White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/247934/

Version: Published Version

Article:
Székely, O., Holmes, N.P., Ashton, J. et al. (2026) Development and validation of the transcranial magnetic stimulation reporting assessment tool (TMS-RAT). Brain Stimulation, 19 (5). 103155. ISSN: 1935-861X

https://doi.org/10.1016/j.brs.2026.103155

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

Brain Stimulation 19 (2026) 103155

Contents lists available at ScienceDirect

Brain Stimulation

journal homepage: www.journals.elsevier.com/brain-stimulation

Development and validation of the transcranial magnetic stimulation reporting assessment tool (TMS-RAT)

Orsolya Székely [a,2], Nicholas P. Holmes [b,*], Jennifer Ashton [c], Friederike Breuer [d,e], Hsin-Yuan Chen [f], Nunzia Valentina Di Chiaro [g], Arnaud Duport [h,i], Polytimi Frangou [j,k,1], Louisa Gwynne [l,m], Umair Hassan [n,o], Cassandra J. Lowe [p], Brian Mathias [q], Nan Peng [r], Jessica L. Pepper [s], Phivos Phylactou [t,u], Malgorzata Anna Szymanska [v], Luigi Tamè [l]

[a] Department of Psychology, University of Bath, Bath, BA2 7AY, United Kingdom
[b] School of Sport, Exercise and Rehabilitation Sciences, University of Birmingham, B15 2TT, United Kingdom
[c] Department of Psychology, University of York, York, YO10 5DD, United Kingdom
[d] Neuroimaging Center (NIC), Focus Program Translational Neuroscience (FTN), Johannes Gutenberg University Medical Center, Mainz, Germany
[e] Leibniz Institute for Resilience Research (LIR), Mainz, Germany
[f] Psychology, Faculty of Health and Life Sciences, Oxford Brookes University, Oxford, OX3 0BP, United Kingdom
[g] Center for Mind/Brain Sciences (CIMeC), University of Trento, Rovereto, Italy
[h] School of Physical Therapy, Faculty of Health Sciences, Western University, London, ON, Canada
[i] The Gray Centre for Mobility and Activity, Lawson Research Institute, London, ON, Canada
[j] Oxford Centre for Integrative Neuroimaging (OxCIN), FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, OX3 9DU, United Kingdom
[k] Medical Research Council Centre of Research Excellence in Restorative Neural Dynamics, United Kingdom
[l] School of Psychology, University of Kent, Canterbury, CT2 7NP, United Kingdom
[m] School of Healthcare Sciences, Cardiff University, University Hospital of Wales, Cardiff, CF14 4XN, United Kingdom
[n] Department of Psychiatry and Behavioral Sciences, Stanford School of Medicine, Stanford, USA
[o] Wu-Tsai Neurosciences Institute, Stanford University, USA
[p] Department of Psychology, University of Exeter, Exeter, EX4 4QG, United Kingdom
[q] School of Psychology, University of Aberdeen, Aberdeen, AB24 3FX, United Kingdom
[r] School of Psychology, University of Sheffield, Sheffield, S1 4DP, United Kingdom
[s] Department of Psychology, Lancaster University, Lancaster, LA1 4YF, United Kingdom
[t] Department of Psychology, University of Nevada, Reno, NV, 89557, USA
[u] Department of Rehabilitation Sciences, Faculty of Health Sciences, Cyprus University of Technology, Limassol, 3036, Cyprus
[v] MRC Cognition and Brain Sciences Unit (MRC CBU), University of Cambridge, Cambridge, CB2 7EF, United Kingdom

ARTICLE INFO

Keywords:
Transcranial magnetic stimulation
Reporting assessment
Reporting guidelines
Reproducibility
Inter-rater reliability
Test-retest reliability

ABSTRACT

Background: A standardised tool for comprehensive reporting can improve transparency, support consistent documentation, and enable comparison across transcranial magnetic stimulation (TMS) studies. To date, the most used reporting checklist lacks definitions of full reporting and was not initially evaluated for usability or reliability.

Methods: We developed the TMS Reporting Assessment Tool (TMS-RAT), a comprehensive reporting framework that provides clear definitions and examples for its items. We tested the usability and reliability of the TMS-RAT by rating all studies published between 1987 and 31st January 2025 using afferent conditioning (n = 333), protocols that encompass many reporting categories identified during tool development. Seventeen independent raters contributed across three development phases, a validation phase, and a retest phase, with naïve raters introduced in each phase. Iterative refinements of the tool were informed by inter-rater reliability, qualitative rater feedback, and consultation with external TMS experts.

This version of the manuscript has been posted as a preprint on BioRxiv, on 18/02/2026.

* Corresponding author.
E-mail address: n.p.holmes@bham.ac.uk (N.P. Holmes).

[1] PF is funded by the Wellcome Trust [grant number 224109/Z/21/Z].
[2] OS is funded by grant MR/W006308/1 for the GW4 BIOMED MRC DTP, awarded to the Universities of Bath, Bristol, Cardiff and Exeter from the Medical Research Council (MRC)/UKRI.

https://doi.org/10.1016/j.brs.2026.103155
Received 19 February 2026; Received in revised form 25 May 2026; Accepted 30 June 2026
Available online 11 July 2026
1935-861X/© 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Results: We present two versions of the tool: the 72-item TMS-RAT v1.0, designed to guide comprehensive reporting, and the TMS-RAT v1.1, a subset of 50 items with the highest inter-rater (overall AC1 = 0.78, range = [0.60–0.99]) and test-retest reliability (overall AC1 = 0.82, range = [0.65–1.0]), intended for retrospective evaluation of reporting in systematic reviews and meta-analyses. The tool has not been validated outside the afferent conditioning literature or on studies not recording motor evoked potentials.
Conclusion: The TMS-RAT is a comprehensive, reliable tool that seeks to improve transparency and reproducibility in TMS research. We aim to extend the tool to cover additional TMS protocols.

1. Introduction

1.1. Background

For transcranial magnetic stimulation (TMS) studies to contribute maximally to scientific progress, their core methods must be reported clearly enough for others to interpret the results accurately, build on the findings, or reproduce experiments across laboratories and over time. For over forty years, TMS has been used to investigate the human nervous system [1], but the field has often struggled with inconsistent effects and difficulties replicating earlier findings [2,3]. A standardised tool for reporting (and assessing the reporting of) core methods can guide the design and documentation of TMS experiments, and enable retrospective assessment and comparison of reporting completeness in systematic reviews and meta-analyses.

The most commonly used TMS reporting checklist was developed by Chipchase et al. [4], using a two-round Delphi consensus process [5] during which 42 nominated experts selected and prioritised items considered essential to report or control. This 30-item checklist provided the foundation for reporting and assessing TMS studies but has several limitations. First, Chipchase's checklist was designed to assess studies of the motor system exclusively and has limited application to other TMS domains. Second, the checklist lacks definitions or examples of what constitutes full reporting. For example, participant age may be reported as 'adults', a range, a group mean (with or without a standard deviation; SD), or as each individual participant's age. Without full, consistent reporting, it is difficult to evaluate age-related effects or confounds across studies in meta-analyses. Third, Chipchase's checklist was not initially evaluated for usability or inter-rater reliability. When this was assessed in later reviews, about one third of checklist items did not achieve acceptable inter-rater agreement or precision, indicating that they are not suitable for retrospective assessment of reporting completeness [6,7].

Nonetheless, since the publication of Chipchase et al. [4], several meta-analytic studies have addressed these limitations by extending the list of items, providing clearer and more concrete definitions for items [6] and/or reporting inter-rater reliability (e.g., Ref. [6,8–11]). However, to date, no unified or updated tool has been developed to integrate these improvements.

1.2. Objectives

The overall aim of this study was to build upon Chipchase et al.'s [4] checklist through an international, community-driven initiative. Specifically, the first aim was to conduct a scoping review to assess current reporting trends, including changes in reporting over time. We examined how frequently the previously identified core TMS methods were reported or controlled, how reliably they could be assessed, and how often the distinction between reported and controlled was used. This ensured that we identified real limitations to address and provided a baseline against which to measure performance. The second aim was to develop a new tool that addresses these limitations. This was done by building on and refining existing items and generating explicit operational definitions and examples to support consistent interpretation and application. The third aim was to refine and validate the tool on a large body of published TMS studies through multiple phases of development, validation, and test-retest assessment. Revisions were informed by qualitative rater feedback and quantitative inter-rater reliability after each phase.

To build, test and validate the tool, we used 333 published studies from the afferent conditioning literature. The afferent conditioning protocol involves a relatively complex paired-pulse conditioning design, including electromyography (EMG), peripheral nerve stimulation, and single-pulse TMS. This reflects the methodological demands of a wide variety of TMS studies, was applied to both healthy and clinical populations, and is therefore ideal for this purpose. In addition, we had access to a recently curated systematic review and meta-analysis of this literature, which provided a defined, likely exhaustive coverage of the field [12]. This set of studies was used because of their relevance to a wide range of protocols, and to minimise potential bias in the selection of studies.

2. Methods

2.1. Scoping review

The study was preregistered on the Open Science Framework (OSF; https://osf.io/tywn8/overview). Searches for the scoping review were conducted in PubMed, Web of Science, and Google Scholar, targeting published studies and grey literature citing Chipchase et al. [4], and studies that adapted it [6,13]. Text-based outputs from the databases were combined and processed using custom MATLAB scripts for bibliographic retrieval (*Supplementary Material [SM]1*, https://osf.io/fpu3k/files/rthnv). The last search date was 31 January 2025. Studies from the database searches were included if they were systematic reviews or meta-analyses that provided item-level and/or study-level reporting based on Chipchase's checklist.

Screening and data extraction were performed by two independent extractors (all articles by the same first extractor and a subset by each of sixteen second extractors). Discrepancies between extractors were resolved by the first extractor, who rechecked the extracted data against the original data. Data was extracted using spreadsheet software, including Microsoft Excel, OpenOffice Calc, and Google Sheets. The extracted data included authors, article title, source title (journal or publication), abstract, publication date, volume, issue, part number, start page, end page, article number, the number of studies reviewed, item and study level reporting of Chipchase's checklist items, and any reported measures of inter-rater agreement. Article level inclusion or exclusion decisions, their reasons, the database for data extraction, and details of how discrepancies were resolved are available in *SM2* (https://osf.io/fpu3k/files/fqm76). The proportions of studies reporting each checklist item were calculated, and bootstrapped 95% confidence intervals (CIs) were generated (see *SM3* for scoping review analysis code; https://osf.io/fpu3k/files/98e5q).

2.2. Tool development

Throughout, we use the term 'the tool' to refer to the TMS-RAT as a whole, including its reporting items, sections, the guidance document, and the rating spreadsheet. Each version of the tool is numbered, from v0.1 to v0.3 (development), and v1.0 to v1.1 (validation).

2.2.1. Tool design

The first, v0.1 of the tool included all 30 items from the Chipchase checklist, two of which were split into two items (13a: Coil shape, 13b: Coil size; 16a: Coil location, 16b: Coil stability), and two of which were combined (18: Stimulation intensity; 26: Test pulse intensity). Two further items (Cognitive state, Time of day) were taken from Pellegrini et al. [13], and 49 other items were added by us (e.g., height, experimental design, neuronavigation, threshold, EMG, afferent stimulation, paired pulse, evoked potentials, repetitive TMS). The scoping review informed several design choices, including removing the distinction between 'reported' and 'controlled' and introducing a distinction between 'partial' and 'full' to assess reporting more precisely. This resulted in three reporting categories (full/part/missing). Whenever available, item definitions were taken or adapted from Rohel et al. [6], and examples for 'full', 'part', and 'missing' reporting were produced. For quantitative variables such as age or stimulation intensity, 'full' reporting was defined as including a measure of both central tendency (e.g., mean or median) and variability (e.g., SD, standard error or interquartile range), to support the quantitative re-use of data in modelling and meta-analysis. A measure of central tendency or variability alone would only be considered 'partial' reporting. We also introduced greater modularity in the tool's structure, and organised items into thematic sections (e.g., participant characteristics, hardware, stimulation intensity) that may be used independently or combined depending on the study's methods.

The list of items in each tool version and their sources is included in *SM4* (https://osf.io/fpu3k/files/af3pj). Two of the v0.1 sections (paired-pulse—later generalised into the *Conditioning* section—and repetitive TMS; rTMS) were included but not tested during the first development phase. The rTMS section will be validated in a future version. The 9 remaining sections included 67 items (see v0.1 of the tool in *SM5*; https://osf.io/fpu3k/files/5d7bt), and an ~18,000-word guidance document (*SM6*; https://osf.io/fpu3k/files/e4ucm). Following v0.1, we developed and evaluated v0.2, v0.3 and then v1.0.

2.2.2. Raters

The 17 authors of this paper served as raters during the development, validation, and retest phases. We were at a range of career stages, including PhD students ($n = 4$), postdoctoral researchers ($n = 8$), senior technical specialists ($n = 1$), and more senior academics ($n = 4$); affiliated with 21 different institutions, with only two authors based at the same university and research team. We are based in the UK ($n = 12$), the USA ($n = 2$), Germany ($n = 1$), Canada ($n = 1$), and Italy ($n = 1$). The concentration of UK-based raters is unlikely to have influenced our assessments of reporting completeness, since only seven of us had English as our first language, the language in which the tool was developed and validated. Our individual experience with TMS ranged from 1 to 24 years ($M = 6.5$, $SD = 5.5$), and the estimated proportion of current research (over the past two years) involving TMS ranged from 15% to 100% ($M = 60.3\%$, $SD = 30.4\%$). We represented a wide range of methodological expertise and practical experience, which aligns with the intended and likely user base of the TMS-RAT—a range of researchers actively involved in conducting original studies and systematic reviews.

2.2.3. Tool development phases

All except two raters (who developed v0.1) were naïve to the tool during their first rating phase. Apart from Phase 1, each further development and validation phase included at least four naïve raters (except Phase 3, when one rater withdrew) to assess if the tool could be used reliably by first-time users, and to compare performance with and without prior experience. Raters did not receive any formal training or calibration sessions but were encouraged to use the examples provided in the tool and the guidance document. Raters flagged ambiguities and noted difficulties during rating. This feedback, alongside inter-rater reliability (Gwet's AC1), informed revisions after each phase. Gwet's AC1 is a chance-corrected measure of agreement, where a value of 1 indicates perfect agreement, and 0 indicates agreement no better than chance [14–16]. Articles were rated in duplicate by pairs of raters matched for experience with previous versions of the tool (naïve vs. not). Within this constraint, rater pairings were rotated across phases to minimise repetition. Table 1 summarises the number of raters and articles at each phase.

Alongside raters' feedback, we sought open-ended comments from the international expert TMS community. We searched for active TMS researchers in each of 143 countries, querying PubMed for TMS articles published since 1 January 2024, including the affiliated country name. Where possible, one or more co-authors from each paper were identified as active TMS researchers by further examining their PubMed and Google Scholar profiles and their institutional websites. We aimed to contact at least one researcher per country, focusing on non-clinical TMS research. We emailed 86 active TMS researchers from 33 countries to request feedback (*SM7*; https://osf.io/fpu3k/files/rz6nb). Items and sections of the tool were revised based on feedback received.

The tool was distributed and completed by raters using a Google Sheet. From v0.2 onwards, a built-in validation page flagged missed items, prompting raters to complete any omissions. Raters recorded item responses on a three-point scale: 0 for not reported, 0.5 for partially reported, and 1 for fully reported, and provided qualitative feedback comments. Raters were asked to rate each paper in one sitting when possible.

After each phase, Gwet's AC1 was calculated using the irrCAC package [17] in R (analysis code provided in *SM8;* https://osf.io/fpu3k/files/pwg68). Mean completion times were estimated using a macro within the rating spreadsheet which recorded the time of each item response; qualitative feedback was reviewed to guide the refinement, combination or removal of items. After applying pre-registered thresholds for inter-rater reliability (Gwet's AC1 < 0.60, and/or $p > .05$, and/or a lower 95% CI bound <0.50), we finalised v1.1 for retrospective reporting assessment.

3. Results

3.1. Scoping review

A total of 513 records citing Chipchase et al. [4] were retrieved across three databases: PubMed (85), Web of Science (187), and Google Scholar (241). After removing duplicates, 267 unique records remained, 32 of which met the inclusion criteria. These 32 systematic reviews or meta-analyses rated a total of 681 individual TMS studies using the Chipchase checklist. Search logs and MATLAB analysis code are provided in *SM9* (https://osf.io/w9rtz). Fig. 1 shows that reporting completeness increased gradually over time, with no clear change following the publication of Chipchase's [4] checklist in *Clinical Neurophysiology*. Reporting practices did not vary substantially across journals. Ten journals published ten or more primary research papers. *Clinical Neurophysiology* published the most (47), but reporting was not significantly higher there than in the other nine journals (e.g., vs. lowest reporting *Neurology*, $t(60) = 1.86$, $p = .084$), and there was no evidence of significantly higher reporting in the five years after versus before the checklist (2013-2017, $M \pm SD = 70.0 \pm 12.3\%$; 2007-2011, *M*

Table 1
Number of raters and articles rated across each phase.

Phase	Version	Raters		Articles Rated	
		Total	Naïve	Per phase	Cumulative
Development 1	v0.1	4	2	28	28
Development 2	v0.2	8	4	40	68
Development 3	v0.3	11	3	72	140
Final Validation	v1.0	17	6	193	333
Retest	v1.0/v1.1	16	0	32	333

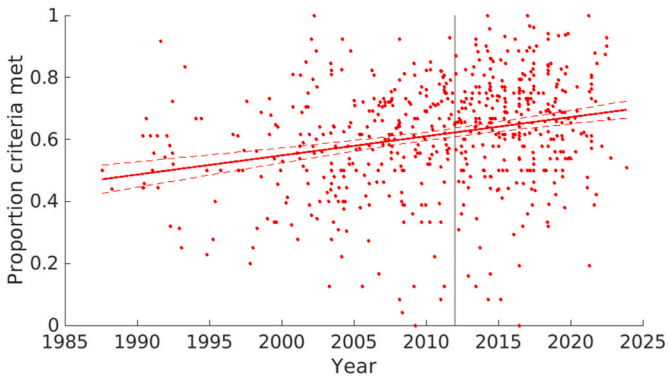


Fig. 1. *Reporting completeness over time based on Chipchase's checklist.* Individual study reporting and linear trend (with 95% CI) of the proportion of Chipchase items reported over time. Grey line: year of Chipchase et al. [4] publication. N = 614 studies with item-level ratings.

$\pm SD = 59.9 \pm 18.3\%$, t (25) = 1.77, $p = .089$; *SM10*; https://osf.io/fpu3k/files/jwrvh). In addition, while the Chipchase checklist distinguished between items being 'Reported' versus 'Controlled', only 31% (10/32) of reviews used this distinction.

The proportion of studies meeting individual checklist criteria varied substantially (Fig. 2). For example, participant age and sex (items #1 and #2), were reported about 90% of the time, while participants' attention and arousal during testing (item #23), were rarely reported (~18%). All studies citing Chipchase et al. [4] and the item-level reporting for all 30 items are provided in *SM11* (https://osf.io/fpu3k/files/jwrvh).

3.2. Tool development

3.2.1. Feedback from TMS researchers

After an initial response from 25 researchers, we received feedback from 12 TMS experts worldwide. The process was open-ended, and contributions varied in scope and depth. Some focused on clarifying item definitions, while others prompted the inclusion and/or removal of items. These contributions played an important role in refining the tool to meet the practical needs of the wider research community. In addition to the feedback that has been used to improve the tool's relevance and usability, a high proportion of consultations focused on the inclusion of items to assess methodological quality or risk of bias. As the intended scope of the TMS-RAT is exclusively a reporting, rather than quality, assessment tool, this type of feedback could not be incorporated. Whenever item definitions or examples were changed to reflect feedback, this is noted in the corresponding Supplementary Material.

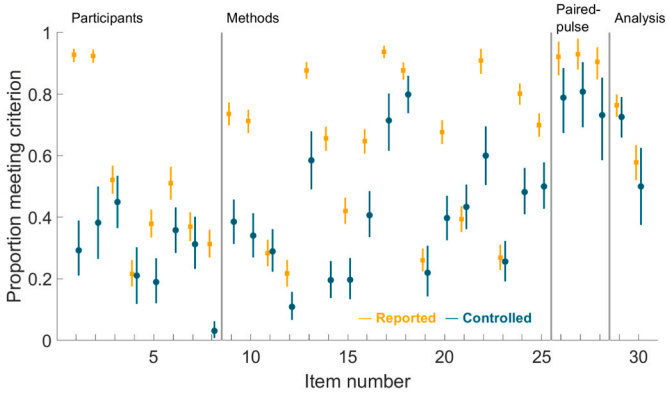


Fig. 2. *Proportions of 614 studies meeting Chipchase criteria 'reported' vs. 'controlled'.*
Note. Each circle shows the proportion of studies that either reported (orange squares) or controlled for (blue circles) each item. Items are grouped into thematic sections: Participants, Methods, Paired-pulse, and Analysis. Error bars represent bootstrapped 95% CIs.

3.2.2. Inter-rater reliability

Inter-rater reliability increased across development phases (Table 2), while the proportion of fully reported items remained stable at approximately 50% (range: 46–56%). Mean completion time per item decreased slightly across phases, indicating some improvement in usability. In the final validation phase, AC1 did not differ between experienced and naïve raters t (13.56) = −1.13, $p = .28$, indicating that the usability was independent of experience with the tool. The number of missing ratings also decreased after introducing the validation page in v0.2.

In the final validation phase, most raters showed comparable performance (mean AC1 = 0.63, $SD = 0.08$). One outlier showed substantially lower agreement (AC1 = 0.40, Fig. 3) alongside a high percentage of missing timestamps (53.8%; no other rater had >1%) and markedly shorter completion times (6.93 s per item vs. group mean of 20.88, $SD = 9.36$ s), indicating that this rater used a different strategy to others. Consequently, this rater was excluded.

3.2.3. Version 1.1: reliable items for retrospective assessment

The validated version of the tool (v1.0) was filtered using preregistered criteria (Gwet's AC1 < 0.60 and/or a p-value comparing agreement to chance of $p > .05$ and/or a lower 95% CI bound <0.50). Twenty-two items did not meet these criteria (see *SM12* [https://osf.io/fpu3k/files/vnw56] for the full list and *SM13* [https://osf.io/q3zck] for discussion of why agreement could not be reached). This indicates that it is not currently possible to provide a single version of the TMS-RAT that both guides reporting of all key methods and includes only items that can be reliably assessed retrospectively. The final version of the tool, excluding these 22 items (v1.1), is appropriate for retrospective reporting assessment. It includes 12 sections and 50 items (excluding the currently unvalidated rTMS section). See Table 3 for the full list of v1.0 and v1.1 items.

3.2.4. Overall reporting trends

Using v1.1 of the tool, we explored overall trends of reporting in the literature. Over time (by publication date), the mean rating per paper showed a gradual improvement $\beta = 0.29\%$, 95% CI: 0.12 to 0.47%, $p = .0013$ (Fig. 4). This resembles the even stronger, positive trend observed for the Chipchase checklist (Fig. 1) over time, $\beta = 0.62\%$, 95% CI = 0.43-0.80%, $p < .001$. However, in the TMS-RAT v1.1, the distribution was far from floor or ceiling effects (range 0.19 to 0.72), as 0% of papers had <15% or >85% of items reported, unlike the distribution for Chipchase's checklist, where 9% of papers had <15% or >85% of items reported.

3.2.5. Test-retest reliability

Test-retest reliability was assessed by re-rating four articles per rater (64 in total) using version 1.0 of the tool, with a minimum interval of two months between the original test and the retest. Analyses were restricted to the subset of items retained in TMS-RAT v1.1. Overall test-retest reliability of the tool was Gwet's AC1 0.82 (95% CI: 0.80 to 0.84). Item-level AC1 values ranged from 0.65 to 1.00, while section-level AC1 values ranged from 0.68 to 0.96.

3.2.6. Documentation of each development and validation phase

All data and documentation for each phase of tool development and validation are available in the SMs, whose links will remain open after publication. For v0.1, this includes the raw rating data (SM14; https://osf.io/fpu3k/files/yfjca), item-level inter-rater reliability (SM15; https://osf.io/fpu3k/files/3m8eh; timing not recorded), and section-level inter-rater reliability statistics (SM16; https://osf.io/fpu3k/files/k4jmp). For v0.2, this includes the tool (SM17; https://osf.

Table 2
Summary of tool performance across development and validation phases.

Tool version	Sections	Items	N ratings		Percentage of items with reporting rated as			Mean reporting across all articles (%)	AC1					Mean time per item rated in seconds (*SD*)[a]
									Overall	Section level		Item level		
			Valid	Missing N (% of total)	0	0.5	1			min	max	min	max	
0.1	9	67	3234	98 (2.94)	45.1	17.1	37.8	46.2	0.57	0.43	0.75	−0.11	1.0	Not recorded
0.2	11	64	4936	26 (0.52)	38.6	21.0	40.4	50.9	0.50	0.12	0.71	−0.2	1.0	22.2 (27.3)
0.3	12	70	9752	40 (0.41)	43.1	15.4	41.6	50.9	0.63	0.22	0.77	0.05	0.99	20.7 (24.6)
1.0	13	72	26,571	149 (0.56)	37.3	14.1	48.6	55.7	0.61	0.43	0.83	0.07	0.98	20.6 (24.7)
1.0[b]	13	72	21,321	103 (0.48)	49.6	15.2	45.2	52.8	0.67	0.46	0.87	0.16	0.99	20.9 (24.3)
1.1	12	50	14,941	55 (0.37)	41.0	12.0	47.0	53.0	0.78	0.67	0.91	0.60	0.99	18.7 (22.0)

[a] Completion times were calculated after excluding values above the upper bound of the 95% CI, to make sure we exclude extreme response time, e.g., when raters took a break.
[b] *Note.* In the final validation phase (v1.0), a second analysis was conducted, excluding one outlier rater (see below for details).

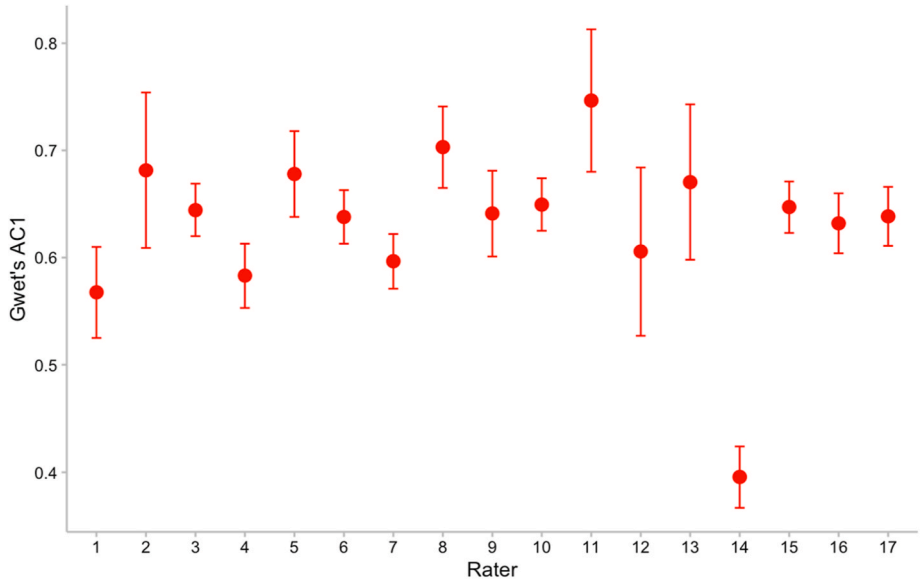


Fig. 3. *Gwet's AC1 values for individual raters.* Error bars represent 95% CIs.

io/fpu3k/files/nxvka), guidance document (SM18; https://osf.io/fpu3k/files/zkhdq), a summary of changes since v0.1 (SM19; https://osf.io/b6ftz), raw rating data (SM20; https://osf.io/fpu3k/files/mzuy6), item-level inter-rater reliability and timing results (SM21; https://osf.io/fpu3k/files/ysvch), and section-level inter-rater reliability results (SM22; https://osf.io/fpu3k/files/vx4cm). For v0.3, this includes the tool (SM23; https://osf.io/fpu3k/files/z4tqf), the guidance document (SM24; https://osf.io/fpu3k/files/9w58f), the summary of changes from v0.2 (SM25; https://osf.io/fpu3k/files/nfqxs), the raw rating data (SM26; https://osf.io/fpu3k/files/jcpqg), the item-level inter-rater reliability and timing results (SM27; https://osf.io/fpu3k/files/bhmn5), and the section-level inter-rater reliability results (SM28; https://osf.io/fpu3k/files/n3qcp). For the final validation version (v1.0), we also provide the tool (SM29; https://osf.io/fpu3k/files/uvby9), the guidance document (SM30; https://osf.io/fpu3k/files/ga9xc), the summary of changes (SM31; https://osf.io/b6ftz), the raw rating data (SM32; https://osf.io/fpu3k/files/e8ms7), the item-level inter-rater reliability and timing results (SM33; https://osf.io/fpu3k/files/26zwe), and the section-level inter-rater reliability results (SM34; https://osf.io/fpu3k/files/ypzsf). Following exclusion of the outlier rater, updated item- and section-level inter-rater reliability results for v1.0 are provided in SM35 (https://osf.io/fpu3k/files/hsd6z) and SM36 (https://osf.io/fpu3k/files/xb3p9), respectively. For the final, retrospective version (v1.1), we provide the tool (SM37; https://osf.io/fpu3k/files/akgwd), item-level inter-rater reliability and timing results (SM38; https://osf.io/fpu3k/files/bxrpq), and section-level inter-rater reliability results (SM39; https://osf.io/fpu3k/files/ugnjm). The test-retest phase raw rating data is available in SM40 (https://osf.io/fpu3k/files/6cm5f) (reliability AC1 values are available in SM41 (https://osf.io/fpu3k/files/spzjc), while section-level AC1 values in SM42 (https://osf.io/fpu3k/files/9pfaj).

4. Discussion

4.1. Summary of contribution

After assessing current reporting trends in TMS research, we developed and validated the TMS-RAT through iterative testing and refinement, across 333 papers rated by 17 independent raters, and consultation with 12 TMS experts worldwide. We established a comprehensive, empirically tested, community-informed tool ready for implementation by TMS researchers. Although the TMS-RAT was developed specifically for TMS studies, it aligns conceptually with broader reporting frameworks such as CONSORT for clinical trials [18], and PRISMA for systematic reviews [19]. TMS-RAT v1.0 and v1.1 together provide a framework for improving reporting and the assessment of reporting across a wide range of study designs, and thereby increasing transparency and reproducibility in TMS studies. To our knowledge, this is the first empirically validated tool designed to assess reporting in TMS research.

Table 3
Full list of v1.0 items. Items excluded from v1.1 are shown in dark grey.

Section	Label	AC1	95%CI	
A. PARTICIPANT BIO			Low	High
	A1. Age QUANTITATIVE	0.88	0.82	0.94
	A2. Sex	0.62	0.52	0.72
	A3. Handedness	0.92	0.87	0.97
	A4. Height QUANTITATIVE	0.97	0.95	1
	A5. Weight QUANTITATIVE	0.99	0.98	1
B. PARTICIPANT STATE				
	B1. Medication	0.49	0.40	0.60
	B2. CNS-active substances	0.69	0.60	0.79
	B3. Neurological, psychiatric or behavioural	0.45	0.33	0.57
	B4. Medical	0.40	0.28	0.53
	B5. Cognitive state	0.67	0.57	0.77
	B6. Sleep	0.94	0.90	0.98
	B7. Posture	0.76	0.67	0.85
C. DESIGN				
	C1. Studies	0.97	0.94	1
	C2. Population	0.97	0.98	1
	C3. Groups	0.89	0.83	0.94
	C4. Protocol	0.89	0.84	0.94
	C5. Time of day	0.93	0.89	0.98
	C6. Session duration and interval	0.56	0.45	0.67
	C7. Attention	0.70	0.61	0.79
	C8. Control condition	0.54	0.43	0.65
	C9. Sensory attenuation	0.80	0.96	1
D. HARDWARE				
	D1. Stimulator	0.94	0.90	0.98
	D2. Coil shape and model	0.76	0.68	0.85
	D3. Coil diameter	0.69	0.60	0.79
	D4. Pulse waveform	0.88	0.81	0.96
E. LOCATIONS				
	E1. Head and brain neuronavigation	0.60	0.50	0.71
	E2. Coil location on skin	0.74	0.69	0.82
	E3. Target location in CNS	0.64	0.54	0.74
	E4. Coil orientation	0.70	0.60	0.80
	E5. Coil current direction	0.81	0.73	0.89
	E6. Coil stability	0.74	0.66	0.83
F. INTENSITY				
	F1. Intensity method	0.89	0.83	0.95
	F2. Machine intensity QUANTITATIVE	0.58	0.47	0.68
	F3. Participant intensity QUANTITATIVE	0.20	0.08	0.33
G. SINGLE PULSE				
	G1. Trial-to-trial interval	0.56	0.46	0.67
	G2. Number of pulses	0.40	0.28	0.52
	G3. Evoked response raw QUANTITATIVE	0.46	0.35	0.57
H. MOTOR THRESHOLD				
	H1. Threshold location	0.61	0.51	0.71
	H2. Threshold type	0.90	0.85	0.95
	H3. Threshold criterion	0.73	0.65	0.82
	H4. Threshold algorithm	0.69	0.60	0.79
	H5. Threshold values QUANTITATIVE	0.64	0.54	0.74

I. MOTOR VARIABLES				
	I1. Motor skills	0.93	0.88	0.97
	I2. Prior motor activity	0.90	0.85	0.95
	I3. Target muscle or body part activity	0.27	0.15	0.40
J. EMG & MUSCLES				
	J1. EMG hardware	0.78	0.70	0.86
	J2. EMG software	0.60	0.50	0.71
	J3. Skin preparation	0.95	0.92	0.99
	J4. Electrode type, material and size	0.78	0.70	0.86
	J5. Electrode location	0.66	0.57	0.76
	J6. Electrode montage	0.58	0.47	0.68
	J7. Sampling and filtering	0.74	0.65	0.83
	J8. EMG measure	0.44	0.33	0.56
	J9. EMG amplitude normalisation	0.19	0.08	0.31
K. AFFERENT STIMULATION				
	K1. Afferent stimulation hardware	0.87	0.80	0.93
	K2. Electrodes or stimulators	0.44	0.32	0.56
	K3. Afferent stimulus location	0.46	0.34	0.57
	K4. Afferent stimulus target	0.89	0.83	0.95
	K5. Duration and waveform	0.74	0.65	0.83
	K6. Threshold type	0.93	0.89	0.98
	K7. Threshold values QUANTITATIVE	0.78	0.70	0.86
L. CONDITIONING				
	L1. Conditioning intensity (machine) QUANTITATIVE	0.76	0.68	0.85
	L2. Conditioning intensity (participant) QUANTITATIVE	0.16	0.05	0.28
	L3. Condition-Test interval(s)	0.78	0.70	0.86
	L4. Conditioned and unconditioned order	0.37	0.25	0.50
	L5. Number of conditioned TMS pulses	0.65	0.56	0.75
	L6. Conditioned response raw QUANTITATIVE	0.51	0.41	0.62
	L7. Conditioned response relative QUANTITATIVE	0.55	0.44	0.66
M. SOMATOSENSORY EVOKED POTENTIALS (SEP)				
	M1. Frequency and repetitions	0.71	0.55	0.86
	M2. EEG montage	0.94	0.88	1
	M3. SEP feature	0.54	0.36	0.72
	M4. SEP values QUANTITATIVE	0.59	0.42	0.76

Note. the *p*-values for all items in v1.0 were <.001.

The TMS-RAT built on the Chipchase et al. [4] checklist by retaining its items selected by Delphi consensus, while addressing its limitations through operational definitions, examples, and empirical evaluation of usability, inter-rater reliability, and test-retest reliability. We introduced a modular structure that allows the tool to be adapted to different research contexts and provides flexibility for extension into further contexts. We removed the 'reported' versus 'controlled' distinction because the limited uptake (see our scoping review) suggested that it is not practical or meaningful to researchers. The distinction is somewhat arbitrary and is difficult to operationalise consistently. For example, Rohel et al. [6] defined 'controlled' by using statistical significance testing (i.e., finding no significant difference between groups indicates that a study is 'controlled' for a specific item). However, this would make smaller or lower-powered studies more likely to be classified as 'controlled', since they are less likely to detect statistically significant differences. Furthermore, distinguishing between full, partial, and missing information allowed reporting to be characterised with greater resolution, avoided floor and ceiling effects, and better reflected the continuum of reporting completeness.

We assessed inter-rater reliability with Gwet's AC1, which is calculated by treating ratings as categorical rather than ordinal, to avoid inflating agreement estimates and to maintain comparability with the categorical scoring used in the Chipchase [4] checklist. This is a conservative comparison, as the TMS-RAT applies a three-level categorisation compared to the Chipchase et al. [4] binary categories. However, all items retained in TMS-RAT v1.1 met predefined reliability criteria, whereas approximately one third of Chipchase items did not meet the same criteria [6,7]. The similarity in reporting trends over time and the overall reporting completeness suggests that TMS-RAT v1.1 reflects the same underlying reporting practices as the Chipchase checklist.

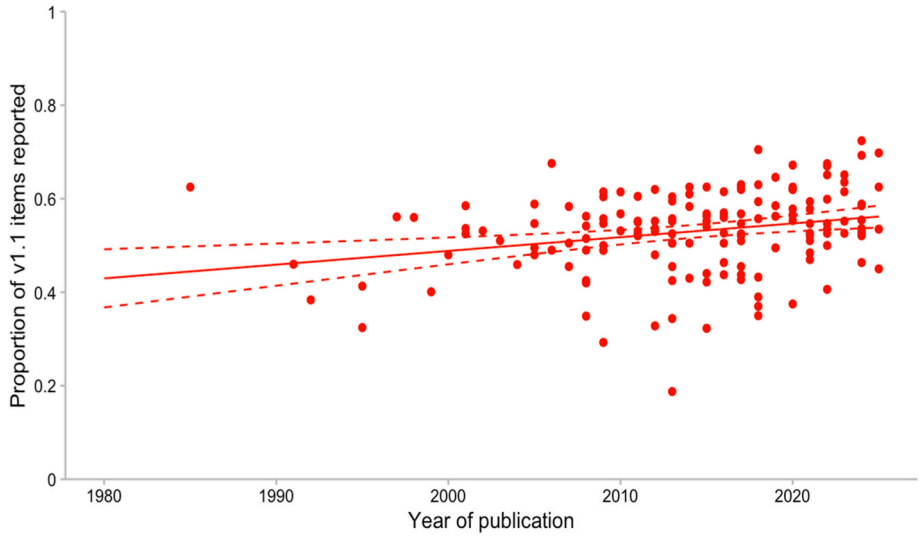


Fig. 4. *Reporting completeness over time, defined as the proportion of TMS-RAT v1.1 items reported* per *study.* The solid line shows the linear trend, and the dashed lines indicate the 95% CI. $N = 153$ studies with double ratings. For each article, the graph shows the mean proportions across two raters' ratings.

4.2. Constraints on reporting assessment

Twenty-two reporting items from v1.0 did not meet our pre-registered reliability threshold and were excluded from v1.1. This included items considered essential, for example, the entire single-pulse section, as well as items such as the EMG electrode montage and the session duration and interval. Despite iterative attempts to produce clearer instructions or examples, raters could not consistently agree on whether the required information was reported. This likely reflects inconsistent reporting of protocols in the literature. Until there is broader alignment in how key methods are reported, a tool cannot be both maximally comprehensive and uniformly reliable, and separate tools for prospective and retrospective use are necessary. The exclusion of these 22 items allows the tool to be useful for retrospective meta-analyses; importantly, removing them does not affect the tool's prospective usefulness for supporting consistent reporting of core TMS methodologies, because these items are still included in v1.0 of the tool, which we recommend for use as reporting guidance (see 4.3).

4.3. The intended use of the TMS-RAT

We recommend using TMS-RAT v1.0 as guidance for reporting methods in original research. Authors are encouraged to complete all relevant sections and to include the completed sheets and the output. csv file as supplementary material when submitting manuscripts for peer review. Reviewers and journal editors are encouraged to assess the completeness of study reporting using the tool as a benchmark.

For retrospective assessments, such as in systematic reviews or meta-analyses, we recommend using TMS-RAT v1.1, which contains only items with high inter-rater and test-retest reliability. This version is intended for evaluating and comparing reporting completeness across studies. Overall reporting completeness can be calculated as the proportion of reported items within the study. This metric could be used as a covariate in meta-regression or subgroup analyses. Importantly, neither v1.0 nor v1.1 of the tool is designed to measure methodological quality or risk of bias, but rather to assess the completeness of reporting, and v1.1 summary measures should be interpreted accordingly.

The inter-rater reliabilities reported here also provide a benchmark for the expected level of agreement that users of the tool can compare against. Even when a tool has acceptable overall reliability, variability between raters can still affect the quality of the ratings. It is always important to assess individual ratings.

To make use of the TMS-RAT easy and intuitive, researchers will be able to use our website (tms-rat.org), to generate a near-complete Methods section in accordance with v1.0; and to retrospectively assess reporting completeness using v1.1.

4.4. Future development

The TMS-RAT is intended as a living resource that will evolve in response to changes in research practice and community feedback. The current version focuses on studies using TMS alone (or in combination with electromyography and somatosensory evoked potentials). While afferent conditioning protocols include a range of different methods, it is important to note that the current versions (1.0 and 1.1) of the tool have not been directly validated outside of this literature, and it is unclear how the tool will perform outside of motor evoked potential studies. It is important that any further development follows the same rigorous testing we implemented.

Future development will include updates and extensions to currently unvalidated protocols, including rTMS, clinical applications, and multimodal approaches such as TMS coregistered with functional magnetic resonance imaging and electroencephalography. Additional items will be needed to guide and assess reporting in these fields. When extending the tool to new areas, a structured consensus (such as that used by Chipchase and colleagues) would be an important consideration. This will greatly increase the tool's applicability for the TMS community. Specifically, rTMS is increasingly important in clinical use, and validating that section will increase the tool's applicability for clinical TMS research.

The latest validated versions of the tool, along with the guidance document, are openly accessible via the project website (tms-rat.org) as well as our GitHub pages (https://github.com/TMSMultiLab/TMSMultiLab/wiki/TMS-Reporting-Assessment-Tool). We encourage the community to contribute to the ongoing refinement of the TMS-RAT by using it both to prospectively generate Methods sections and to retrospectively assess reporting completeness. Real-world usage feedback and data will be incorporated into future versions of the tool.

CRediT authorship contribution statement

Orsolya Székely: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Visualization, Writing – original draft, Writing – review & editing. **Nicholas P. Holmes:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision,

Visualization, Writing – review & editing. **Jennifer Ashton:** Data curation, Investigation, Writing – review & editing. **Friederike Breuer:** Data curation, Investigation, Writing – review & editing. **Hsin-Yuan Chen:** Data curation, Investigation, Writing – review & editing. **Nunzia Valentina Di Chiaro:** Data curation, Investigation, Writing – review & editing. **Arnaud Duport:** Data curation, Investigation, Writing – review & editing. **Polytimi Frangou:** Data curation, Investigation, Writing – review & editing. **Louisa Gwynne:** Data curation, Investigation, Writing – review & editing. **Umair Hassan:** Data curation, Investigation, Writing – review & editing. **Cassandra J. Lowe:** Data curation, Investigation, Writing – review & editing. **Brian Mathias:** Data curation, Investigation, Writing – review & editing. **Nan Peng:** Data curation, Investigation, Writing – review & editing. **Jessica L. Pepper:** Data curation, Investigation, Writing – review & editing. **Phivos Phylactou:** Data curation, Investigation, Writing – review & editing. **Malgorzata Anna Szymanska:** Data curation, Investigation, Writing – review & editing. **Luigi Tamè:** Data curation, Investigation, Writing – review & editing.

Declaration of competing Interest

The authors declare that they have no competing financial interests or personal relationships that could be perceived to have influenced the work reported in this paper.

Acknowledgements

This work was supported by the Wellcome Trust [grant number 224109/Z/21/Z to PF]; and the Medical Research Council (MRC)/UKRI [grant number MR/W006308/1, for the GW4 BIOMED MRC DTP awarded to the Universities of Bath, Bristol, Cardiff and Exeter and supporting OS].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.brs.2026.103155.

References

[1] Barker AT, Jalinous R, Freeston IL. Non-invasive magnetic stimulation of human motor cortex. Lancet 1985;1(8437):1106–7. https://doi.org/10.1016/s0140-6736(85)92413-4.

[2] Hamada M, Murase N, Hasan A, Balaratnam M, Rothwell JC. The role of interneuron networks in driving human motor cortical plasticity. Cerebr Cortex 2013;23(7):1593–605. https://doi.org/10.1093/cercor/bhs147.

[3] Héroux ME, Taylor JL, Gandevia SC. The use and abuse of transcranial magnetic stimulation to modulate corticospinal excitability in humans. PLoS One 2015;10(12):e0144151. https://doi.org/10.1371/journal.pone.0144151.

[4] Chipchase L, Schabrun S, Cohen L, Hodges P, Ridding M, Rothwell J, Taylor J, Ziemann U. A checklist for assessing the methodological quality of studies using transcranial magnetic stimulation to study the motor system: an international consensus study. Clin Neurophysiol 2012;123(9):1698–704. https://doi.org/10.1016/j.clinph.2012.05.003.

[5] Hsu CC, Sandford BA. The Delphi technique: making sense of consensus. Practical Assess Res Eval 2007;12(1):10. https://doi.org/10.7275/pdz9-th90.

[6] Rohel A, Bouffard J, Patricio P, Mavromatis N, Billot M, Roy J-S, Bouyer L, Mercier C, Masse-Alarie H. The effect of experimental pain on the excitability of the corticospinal tract in humans: a systematic review and meta-analysis. Eur J Pain 2021;25(6):1209–26. https://doi.org/10.1002/ejp.1746.

[7] Desmons M, Cherif A, Rohel A, de Oliveira FCL, Mercier C, Massé-Alarie H. Corticomotor control of lumbar erector spinae in postural and voluntary tasks: the influence of transcranial magnetic stimulation current direction. eNeuro 2024;11(2).

[8] Bai Z, Zhang J, Fong KNK. Effects of transcranial magnetic stimulation in modulating cortical excitability in patients with stroke: a systematic review and meta-analysis. J NeuroEng Rehabil 2022;19(1):24. https://doi.org/10.1186/s12984-022-00999-4.

[9] Gunduz ME, Pinto CB, Saleh Velez FG, Duarte D, Pacheco-Barrios K, Lopes F, Fregni F. Motor cortex reorganization in limb amputation: a systematic review of TMS motor mapping studies. Front Neurosci 2020;14:314. https://doi.org/10.3389/fnins.2020.00314.

[10] Lanza G, Mogavero MP, Ferri R, Pani T. Motor cortex excitability in restless legs syndrome: a systematic review and insights into pathophysiology via transcranial magnetic stimulation. Sleep Med Rev 2025;79:102027. https://doi.org/10.1016/j.smrv.2024.102027.

[11] Snow NJ, Wadden KP, Chaves AR, Ploughman M. Transcranial magnetic stimulation as a potential biomarker in multiple sclerosis: a systematic review with recommendations for future research. Neural Plast 2019;2019(1):6430596. https://doi.org/10.1155/2019/6430596.

[12] Holmes NP, Székely O, Araújo LM, Dyke K, Tamè L. Short- and long-latency conditioning of corticospinal excitability in hand muscles: comprehensive meta-analysis reveals distinct response functions for mixed versus cutaneous afferents. Preprint available at, https://osf.io/39axj/; 2026.

[13] Pellegrini M, Zoghi M, Jaberzadeh S. A checklist to reduce response variability in studies using transcranial magnetic stimulation for assessment of corticospinal excitability: a systematic review of the literature. Brain Connect 2020;10(2):53–71. https://doi.org/10.1089/brain.2019.0715.

[14] Gwet K. Kappa statistic is not satisfactory for assessing the extent of agreement between raters. Statistical methods for inter-rater reliability assessment 2002;1(6):1–6.

[15] Gwet KL. Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters. In: Advanced analytics. LLC; 2014.

[16] Vach W, Gerke O. Gwet's AC1 is not a substitute for Cohen's kappa–A comparison of basic properties. MethodsX 2023;10:102212. https://doi.org/10.1016/j.mex.2023.102212.

[17] Gwet KL. In: Package 'irrCAC'. Computing chance-corrected agreement coefficients (CAC), 1; 2019.

[18] Hopewell S, Chan A-W, Collins GS, Hróbjartsson A, Moher D, Schulz KF, Tunn R, Aggarwal R, Berkwits M, Berlin JA, Bhandari N, Butcher NJ, Campbell MK, Chidebe RCW, Elbourne D, Farmer A, Fergusson DA, Golub RM, Goodman SN, Boutron I. CONSORT 2025 statement: updated guideline for reporting randomised trials. Br Med J 2025;389:e081123. https://doi.org/10.1136/bmj-2024-081123.

[19] PRISMA-P Group, Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart LA. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 2015;4(1):1. https://doi.org/10.1186/2046-4053-4-1.