Quality & Quantity
https://doi.org/10.1007/s11135-024-01969-9

Guidelines for reporting research using systematic coding of observed human behaviour (SCOBe)

Marij A. Hillen[1] · Evelien M. Hoeben[2] · Rianne Kok[3] · Calum McHale[4] · Melissa M. Sexton[5] · Sabine van der Asdonk[6] · Daphne van de Bongardt[3] · Brett Bowman[7] · Richard Brown[8] · Susan Branje[9] · Jill Chorney[10] · Monika H. Donker[9] · Peter Ejbye-Ernst[2] · Sanne Geeraerts[9] · Danielle Van der Giessen[11] · Gerald Humphris[4] · Helle Larsen[12] · Richard Philpot[13] · Mariona Portell[14] · Siyang Yuan[15] · Chiel van der Veen[16] · Janneke Noordman[17]

Accepted: 24 August 2024

Abstract
Systematic coding of observed human behaviour (SCOBe) is used across disciplines and topics but methodological reporting is often incomplete. We developed internationally generated, interdisciplinary guidelines for methodological reporting of such research. Using Delphi methodology, a working group of 22 experts sought group consensus in three rounds. Participants first assessed an initial set of reporting criteria (round 1). Next, in interactive meetings participants revised these criteria and reached consensus on reporting content (rounds 2 & 3). We present 20 criteria constituting the first comprehensive reporting guidelines for SCOBe research using existing, newly developed, or modified coding systems. The criteria encompass three procedural domains: 1. Research context; 2. Properties of the coding scheme; and 3. Application of the coding scheme. The presented guidelines will assist in substantiating and assessing the quality of SCOBe research. We encourage researchers to adopt these guidelines, to enhance quality of mono- and interdisciplinary research.

Keywords Observational research · Systematic coding · Interdisciplinary research · Methodological reporting · Delphi methodology

1 Introduction

Systematic coding of observed human behaviour (SCOBe) is a method used to interpret, quantify, and analyse both verbal and non-verbal human behaviour.[1] Observations can occur either in real-time or involve audio- or video-recorded data. Researchers may observe specific behaviours at specific times or during longer periods using more global ratings (Bakeman and Quera 2011; Yoder and Symons 2010). This type of research allows

[1] We use the term 'behaviour' in a broad sense (see Online Appendix B, Glossary). Hence, whenever we use the term 'behaviour', this may also refer to an individual's emotions, states, or attributes.

Extended author information available on the last page of the article

reliable and observably verifiable investigation of naturally occurring behaviours or mechanisms exhibited during interactions that cannot always be validly reported by participants themselves, as many of these behaviours are unconscious and brief, or are subject to memory biases and effects (Philpot et al. 2019; van Dulmen et al. 2012). Systematic coding of observed human behaviour is used for a large variety of research goals across disciplines, including family studies (e.g., Gardner 2018), criminology (e.g., Lindegaard and Bernasco 2018), psychology (e.g., Hiller et al. 2018), education (e.g., Terroba et al. 2022), sports science (e.g., Castañer et al. 2017), and health communication (e.g., Bensing et al. 2006). Examples of behaviours under study are patient involvement in medical decision-making (Menichetti et al. 2021), the public's social distancing compliance behaviours (Hoeben et al. 2021), bystander behaviour during public violence (Ejbye-Ernst et al. 2022), emotional flexibility of parent–child dyads (Van der Giessen et al. 2015), and police-citizen interactions (Philpot et al. 2021). Although the focus of such research using SCOBe may differ substantially across and even within disciplines, methodological quality and transparency are always of utmost importance to ensure reliable, valid, and replicable results.

Across these various disciplines and topics, numerous systems for coding behaviour have been developed. Although some have been validated for wider use (e.g., Biringen et al. 2000; Zimmermann et al. 2011), many are tied to specific studies or projects (e.g., Hiller et al. 2018). The quality and detail of scientific reporting about SCOBe systems and their application varies widely. For example, many reports on coding schemes (Online Appendix B) lack sufficient detail about the development process and application of the codes (Online Appendix B), and on validity and/or reliability testing. It is therefore often difficult to assess the quality of this type of research, limiting the replicability, visibility, secondary use, and further refinement of existing coding systems, and impeding the field's progression.

Efforts to develop methodological reporting guidelines (e.g., CONSORT for randomised controlled trials) have shown that such guidelines can enhance consistency in reporting and facilitate the evaluation of research underlying multiple studies (Kearney 2014). Similarly, we believe that guidelines will advance the methods used in these studies by increasing the transparency and consistency of scientific reporting. Such guidelines could not only support researchers in reporting about their research, but also facilitate its evaluation for reviewing or replication purposes. Although initial reporting criteria or 'rules of thumb' have been proposed within specific disciplines (e.g., Chorney et al. 2015; Portell et al. 2015), no comprehensive reporting guidelines exist.

We therefore aimed to compile interdisciplinary guidelines for methodological reporting of research using systematic coding of observed human behaviour (SCOBe). Specifically, we sought to develop such guidelines for reporting the use of existing systems, newly developed systems, and modification of existing coding systems. Recognising that researchers may use SCOBe in combination with a variety of other methods, we have not considered broader methodological considerations (e.g., justification of the sample size) in the development of our guidelines.

2 Methods

An interdisciplinary and international working group of experts in SCOBe methodology participated in a multi-day hybrid workshop funded and supported by the Lorentz Center (www.lorentzcenter.nl/about-us.html). The collaborative effort and input from experts in

different fields and countries allowed for the robust, interdisciplinary approach required to develop comprehensive guidelines (Moher et al. 2010).

2.1 Design

We used Delphi techniques to establish consensus among the involved experts. In Delphi methodology, experts typically evaluate an issue in multiple 'rounds' –in our case three–interspersed with controlled feedback (Niederberger et al. 2021; Powell 2003). Delphi methods were deemed particularly suitable for our specific purpose of creating broadly supported guidelines among a diverse group of experts, because of their ability to structure and organize group communication (Powell 2003). Prior to the workshop, participants assessed an initial set of reporting criteria (round 1), based on existing preliminary and mono-disciplinary guidelines (see 'preparatory steps'). Next, participants revised the initial reporting criteria and collaboratively reached consensus on reporting content for each criterion, using in-person meetings (round 2) and online meetings (round 3; see 'group work').

2.2 Workshop participants

The workshop brought together 22 participants who were experienced in systematic coding of observed human behaviour (SCOBe). All participants were intricately involved throughout the development of the guidelines and the writing process, and collaboratively co-authored the current manuscript. For the selection of participants, the organisers (M.H., J.N., E.H., C.M., and M.S.) used a snowballing approach to seek representation from different scientific disciplines, levels of professional seniority, geographical location, and gender (see Online Appendix A). The workshop consisted of a three-day in-person meeting held in February 2023, followed by two online meetings in March and May 2023. To maximise global participation, plenary sessions of the in-person workshop were recorded and made available to non-European online participants.

2.3 Preparatory steps–Delphi round 1

Prior to the workshop, the organisers searched the literature using Google Scholar with variations of the keywords 'behavioural coding' and 'coding scheme', to identify and compare existing SCOBe reporting guidelines. Additionally, group members were asked to suggest existing guidelines, resulting in two articles (Chorney et al. 2015; Portell et al. 2015) that were utilised for a forward/backward citation search to identify other source materials. Useful methodological guidance was identified in seven additional publications (Bakeman et al. 2008; Blanch-Hartigan et al. 2018; Harris and Lahey 1982; Heyman et al. 2014; Jakubauskaite 2021; Jones and Raymond 2012; Tschan et al. 2018). Based on the literature, two preliminary sets of criteria were created, focusing on *conducting* and *reporting on* SCOBe research, respectively. Group members provided feedback on the preliminary criteria through a questionnaire, addressing aspects such as order, in/exclusion, and terminology. Based on the input, the organisers agreed to narrow the workshop focus to creating *reporting* guidelines (as opposed to guidelines for *conducting* research).

After integrating feedback, the organisers developed a second version of the reporting guidelines. The guidelines were developed in a non-sequential checklist format and divided

into three procedural domains: Context (Domain 1); Properties of coding scheme (Domain 2); and Application of coding scheme (Domain 3). Additionally, the organisers created a glossary (see Online Appendix B) to facilitate interdisciplinary consensus on definitions, and illustrate links between the terms used throughout the guidelines and other common terms in the disciplines represented. Lastly, the project was registered on Equator to ensure transparency, raise awareness, and prevent duplication (see https://www.equator-network.org/library/reporting-guidelines-under-development/reporting-guidelines-under-development-for-other-study-designs/#SCOBE).

2.4 Group work–Delphi rounds 2 & 3

During the three in-person workshop days, small multi-disciplinary groups updated and refined the criteria within all three domains. Subsequently, plenary discussions resolved remaining issues, such as terminology, prioritisation, and allocation of criteria to domains. Day 1 focused on determining the scope and focus of each domain. Day 2 entailed reviewing the results of day 1. Day 3 entailed creating detailed descriptions and instructions per criterion. During days 1–3, considerable time was spent discussing and aligning terminology regarding the exact focus of these guidelines to be universally applicable across disciplines. By the end of day 3, we reached consensus on the most accurate terminology and agreed on the terms 'Systematic Coding of Observed human Behavior' (abbreviated as SCOBe). This combination of terms was agreed to be sufficiently specific—for example, the terms 'observational research' were considered too broad, and 'human' was added because we cannot speak to the applicability of the guidelines. The SCOBe terms were also determined to be sufficiently inclusive—for example, the term 'video' was intentionally left out, as also audio data or transcriptions are used in some disciplines. After day 3, revised reporting guidelines were shared with the online participants.

Online participants reviewed the revised guidelines and provided feedback prior to the online meeting days (workshop days 4 and 5; Delphi round 3). During day 4, online participants reviewed and revised each domain in online breakout rooms. Participants from the in-person workshop joined to respond to queries from online participants and to facilitate consensus on any additional changes. After the workshop, the organisers compiled an initial manuscript draft, which they distributed to all participants for feedback.

Day 5 addressed unresolved issues about the manuscript draft, facilitated collaborative real-time drafting of the discussion section, and encouraged discussions of guideline dissemination plans. Afterwards, the organisers consolidated all suggested edits and additions to create an updated draft, which all participants reviewed before finalizing the manuscript. Consensus on the revised set of criteria was reached among all participants.

3 Results

The results section comprises two elements: (1) a checklist of criteria recommended for inclusion in any scientific reporting of systematic coding of observed human behaviour (SCOBe) research, categorised into the three procedural domains (Table 1) elaboration and further clarification of each criterion. A glossary of essential terms (marked in text using underlining) is provided in Online Appendix B. An online checklist version of the guidelines can be found in Online Appendix C. The reporting guidelines describe only those elements specific to SCOBe research. Researchers who combine SCOBe with other methods

Table 1 Checklist of reporting guidelines

Item	Domains and criteria	Explanation
	Domain 1: Research context	
1	Concept(s) under study	Describe the research concept(s) on which the systematic observational coding focuses, including any potential sub-concepts, preferably including empirical and theoretical foundation.
2	Rationale for use of systematic observational coding	Explain why systematic observational coding was chosen as the most appropriate method for the study and how it links to the theoretical framework.
3	Context of research	Provide a description of the context of the observational research and report where and when source data were obtained.
4	Research population	Specify who were present in the observation and whose behaviours were being coded. Report relevant characteristics of the sample.
5	Acquisition of source data	Describe observational data sources, how these were obtained (e.g., video, audio, in situ, or transcription; level of experimental manipulation; level of researcher participation), and whether the data were collected for the current research aim (primary or secondary data).
6	Sampling within available source data	Describe the sampling approach, for example, the designated periods or events within the source data in which the concepts were coded.
	Domain 2: Properties of coding scheme	
	Domain 2a: General description of coding scheme	
7	Name and availability of the coding scheme	Provide the name of the coding scheme. Ideally, provide the full coding scheme and codebook in the manuscript or supplementary materials. If that is not possible, provide information on how readers can access those materials.
8	Rationale for use of this specific coding scheme	Provide an explanation for the selection and use of this specific coding scheme (versus other available schemes).
9	Validity and reliability of the coding scheme as established in previous research	Provide the validity and reliability of the original coding scheme as it has been established in previous research.
10	Specification of codes	
10a	Description of codes	Provide a description and explanation of all the codes that were used in the study.
10b	Connections between codes	Describe the interrelatedness of the different codes, including potential hierarchy between the codes, whether codes were mutually exclusive and exhaustive, and any prioritisation in coding procedures.
10c	Granularity of codes	Specify the coding unit, i.e., the level of specificity at which codes were assigned (e.g., micro or macro level).
	Domain 2b: General description of newly developed or modified coding scheme	
11	Justification for coding scheme modification or development	Describe why a coding scheme was modified or developed. That is, describe why existing coding schemes could not be used and elaborate on the theoretical and empirical framework of the new or modified coding scheme.
12	Modification or development process	Specify how the coding scheme was developed or modified.
13	Validity assessment for the newly developed or modified coding scheme	Describe how validity of the newly developed or modified coding scheme was established, and which types of validity were assessed.
14	Reliability assessment of the newly developed or modified coding scheme	Describe how reliability of the newly developed or modified coding scheme was established, and which types of reliability were assessed.
	Domain 3: Application of coding scheme	
15	Coder selection and training	Describe the process of training individuals to apply the coding scheme to the data correctly and reliably in relation to the study purpose. Additionally, describe any relevant information related to the selection, eligibility, and/or suitability of coders.
16	Coding procedures	Provide a comprehensive description of how and by whom the codes were practically applied to the raw observational data, including the calibration process.
17	Reliability preservation throughout coding scheme application	Describe the steps taken to maintain reliability during the coding process, i.e., procedures to minimise coder drift and reliability outcomes achieved when applying the coding scheme to source data.
18	Ethical considerations	Describe ethical considerations and aspects of the approval process relevant or pertinent to the observations undertaken.
19	Facilitating software	Describe which software was used to view, code, and analyse the data.
20	Data preparation and analysis	Describe how coded data were prepared pre-analysis and which descriptive and statistical tests were applied.

[§]Reporting order may vary depending on disciplinary standards and/or journal requirements.

are advised to additionally consult available guidelines for other methodologies through the Equator database for reporting guidelines (www.equator-network.org).

Because the reporting guidelines are intended for application within multiple disciplines and for multiple types of systematic coding of observed human behaviour, not all criteria may be applicable to each individual study. Moreover, the order of reporting required by the conventions of any specific discipline may be different from the order in which the criteria are listed. What information is considered essential to report, and in which section of the manuscript it should be reported, may vary depending on disciplinary standards or practices. We recommend that where it is not possible to report on certain criteria, authors briefly explain this in the appropriate section of their manuscript.

3.1 Domain 1: Research context

This domain includes the reporting of the theoretical background, context, and focus of the research, and the procedures of data collection.

3.1.1 Criterion 1: Concept(s) under study

Describe the research concept(s) underpinning the systematic coding of observed human behaviour (SCOBe), preferably including their associated empirical and theoretical foundations. Make sure that the sub-concepts are described, when applicable (e.g., 'caregiving skills' as the main concept with sub-concepts 'instrumental and emotional caregiving'). If the study combines systematic observational coding with other methods, specify how the concepts captured through the SCOBe design relate to the objectives of the overall study.[2]

3.1.2 Criterion 2: Rationale for use of systematic observational coding

Explain why systematic coding of observed human behaviour was the most appropriate method to address the specific research question(s), research concept(s), research population, and (if relevant) how it aligns with the theoretical framework.

3.1.3 Criterion 3: Context of research

Provide a description of the context of the source data (Online Appendix B). Specifically, describe *where* the data were collected, both in terms of geographical location (e.g., city, country) and observation sites (e.g., classroom, paediatric out-patient clinic, lab, home), and *when* the data were collected (e.g., October 2022–March 2023). Where relevant, include contextual information related to the timing of data collection (e.g., during a pandemic lockdown) or the situation that was observed (e.g., mother-infant interactions during nappy change). Consider data protection and privacy conventions and regulations when reporting this information–overly high levels of specificity may jeopardise confidentiality.

[2] Using a framework that distinguishes different types of (observational) methodologies might be helpful, e.g., Portell et al. 2015.

3.1.4 Criterion 4: Research population

Specify who were present in the observation and whose behaviours were coded (i.e., sample), including whether the observation included individuals, dyads, triads, or larger groups. Make sure to clearly distinguish between the people present in the observation and those whose behaviours were coded. For example, a recorded medical consultation could involve a physician, nurse, patient, and informal caregiver, whereas the coding focuses exclusively on physician behaviour.

3.1.5 Criterion 5: Acquisition of source data

Describe how the source data were collected (e.g., video, audio, in situ, or transcription). Specifically, describe recording procedures, such as the use of hard- and software, set-up of cameras (e.g., camera angle), and related technical aspects (e.g., recording of sound). Address the level of 'naturalness' of the observation, both regarding participants' behaviour and the setting (Tunnell 1977). These aspects may vary in the degree to which behaviour was naturalistically observed *vs.* experimentally manipulated and observed in a naturalistic environment *vs.* controlled (laboratory) setting. Elaborate on the level of researcher participation and corresponding reactivity risks, that is, whether actively monitored cameras, unmonitored cameras, participatory filming, or in situ coding were used. Also, describe whether source data were collected for the current research aim or for another purpose (i.e., primary or secondary data). In the case of secondary data, disclose the data origin, owner, and initial purpose of collection (e.g., CCTV footage of public spaces in downtown Cape Town, collated by municipality operatives for local safety initiatives), including whether and where these data may be accessed. Explain why these source data were selected, addressing the fit between the task or context and the research aims.

3.1.6 Criterion 6: Sampling within available source data

Describe the sampling approach, specifically in which designated periods or events within the available source data the concepts were coded. This may include clarifying whether data were sampled in a single or in multiple sessions. Describe whether all raw source data were used (continuous sampling, i.e., using the whole length of the observation period) or whether and how random or non-random sampling was applied. For example, describe the selection of specific time points (e.g., video-still; instantaneous sampling), time intervals or segments (interval or scan sampling), or the non-random selection of segments in which specific behaviours or situations of interest occurred. Provide a rationale for the sampling strategy in the context of a theoretical framework or overall methodological parameters of the study.

3.2 Domain 2–Properties of coding scheme

This domain covers the reporting of the coding scheme characteristics. Domain 2a includes criteria that describe the general characteristics of coding schemes. We recommend researchers to report these for all studies applying systematic coding of observed human behaviour methodologies, regardless of whether it involves the application of an

existing coding scheme, development of a new scheme, or modification of an existing scheme. Domain 2b contains additional criteria that apply specifically to newly developed or modified coding schemes. These need *not* be reported for studies that exclusively use existing coding schemes.

3.3 Domain 2a—General description of coding scheme

3.3.1 Criterion 7: Name and availability of the coding scheme

Provide the name of the coding scheme that was used. If possible, display the full coding scheme either in the manuscript itself, in supplementary materials, or as reference to its public location elsewhere (e.g., website or online repository). Additionally provide information on how and where to access the codebook (Online Appendix B). For existing coding schemes, refer to an open repository or prior paper that provides the codebook, or list the developers' contact information. Report whether the coding scheme is freely available and/or whether coder training is available and/or required.

3.3.2 Criterion 8: Rationale for use of this specific coding scheme

Explain why this specific coding scheme was used. Specifically, describe the theoretical and empirical framework underpinning the original coding scheme, such as its research population parameters, or the cultural and historical context in which it was developed. Reflect on any relevant differences between the original and the current study, for example, in terms of context, sample, or research procedures. Discuss transferability if a coding scheme was used outside of the originally validated (geographical, cultural) context or research population. For example, if a coding scheme developed for children was used to study adults, reflect on its applicability in this population.

3.3.3 Criterion 9: Validity and reliability of the coding scheme as established in previous research

For studies using an existing or modified coding scheme, provide brief information on any previous research that illustrates its validity and reliability (Mokkink et al. 2012). Different types of validity and reliability may be more or less relevant to report, depending on the context or study characteristics. For example, if a coding scheme is based on a reflective model, –where multiple sub-codes jointly reflect or form one measure– assessing structural validity and internal consistency is crucial. Examples of types of validity that may be relevant include:

- *Content validity*: for example, clinical and teaching experts' evaluation of the degree to which a 'patient-centred communication' coding scheme reflects the intended behaviour;

- *Construct validity*: for example, whether observationally coded child compliance positively relates to parent-reported child compliance (convergent), and the degree to which the codes of child compliance relate to an adjacent concept, such as child aggression (divergent);
- *Structural validity*: for example, the degree to which scores on a measure to assess caregiving skills adequately reflect the dimensionality of caregiving skills;
- *Cross-cultural validity*: for example, whether the codes are equally applicable in a Dutch vs. South-African setting; and
- *Cross-contextual validity*: for example, whether the codes are equally applicable in street fights vs. robbery, general population vs. autistic children, 30 years prior vs. now.

Information on reliability may include:

- *Inter-rater reliability*: for example, to what extent two independent raters agreed on their application of codes to assess empathic communication across a set of observations;
- *Intra-rater reliability:* for example, to what extent an individual coder's application of codes was consistent over a large dataset of 80 observations, conducted over 3 months; and
- *Internal consistency:* for example, degree of interrelatedness across sub-codes within a measure to assess teachers' behaviours to activate students.

For reporting on reliability and validity regarding newly developed or modified coding schemes, please refer to criteria 13, 14, and 15.

3.3.4 Criterion 10: Specification of codes

a. Description of codes. Name and globally describe each code and sub-code. For complex or extensive coding schemes or detailed descriptions of the codes, refer to the codebook or supplemental materials. If a code contains different values, outcomes, or answer categories, provide those too. For example, the code 'level of maternal warmth towards the child' could consist of a scale ranging from 1 (low) to 5 (high). Consider the use of examples, anchor points, video stills, or pictures to enhance your description.

b. Connections between codes. Describe the interrelatedness of the different codes. Specifically, address hierarchies between codes: distinguish overarching from more specific sub-codes. Discuss mutual exclusivity of codes, that is, the extent to which multiple (sub-) codes from the same or different categories could be assigned to the same coding unit. Describe whether codes were exhaustively or selectively applied, that is, whether all observed behaviours were coded or whether it was possible that none of the codes described the observations in particular moments. Finally, if applicable, describe rules for the prioritisation of codes. Specifically, describe the selection procedure and rationale for situations in which multiple codes were applicable, but only one code could be assigned. To enhance comprehensibility, consider including a flowchart or schematic diagram to support the description.

c. Granularity of codes. Specify the coding unit, i.e., the level of specificity at which codes were assigned (e.g., micro or macro level). Particularly, describe the use of fine-grained or micro level coding (e.g., applying codes to specific elements within interaction(s) or recording(s)) vs. global coding (applying codes to entire interaction(s) or

recording(s)). When applying 'fine-grained' coding, describe whether codes were assigned to behaviours that occurred at specific time points (point coding, time stamped coding), over some amount of time (state coding, duration coding), or a combination. Describe whether temporality was considered, i.e., whether the sequencing or exact timing of behaviours was recorded (e.g., to establish latency to behaviour or sequential processes). When applying global coding, specify the approach to evaluating the full interaction (e.g., 'what was the average affection shown between mother and child during their entire interaction', 'did shouting take place at any time during the conversation'). Attributes of individuals that endure over the course of the interaction (e.g., gender), should be considered as global codes.

3.4 Domain 2b–General description of newly developed or modified coding scheme

For newly developed or modified coding schemes, we recommend that researchers report both Domains 2a and 2b. Some modifications to existing coding schemes may require all criteria outlined in Domain 2b, while others may only require some. The reporting required will depend on the type and extent of modification as well as on the research question. For extensive modifications (e.g., addition of multiple codes, alteration/removal of an entire dimension), all criteria outlined in Domain 2b should be reported. In any case, researchers should be transparent about any modifications they make (regardless of perceived significance) and reflect on whether these changes warrant further validity and reliability testing.

3.4.1 Criterion 11: Justification for coding scheme modification or development

Describe why a coding scheme was modified or developed and why existing coding schemes could not be used in their original form. Potential reasons could include: no existing appropriate coding scheme was available, existing coding schemes did not fit the current research question, or existing schemes were of insufficient quality. If this conclusion was based on any applicability or pilot testing, describe this process, and substantiate why the existing coding scheme was not deemed fit for purpose (in supplementary materials, if necessary). Finally, describe the theoretical and empirical framework behind the newly developed or modified coding scheme (see criterion 8).

3.4.2 Criterion 12: Modification or development process

Describe the steps in the modification or development process of the coding scheme. Specifically, indicate which people were involved, and how and when in the process. Describe how decisions about the codes and coding unit(s) were made. Specify which modifications were made throughout the development process (for new coding schemes), or to the existing scheme (for modified coding schemes). A schematic figure or table may help visualise the iteration process. If the newly developed or modified coding scheme was tailored to the study and not necessarily intended for more general use, then it is sufficient to incorporate all information about the development of the scheme within the manuscript or supplementary materials. If the coding scheme is intended for future use by other researchers, then it is recommended that the description of development, reliability, and validity testing is

reported more elaborately in a separate methods paper (e.g., Choenni et al. 2022) or in supplementary materials.

3.4.3 Criterion 13: Validity assessment for the newly developed or modified coding scheme

Describe the steps followed during the modification or development of the coding scheme to assess validity, and report on the validity of its application. Describe which types of validity were assessed for the current study, how, and why (see criterion 9 for various types of validity to report and examples).

3.4.4 Criterion 14: Reliability assessment of the newly developed or modified coding scheme.

Describe the steps followed during the modification or development of the coding scheme to establish reliability. Specify which types of reliability were assessed (see criterion 9). Report and substantiate the percentage of data that were double coded to calculate inter-rater or intra-rater reliability. Elaborate on the statistical analysis of the reliability indicators. Address which coders were involved in establishing reliability in the modification or development phase, and report their roles (e.g., coding scheme developers, students). Specify at which levels inter-rater and intra-rater reliability were calculated, that is, for individual codes, aggregated groups of codes, subscales, across measurement time-points, or across interaction type. The relationship between the specificity level at which reliability was established and the specificity level of the variables used in the analyses should always be clear. Report reliability for each measurement time point in case of repeated observations. Consider reporting reliability separately for different subgroups or conditions when relevant. Note that assessment of reliability should be maintained during the coding process (see criterion 17).

3.5 Domain 3 Application of coding scheme

Domain 3 covers the reporting of how the coding scheme was applied in the study, including a description of coding procedures and software, ethical considerations, and analysis.

3.5.1 Criterion 15: Coder selection and training

Describe the process of coder training to apply the coding scheme to the source data in relation to the study question. Describe the nature of the training itself, including trainers, duration, frequency, and any post-training support (e.g., consensus sessions). Outline which observations (e.g., from the current study or from previously collected data) and how many were used to align the coders. Describe how inter-rater reliability was established prior to application of the coding scheme. Also, describe any relevant information related to the selection, eligibility, and/or suitability of coders, for example, educational background, skills, vetting or security clearance levels.

3.5.2 Criterion 16: Coding procedures

Provide a comprehensive description of how and by whom the codes were applied to the source data. Specifically, clarify who prepared the data for the coding process and how (e.g., selecting relevant segments for coding, anonymising, allocating data to coders). Report whether coding was done in vivo, from video, from transcription, or some combination thereof. Clarify how many coders were involved and how much of the data were coded by each coder (e.g., were repeated data of the same person coded by the same coder, were both members of a dyad in dyadic interactions coded by the same coder). Report whether the subset of data used for coder training purposes or to establish initial reliability of newly developed or modified schemes was recoded or included in a later stage of the project, and whether these subsets of data were included in the final analyses.

3.5.3 Criterion 17: Reliability preservation throughout coding scheme application

Describe the steps taken to maintain reliability throughout the coding process. Describe any procedures followed to minimise coder drift (Online Appendix B) and the reliability outcomes achieved when applying the coding scheme to source data. Particularly, describe how the coding process evolved over time, frequency of consensus meetings to enhance coder alignment, and what was discussed at these meetings. Clarify how differences and the outcomes of any consensus meetings were managed. Report whether and how coder notes were incorporated in subsequent coding. Also report which types of reliability were assessed (see criterion 8) and elaborate on the statistical assessment of these indicators. Report and substantiate the percentage of data that were coded for reliability. Address the timing of the reliability assessments, such as whether reliability was assessed at one or multiple time points throughout the project and specify timings. Describe the level of specificity at which reliability was assessed (e.g., for individual codes or aggregated groups of codes). Report reliability for each measurement time-point in case of repeated observations. Consider reporting reliability separately for different subgroups or conditions when relevant.

3.5.4 Criterion 18: Ethical considerations

Describe ethical considerations specific to the observations undertaken (Levine et al. 2023), in addition to reporting more generally relevant ethical considerations. Discuss potential risk of distress or harm to participants and to researchers viewing the data, and the protocols and procedures to mitigate this (e.g., protocols to handle coder distress). Examples of relevant ethical considerations are:

- How any sensitive information (e.g., personal, ethical, legal) yielded by the observations was handled, as such information may be particularly difficult to anonymise during systematic coding of observed human behaviour;
- The nature of obtaining informed consent of those captured in the observations, participants' rights to withdraw or view data, proportionality. For example, informed consent for video recordings may have specified that participants can view recordings only after permission from all parties;

- Whether coders were also involved in collecting the observational data or had access to background information on the observed participants (e.g., blinded or non-blinded coding). For example, in a study where parents with psychiatric illness are compared to healthy controls on caregiving skills, coders could be blind to diagnostic information to objectively code caregiving skills; and
- How the researchers handled security of private and/or public storage of data, including non-anonymised recordings, and protocols to delete data.

3.5.5 Criterion 19: Facilitating software

Describe which software was used to view, code, and analyse the data. Include a brief description of how it was used—for example, only to code directly from recordings, or additionally to calculate inter-rater reliability or descriptive statistics.

3.5.6 Criterion 20: Data preparation and analysis

Describe how the coded data (Online AppendixB) were prepared for the analysis and which descriptive and statistical tests were applied. Include the steps prior to formal analysis, specifically, procedures for data cleaning, transformation of raw codes for analysis, and potential (dis)aggregation of codes. If sequential coding was applied, describe how sequential analysis (i.e., examination of the relation between antecedent and target behaviours) was approached (Bakeman and Quera 2011). In describing statistical or descriptive tests applied to the data, be transparent about any coded elements or data that were *not* used in analysis and explain why this was the case (e.g., coding was not reliable; source data were unusable, withdrawal of participants). If ethically possible, provide anonymised or sanitised data and analysis scripts, to enable others to replicate the work.

4 Discussion

Using Delphi methodology, we developed international and interdisciplinary guidelines for reporting about research involving systematic coding of observed human behaviour (SCOBe), to enhance the visibility and secondary use of coding schemes. Adherence to these reporting guidelines is essential to the quality, comparability, and replicability of research by increasing general awareness of key methodological issues. We created these guidelines and a unified terminology based on the expertise and approaches available in various fields. Thereby, the guidelines constitute the combined strength of best practices across multiple disciplines.

Various disciplines have their own research traditions and terminology regarding SCOBe methodology. Whereas some research areas rely predominantly on a body of established and extensively validated coding schemes, others tend to develop specifically tailored coding schemes for each new research question. Data acquisition also varies across disciplines, ranging from working with existing footage (e.g., CCTV footage) to collecting recordings in a controlled or experimentally manipulated setting. Moreover, coding techniques range from very broad to highly specific. Some, but not all, of these differences can be explained through variation between disciplines in the extent and duration in which SCOBe methodology has been used.

Through involving a wide range of disciplines, our guidelines are specific enough to provide researchers with concrete tools for optimal reporting of studies using systematic coding of observed human behaviour, while being sufficiently broad to fit with research practices across disciplines. The guidelines enable increasing inter-disciplinarity by facilitating and simplifying multi- and inter- disciplinary collaborations. With this integrative focus, our endeavour is in line with the current international trend of increasing interdisciplinary research. Many societal problems are complex and therefore require such interdisciplinary solutions. Research collaborations that cross and combine disciplinary and thematic boundaries are better equipped to deal with this global reality (Wilthagen et al. 2018), and such collaborations will benefit from the guidelines and terminology we have developed.

The guidelines can assist researchers, reviewers, and editors in ensuring that sufficient information is provided in any SCOBe research report, thereby ensuring transparency and replicability. All this is essential to enable the scientific community in critically appraising and evaluating studies using systematic coding of observed human behaviour. Ultimately, this may also improve the quality of such research. However, our guidelines do not describe procedures for *conducting* research, nor do they represent indicators for quality of the underlying scholarship. Rigorous reporting is essential to high-quality, yet not sufficient to conduct meaningful research (Kearney 2014). We realise that, in addition to our reporting guidelines, there is also a need for an interdisciplinary how-to guide on how to design and conduct SCOBe research. Therefore, our group will work on this in a follow-up collaborative project. In the meantime, researchers can already use our reporting guidelines when setting up studies, as a reminder of which components they will later be expected to describe. Additionally, we refer interested readers to general and discipline-specific guidance and best practices that have been described elsewhere, which are summarised in Online Appendix C.

We acknowledge that SCOBe is an evolving approach. As such, our reporting guidelines may not prove exhaustive nor immutable. For example, the exponential advances in machine learning and AI will likely result in increased use of automated video coding procedures (Bernasco et al. 2021). Moreover, the development and use of hardware (e.g., smartphones) that enables audio and video recording is rapidly growing, as is dissemination of data through social media platforms. As a result, the proliferation of the kinds of data that are amenable to SCOBe is inevitable. To cater to this vastly changing landscape, we created an openly accessible directory at https://osf.io/tdnmq/?view_only=f8f504573f8645b3999c0bace76ca96c. We invite any interested researchers to further discuss, comment on and update our guidelines.

While this study addresses a clear gap in both the methodological and substantive disciplinary literature, some limitations deserve acknowledgement. Most importantly, the selection of the expert participants through the network of the organisers resulted in an overrepresentation of scholars from the Netherlands. Also, the variety of experts included in the Delphi process may have introduced biases in the communication process due to differences in authority, personality, or reputation and corresponding issues of hierarchy. We aimed to reduce these problems and maximise the group's potential by structuring the information flow, varying group compositions, and explicitly inviting differing opinions while encouraging open communication and constructive critique.

Despite these limitations, we believe we have succeeded in developing an interdisciplinary organizing framework, building upon and significantly enriching preliminary monodisciplinary guidelines. These newly developed guidelines can assist in advancing and appraising the methodological rigor of systematic observational research. We hope these

guidelines will help shape future work involving systematic coding of observed human behaviour and encourage researchers to adopt this approach in their interdisciplinary collaborations.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11135-024-01969-9.

Author contributions Study conception, acquisition of funding for the workshop, and material preparation were performed by Marij Hillen, Evelien Hoeben, Calum McHale, Melissa Sexton, and Janneke Noordman. All authors contributed to the study design, conceptualization, and development of reporting criteria. The first draft of the manuscript was written by all authors collaboratively, led by Marij Hillen, Evelien Hoeben, Rianne Kok, Calum McHale, Melissa Sexton, and Janneke Noordman. All authors commented on previous versions of the manuscript, read and approved the final manuscript.

Funding Funding for this work was obtained from the Lorentz Center (Leiden, the Netherlands; Grant number 22537), specifically to organize the in-person and online workshop days. Lorentz Center had no influence on the workshop focus or manuscript contents.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Bakeman, R., Quera, V.: Sequential analysis and observational methods for the behavioral sciences. Cambridge University Press. (2011)

Bakeman, R., Deckner, D. F., Quera, V.: Analysis of behavioral streams. In: G. Adams & M. Berzonsky (Eds.) Handbook of research methods in developmental science, p. 394–420. Blackwell Publishing. (2008)

Bensing, J.M., Tromp, F., Van Dulmen, S., Van Den Brink-Muinen, A., Verheul, W., Schellevis, F.G.: Shifts in doctor-patient communication between 1986 and 2002: a study of videotaped general practice consultations with hypertension patients. BMC Fam. Pract. **7**, 1–7 (2006)

Bernasco, W., M. Hoeben, E., Koelma, D., Liebst, L. S., Thomas, J., Appelman, J., Snoek, C. G., Lindegaard, M. R.: Promise into practice: Application of computer vision in empirical research on social distancing. Sociol. Methods Res., 00491241221099554 (2021)

Biringen, Z., Robinson, J.L., Emde, R.N.: Appendix B: the emotional availability scales (an abridged infancy/early childhood version). Attach. Hum. Dev. **2**(2), 256–270 (2000)

Blanch-Hartigan, D., Ruben, M.A., Hall, J.A., Mast, M.S.: Measuring nonverbal behavior in clinical interactions: a pragmatic guide. Pat. Educ. Couns. **101**(12), 2209–2218 (2018)

Castañer, M., Barreira, D., Camerino, O., Anguera, M.T., Fernandes, T., Hileno, R.: Mastery in goal scoring, T-pattern detection, and polar coordinate analysis of motor skills used by Lionel Messi and Cristiano Ronaldo. Front. Psychol. **8**, 741 (2017)

Choenni, V., Kok, R., Verhulst, F. C., van Lier, M. H., Lambregtse-van den Berg, M. P: The Dutch Infant Caregiving Assessment Scales: Psychometric properties in mothers with and without a severe psychiatric disorder. Int. J. Methods Psychiatr. Res. 31(2):e1902 (2022)

Chorney, J.M., McMurtry, C.M., Chambers, C.T., Bakeman, R.: Developing and modifying behavioral coding schemes in pediatric psychology: a practical guide. L Pediatr. Psychol. **40**(1), 154–164 (2015)

Ejbye-Ernst, P., Lindegaard, M.R., Bernasco, W.: How to stop a fight—A qualitative video analysis of how third-parties de-escalate real-life interpersonal conflicts in public. Psychol. Violence **12**(2), 84–94 (2022)
Gardner, F.: Observational Studies of Parent-Child Interaction and Behaviour Problems: Their Implications for Parenting Interventions. In A. Buchanan, B. L. Hudson (Eds.): Parenting, Schooling and Children's Behaviour, pp. 53–72. Routledge (2018)
Harris, F.C., Lahey, B.B.: Subject reactivity in direct observational assessment: a review and critical analysis. Clin. Psychol. Rev. **2**(4), 523–538 (1982)
Heyman, R. E., Lorber, M. F., Eddy, J. M., West, T. V.: Behavioral observation and coding. In: H. T. Reis, C. M. Judd (Eds.): Handbook of research methods in social and personality psychology, pp. 345–372. Cambridge University Press (2014)
Hiller, R.M., Meiser-Stedman, R., Lobo, S., Creswell, C., Fearon, P., Ehlers, A., Murray, L., Halligan, S.L.: A longitudinal investigation of the role of parental responses in predicting children's post-traumatic distress. J. Child Psychol. Psychiatr. **59**(7), 781–789 (2018)
Hoeben, E.M., Bernasco, W., Suonperä Liebst, L., Van Baak, C., Rosenkrantz Lindegaard, M.: Social distancing compliance: a video observational analysis. PLoS ONE **16**(3), e0248221 (2021)
Jakubauskaite, V.: Development of an observational coding scheme for parental behaviour in a play-based instructional context with 5-to 7-year-olds. [Doctoral Thesis, University of Kent]. ProQuest Dissertation Publishing (2021)
Jones, N., Raymond, G.: "The camera rolls" using third-party video in field research. Ann. Am. Acad. Pol. Soc. Sci. **642**(1), 109–123 (2012)
Kearney, M.H.: Hoping for a TREND toward PRISMA: the variety and value of research reporting guidelines. Res. Nurs. Health **37**(2), 85–87 (2014)
Levine, M., Philpot, R., Nightingale, S., Kordoni, A.: Visual Digital Data, Ethical Challenges and Psychological Science. Am. Psychol. (2023)
Lindegaard, M.R., Bernasco, W.: Lessons learned from crime caught on camera. J. Res. Crime Delinq. **55**(1), 155–186 (2018)
Menichetti, J., Henselmans, I., Noordman, J., Landmark, A. M., Gerwing, J., Hillen, M., Pieterse, A.: Methods for analyzing actual communication: A shared decision making consultation under multiple methodological lenses. Pat. Educ. Couns. (2021)
Moher, D., Schulz, K.F., Simera, I., Altman, D.G.: Guidance for developers of health research reporting guidelines. PLoS Med. **7**(2), e1000217 (2010)
Mokkink, L.B., Terwee, C.B., Patrick, D.L., Alonso, J., Stratford, P.W., Knol, D.L., Bouter, L.M., de Vet, H.C.: COSMIN checklist manual. University Medical Center, Amsterdam (2012)
Niederberger, M., Köberich, S., Network, D.: Coming to consensus: the Delphi technique. Oxford University Press (2021)
Philpot, R., Liebst, L.S., Møller, K.K., Lindegaard, M.R., Levine, M.: Capturing violence in the night-time economy: a review of established and emerging methodologies. Aggress. Violent Behav. **46**, 56–65 (2019)
Philpot, R., Levine, M., Acre-Plata, C., Elphick, C., Zhang, M., Stuart, A., Walkington, Z., Pike, G., Frumkin, L., Popple, D.: No Procedural Justice, No Peace? Judgements of Police Legitimacy in 'Real-time'Interactions Captured on Camera-Registered Report Stage 1. R. Soc. Open Sci. (2021)
Portell, M., Anguera Argilaga, M.T., Chacón Moscoso, S., Sanduvete Chaves, S.: Guidelines for reportingevaluations based on observational methodology. Psicothema **27**(3), 283–289 (2015)
Powell, C.: The Delphi technique: myths and realities. J. Adv. Nurs. **41**(4), 376–382 (2003)
Terroba, M., Ribera, J.M., Lapresa, D., Anguera, M.T.: Observational analysis of the development of computational thinking in early childhood education (5 years old) through an intervention proposal with a ground robot of programmed directionality. Eur. Early Child. Educ. Res. J. **30**(3), 437–455 (2022)
Tschan, F., Zimmermann, J., Semmer, N. K.: Rules for coding scheme development. In: E. Brauner, M. Boos, M. Kolbe (Eds.): The Cambridge handbook of group interaction analysis, p. 191–207. Cambridge University Press (2018)
Tunnell, G.B.: Three dimensions of naturalness: an expanded definition of field research. Psychol. Bull. **84**(3), 426 (1977)
Van der Giessen, D., Hollenstein, T., Hale, W.W., III., Koot, H.M., Meeus, W., Branje, S.: Emotional variability in mother-adolescent conflict interactions and internalizing problems of mothers and adolescents: dyadic and individual processes. J. Abnorm. Child Psychol. **43**(2), 339–353 (2015)
van Dulmen, S., Humphris, G., Eide, H.: Towards a guideline for person-centered research in clinical communication; lessons learned from three countries. Int. J. Pers. Cent. Med. **2**(1), 58–63 (2012)
Wilthagen, T., Aarts, E. H., Valcke, P.: A time for interdisciplinarity-an essay on the added value of collaboration for science, university, and society. (2018)
Yoder, P., Symons, F.: Observational measurement of behavior. Springer Publishing Company (2010)

Zimmermann, C., Del Piccolo, L., Bensing, J., Bergvik, S., De Haes, H., Eide, H., Fletcher, I., Goss, C., Heaven, C., Humphris, G.: Coding patient emotional cues and concerns in medical consultations: the Verona coding definitions of emotional sequences (VR-CoDES). Pat. Educ. Couns. **82**(2), 141–148 (2011)

Authors and Affiliations

Marij A. Hillen[1] · Evelien M. Hoeben[2] · Rianne Kok[3] · Calum McHale[4] · Melissa M. Sexton[5] · Sabine van der Asdonk[6] · Daphne van de Bongardt[3] · Brett Bowman[7] · Richard Brown[8] · Susan Branje[9] · Jill Chorney[10] · Monika H. Donker[9] · Peter Ejbye-Ernst[2] · Sanne Geeraerts[9] · Danielle Van der Giessen[11] · Gerald Humphris[4] · Helle Larsen[12] · Richard Philpot[13] · Mariona Portell[14] · Siyang Yuan[15] · Chiel van der Veen[16] · Janneke Noordman[17]

✉ Marij A. Hillen
m.a.hillen@amsterdamumc.nl

1 Department of Medical Psychology, Amsterdam Public Health, Amsterdam UMC, Amsterdam, the Netherlands

2 Netherlands Institute for the Study of Crime and Law Enforcement, Amsterdam, the Netherlands

3 Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, Rotterdam, the Netherlands

4 School of Medicine, University of St Andrews, St Andrews, UK

5 School of Business and Economics, KIN Center for Digital Innovation, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

6 Institute of Education and Child Studies, Leiden University, Leiden, the Netherlands

7 Department of Psychology, University of the Witwatersrand, Johannesburg, South Africa

8 Department of Health Behavior and Policy, Virginia Commonwealth University School of Population Health, Richmond, USA

9 Department of Youth and Family, Utrecht University, Utrecht, the Netherlands

10 Department of Psychiatry, Dalhousie University, Halifax, Canada

11 Department of Child Development and Education, University of Amsterdam, Amsterdam, the Netherlands

12 Developmental Psychology, University of Amsterdam, Amsterdam, the Netherlands

13 Psychology Department, Fylde College, Lancaster University, Lancaster LA1 4YF, UK

14 Department of Psychobiology and Methodology of Health Sciences, Autonomous University of Barcelona, Cerdanyola del Vallès, Spain

15 School of Dentistry, University of Dundee, Dundee, UK

16 Windesheim University of Applied Sciences, Almere, the Netherlands

17 Nivel (Netherlands Institute for Health Services Research), Utrecht, the Netherlands