. M. Buchanan) volunteered to proofread the items, especially focusing on clarity and conciseness. Their feedback and minor remarks of U. S. Tran on the order of items were included

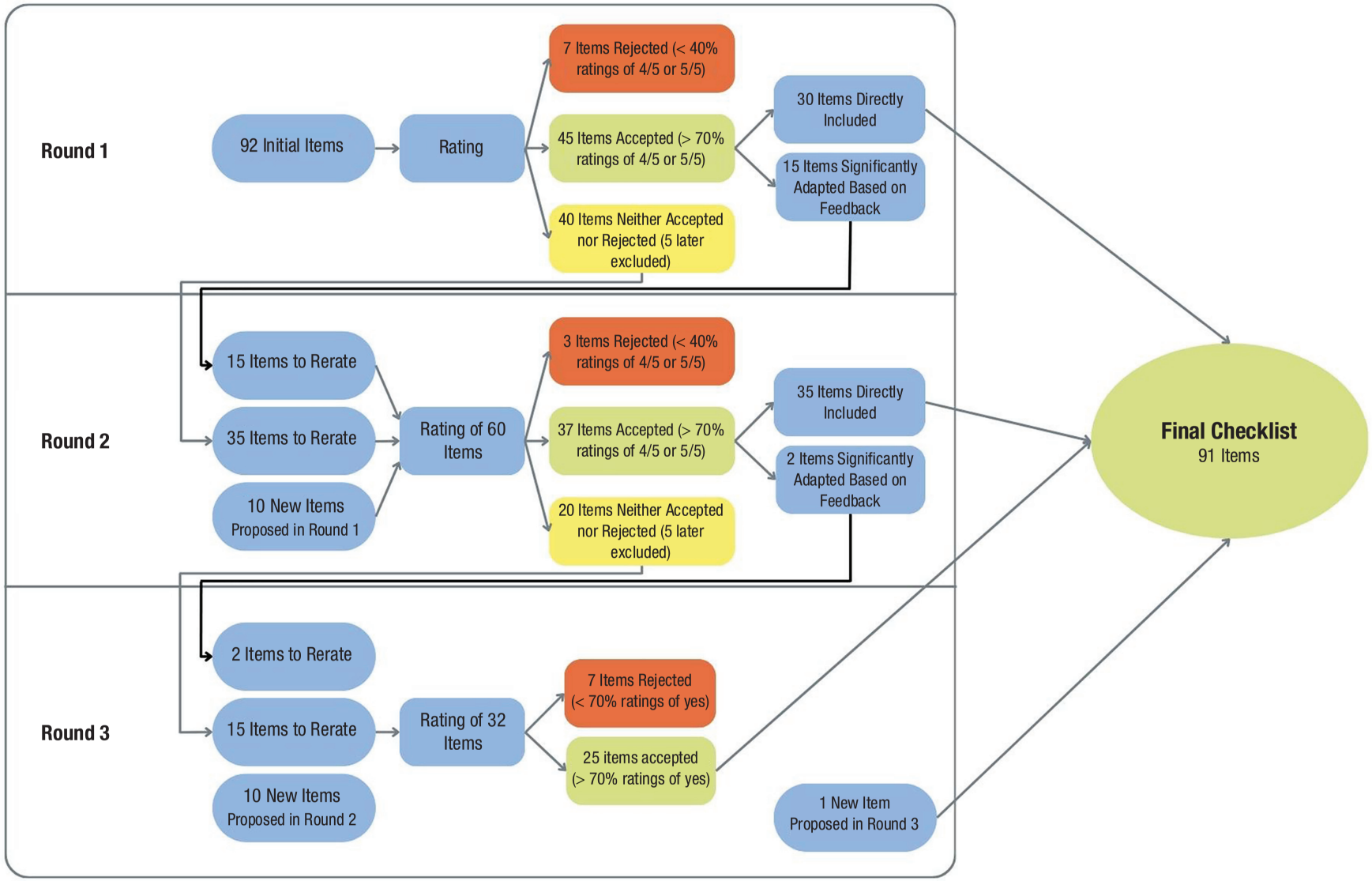


Fig. 5. Overview of the Delphi study results by round.

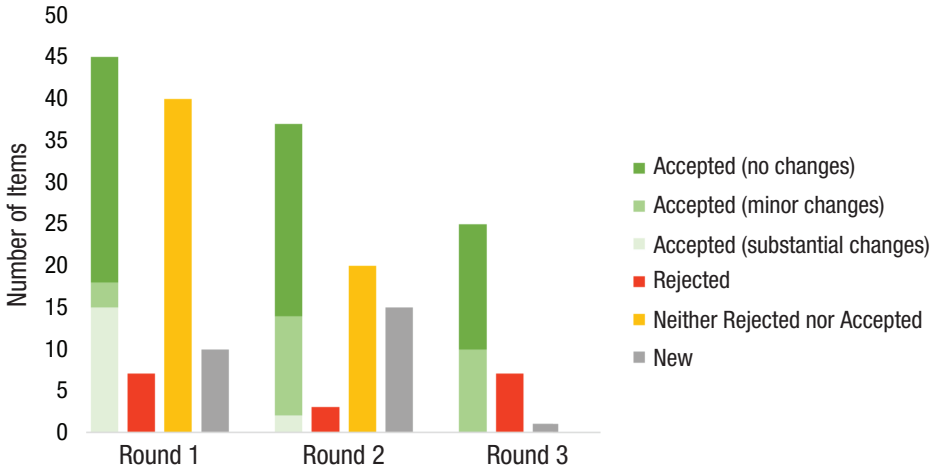


Fig. 6. Overview of the results.

in the final version of the item list. For a detailed explanation of all changes made after the Delphi study's end, see Appendix E in the Supplemental Material. All changes were documented and are traceable in the corresponding document "Item trajectories" on OSF (https://osf.io/pz8mr).

Final form of the VALID checklist

The checklist's adaptive form was already considered during the rating rounds and was finalized based on the developed item list. The grouping of items in categories (e.g., items relevant for all studies, items relevant only for designs using groups and/or conditions) allows for the inclusion of suitable items in each checklist (for the full list of categories, see Table 3).

The categorization of the items allows for the automatic generation of a checklist tailored to the study design of interest based on the categories applicable to the individual research project. The base version of the checklist, without any additional categories added, contains 57 items. For the number of items added to each stage based on the categories relevant to the individual project, see Table 4. For the final, full version of the VALID checklist, see Appendix E in the Supplemental Material.

Generation of a tailored VALID checklist

The generation of a tailored VALID checklist based on an individual study's needs can be done on the website of the tool (http://www.validchecklist.com/). As a result

Table 3. Item Categories

Category	Example item
All studies	Careful specification/definition of the theoretical constructs
Studies containing confirmatory research	Articulation of how the theory is translated into the (statistical) hypotheses
Designs using groups and/or conditions	Random assignment of participants to conditions
Designs without random assignment of participants to groups/conditions (relevant only for designs that are using groups and/or conditions)	Use of baseline comparisons to compare groups on relevant criteria
Studies adjusting measures, developing their own, or validating existing measures	Use of enough items to fully capture the concept
Studies employing frequentist data analysis	If the sample size, *N*, is not fixed, use of statistical methods that are specifically suited for repeated testing/continuous sampling
Studies employing Bayesian data analysis	Use of posterior predictive checks to validate and improve the model

Table 4. Item Breakdown per Category

Stage	Base version	Confirmatory research	Groups and/or conditions	No random assignment	Developing measures	Frequentist	Bayesian
Planning	16	+8	+3 (4[a])	—	+6	+1[b]	+4[b]
Data collection	7	+2/+1[c]	—	—	—	—	—
Data analysis	16	+1	—	+1	—	+1	+1 (3[b])
Reporting	18	—	—	—	—	—	+1 (2[b])

[a]Under the condition that the checkbox "No random assignment" has not been selected.
[b]Under the condition that "confirmatory" has been selected as well.
[c]If "confirmatory" has been selected, two items are added; if this is not the case (i.e., exploratory research), a different item is added to the base version instead.

of its adaptiveness, the tool does not merely provide a single form but, rather, allows generation of 331 distinct versions of the VALID checklist, purposefully tailored to assist researchers according to their individual study's requirements. The creation of a checklist consists of three short steps; first, users provide information on their project, then make a choice on which project stages to include, and finally, download the checklist in PDF format.

For a detailed overview of all 331 possible combinations of research stages and project characteristics, see Appendix F in the Supplemental Material. Figure 7 demonstrates the process of generating a custom VALID checklist.

At the top of the checklist, users can also provide additional information, such as the date and the name of their project, to help them organize their work. To help users keep track of the choices made during the generation of the checklist, the choices are always included at the top of the VALID checklist as well.

The source code for the VALID checklist tool is openly accessible to the public. Interested users can find the

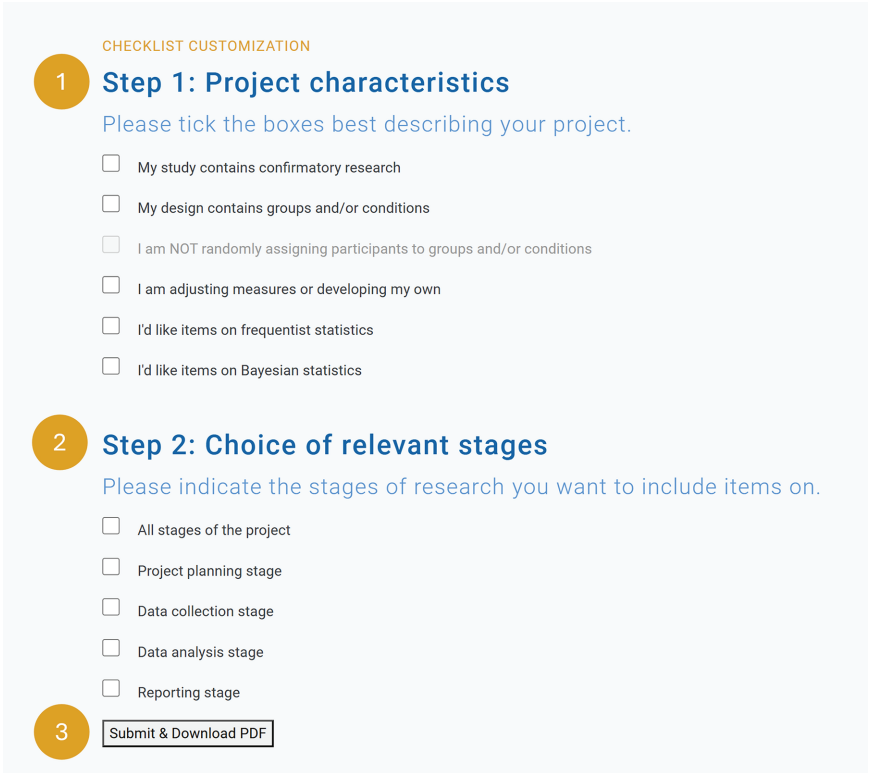


Fig. 7. VALID checklist steps.

code repository on GitHub (https://github.com/susik98/VALID-Checklist), where they can review, contribute to, and collaborate on the tool's development.

Discussion

In this study, we developed a comprehensive checklist that aims to assist researchers in enhancing and monitoring the validity of their own research in all its stages. With the collaboration of more than 30 experts in psychological science and related fields, a collection of 91 items was created, the VALID checklist. One notable feature of the VALID checklist is its adaptability and applicability to a wide range of research projects. Additional items can be added to the base version of the checklist depending on the researcher's specific needs. Personalization occurs through two successive steps using checkboxes in the online tool, where the user indicates the relevant research stages and provides additional information on the project.

The name "VALID checklist" for our tool primarily signifies our commitment to enhancing the validity of research projects. Yet it also reflects our hope for the tool to possess validity itself. We aimed to create a tool that not only serves as a helpful and well-grounded resource for researchers but also upholds a meaningful and thoughtful development process. The VALID checklist stands out as the first checklist explicitly designed with the researcher's needs in mind and the stages of the research process as its underlying structure. In contrast, the previously mentioned Seaboat online tool by Schiavone et al. (2023) focuses on guiding reviewers during the evaluation of submitted articles. Table 5 provides an overview of both tools to facilitate comparisons between them and to highlight their similarities and differences.

The key distinction between these two tools lies in their target audience (evaluators of empirical research for Seaboat, all types of researchers for the VALID checklist), which consequently influences the implementation, the output, and the different scope of these two tools. In addition, the tools notably differ in their areas of application and the way they work. Apart from the account of Schiavone et al. (2023), no other projects we are aware of so far have attempted to create an overarching tool that would provide guidance in enhancing validity in research.

Table 5. Comparison of the Seaboat Tool With the VALID Checklist

Characteristic	Seaboat (Schiavone et al., 2023)	VALID checklist
Main target audience	Reviewers of empirical research papers submitted to journals	All types of researchers
Implementation	https://www.seaboat.io/ (Shiny app)	http://www.validchecklist.com/ (interactive website)
Adaptability	Fixed set of items on validity threats	The selected set of items varies, based on individual user needs, with 331 potential checklist versions
Output	Downloadable validity report on the validity of the study	Downloadable customized checklist
Focus	Identification of the most common validity threats	A wide range of suggestions for improvements of validity
Types of validity	Internal, external, construct, and statistical-conclusion validity	Internal, external, construct, and statistical-conclusion validity
Time of use	After a study's submission for publication	Accompanying the research process (from planning to reporting)
Areas of application	Peer-review process, postpublication peer review, training for editorial boards and grant panels, review of own research, use in university courses or other training contexts	Research project planning and implementation, evaluation of research projects, metaresearch implementation, support during project preregistration, tool for didactic purposes
Development of the tool	Delphi study	Delphi study
Number of items	32 items	91 items
How it works	Guided step-by-step identification of potential validity threats present in the research that can then be used to rate each of the four validities	Selection of relevant research stages (planning, data collection, data analysis, reporting) and project characteristics to create an individual, tailor-made checklist

Strengths and limitations

One important strength of the VALID checklist is the involvement of a large, international group of researchers from diverse fields of expertise in its creation. Although most of the experts identified themselves as psychologists specializing in various areas, the group also included statisticians, mathematicians, and health professionals. This diversity allowed for the incorporation of opinions from scientists conducting empirical research in vastly different fields and building on their high-level but also distinctive methodological expertise.

The use of the Delphi method allowed for the collaborative development of the checklist. Because there is currently a lack of clear recommendations on some aspects of the method (e.g., the number of rounds and participants), which we mentioned in the methodological considerations for Stage 2, we propose discussing these criteria within or outside of the team to reflect on the choices made and explore possible improvements before recruiting experts. An important consideration regarding the ideal number of experts and rounds also concerns the possibility of dropouts (Humphrey-Murto & de Wit, 2019). Besides using reminders during the study (Belton et al., 2019; Humphrey-Murto & de Wit, 2019), we also recommend using common sense in carefully examining the study's goals and the anticipated capacities of the experts involved and the researchers organizing this endeavor.

Although the Delphi panel included a large number of experts, the pragmatic decision to invite experts based exclusively on their authorship in two specific methods-focused journals limited the pool of eligible candidates. This may have excluded other experts on validity, particularly from fields such as educational psychology, from which journals such as the *Journal of Educational Psychology* and *Educational and Psychological Measurement* were not considered for recruitment.

Thanks to its adaptive nature, the VALID checklist can be applied across a wide range of study designs and research projects, including planning stages, evaluation, metaresearch implementation, support for study preregistration, and as a tool for didactic purposes (e.g., planning course curriculums, guidance for students). Our aim was to design a flexible tool that would serve as a comprehensive solution for research in psychology and related fields by establishing an initial consensus on important factors to consider when developing, implementing, analyzing, and reporting valid studies.

To make full use of the VALID checklist, users should be proficient in statistical data analysis and possess sufficient knowledge in research design and implementation. Some users may first need to improve their methodological skills to fully understand all the checklist's items. For researchers currently extending their methodological expertise, the VALID checklist could function as a guide and highlight areas in which further learning might be beneficial. Even though some items, especially in the data-analysis stage, may well not be immediately recognizable to lesser experienced scientists, these items, at a minimum, can encourage researchers to investigate these topics further. Moreover, many items (e.g., preregistration of studies, a priori specification of result interpretation, power analysis, and consideration of inflated effect sizes in publications) cover steps that most researchers are familiar with but sometimes do not think of, thus leading to the possibility of less than optimal (or even questionable) research practices, as outlined in the introduction section. Such items should be accessible to all users and should support them in avoiding questionable research practices.

In addition, whereas the checklist was developed rigorously, following a comprehensive literature review and expert guidance, it does not claim to encompass all conceivable aspects of research validity. The decision to limit the literature search to the first 50 results per keyword without using databases such as Web of Science or employing a more comprehensive search strategy with Boolean operators may have led to the omission of relevant works. Furthermore, the lack of consideration for more diversity, equity, and inclusion (DEI) in our recruitment process may have restricted the possible range of perspectives in the item-development process, thus potentially affecting the comprehensiveness of the VALID checklist.

In addition, the VALID checklist does not yet incorporate key ideas from recent theoretical advancements in research validity that emphasize the importance of DEI. For example, movements such as QuantCrit (Gillborn et al., 2018), which integrate critical race theory into quantitative research, and the matrix approach of evidence for validity argumentation of Solano-Flores (2019), which highlights cultural responsiveness in large-scale assessments, underscore the need for a more inclusive approach to research practices.

Directions for future research

In the future, it would be desirable to design an educational version of the checklist (possibly shortened in its contents but more elaborate in its explanations), providing users with additional explanations for procedures and methods that may not be fully known to them. In addition, because the VALID checklist represents (a part of the) current opinions and trends in enhancing validity, its ongoing validation should be a primary goal of future studies. During the last round, the open-ended comment section revealed various alternative paths that could have been chosen for the final design of the checklist. The VALID checklist thus understands itself as one roadmap

supporting researchers in carefully conducting research, not as a definitive set of rules. The adoption and use of the VALID checklist by researchers may lead to future improvements to make the tool as useful and user-friendly as possible.

Future versions of the VALID checklist should also incorporate items promoting DEI, such as the use of inclusive language and citation practices, sample diversity and sample justifications, collaborative research models, and data sharing to improve access for all (American Psychological Association, 2024; CellPress, n.d.; The Wiley Network, n.d.). By integrating these considerations, the VALID checklist can better support validity of research practices across diverse populations, thereby aligning with the evolving understanding of validity.

To sum up, in this study, we presented a checklist specifically focused on increasing validity in all stages of the research process, created and developed in collaboration with a diverse group of experts. The accompanying interactive website and tool (www.validchecklist.com) allows for the automatic generation of tailor-made validity checklists for a wide range of research designs. The VALID checklist is a comprehensive tool designed to help researchers assess the validity of their studies, with a focus on psychology and social sciences. It is our hope that this checklist will improve the quality and transparency of research in these fields.

Transparency

Action Editor: David A. Sbarra
Editor: David A. Sbarra

Author Contributions

Susanne Kerschbaumer: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Software; Validation; Visualization; Writing – original draft; Writing – review & editing.
Martin Voracek: Conceptualization; Investigation; Methodology; Supervision; Writing – review & editing.
Balazs Aczél: Investigation; Writing – review & editing.
Samantha F. Anderson: Investigation; Writing – review & editing.
Brandon M. Booth: Investigation; Writing – review & editing.
Erin M. Buchanan: Investigation; Software; Writing – review & editing.
Rickard Carlsson: Investigation; Writing – review & editing.
Daniel W. Heck: Investigation; Writing – review & editing.
Anu Pauliina Hiekkaranta: Investigation; Writing – review & editing.
Rink Hoekstra: Investigation; Writing – review & editing.
Julian D. Karch: Investigation; Writing – review & editing.
Ginette Lafit: Investigation; Writing – review & editing.
Zhicheng Lin: Investigation; Writing – review & editing.
Siwei Liu: Investigation; Writing – review & editing.
David P. MacKinnon: Investigation; Writing – review & editing.
Emma L. McGorray: Investigation; Writing – review & editing.
David Moreau: Investigation; Writing – review & editing.
Marietta Papadatou-Pastou: Investigation; Writing – review & editing.
Helena Paterson: Investigation; Writing – review & editing.
Robert A. Perera: Investigation; Writing – review & editing.
Daniel J. Schad: Investigation; Writing – review & editing.
David K. Sewell: Investigation; Writing – review & editing.
Moin Syed: Investigation; Writing – review & editing.
Louis Tay: Investigation; Writing – review & editing.
Jorge N. Tendeiro: Investigation; Writing – review & editing.
Michael D. Toland: Investigation; Writing – review & editing.
Wolf Vanpaemel: Investigation; Writing – review & editing.
Don van Ravenzwaaij: Investigation; Writing – review & editing.
Lieke Voncken: Investigation; Writing – review & editing.
Ulrich S. Tran: Conceptualization; Investigation; Methodology; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Open Practices

This article has received the badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.



ORCID iDs

Susanne Kerschbaumer https://orcid.org/0009-0006-8224-3154
Martin Voracek https://orcid.org/0000-0001-6109-6155
Balazs Aczél https://orcid.org/0000-0001-9364-4988
Samantha F. Anderson https://orcid.org/0000-0003-4451-5295
Brandon M. Booth https://orcid.org/0000-0002-5780-8882
Erin M. Buchanan https://orcid.org/0000-0002-9689-4189
Daniel W. Heck https://orcid.org/0000-0002-6302-9252
Anu Pauliina Hiekkaranta https://orcid.org/0000-0001-9789-8539
Rink Hoekstra https://orcid.org/0000-0002-1588-7527
Julian D. Karch https://orcid.org/0000-0002-1625-2822
Ginette Lafit https://orcid.org/0000-0002-8227-128X

Zhicheng Lin https://orcid.org/0000-0002-6864-6559
Siwei Liu https://orcid.org/0000-0002-2972-426X
David P. MacKinnon https://orcid.org/0000-0003-0866-6010
Emma L. McGorray https://orcid.org/0000-0001-5761-7501
David Moreau https://orcid.org/0000-0002-1957-1941
Helena Paterson https://orcid.org/0000-0001-7715-5973
Robert A. Perera https://orcid.org/0000-0002-0375-0427
Daniel J. Schad https://orcid.org/0000-0003-2586-6823
David K. Sewell https://orcid.org/0000-0002-9966-8232
Moin Syed https://orcid.org/0000-0003-4759-3555
Louis Tay https://orcid.org/0000-0002-5522-4728
Jorge N. Tendeiro https://orcid.org/0000-0003-1660-3642
Michael D. Toland https://orcid.org/0000-0002-9210-4012
Wolf Vanpaemel https://orcid.org/0000-0002-5855-3885
Don van Ravenzwaaij https://orcid.org/0000-0002-5030-4091
Lieke Voncken https://orcid.org/0000-0002-6710-271X
Ulrich S. Tran https://orcid.org/0000-0002-6589-3167

Acknowledgments

We thank Sebastian Castro-Alvarez, Denis Cousineau, John J. Dziak, Tobias Ebert, Peder Mortvedt Isager, Daniël Lakens, and Trang Quynh Nguyen for their participation in the Delphi study and their valuable input.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/supp10.1177/25152459241306432.

Note

1. Two of the original items (use of groups/conditions, development of own measures) were directly implemented as checkboxes in the survey and later in the checklist.

References

Aczél, B., Hoekstra, R., Gelman, A., Wagenmakers, E., Klugkist, I., Rouder, J. N., Vandekerckhove, J., Lee, M., Morey, R. D., Vanpaemel, W., Dienes, Z., & Van Ravenzwaaij, D. (2020). Discussion points for Bayesian inference. *Nature Human Behaviour*, *4*(6), 561–563. https://doi.org/10.1038/s41562-019-0807-z

American Psychological Association. (2024). *Equity, diversity, and inclusion in APA journals*. https://www.apa.org/pubs/authors/equity-diversity-inclusion

Belton, I., MacDonald, A., Wright, G., & Hamlin, I. (2019). Improving the practical application of the Delphi method in group-based judgment: A six-step prescription for a well-founded and defensible process. *Technological Forecasting and Social Change*, *147*, 72–82. https://doi.org/10.1016/j.techfore.2019.07.002

Boulkedid, R., Abdoul, H., Loustau, M., Sibony, O., & Alberti, C. (2011). Using and reporting the Delphi method for selecting healthcare quality indicators: A systematic review. *PLOS ONE*, *6*(6), Article e20476. https://doi.org/10.1371/journal.pone.0020476

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297–312. https://doi.org/10.1037/h0040950

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton, Mifflin.

CellPress. (n.d.). *Cell Press inclusion and diversity statement FAQs*. https://www.cell.com/inclusion-diversity-statement-faqs

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Houghton Mifflin.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Diamond, I. R., Grant, R. C., Feldman, B. M., Pencharz, P. B., Ling, S. C., Moore, A. M., & Wales, P. W. (2014). Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies. *Journal of Clinical Epidemiology*, *67*(4), 401–409. https://doi.org/10.1016/j.jclinepi.2013.12.002

Donohoe, H., & Needham, R. D. (2009). Moving best practice forward: Delphi characteristics, advantages, potential problems, and solutions. *International Journal of Tourism Research*, *11*(5), 415–437. https://doi.org/10.1002/jtr.709

Dyrvig, A.-K., Kidholm, K., Gerke, O., & Vondeling, H. (2014). Checklists for external validity: A systematic review. *Journal of Evaluation in Clinical Practice*, *20*(6), 857–864. https://doi.org/10.1111/jep.12166

Fabrigar, L. R., Wegener, D. T., & Petty, R. E. (2020). A validity-based framework for understanding replication in psychology. *Personality and Social Psychology Review*, *24*(4), 316–344. https://doi.org/10.1177/1088868320931366

Gillborn, D., Warmington, P., & Demack, S. (2018). QuantCrit: Education, policy, 'big data' and principles for a critical race theory of statistics. *Race Ethnicity and Education*, *21*(2), 158–179. https://doi.org/10.1080/13613324.2017.1377417

Hartig, J., Frey, A., & Jude, N. (2020). Validität von Testwertinterpretationen [Validity of test scores interpretations]. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* (pp. 529–545). Springer. https://doi.org/10.1007/978-3-662-61532-4_21

Henderson, V. C., Kimmelman, J., Fergusson, D., Grimshaw, J. M., & Hackam, D. G. (2013). Threats to validity in the design and conduct of preclinical efficacy studies: A systematic review of guidelines for in vivo animal experiments. *PLOS Medicine*, *10*(7), Article e1001489. https://doi.org/10.1371/journal.pmed.1001489

Hohmann, E., Cote, M. P., & Brand, J. C. (2018). Research pearls: Expert consensus-based evidence using the Delphi method. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, *34*(12), 3278–3282. https://doi.org/10.1016/j.arthro.2018.10.004

Humphrey-Murto, S., & de Wit, M. (2019). The Delphi method: More research please. *Journal of Clinical Epidemiology*, *106*, 136–139. https://doi.org/10.1016/j.jclinepi.2018.10.011

Kenny, D. A. (2019). Enhancing validity in psychological research. *American Psychologist*, *74*(9), 1018–1028. https://doi.org/10.1037/amp0000531

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Krantz, J. H., & Dalal, R. (2000). Validity of web-based psychological research. In M. H. Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 35–60). Academic Press. https://doi.org/10.1016/B978-012099980-4/50003-4

Krosnick, J. A. (2018). Improving question design to maximize reliability and validity. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave handbook of survey research* (pp. 95–101). Springer. https://doi.org/10.1007/978-3-319-54395-6_13

Kruschke, J. K. (2021). Bayesian analysis reporting guidelines. *Nature Human Behaviour*, *5*(10), 1282–1291. https://doi.org/10.1038/s41562-021-01177-7

Kuhlmann, J. (2007). Ausgewählte Verfahren der Holdout- und Kreuzvalidierung [Selected holdout- and cross-validation methods]. In S. Albers, D. Klapper, U. Konradt, A. Walter, & J. Wolf (Eds.), *Methodik der empirischen Forschung* (pp. 407–416). Gabler. https://doi.org/10.1007/978-3-8349-9121-8

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLOS Medicine*, *6*(7), Article e1000097. https://doi.org/10.1371/journal.pmed.1000097

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

Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods*, *18*(3), 301–319. https://doi.org/10.1037/a0032969

Niederberger, M., & Renn, O. (2023). *Delphi methods in the social and health sciences: Concepts, applications and case studies*. Springer Nature.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T., Mulrow, C. D., Shamseer, L., Tetzlaff, J., Akl, E. A., Brennan, S., Chou, R., Glanville, J., Grimshaw, J., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *The BMJ*, *372*, Article n71. https://doi.org/10.1136/bmj.n71

Panić, N., Leoncini, E., De Belvis, G., Ricciardi, W., & Boccia, S. (2013). Evaluation of the endorsement of the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement on the quality of published systematic review and meta-analyses. *PLOS ONE*, *8*(12), Article e83138. https://doi.org/10.1371/journal.pone.0083138

Schiavone, S. R., Quinn, K. A., & Vazire, S. (2023). *A consensus-based tool for evaluating threats to the validity of empirical research*. PsyArXiv. https://doi.org/10.31234/osf.io/fc8v3

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Sharma, A., Minh Duc, N. T., Luu Lam Thang, T., Nam, N. H., Ng, S. J., Abbas, K. S., Huy, N. T., Marušić, A., Paul, C. L., Kwok, J., Karbwang, J., de Waure, C., Drummond, F. J., Kizawa, Y., Taal, E., Vermeulen, J., Lee, G. H. M., Gyedu, A., To, K. G., . . . Karamouzian, M. (2021). A consensus-based checklist for reporting of survey studies (CROSS). *Journal of General Internal Medicine*, *36*(10), 3179–3187. https://doi.org/10.1007/s11606-021-06737-1

Slaney, K. (2017). *Validating psychological constructs*. Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-38523-9

Solano-Flores, G. (2019). Examining cultural responsiveness in large-scale assessment: The matrix of evidence for validity argumentation. *Frontiers in Education*, *4*, Article 43. https://doi.org/10.3389/feduc.2019.00043

Taylor, E. (2019). We agree, don't we? The Delphi method for health environments research. *HERD: Health Environments Research & Design Journal*, *13*(1), 11–23. https://doi.org/10.1177/1937586719887709

Tunis, A. S., McInnes, M. D. F., Hanna, R., & Esmail, K. (2013). Association of study quality with completeness of reporting: Have completeness of reporting and quality of systematic reviews and meta-analyses in major radiology journals changed since publication of the PRISMA statement? *Radiology*, *269*(2), 413–426. https://doi.org/10.1148/radiol.1313027

The Wiley Network. (n.d.). *How to create a journal diversity, equity & inclusion statement*. https://www.wiley.com/en-us/network/publishing/research-publishing/editors/how-to-create-a-journal-diversity-equity-inclusion-statement