



ORIGINAL RESEARCH

Applying the GREET checklist for assessing reporting of evidence-based practice educational interventions and teaching showed inter-rater discrepancies and item-level challenges

Tea Andabaka^{a,*}, Mersiha Mahmić-Kaknjo^b, Livia Puljak^c

^aResearch Department, DGT Remedy d.o.o., Zagreb, Croatia

^bDepartment of Clinical Pharmacology, Zenica Cantonal Hospital, Zenica, Bosnia and Herzegovina

^cDepartment of Nursing, Catholic University of Croatia, Zagreb, Croatia

Accepted 16 October 2025; Published online 23 October 2025

Abstract

Objectives: This study aimed to analyze in-depth inter-rater discrepancies in Guideline for Reporting Evidence-Based Practice Educational Interventions and Teaching (GREET) item ratings between two first-time GREET users (GREET-naïve) data extractors, and to explore challenges in using the GREET checklist as a tool to assess the completeness of reporting.

Study Design and Setting: This was a secondary analysis, conducted on the literature synthesized in a prior scoping review. Two independent raters, first time users of the GREET checklist, evaluated the trials using a modified version of the 17-item checklist. Item 4 (evidence-based practice [EBP] content) was excluded as inapplicable, while item 5 (educational materials) was subdivided to assess description and accessibility separately. Discrepancies in GREET item ratings and inter-rater agreement were analyzed using descriptive statistics and Cohen's Kappa. We noted challenges in assessing individual GREET items.

Results: We analyzed 161 randomized controlled trials. Initial assessments yielded discrepancies in 20% of item ratings ($n = 561/2737$), which were reduced to 14% and 2% through successive consensus rounds, ultimately achieving full agreement. The mean number of discrepancies per trial was 3.5. Inter-rater agreement was substantial ($\kappa = 0.616$; 95% CI: 0.590–0.642). Highest disagreement rates were observed for items addressing “Environment,” “Materials included,” “Attendance,” and “Adaptations.” Detailed analysis revealed that ambiguity in item phrasing and variability in manuscript reporting contributed to inconsistency. Several GREET items were identified as candidates for future refinement.

Conclusion: Although the GREET checklist provides a valuable framework for assessing reporting quality in EBP educational interventions, its application may yield substantial discrepancies between GREET-naïve raters. Clearer item definitions, improved guidance materials, and training could enhance its reliability. Findings support the need for continued refinement of the checklist and underscore the importance of comprehensive and transparent reporting in educational research. © 2025 Elsevier Inc. All rights reserved, including those for text and data mining, AI training, and similar technologies.

Keywords: Evidence-based practice; Guideline adherence; Reproducibility of results; Observer variation; Checklist; Quality assurance; Data accuracy

Funding: This study was conducted as a secondary follow-up to a scoping review conducted within the Erasmus+ project “Synergistic education of parents with children with developmental disabilities” (SynergyEd). The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The study described in this manuscript was not planned within the SynergyEd project and the

authors did not receive any funding to conduct this study or write this manuscript.

* Corresponding author. Research Department, DGT Remedy d.o.o., Podbrezje XI. 8, Zagreb 10000, Croatia.

E-mail address: tea.andabaka@gmail.com (T. Andabaka).

Plain Language Summary

Complete and clear descriptions of teaching interventions help others repeat and build on research. We studied how consistently two first-time users could apply the Guideline for Reporting Evidence-Based Practice Educational Interventions and Teaching (GREET) checklist. The GREET is a tool meant to help authors better report evidence-based practice education. In this study, two independent raters who did not have previous experience with the GREET assessed 161 randomized controlled trials about education for parents/caregivers of children with disabilities. They checked whether those trials adhered to the GREET in reporting various aspects of analyzed educational interventions. For each checklist item, raters marked trials with “Yes/No/Unclear” and copied the text that supported their decision. Disagreements were tracked and then discussed in several rounds to reach a consensus. At first, the raters disagreed on 20% of all item ratings (561 of 2737). After two discussion rounds, disagreements dropped to 14% and then to 2%, and full agreement was achieved after a final consensus step. On average, each trial had 3.5 initial discrepancies. Overall agreement between the two raters was substantial. The most disagreement occurred for items about the learning environment, materials included (what was actually provided and whether it was accessible), attendance, and adaptations (planned changes). Clearer GREET items, such as basic intervention description, theory, and educational strategies, had far fewer discrepancies. Many rating challenges were due to vague wording in GREET and inconsistent reporting in manuscripts. Some GREET items seemed to overlap, and several could benefit from being split into multiple items and having more precise definitions. What does this mean? The GREET checklist is useful, but first-time users can interpret its items differently. Short training and calibration, clearer item definitions, better examples, and guidance on what counts as “materials” or “attendance” would likely improve rating reliability. Authors of trials testing educational interventions can also help by reporting settings, educational materials (and how to access them), attendance, and any planned or unplanned changes with more detail. These steps would make education research more transparent and easier to reproduce.

1. Introduction

Transparent and complete reporting of evidence-based practice (EBP) educational interventions is essential for ensuring reproducibility, facilitating knowledge transfer, and enabling critical appraisal of research outcomes. Lack of transparency in reporting can lead to unnecessary duplication of research efforts, inefficient use of resources, and ultimately hinder progress in improving healthcare education [1].

A review by Albarqouni et al. examined the completeness of reporting in EBP educational interventions using the Template for Intervention Description and Replication (TIDieR) checklist. Their findings revealed that none of the 83 included studies completely reported all main items of the educational intervention within the original publication or in additional sources. Even after contacting study authors for missing information, complete details were available for only 20% of the interventions. The item most frequently missing was “intervention materials,” which was absent in 96% of the original publications [1].

The Guideline for Reporting Evidence-Based Practice Educational Interventions and Teaching (GREET) checklist was developed to improve the transparency and completeness of reporting in studies of EBP educational interventions [2]. To ensure consistency between these reporting guidelines, the TIDieR framework was adopted as a foundational structure for the GREET checklist [3].

Despite the availability of the GREET checklist for nearly a decade, its practical application and utility in assessing reporting quality remain relatively unexplored in

the literature. Based on the literature search, only a study by Bała et al. from 2024 has reported using this checklist to evaluate the reporting of evidence-based healthcare e-learning interventions [4]. They revealed substantial variability in adherence to the GREET checklist across published studies, highlighting potential challenges in its application or understanding among researchers. However, Bała et al. [4] did not assess and analyze inter-rater agreement in applying the checklist.

The GREET checklist is a valuable tool for improving the reporting of EBP educational interventions, but researchers unfamiliar with it may face challenges. Varying interpretations between raters can cause inconsistencies in its application. Recognizing these issues is key to refining the checklist and enhancing its usefulness for reporting educational interventions.

The present study aims to address these gaps by conducting an in-depth analysis of discrepancies in GREET ratings between two assessors without prior experience with the checklist, and by systematically documenting challenges encountered during the assessment process.

2. Methods

2.1. Included studies

This study included randomized controlled trials (RCTs) identified through a scoping review of educational interventions targeting parents/caregivers of children with developmental disabilities (registered at: <https://osf.io/fz5sq/>),

What is new?

Key findings

- In analysis of 161 randomized controlled trials, two first-time users of the GREET checklist achieved substantial inter-rater agreement ($\kappa = 0.616$; 95% CI 0.590–0.642), with initial item-level discrepancies reduced to zero after four cycles of assessments.
- Disagreements clustered around GREET items on Environment, Materials included (description vs accessibility), Attendance, and Adaptations, largely driven by ambiguous item phrasing and variable manuscript reporting.

What this adds to what is known?

- The study provides an item-level map of where GREET-naïve raters struggle, identifying overlap and candidate items for refinement (including subdivision of the “educational materials” item) and the need for more precise definitions and examples.
- Calibration/consensus exercises can markedly improve reliability for first-time GREET users, informing practical guidance for data extraction teams.

What is the implication/what should change now?

- We recommend refining GREET wording and guidance, with specific examples, and encourage authors/reviewers to report setting, provided materials and how to access them, attendance/engagement, and planned/unplanned adaptations. Authors using GREET for assessment should implement calibration/consensus to enhance reliability.

which is currently under review for publication. The present analysis constitutes a secondary study of the original scoping review.

2.2. Outcomes

We analyzed the number and percentage of discrepancies in item ratings between the two raters and assessed inter-rater agreement for items on the GREET checklist. We documented challenges encountered during the assessment of specific GREET items.

2.3. Raters

Two authors (TA and MMK), experienced in evidence synthesis but naïve to GREET methodology, independently

applied the checklist and its accompanying explanation and elaboration document to assess a sample of RCTs. We used two raters to mirror the typical real-world application of reporting checklists in systematic review teams. While designs with more raters can further reduce rater-specific influence, they were beyond the scope of the present study, which aimed to evaluate GREET’s usability and interpretability under the common two-rater workflow.

2.4. Evaluation of reporting against the Guideline for Reporting Evidence-Based Practice Educational Interventions and Teaching checklist

We evaluated the included trials using the GREET checklist, which comprises 17 items [3]. Raters assigned one of three possible ratings to each item: “Yes,” “No,” or “Unclear.” For every item, raters were required to extract verbatim text from the manuscript to substantiate their rating or indicate that the topic addressed by the item was not reported.

All rating discrepancies were adjudicated through a sequential, multistep process, which began with a review by one rater (MMK), followed by a secondary review of unresolved items by the second rater (TA), and culminated in iterative consensus discussions until 100% agreement was reached.

For this analysis, we excluded item 4 and divided item 5 into two distinct subitems. The inter-rater calibration process was undertaken on the initial seven manuscripts using a methodical approach.

GREET item 4 (“identify the EBP content,” ie, the EBP steps targeted) was not scored because our a priori objective was to assess reporting completeness and replicability, whereas item 4 chiefly addresses curricular scope (what content was taught). In addition, item 4 overlaps conceptually with other GREET domains we scored (eg, learning objectives, materials, procedures), and including it would risk double-counting content-related information without improving the precision of our reporting assessment.

Item 5 was subdivided to independently assess its two distinct components: the adequacy of the description of educational materials (item 5a) and the provision of the materials themselves to ensure replicability (item 5b). Extensive details about the study methods can be found in [Supplementary file 1](#).

2.5. Data synthesis and analysis

We used descriptive statistics to summarize and present the rating discrepancies between the two raters as numbers and percentages. Adherence to the GREET checklist was expressed with the median and IQR of ratings. Agreement was defined as exact concordance between raters on three nominal categories (Yes/No/Unclear); any nonidentical pair was counted as a disagreement. Inter-rater agreement was

Table 1. Discrepancies in the assessment of reporting adherence to the GREET checklist between two raters ($N = 161$)

No.	GREET item	Description of the item	Discrepancies in ratings between two raters after initial round of assessment, N	Discrepancies in ratings between two raters after the second round of assessment, N	Discrepancies in ratings between two raters after the third round of assessment, N
			(%) ^a	(%) ^a	(%) ^a
1	Intervention	Provide a brief description of the educational intervention for all groups involved [eg, control and comparator(s)].	1 (1)	1 (1)	0 (0)
2	Theory	Describe the educational theory(ies), concept or approach used in the intervention.	1 (1)	1 (1)	0 (0)
3	Learning objectives	Describe the learning objectives for all groups involved in the educational intervention.	17 (11)	10 (6)	0 (0)
5a	Materials described	Describe the specific educational materials used in the educational intervention.	37 (23)	21 (13)	3 (2)
5b	Materials included	Include materials provided to the learners and those used in the training of educational intervention providers. Is educational material publicly available?	53 (33)	47 (29)	8 (5)
6	Educational strategies	Describe the teaching/learning strategies (eg, tutorials, lectures, online modules) used in the educational intervention.	9 (6)	6 (4)	0 (0)
7	Incentives	Describe any incentives or reimbursements provided to the learners.	15 (9)	9 (6)	4 (2)
8	Instructors	For each instructor(s) involved in the educational intervention describe their professional discipline, teaching experience/expertise. Include any specific training related to the educational intervention provided for the instructor(s).	27 (17)	19 (12)	1 (1)
9	Delivery	Describe the modes of delivery (eg, face-to-face, internet or independent study package) of the educational intervention. Include whether the intervention was provided individually or in a group and the ratio of learners to instructors.	24 (15)	10 (6)	0 (0)

(Continued)

Table 1. Continued

No.	GREET item	Description of the item	Discrepancies in ratings between two raters after initial round of assessment, N (%) ^a	Discrepancies in ratings between two raters after the second round of assessment, N (%) ^a	Discrepancies in ratings between two raters after the third round of assessment, N (%) ^a
10	Environment	Describe the relevant physical learning spaces (e.g., conference, university lecture theatre, hospital ward, and community) where the teaching/learning occurred.	68 (42)	41 (25)	3 (2)
11	Schedule	Describe the scheduling of the educational intervention including the number of sessions, their frequency, timing, and duration.	29 (18)	21 (13)	1 (1)
12	Time spent	Describe the amount of time learners spent in face-to-face contact with instructors and any designated time spent in self-directed learning activities.	32 (20)	17 (11)	0 (0)
13	Adaptations	Did the educational intervention require specific adaptation for the learners? If yes, please describe the adaptations made for the learner(s) or group(s).	54 (34)	40 (25)	7 (4)
14	Modifications	Was the educational intervention modified during the course of the study? If yes, describe the changes (what, why, when, and how).	35 (22)	25 (16)	4 (2)
15	Attendance	Describe the learner attendance, including how this was assessed and by whom. Describe any strategies that were used to facilitate attendance.	64 (40)	50 (31)	13 (8)
16	Delivered as planned check	Describe any processes used to determine whether the materials (item 5) and the educational strategies (item 6) used in the educational intervention were delivered as originally planned.	46 (29)	27 (17)	1 (1)
17	Delivered as scheduled	Describe the extent to which the number of sessions, their frequency, timing, and duration for the educational intervention was delivered as scheduled (item 11).	49 (30)	34 (21)	0 (0)
TOTAL		561 (20)	379 (14)	45 (2)	

GREET, Guideline for Reporting Evidence-Based Practice Educational Interventions and Teaching.

Disagreement is defined as any nonidentical pair of ratings across the three nominal categories (Yes/No/Unclear).

^a Analyses were conducted following the initial rating performed independently by two assessors. Percentages were calculated based on the total number of analyzed trials: 161.

calculated using Cohen's Kappa. To interpret inter-rater agreement, we utilized the following scale in line with Landis et al [5]: Kappa < 0: no agreement, Kappa 0.00–0.20: slight agreement, Kappa 0.21–0.40: fair agreement, Kappa 0.41–0.60: moderate agreement, Kappa 0.61–0.80: substantial agreement, and Kappa 0.81–1.00: almost perfect agreement. For these analyses, we used MedCalc Statistical Software version 23.1.3 [6].

3. Results

We analyzed 161 RCTs. The list of included trials is provided in [Supplementary file 2](#).

3.1. Discrepancies between raters

Inter-rater calibration was conducted using seven manuscripts selected alphabetically by first author surname, with two raters independently evaluating each manuscript sequentially and documenting assessments with verbatim extracts in predefined spreadsheets. Following each manuscript evaluation, raters exchanged spreadsheets to facilitate secondary review and iterative discussion, repeating this process across multiple cycles until 100% consensus was achieved on all GREET checklist items.

Following a calibration process, the initial round of independent assessments by the two raters revealed discrepancies in 561 (20%) of the 2737 GREET items across 161 included studies. In the second round of assessments, one rater (MMK) analyzed all 561 initial GREET items discrepancies, leading to 379 (14%) unresolved items for further discussion and revision. The third round of assessments was done by the second rater (TA). After the third round, 45 items (2%) remained unresolved, necessitating further discussion between the two raters. Ultimately, within four cycles of discussion, complete alignment (100%) was achieved, with all discrepancies resolved through consensus between assessors MMK and TA ([Table 1](#); [Figs 1 and 2](#)). On average, each trial exhibited 3.5 discrepancies, with a median of 3 discrepancies per trial (range: 0–8).

3.2. Inter-rater agreement in assessing reporting using the Guideline for Reporting Evidence-Based Practice Educational Interventions and Teaching checklist

Inter-rater agreement yielded a value of 0.616 (SE = 0.013, 95% CI = 0.590–0.642), indicating substantial agreement between the two raters.

The highest levels of disagreement were observed in the domains of "Environment" (43%), "Materials included" (40%), "Adaptations" (40%), and "Attendance" (40%). Conversely, greater consensus with minimal discrepancies was noted in areas such as "Intervention" (1%), "Theory" (1%), "Educational strategies" (6%), and "Incentives" (9%).

3.3. Challenges applying the Guideline for Reporting Evidence-Based Practice Educational Interventions and Teaching checklist and suggestions for the checklist revision

Raters faced multiple challenges when evaluating RCTs against the GREET checklist.

3.3.1. Methods for scoring items

When designing this study, our goal was to assess whether published studies could be replicated based on manuscript information. Results for items 5a and 5b revealed distinct outcomes for educational material description and provision, respectively.

During the study, we observed that more items could be further subdivided or more precisely defined. Many GREET items encompass multiple concepts that may be considered subitems, as detailed in this document. Such items required raters to reach consensus on whether partial information provided by authors justified a "Yes" rating, or whether an "Unclear" rating was more appropriate when details for certain subitems were absent.

3.3.2. Availability of educational materials (item 5b)

Raters agreed to assign a "Yes" to item 5b if educational materials were referenced in the publication, such as citing a manual, since this could support replication. However, such manuals are rarely published and are often behind paywalls, limiting accessibility. Uploading materials online is another option, but links may expire, and websites lack permanence. In some cases, materials that claimed to be "freely available" were paywalled, yet item 5b was still rated as "Yes." The GREET checklist does not clarify whether materials must be freely accessible, posing challenges in consistent rating.

3.3.3. Defining educational materials: material videotaped during the trial and telehealth interventions (item 5b vs item 9)

Raters debated whether videotaped parent-child interactions used in trials qualified as educational materials (item 5). They concluded that such videos, while informative, lack structure and reproducibility and thus do not meet GREET criteria for educational materials.

Item 9 of the GREET checklist addresses modes of delivery for educational interventions. For telehealth studies, technologies like videoconferencing were classified under item 9 as delivery modes. However, item 5a was rated "Yes" only when authors clearly described the educational content delivered through these platforms.

3.3.4. Incentives (item 7)

Item 7 of the GREET checklist asks whether incentives were provided. Raters deliberated whether verbatim "no incentives were given" should be scored as "No" (with an explanation "no incentives were given") or as "Yes" (with

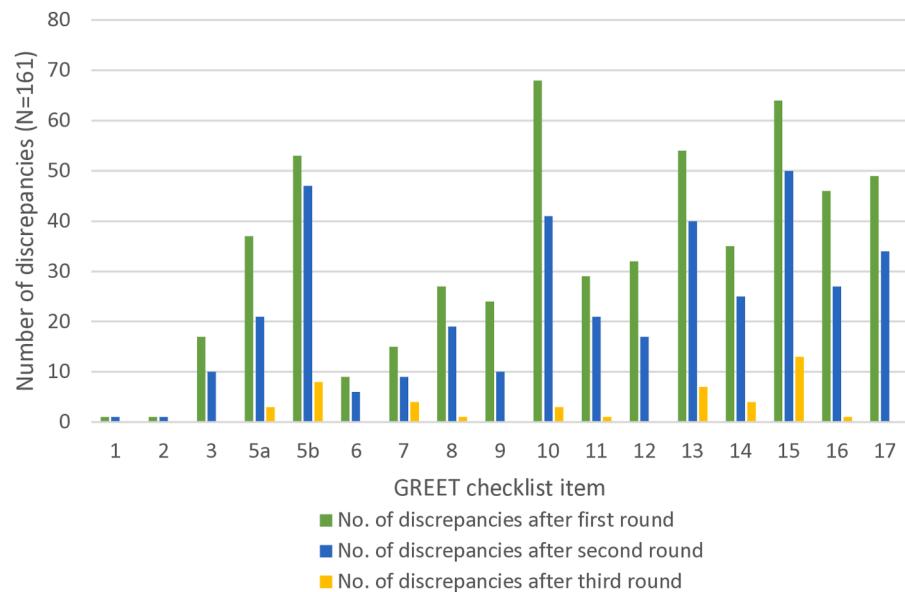


Figure 1. Evolution of rater discrepancies across three rounds of GREET checklist assessments for 161 trials. GREET, Guideline for Reporting Evidence-Based Practice Educational Interventions and Teaching.

an explanation “yes, the authors reported this item”). Since the checklist was used to assess reporting quality, not intervention content, item 7 was scored “No” if the information was not reported, regardless of whether incentives were provided.

3.3.5. Incentives vs. strategies to facilitate attendance (item 7 vs item 15)

In some trials, telephone calls were utilized as weekly reminders or fortnightly check-ins (phone, text, or email). In these instances, raters classified these as strategies to facilitate attendance, described in item 15, rather than incentives (item 7).

3.3.6. Mode of delivery of educational intervention (item 9)

In studies where the mode of delivery was not clearly stated, raters assumed face-to-face delivery unless digital methods were explicitly mentioned. This assumption applied especially to studies published before 2019, as digital delivery was less common pre-COVID-19, and authors may have considered in-person delivery the default, requiring no clarification.

3.3.7. Environment (item 10)

When authors reported that the educational intervention was delivered in a home-based setting, raters assigned a

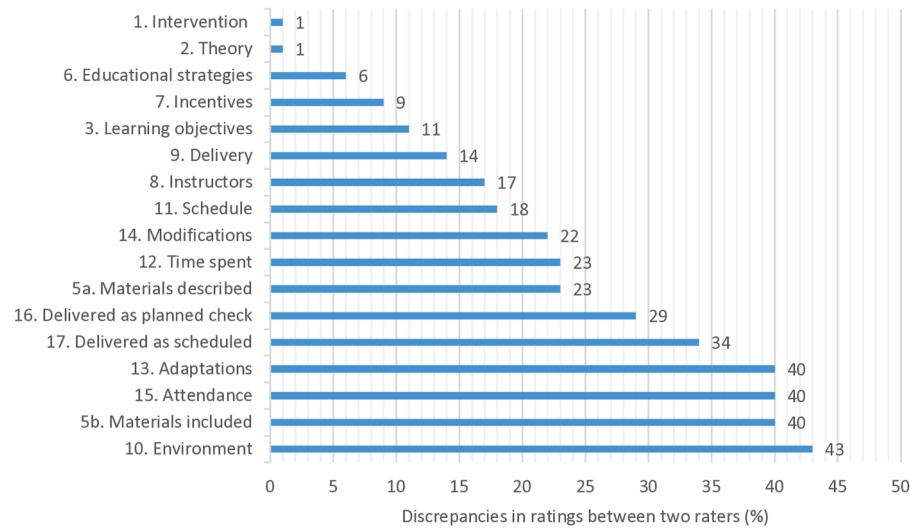


Figure 2. Discrepancies in ratings between the two raters for each individual GREET checklist item, expressed as the percentage of trials in which the raters initially disagreed on their assessments ($N = 161$). GREET, Guideline for Reporting Evidence-Based Practice Educational Interventions and Teaching.

“Yes” rating to item 10, as the participants’ home was explicitly identified as the environment for delivery. However, when trials had multiple treatment sessions, some authors specified the environment only for the initial and post-treatment follow-up sessions, leaving the settings for the remaining sessions unreported. This ambiguity led to a debate among raters. One perspective argued for a “Yes” rating, reasoning that partial reporting of the environment could be considered sufficient. Conversely, “Unclear” rating can be found more appropriate due to incomplete information. Ultimately, the raters decided to assign a “Yes” rating in such cases, opting for a less stringent interpretation of reporting requirements.

For telehealth studies, a “Yes” was assigned where authors explicitly stated participants’ locations during telehealth sessions—such as living rooms, offices, or community center quiet rooms.

3.3.8. Interchangeability of items 11 and 12

Items 11 and 12 of the GREET checklist both address time-related aspects of educational interventions, leading raters to view them as overlapping. Item 11 covers scheduling (eg, session timing, frequency, and duration), while item 12 focuses on time spent in face-to-face or self-directed learning. In many cases, raters scored both items identically, using the same excerpts. A clearer distinction might be achieved by refining item 12 to focus specifically on self-directed learning outside scheduled sessions. Alternatively, requiring authors to report total time spent in these activities could reduce redundancy and improve clarity.

3.3.9. Adaptations (item 13) and modifications (item 14)

The GREET checklist distinguishes between adaptations (item 13), which are planned changes made before a study starts, and modifications (item 14), which are unplanned changes during implementation. Item 13 was rated “Yes” when prestudy changes to an existing program were clearly reported. Item 14 was rated “Yes” if changes occurred after the intervention began. However, educators rarely use programs without some adaptation, making the boundary blurry. Tailoring content to individual learners (eg, selecting optional assignments) was not considered a modification unless elaborated as systematic. Lack of detail led to “Unclear” ratings for item 14.

3.3.10. Attendance (item 15)

To score “Yes” on item 15 regarding attendance, authors had to explicitly report learner attendance (eg, percentage of sessions attended). Notably, data presented in Consolidated Standards of Reporting Trials (CONSORT) diagrams, which typically outline the flow of study participants, were not considered sufficient for this item, as they do not specifically address attendance at educational intervention sessions. The trials were scored for item 15 with “No” when the authors did not directly describe or quantify learner attendance.

3.3.11. Processes used to determine whether the intervention was delivered as originally planned (item 16)

Item 16 assesses whether materials and strategies were delivered as planned. A “Yes” was given when structured fidelity checks (eg, checklists, videotaped reviews) were clearly reported. General mentions of supervision or vague references to compliance led to “No” ratings. When only parent adherence was assessed, but not therapist delivery, the rating was Unclear.

In a trial that analyzed videotaped sessions, one rater assigned a “Yes” score, reasoning that analyzing videotaped materials for treatment fidelity constituted a valid assessment process, albeit prone to errors and inconsistencies. However, another rater scored item 16 as “No,” arguing that analyzing videotaped materials without structured assessment forms, charts, or processes cannot be considered a reliable or reproducible method. Ultimately, raters reached consensus and assigned a “Yes” score for this trial and similar cases where sessions were videotaped and checked for treatment fidelity.

In instances where authors explicitly stated that structured checklists or procedures were employed to verify fidelity to treatment, assessors consistently rated item 16 as “Yes.” When studies lacked detail, focused only on learner adherence (eg, parents), or used vague terms like “compliance”, raters assigned “Unclear” or “No” ratings. Discussions highlighted the need to clarify whether fidelity should refer to educators, learners, or both, suggesting potential refinement of the checklist.

3.3.12. Intervention fidelity assessment: item 16 vs item 17

Item 17 of the GREET checklist assesses whether the intervention schedule (sessions, timing, frequency, duration) was followed as planned. However, most studies failed to clearly report this, often conflating it with item 16, which covers fidelity to materials and strategies. Only 4 of 161 trials (2.5%) explicitly addressed adherence to the schedule. This reporting gap suggests a need to revise item 17 to clearly require details on protocol deviations, especially in self-directed interventions. Clearer wording and distinction from item 16 would improve fidelity reporting and study reproducibility.

Rigorous adherence to the GREET checklist, with explicit reporting of both the processes used to assess fidelity (item 16) and the extent to which the intervention schedule was followed (item 17), is essential for improving the transparency and reproducibility of educational intervention studies.

Extensive details about the study results and detailed description of the challenges encountered, with references indicating challenging examples, are provided in [Supplementary file 3](#).

4. Discussion

Our study revealed major challenges and high inter-rater discrepancies in applying the GREET checklist for assessing the reporting quality of educational interventions. Despite the calibration process, we observed considerable initial disagreement between raters, with discrepancies in 20% of the evaluated GREET items across 161 trials. Through an iterative consensus process, these discrepancies were gradually resolved over four rounds of ratings, highlighting the complexity of consistently interpreting and applying the GREET checklist items.

The highest levels of disagreement between raters were observed for items such as “Adaptations,” “Materials included,” “Attendance,” and “Environment.” These findings suggest that these particular aspects of educational intervention reporting may be inherently more challenging to evaluate or that the GREET checklist guidance for these items may benefit from further clarification. The substantial disagreement on these items aligns with previous research, which indicated that intervention materials were frequently missing in educational intervention reports, being absent in 96% of original publications they reviewed [1].

Conversely, in our study, foundational areas for intervention reporting—such as “intervention,” “theory,” and “educational strategies”—exhibited minimal disagreement. This pattern suggests that some aspects of educational interventions are more consistently reported or more straightforward to evaluate than others.

Our findings complement the work of Bała et al., who found that studies assessing evidence-based healthcare e-learning interventions poorly adhered to the GREET checklist [4]. Similar to our findings, they noted that certain critical items were consistently missing across studies, such as information about modifications of educational interventions (item 14) and details about processes to determine whether materials and educational strategies were delivered as planned (item 16). While Bała et al. focused on adherence to the checklist, our study provides deeper insights into the methodological challenges of applying the checklist itself. The difficulties we encountered in distinguishing between certain items (such as items 11 and 12 regarding scheduling and time spent) and in interpreting others (such as items 13 and 14 on adaptations and modifications) suggest that the practical application of the GREET checklist may present challenges even for experienced researchers.

4.1. Implications for practice and research

This study underscores the need to recognize and address inter-rater discrepancies that arise when reporting guidelines, developed primarily as author guidance, are repurposed as rating instruments. While collaborative discussion significantly improved inter-rater agreement in our study, the findings highlight the importance of thorough training and calibration among the checklist’s users.

Multiple GREET items were found to be especially prone to disagreement, suggesting a need for clearer definitions and more detailed guidance in future revisions. In addition, the overlap and ambiguity between certain items, such as those assessing schedule, duration, and intervention fidelity, point to opportunities for refinement to enhance clarity and reduce redundancy.

Although the GREET checklist has been available since 2016 [3], its application in evaluating reporting quality remains underexplored, and this study offers valuable insights for researchers, editors, and peer reviewers. This study highlights several important directions for future research. First, methodological research is needed to explore the optimal sample size of included trials for calibration exercises in meta-research studies assessing adherence to reporting checklists, as current literature lacks guidance on this issue. Second, future research should examine whether additional guidance materials or training programs could improve inter-rater reliability when applying the GREET checklist. Third, the GREET checklist may benefit from revisions to improve clarity and reduce ambiguity, including subdivisions of complex items.

We acknowledge that the original aim of reporting guidelines is to help authors write better reports. This has already been questioned in the literature, with warnings that reporting checklists should not be used as an assessment tool for reporting quality, as they are not a measurement tool [7]. Few checklists, such as Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) [8], have designed an additional appraisal tool to help investigators use a checklist as an assessment tool. However, many methodological studies (including ours) are using reporting checklists for assessing reporting completeness [9], even though they have not been validated for such use [10]. This is easy to understand, since there are no other reporting appraisal tools for study reports. For this reason, it is important to conduct studies such as ours. We hope that our results will be useful not only to help other authors use the GREET for appraisal, but also the GREET authors can take these findings into account to consider revising the GREET checklist and its accompanying elaboration document. If raters struggle with understanding and applying GREET to appraise studies, then authors who use GREET may also face challenges in understanding what is expected of them when they aim to use it for improved reporting.

GREET’s developers followed Enhancing the Quality and Transparency of Health Research (EQUATOR) guidance, using a staged process (systematic review, four-round Delphi study and international consensus meetings), produced an Explanation & Elaboration (E&E) paper with item-level guidance, and undertook psychometric testing (intraclass correlation coefficient for inter-rater reliability and criterion validity) with tertiary health professional students applying GREET to a published study [3,11]. A planned pilot by Delphi panelists during manuscript

drafting did not proceed [11], as no panelists accepted [3]. The authors flagged this as a limitation and motivation for further validation with expert users. In GREET's testing, agreement was lowest for steps of EBP and Materials, partly mirroring our finding that "Materials included" was disagreement-prone. Our study additionally highlights Environment, Attendance, and Adaptations as high-friction items for naïve raters. These observations support our suggestion to refine the definitions and examples for these items in future GREET updates.

4.2. Comparison with other reporting guidelines: supports for adherence

Several established reporting guidelines pair their checklists with implementation supports that facilitate author adherence. CONSORT and Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) publish comprehensive E&E papers with item-by-item rationale and exemplars; PRISMA also provides standardized flow-diagram templates and an online Shiny app to generate them [12–15]. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) and Standards for Quality Improvement Reporting Excellence (SQUIRE) likewise offer E&E companions with worked examples [16–19]. Beyond materials, journal endorsement of CONSORT/PRISMA has been associated with more complete reporting in multiple evaluations. GREET includes an E&E document, but (to our knowledge) lacks author-facing templates or interactive tools. Our results point to where GREET's guidance could most profitably be expanded (eg, decision rules, exemplar text, and minimal data elements).

4.3. Limitations

Our study had several limitations. Our focus on educational interventions for parents/caregivers of children with developmental disabilities may not capture the full range of challenges that might arise when applying the GREET checklist to other types of educational interventions. Also, our decision to exclude item 4 and subdivide item 5, while methodologically justified, represents a deviation from the standard application of the GREET checklist that may affect comparability with other studies using the checklist in its original form.

Our choice of two raters increases ecological validity by reflecting how GREET is commonly used in real-world appraisal workflows. We acknowledge that including more raters would be useful for isolating checklist- vs. rater-level variance and for improving the precision of item-level estimates. Future studies should consider multirater, pair-randomized assignments (eg, two raters per paper with raters rotated across papers) and modeling approaches that partition rater, item, and study effects.

We acknowledge that a larger number of raters would be preferable and would likely produce more stable reliability estimates and further minimize rater-specific influence. Our study was designed to test GREET under the prevalent two-rater workflow; therefore, the results should be interpreted in that context. Evaluations with larger multirater designs are warranted.

5. Conclusion

Although the GREET checklist provides a valuable framework for assessing reporting quality in EBP educational interventions, its application may yield substantial discrepancies between raters without prior experience. Clearer item definitions, improved guidance materials, and training could enhance its reliability. Findings support the need for continued refinement of the checklist and underscore the importance of comprehensive and transparent reporting in educational research.

Ethics statement

The study analyzed published research articles. Thus, approval from an ethics committee is not applicable.

CRediT authorship contribution statement

Tea Andabaka: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Mersiha Mahmic-Kaknjo:** Writing – review & editing, Methodology, Data curation. **Livia Puljak:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare no conflicts of interest.

Acknowledgments

We are grateful to the anonymous reviewers for their time and insightful comments, which helped us to improve the manuscript.

Supplementary data

Supplementary data related to this article can be found at <https://doi.org/10.1016/j.jclinepi.2025.112020>.

Data availability

The raw data set generated within this study is available from the corresponding author on request.

References

- [1] Albarqouni L, Glasziou P, Hoffmann T. Completeness of the reporting of evidence-based practice educational interventions: a review. *Med Educ* 2018;52:161–70. <https://doi.org/10.1111/medu.13410>.
- [2] Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: Template for Intervention Description and Replication (TIDieR) checklist and guide. *BMJ* 2014;348:g1687.
- [3] Phillips AC, Lewis LK, McEvoy MP, Galipeau J, Glasziou P, Moher D, et al. Development and validation of the Guideline for Reporting Evidence-based practice Educational interventions and Teaching (GREET). *BMC Med Educ* 2016;16:237. <https://doi.org/10.1186/s12909-016-0759-1>.
- [4] Baća MM, Peričić TP, Žuljević MF, Bralić N, Zajac J, Motaze NV, et al. Adherence to the guideline for reporting Evidence-based practice educational interventions and Teaching (GREET) of studies on evidence-based healthcare e-learning: a cross-sectional study. *BMJ Evidence-Based Med* 2024;29:229–38.
- [5] Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
- [6] Schoonjans F. MedCalc statistical software - free trial available. MedCalc n.d. Available at: <https://www.medcalc.org/>. Accessed April 29, 2025.
- [7] Logullo P, MacCarthy A, Kirtley S, Collins GS. Reporting guideline checklists are not quality evaluation forms: they are guidance for writing. *Health Sci Rep* 2020;3:e165. <https://doi.org/10.1002/hsr2.165>.
- [8] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015;350: g7594. <https://doi.org/10.1136/bmj.g7594>.
- [9] Plenkovic M, Civljak M, Puljak L. Authors arbitrarily used methodological approaches to analyze the quality of reporting in research reports: a meta-research study. *J Clin Epidemiol* 2023;158:53–61. <https://doi.org/10.1016/j.jclinepi.2023.03.008>.
- [10] Puljak L. Reporting checklists are used as measurement tools for assessing quality, even though they have not been validated for such use. *Trials* 2019;20:676. <https://doi.org/10.1186/s13063-019-3858-6>.
- [11] Phillips AC, Lewis LK, McEvoy MP, Galipeau J, Glasziou P, Hammick M, et al. Protocol for development of the guideline for reporting evidence based practice educational interventions and teaching (GREET) statement. *BMC Med Educ* 2013;13:9. <https://doi.org/10.1186/1472-6920-13-9>.
- [12] Hopewell S, Chan A-W, Collins GS, Hróbjartsson A, Moher D, Schulz KF, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. *BMJ* 2025;389:e081123. <https://doi.org/10.1136/bmj-2024-081123>.
- [13] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71. <https://doi.org/10.1136/bmj.n71>.
- [14] Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* 2021;372:n160. <https://doi.org/10.1136/bmj.n160>.
- [15] Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: an R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. *Campbell Syst Rev* 2022;18: e1230. <https://doi.org/10.1002/cl2.1230>.
- [16] Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening the reporting of observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol* 2008;61:344–9. <https://doi.org/10.1016/j.jclinepi.2007.11.008>.
- [17] Vandenbroucke JP, Von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *Plos Med* 2007; 4:e297. <https://doi.org/10.1371/journal.pmed.0040297>.
- [18] Ogrinc G, Davies L, Goodman D, Batalden P, Davidoff F, Stevens D. SQUIRE 2.0 (*Standards for QUality Improvement Reporting Excellence*): revised publication guidelines from a detailed consensus process. *BMJ Qual Saf* 2016;25:986–92. <https://doi.org/10.1136/bmqs-2015-004411>.
- [19] Goodman D, Ogrinc G, Davies L, Baker GR, Barnsteiner J, Foster TC, et al. Explanation and elaboration of the SQUIRE (*Standards for Quality Improvement Reporting Excellence*) guidelines, V.2.0: examples of SQUIRE elements in the healthcare improvement literature. *BMJ Qual Saf* 2016;25:e7. <https://doi.org/10.1136/bmqs-2015-004480>.