Judgment and Decision Making (2025), **vol. 20**:e39 1–22
doi:10.1017/jdm.2025.10010

THEORY ARTICLE

'Please explain your response': A guide to uncovering cognitive processes from open-text box data using pragmatic and reflexive content analysis

Stephen H. Dewitt[1], Alice Liefgreen[2], Nine Adler[1] and Laura Elaine Strittmatter[3]

[1]Department of Experimental Psychology, University College London, 26 Bedford Way, WC1H 0AP; [2]Department of Language and Cognition, University College London, 2 Wakefield St, WC1N 1PF and [3]ETH Zürich, Department of Management, Technology, and Economics, Weinbergstrasse 56/58, 8092 Zürich, Switzerland

Corresponding author: Stephen Dewitt; Email: dewitt.s.h@gmail.com

Received: 26 September 2024; **Revised:** 26 June 2025; **Accepted:** 4 August 2025

Keywords: content analysis; reflexivity; mixed methods; open-text box; survey

Abstract

This guide provides a philosophical framework and practical advice for gathering, analyzing, and reporting a particular type of qualitative data. These data are obtained from including an open-text box following the key quantitative question in survey-style studies with the request to 'Please explain your response'. While many studies currently collect such data, they often either fail to report or analyze it, or they conduct unstructured analyses with limited detail, often mistakenly referring to it as 'thematic analysis'. Content analysis provides a well-established framework for analyzing such data, and the simplicity of the data form allows for a highly pragmatic and flexible approach. The guide integrates the concept of reflexivity from qualitative research to navigate the large number of researcher degrees of freedom involved in the process, particularly in working with the second coder. It begins by arguing for the value of this data, before outlining the guide's philosophy, offering advice on maximizing the validity of your data, and addressing the common concern of confabulation. It then provides advice on developing a coding scheme, recruiting and collaborating with a second coder, and writing your report, considering the potential role of large language models at these various stages. Additionally, it provides a checklist for reviewers to evaluate the quality of a given analysis. Throughout the guide, a running example is used to demonstrate the implementation of the provided advice, accompanied by extensive example materials in the online repository, which can be used to practice the method.

1. Why add open-text boxes?

> . . .open-ended responses often capture the 'why' that complements quantitative results, helping to tell a more nuanced story with the data.
>
> —Rouder et al., 2021, pp.1

This guide advocates for an exceedingly simple 'bolt on' research method for judgment and decision making (JDM) and related disciplines: in any experimental online survey study (Reips, 2021), include an open-text box following your key quantitative question, asking the participant to explain their response (Figure 1). This addition provides you with written qualitative data regarding the participant's cognitive processes that led to their quantitative response. Within JDM, there is a

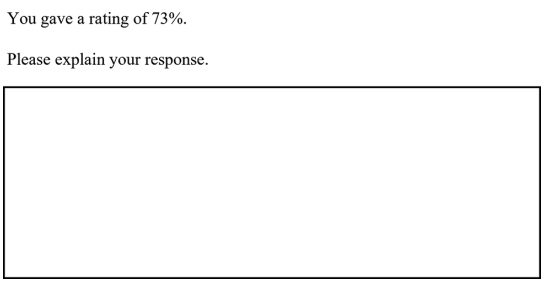

Figure 1. *An example of a simple open-text box asking the participant to explain their previous quantitative response. As the open-text box should ideally be on a separate page, it can be a good idea to use a 'piping' function to remind the participant of their quantitative response.*

A case study in the value of open text data (Dewitt et al., 2019)
The first author's initial experience with incorporating this kind of data into experimental surveys provides a neat illustration of its potential value. He was working within the natural frequency / nested sets debate (e.g. Gigerenzer & Hoffrage, 1995; Hoffrage et al., 2002; Macchi, 2000) regarding which framings help people solve simple Bayesian word problems more accurately, and why. Within a complex debate largely about internal cognitive processes one claim by the natural frequencies side was that the nested sets approach only worked because it encouraged people to convert the problem from its original percentage format to a frequency format, and that it was this conversion that was the ultimate reason for the greater accuracy seen. Both sides of this debate had produced their own papers which they believed supported their position but were dismissed by the other as confounded in some way. The first author spent a considerable amount of time trying and failing to construct a purely quantitative experiment which would definitively provide a ruling on this issue. Eventually, inspired by some of the work in this field, he added a 'Please explain your response' text box after the quantitative question. Pleasingly, many participants clearly reported converting the percentages in the problems to frequencies before solution, even though it did not affect the mathematics (e.g. from 17% to 17/100). For example, P105 said 'To make my math easier, I am going to assume there are 100 women...' Furthermore, by coding these participants as 'conversion to frequencies' we established that this was more common in the nested sets condition than the control condition, and that it was also associated with success on the problem. While not perhaps definitive, this provided some support for the natural frequencies claim.

Figure 2. *A textbox providing a case study in the value of open-text data.*

stronger history of collecting spoken word data, for example, 'think aloud' (e.g., Cokely & Kelley, 2009; Schulte-Mecklenbeck et al., 2011; Schulte-Mecklenbeck et al., 2019). This is unsurprising given that the majority of that history has involved physical in-person experiments. However, now that a large amount of research is conducted via online surveys, collecting think aloud data is more challenging. Simultaneously, online surveys lend themselves very well to collecting typed text data, which requires no transcription from audio or digitization from, for example, hand-written notes (Decorte et al., 2019)[1]. This kind of data is therefore exceptionally easy and low-cost to collect and prepare for analysis in the kind of experiments routinely being conducted by modern JDM researchers (Singer & Couper, 2017). A case study illustrating the potential value of this data can be seen in Figure 2.

All four authors are JDM researchers who received a typical scientific/quantitative education but have come to see the complementary value of collecting and analyzing this kind of data for the difficult challenge we often face of developing and testing theories about unobservable cognitive processes

[1] And also allows the potential collection of meta data such as time spent typing etc.

(Glöckner & Betsch, 2011; Oberauer & Lewandowsky, 2019; Weiss & Shanteau, 2021). Over the past ten years, the authors have encountered many researchers, across various levels of seniority, who have, at one time or another, collected such data and found it personally insightful in understanding participants' cognitive processes. It seems to be frequently collected during piloting, where researchers use it to understand how their participants are experiencing their stimuli and questions. However, it is then typically either not collected for the main paper or collected but not analyzed or reported (Rouder et al., 2021). This seems to be because, having been trained within a quantitative paradigm, many researchers struggle with how to analyze the data in a way that satisfies their scientific principles, and how to present any insights drawn from this data effectively (Decorte et al., 2019). When it is presented, as Rouder et al. lament, 'approaches may be perfunctory at best—such as selecting a few quotes to exhibit, or word clouds' (Rouder et al., 2021, pp. 2). The aim of the present paper is to provide a comprehensive guide to analyzing and presenting this kind of data, both philosophically and practically.

The analytic approach advocated here is straightforward, using a simple and pragmatic form of content analysis (Bengtsson, 2016; Hsieh & Shannon, 2005). There are many forms of content analysis used across a variety of fields, with a large amount of inconsistency in almost all aspects (Nicmanis, 2024). However, core elements involve the creation of categories called 'codes' (e.g., Saldaña, 2021), the sorting of participants' responses into these categories, and the validation of that categorization by a second human coder, including the calculation of inter-coder agreement. In some forms of content analysis, text data are broken down further into 'meaning units' before coding; however, as open-text data of the type advocated here is already quite short (typically one or two sentences), we have never felt the need to do this, and this simplifies the analysis considerably. The approach advocated here is also relatively interpretative/latent, moving beyond the surface-level text to infer cognitive processes, and relies on reflexivity and inter-coder agreement to combat issues of researcher bias. Despite some degree of 'reading between the lines', because the data are typically short and concerning a relatively 'contained' psychological process (thought processes across a short task), we can usually achieve high levels of inter-coder agreement. While this is a simple concept, there are many 'researcher degrees of freedom' (Wicherts et al., 2016) in this process, in particular when working with a second coder. The present paper provides a large amount of guidance on navigating these to minimize the impact of researcher bias and guide advocates for a highly open and transparent approach, which sits well with current open science principles.

2. Reflexivity

While content analysis provides the basic analytical approach, this guide borrows the concept of reflexivity from the qualitative world as its deeper guiding framework (e.g., Olmos-Vega et al., 2023). This approach to research was recently promoted as a framework for quantitative work by Jamieson et al. (2023), which we wholly recommend reading in combination with this paper.

Although the term 'reflexivity' is not well-known within quantitative science, the concept aligns well with traditional scientific values and all scientists already practice it to some extent without labelling it as such (e.g., Glockner et al., 2024). However, we believe it should be taught more explicitly, in line with Jamieson et al. (2023). The late physicist Richard Feynman (Feynman et al., 2007) shared this perspective:

> That is the idea that we all hope you have learned in studying science in school—we never explicitly say what this is, but just hope that you catch on by all the examples of scientific investigation. It is interesting, therefore, to bring it out now and speak of it explicitly. It's a kind of scientific integrity, a principle of scientific thought that corresponds to a kind of utter honesty. . . The first principle is that you must not fool yourself—and you are the easiest person to fool. So, you have to be very careful about that. (Richard Feynman, Surely You're Joking, Mr. Feynman!: Adventures of a Curious Character)

At first, Feynman discusses integrity, but he then moves on to personal, internal honesty, which closely aligns with the concept of 'reflexivity'. Reflexivity is a metacognitive skill that involves maintaining awareness of the personal baggage that you bring to the research process and its impact on your decisions. In the language of quantitative science, it involves being aware of the choices you make at each 'researcher degree of freedom' (or, as you navigate the 'garden of forking paths'), and crucially, understanding why you make those choices. Jamieson et al. (2023), as personality and social psychologists, focus on topics such as gender and race, making it easier to grasp this concept. When researching these topics, it is likely obvious how our own gender or race could influence the research process. Our personal motivations and desires about what we may want to be true about society or human psychology are likely to influence our entire study design and implementation (Haidt & Jussim, 2016; Jussim et al., 2015). However, a more ever-present influence, regardless of discipline, is our desire for (or perhaps just expectation of) particular results (usually significant[2] if defending a theory, nonsignificant if attacking) due to the need to publish to facilitate career progression or the desire to support or attack previously proposed theories (Bishop, 2019).

Recently, the importance of the baggage researchers bring into their research has also been recognized in the area of applied behavioral science. Hallsworth (2023), in a section titled 'No "view from nowhere"' wrote the following:

> Taking such a [view from nowhere] stance may not be possible for behavioural scientists. We bring certain assumptions and ways of seeing to what we do; we are always situated in, embedded in and entangled with ideas and situations. We cannot assume that there is some set-aside position from which to observe the behaviour of others; no objective observation deck outside society exists. (Hallsworth, 2023, pp. 317).

Integrating this with the quantitative concept of researcher degrees of freedom, these moments should become a habitual cue to 'switch on' our reflexive awareness. This means asking ourselves what choice that part of us that desires a particular result wants us to make, and why. Does that part suspect that this path is more likely to produce the desired outcome? Why do we want that result? Now, the goal is not just to observe these impulses, but to resist them and choose the path that is, in our best judgment, the right one to follow. If young researchers need more motivation to take the 'right' path, we encourage them to think long term. The pressures to publish may make you want a certain result right now, but consider in a couple of decades the possibility of large amounts of your career's work being found to be false, or worse, accused of falsification (e.g., O'Grady, 2021).

3. Maximizing data validity

> No one, save the strictest of behaviourists, argues that verbal reports are completely invalid. Nor does anyone argue that verbal reports provide a perfect window into the mind.
>
> Wilson, 1994.

In this section, we will consider the validity of open-text data (i.e., how accurately it represents the participants' real cognitive processes while undertaking your study) and how we can increase this through our study design. This guide focuses on typed open-text box data rather than 'verbal report' data, yet Wilson's (1994) balanced position quoted above still applies: such data carries valuable information not easily obtained from quantitative data but must be analyzed with an appropriate dose of skepticism. Throughout the paper, a running example will be used to illustrate key ideas. An introduction to this example can be seen in Figure 3.

To obtain high-quality open-text data from your participants, you need them to put in effort to introspect and report the contents of their minds. This is potentially problematic as you will typically collect this kind of data at the end of your experiment. Factors that will affect participants' level

[2]Other statistical paradigms are available.

Introduction to the 'legal detainment' running example:
In Dewitt, Glatzel & Lagnado (2023a) we conducted an online survey study looking at how participants' estimates of the guilt of a defendant were affected by being told that the defendant had been detained (not granted bail) prior to the trial. Further, we wanted to know whether this potential effect of detainment depended upon whether that detainment decision was (1) based on an absolute iron clad rule for the type of crime or (2) based on an assessment of the case by a bail judge. This produced a 2 (detained vs not-detained) x 2 (explained vs unexplained) between-subjects design with a quantitative response from 0%-100% guilty. We added an open text box asking participants to explain their quantitative response.

Figure 3. *Introduction to the 'legal detainment' running example.*

of effort include how engaged they are in your study, how much time pressure they feel to finish, and their feelings toward you as the researcher. To this end, we are continually trying to make our research more engaging, using interesting scenarios, clear, simple, and minimal 'everyday' language, and images where possible to maintain interest. When piloting, we recommend also asking your pilot participants if they enjoyed the study, as engaged participants are more likely to want to share their thoughts. Furthermore, we typically slightly over-pay participants and tell them this clearly at the start of the experiment, for example, 'We have paid you for six minutes of your time when the study only takes an average of five, so please take your time to carefully answer the questions'. By combining this with a progress indicator, participants can see how much of the experiment remains versus how much time they have spent. This transparency ensures they know you have been honest or generous in your time estimate and payment. If participants feel the study has taken longer than expected, they may rush through the final open-text box, providing poor-quality data.

We strongly advise against using coercive approaches such as enforcing minimum word counts on the text box or imposing a minimum time spent on the page (although we do remind participants if they do not answer the open-text question at all). Our experience suggests that these methods are more likely to frustrate participants, setting up a feeling of opposition between them and you, and leading them to either wait out the clock or provide irrelevant content just to meet the word count requirement. What you truly need is meaningful engagement and introspection from your participants regarding their cognitive processes, and this level of effort cannot be coerced. It is far better to prioritize building rapport and goodwill with your participants to yield more genuine and insightful responses (Mayser & von Wangenheim, 2013).

3.1. Concurrent vs retrospective collection

We will now explain why we recommend collecting open-text data retrospectively, after collecting the quantitative response (and probably on a separate survey page). The more historically widely used 'think aloud' protocol involves participants verbally explaining their thought processes concurrently, that is, while they work through an experiment (Eccles & Arsal, 2017; Sela et al., 2000). The process we describe here not only asks participants to write their thought process rather than speak it, but also to do this retrospectively (albeit only seconds later), after they have already completed that part of the task and provided their quantitative response. Due to some high-profile studies (e.g., Haidt, 2007; Johansson et al., 2008; Nisbett & Wilson, 1977), there is a prevalent view within cognitive science and JDM in particular that if you ask people why they gave a particular response after the fact, they will simply 'confabulate' or make up reasons. We believe this is an erroneous leap of logic, unfortunately common in work on human biases: just because some experiments show that some people confabulate sometimes, in certain situations, it does not follow that all people will do this always, in all situations. More generally, we agree with a growing view that our discipline has exaggerated the level of bias in human reasoning (e.g., Lejarraga & Hertwig, 2021; Madsen et al., 2024; Weiss & Shanteau, 2021).

We typically conduct our experiments in contexts where participants have little reason to feel defensive or 'on the spot' (unlike the above-cited work), with minimal risk to their ego or identity if they cannot explain their reasoning. Consequently, we have seen little reason to be concerned about this issue. Many of our participants are open to admitting complete uncertainty about their choice or response (e.g., 'I don't know') or reporting intuitive responses (e.g., 'I went with my gut') without seeming to feel the need for a well-articulated answer.

However, if one wishes to study phenomena where this might be an issue, an additional layer of skepticism should be applied by both the researcher and any reviewer or reader. For example, one might face such an issue if studying decision-making by professionals such as doctors, asking them to explain their reasons for making a medical decision, especially if it led to a bad outcome. While much of their decision-making is likely to be intuitive and based on extensive experience (e.g., Hall, 2002), they may feel social pressure to provide a well-thought-out explanation. This example highlights that confabulation may, in fact, be less of an issue with written reports collected via an online survey, which could be a benefit over a verbal report approach in some cases. It is hard to imagine a lower-pressure situation than conducting an online survey at home pseudonymously for a faceless researcher (Reips, 2002).

However, there are good reasons to think that writing your thought process concurrently, that is, while solving a task, is likely to change cognitive processes, possibly more so than verbalizing. Writing is slower, and it is well known within the therapeutic world that expressing cognitive processes through methods like journaling opens up a complex range of metacognitive processes, allowing individuals to examine and reconsider their thinking (e.g., Hubbs & Brand, 2005). This might improve performance in a range of cognitive tasks. Tentatively supporting this, the first author has unpublished data showing that accuracy rates on a simple Bayesian problem solving task rose from 5.1% (n = 99) to 16.7% (n = 60) by providing brief 'think aloud' training plus a 'concurrent' open-text box (using binary logistic regression: $Z = 2.3$, $p = .02$), which required participants to explain their reasoning before being allowed to proceed to the next page to provide their numerical response. Details of this simple study, including the dataset and materials, can be found at the online repository for this paper: https://osf.io/cyjhd/. This runs in contrast to the largest review of the effect of concurrent 'think aloud' protocols, which Fox et al. (2011) found did not typically affect performance, possibly supporting a difference between verbalizing and writing here.

This is the primary reason we recommend retrospective collection of qualitative data after the quantitative response has been made: collecting it concurrently may affect your quantitative data in unpredictable ways and make it less comparable with other purely quantitative studies. In many previous studies, we have included the open-text box on the same page immediately below the quantitative response. However, Lerner and Tetlock (1999) and Ranyard and Svenson (2011) make the good point that if participants anticipate having to explain their thought process, this could still influence their reasoning (albeit likely smaller than being forced to write out their thinking beforehand). For a similar reason, they recommend not advertising to your participants that they will be providing qualitative data. However, there is a tension and trade-off here. If the open-text box, which many participants experience as an effortful task, comes as a surprise, they may be less willing to engage with it, and you may get poor-quality, low-effort responses, or they may even drop out. You are also increasing the time between quantitative collection and their explanation, potentially allowing more time for memories to fade or crystallize. We therefore do not have a definitive position on this, and your choice may depend on whether you want to minimize the possible impact on the quantitative data or maximize the quality of your open-text data. Future studies comparing both methods would be valuable, and researchers starting out with these methods may want to use both during piloting to see if there are any obvious impacts on either kind of data.

3.2. Frameworks for judging the validity of the data

Generally, it is crucial to be aware of and think deeply about the type of cognitive process you are attempting to study and the likely limitations of written, retrospective reports for this type of process.

Indicative quote from the 'legal detainment' running example:
Here is an indicative quote from the detainment study. Participants in the 'Unexplained-Detained' condition, where the defendant had been detained before trial based on a bail judge's decision, frequently made the inference that the fact that the judge detained the defendant means they must know something about them that makes them more likely to be guilty. This 'backdoor' inference about the defendant's guilt via an inference about the judge's reasons for their decision was the exact cognitive process we theorised would occur, and so seeing these articulated provided additional support to the quantitative results.

P120: "Because [the] judge decided to detain him, so there is [a] higher chance that he is guilty, maybe he said something that gave [the] judge a reason to keep him or maybe he has no alibi."

Figure 4. *Indicative quote from the 'legal detainment' running example.*

We tend to operate within a pragmatic framework broadly in line with Ericsson and Simon's (1980) model. Participants will be able to report the cognitive process of interest, if they have (1) insight into it (i.e., if it is available to their conscious mind), if they (2) put in the required effort to introspect, and if they (3) choose to be honest in reporting it. These are the three key principles we use to consider potential limitations in our data. The first author has conducted work where the cognitive process of interest simply and clearly emerged directly from the open-text data (e.g., Dewitt et al., 2019, the example in Figure 2), while in other work, participants' responses did not mention the target cognitive process (e.g., Dewitt, Adler, et al., 2023). Whether participants can report the cognitive process of interest can be informative in itself, providing clues about the nature of the process. For example, the first author concluded from that latter work that there was some hidden assumption at play, informing the design of future experiments.

Gricean principles of conversation (Grice, 1975) have also been valuable in understanding the limitations of this kind of data. Participants engaging in these open-text responses are effectively having a 'one-shot' dialogue with us as researchers, aiming to communicate what they believe are the most critical aspects of their reasoning within the constraints of time and effort they are willing to invest. Consequently, much of their thinking might remain unexpressed, and they often write only one or two sentences. In our research area, if a participant mentions a cognitive process, we typically consider it to be approximately true, unless there is a reason to suspect confabulation. However, it would be a mistake to infer that the absence of a mention implies the absence of an experience or thought. Participants might omit certain details, assuming they are self-evident or not crucial to communicate at that moment, or the cognitive process may have occurred but be less than fully conscious. This creates an asymmetry in their evidential value: they can be very powerful evidence for the presence of a cognitive process (e.g., see Figure 4) but are generally a poor form of evidence for the absence of one (although see Figure 6 for a counterexample). Concurrent collection may be less prone to this asymmetry (Kuusela & Paul, 2000).

In general, we must remain aware that open-text data, like quantitative measures (e.g., Toomela, 2008), are an indirect representation of the individual's cognitive process mediated by a communication process. As Ranyard and Svenson (2011) also remind us, the mental representations of a person engaged in a decision are likely in continuous flux during the decision process, and even after. We likely capture a relatively crystallized and 'cleaned' version of what actually happened in their mind using this method. The benefits are that these data are easy to obtain and do not compromise the validity of the quantitative data, but if one wants to fully map out those changing representations during a decision process, they may wish to turn to the concurrent 'think aloud' approach or other process tracing methods (Schulte-Mecklenbeck et al., 2019). Generally, inferences from such data must be analyzed and communicated to readers critically and tentatively, with consideration to how likely these issues are given the phenomenon and participants under study and the method used to collect the data.

4. Analysis

> . . .a researcher attempting to qualitatively analyze the data from general open questions may struggle with a lack of context, and a lack of conceptual richness typical of qualitative data, because individual responses often consist of a few words or sentences.
>
> Decorte et al. (2019, pp. 2)

The following section details a three-stage process for analyzing open-text data involving a first coder (usually the first author) and an independent second coder whose role is to provide some degree of validation of the coding scheme and coded responses. A flow chart of this process can be seen in Figure 5, and we hope it will provide a useful reminder once this section has been read and digested.

4.1. Regarding the use of large language models

Before we explain this process, which at its core involves two human beings, we are aware, we may need to address the potential involvement of large language models (LLM's). LLM's are a valuable tool for researchers (e.g., Demszky et al., 2023), including for analyzing very large amounts of text data that would be unfeasible for a human being (Bhatia et al., 2021; Feuerriegel et al., 2025). However, there is also much disagreement over the role they should have within science (Binz et al., 2025). They have the potential to reduce workload and even provide an alternative perspective; however, while we will give suggestions for how they could be integrated at various stages, we argue that they should never replace the two human coder roles due to the need for independent and accountable oversight.

The use of an LLM to replace the second coder, that is, to 'validate' the first coders' scheme and coded responses, which Bijker et al. (2024) suggest as a future possibility, would be, we think, highly problematic. As will become clear throughout the following sections, the role of the second coder is not as a passive automaton, but as a curious person with an independent will, tasked with pushing back against the first coder and questioning and critiquing their coding. Much of the advice in the following sections will involve ways to maximize the independence of the second coder to increase your reader's confidence that they provided this robust check. Second coders also provide a second person 'in the

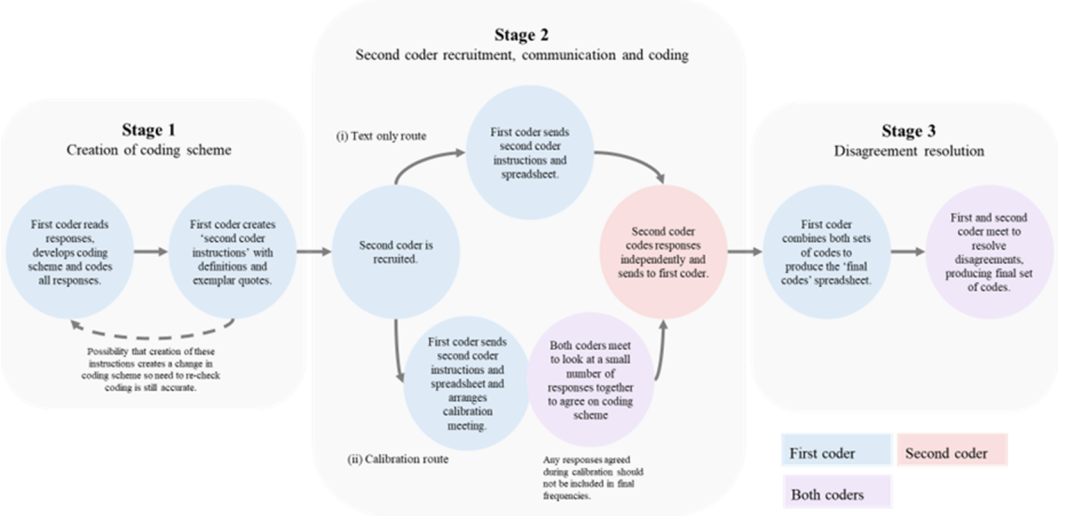

Figure 5. *A diagram depicting the three-stage process of developing a coding scheme and working with a second coder.*

room' that should be named in the paper (and ideally, a named author) and could be approached by an inquisitive reader to check that the first coder/author's depiction of the process is accurate.

LLM's are not well-suited to this role. First, considering off-the-shelf versions, there is, of course huge scope for iterative prompt engineering to achieve a desired outcome. LLM's passivity (and even sycophancy) makes them ripe for intentional or unintentional manipulation by the first coder. While a researcher can provide extensive documentation on their use of prompts, this is a huge amount of work for a reader/reviewer to process, and it is easy to conceal result-hacking behavior. While an LLM specifically developed or trained for this kind of purpose (e.g., Alaswad et al., 2023; Binz et al., 2024) is likely to be more accurate at categorizing responses, this provides even more opportunity for the primary researcher's desired outcomes to become embedded in the model. This would require a huge level of scrutiny and expertise by readers and reviewers to be confident that training did not in some way produce the desired result.

However, there are two ways in which we can see LLM involvement being valuable. First, the LLM could be involved in the development of the coding scheme in tandem with the first coder. We add a cautionary note here, as creating the coding scheme (the role of the first coder) is an act of directed creativity driven by a deep understanding of the purpose of the program of research, and so, while they could prove useful as a collaborative partner, we would not recommend just handing this over to an LLM.

As a second route, and one we would prefer, for large datasets, the method advocated in this guide could be adapted, as Rodger et al. (2025) did, by the first coder creating the coding scheme in a random subset of the data, then using an LLM to apply it to the whole dataset. This introduces a new issue beyond the first coder's potential researcher bias: the LLM's accuracy. These need to be checked separately. For the former issue, we would recommend having the second coder independently validate the scheme in that initial subset before it is handed over to the LLM. Following this, as Rodger et al. did, both first and second coders should then check a new random subset of the LLM's coding as a validation check on its accuracy.

One final reason to advise minimizing the handing of this analysis over to an LLM is that, while the process can be fairly time-consuming (typically half an hour per 100 responses), we strongly believe that this is time well spent. We have learned a huge amount about our domain of expertise by deeply engaging with participants' explanations of their thought processes. We do our best to find patterns in the data and present these to our readers, but they will never learn as much from our report as we learned from those hours spent reading them[3]. This applies to second coders too, who are typically students and, in our experience, learn a lot from the process.

4.2. Stage one—Creation of the coding scheme by the first coder

Stage one involves creating a coding scheme, undertaken by the first coder. While this task is usually performed by an individual, it can also be completed by a team acting as a collective first coder (or as mentioned above, a human-LLM team). In this case, none of the team members should subsequently assume the role of the second coder.

There are no 'rules' governing stage one. This phase can be compared to the 'theory generation' phase in the scientific method, where the focus is on idea generation without predefined constraints. Any strong guidelines provided in this section cover the process for validating that coding scheme using a second coder (akin to the 'testing phase' within science). Figure 6 contains some information on how the coding scheme was developed for the running example.

However, some advice can be offered. As with any analysis, it is crucial to keep your broad research question at the forefront, as it provides the purpose and direction for your work. We typically do this analysis in a simple spreadsheet program, creating new columns to represent codes (see Figure 7 for a

[3]This is quite a difference with quantitative analysis. A single participant's open-text response can transform your understanding of a cognitive phenomenon.

An example of a 'no information / evidence' quote from the running 'legal detainment' example:
In the detainment study, the first author of the present paper developed the coding scheme. They were driven by the research question 'Do participants make the inference from the detainment decision to the likelihood of guilt in the main trial? More specifically, do they do this more often when that decision is unexplained?' The expectation driving the work was that participants would only make this inference when the decision was 'unexplained' i.e. when it was 'valid'. The quote from Figure 4 is an example of a response which was assigned the code 'Guilt more likely', and this was found to be most common in the 'detained-unexplained' condition. In the 'detained-explained' condition however, the code 'No information / evidence' was most common, and here is an example:

P52: "There is absolutely no relevant information to go on. Therefore, it is impossible to tell either way."

This also supported the quantitative results, which found an interaction effect (larger effect of detainment on guilt when unexplained) and made for a stronger overall claim that that interaction effect was due to participants making this inference more often in that condition. Relying solely on the quantitative data would leave the claim more open to alternative explanations such as methodological flaws / confounds.

The materials for this study including the paper itself can be found at the online repository.

Figure 6. *An example of a 'no information/evidence' quote from the running 'legal detainment' example.*

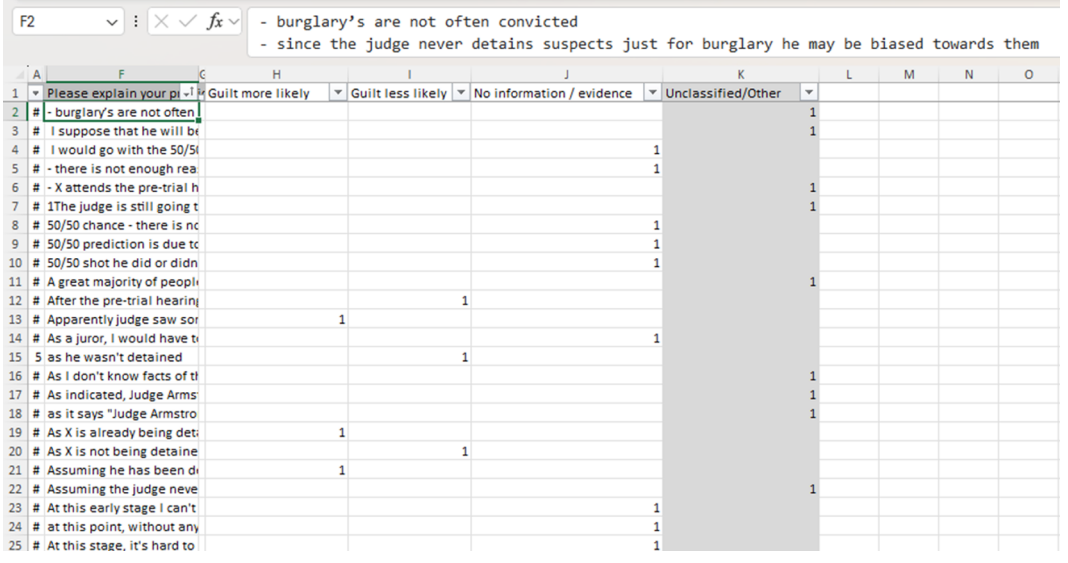

Figure 7. *A screenshot of a coding scheme in development with main codes 'Guilty more likely', 'Guilt less likely', and 'No information/evidence' as well as 'Unclassified/other'. The final scheme ended up with five main codes.*

depiction of a coding scheme in development). In the creation of codes, utilizing existing theoretical frameworks can be helpful. For example, the second author collected open-text data during a study examining information search strategies. By considering known search strategies, such as confirmation bias, during the coding process, participants' responses were mapped onto well-established cognitive processes (Liefgreen et al., 2020). Of course, you might be studying something novel without previous

literature to use as a guide. Either way, we recommend you initially simply read all your qualitative data with your research question in mind, either making notes on the key categories you observe or creating columns to represent them (you can modify/delete these later). Your aim is to identify similar cognitive processes or representations across participants' responses that are relevant to your research question/might be interesting to tally. Your label for each code should capture the core essence of the cognitive process you are aiming to catalogue. Typically, we only end up with 5–10 codes, which is very different from a typical content analysis on larger forms of data, which can easily stretch to 100+. Some of our more complicated work has stretched to 20 codes, but this can become unwieldy and hard to communicate to your reader, and we recommend aiming toward the lower end of this scale (around five). If more complicated code structures are necessary to represent the data, you can organize them into nested groups. However, we recommend avoiding using the word 'theme' to prevent confusion with thematic analysis.

While in early versions of implementing this process, we thought that the first coder should define their coding scheme before doing any coding, we have found over the years that this process is often more iterative than linear. Applying your developing coding scheme to your data, that is, simultaneously creating code columns and assigning responses to those codes, helps you identify what works and what does not, allowing you to refine it as you go. It is crucial, however, that once the coding scheme is finalized that the first coder revisits all the data coded during the development phase to ensure that every participant has been coded according to that final scheme. This is typically the process that we use: working through the qualitative data within a spreadsheet, creating columns to represent codes and assigning responses as we go, changing the codes' titles, removing, or combining them as seems fit. The process is not complete until the first coder is happy with all the codes and that all the responses have been coded with the final version of the coding scheme. Many examples of such spreadsheets can be seen on the online repository: https://osf.io/cyjhd/. As we will see in a future section, the process of communicating this to the second coder may also prompt further changes.

4.2.1. Unclassified/other

While your 'main' codes will differ across studies, we recommend you always include a special code labelled 'unclassified/other' (Figure 7: far right, highlighted grey). While this code is somewhat like a 'bin' it is not simply for junk responses. If a person writes anything other than the research language (English for us), including spam responses such as '*dahsdiowqn*', we will delete all their data[4]. However, as long as the response is in English, we do not remove them from the study even if their response indicates low engagement, and we advise against deleting participants based on their qualitative data, other than clear spam like above. Generally, data removal on this basis is too subjective and could lead to p-hacking if these participants are also excluded from the quantitative analysis. Examples of responses coded as unclassified/other can be seen in Figure 8.

So, the 'unclassified/other' code is not just for junk responses and is separate from the decision to remove data. Many participants might say things like 'It just felt right', which is a perfectly reasonable response, but provides little insight into their cognitive process and is unlikely to be assigned to a main code. Alternatively, they might express a highly unique view, which, while interesting, is not shared by anyone else. Coding is a process of identifying common cognitive processes, but if you wish to include a paragraph on unique but interesting responses at the end of your qualitative section, this is perfectly acceptable.

Once you have decided on your main codes of interest, the 'unclassified/other' code serves as a place for everything else, including some perfectly valid and interesting responses. There will always be more interesting elements in a dataset than you can condense into a readable format for your audience, and qualitative coding is already quite time-consuming. Therefore, you need to pick a focus and stick to it. It is important to note that the coding scheme presented to the reader reflects what the first author found interesting with respect to their research question. The second coder (more on this soon) provides some

[4]In this way the open-text box data also serves as a very basic engagement check (Reips, 2021).

An example of some quotes coded as 'unclassified/other' in the running 'legal detainment' example:
Here are a few responses coded 'unclassified / other' in the detainment study:

P8 "Not being detained until trial, gives me that impression."

P17 "Trusting a suspected burglar goes to far."

P50 "It depends on the jury, on the advocate and on the info they have about his fellony."

Figure 8. *An example of some quotes coded as 'unclassified/other' in the running 'legal detainment' example.*

Am I confident enough?
Recently, while the first author was training a student in this type of coding, a situation emerged where the student wanted to assign a particular response to a main code. The first author wasn't convinced and asked the student 'Are you confident?'. The student responded, 'fairly confident ... somewhat confident'. Upon further discussion, we agreed that there were other legitimate interpretations of the response and the student was likely reading between the lines too much. So, asking yourself 'Am I confident?' seems to be a helpful personal prompt. If you feel compelled to add a qualifier such as 'fairly' or 'somewhat' to 'confident', that is probably not confident enough, and the response should go in 'unclassified/other'.

Figure 9. *Am I confident enough?*

validation that these elements are genuinely present in the data. However, this does not imply that there is not something else interesting in the data, and providing your data in an open repository could allow other researchers to use your data to answer a different research question.

4.2.2. Confidence threshold for assigning main codes

We recommend that you adopt a conservative approach to coding, placing a response in 'unclassified/other' if you do not feel (subjectively) 'confident' that it should be assigned one of your main codes (see Figure 8 for examples and Figure 9 for further thoughts on the 'confident' threshold). While it may be tempting, especially when some part of you wants a particular result, to include more responses in your main codes, doing so actually just introduces more 'noise'. By restricting the responses to those you are confident about, your relative frequencies between codes (or across conditions), which will be your main focus of analysis, should be less noisy. If both you and your second coder adopt this approach, it also increases the likelihood of achieving decent inter-coder agreement, avoiding potential complications (see stage three for when things go wrong).

However, this is a relatively 'latent' form of content analysis, with the coders using their judgement as to whether they think a given cognitive process, represented by a code, occurred. Participants will often express themselves in a casual or rushed way and so you should not only be assigning a code if the participant literally writes it out word for word. Your subjective confidence as an intelligent human being with expertise in the domain should be your guide, and the independent second coder provides a check on you reading between the lines too much. Furthermore, it is important to recognize that this process is fairly time-consuming (very roughly half an hour per one hundred responses, but one tends to get quicker as the analysis continues), and so we do not advise spending minutes agonizing over whether you are confident enough on a particular response. Instead, if you cannot decide within around 10 seconds, you are probably not confident enough, and you should just put it in unclassified/other. If you spend minutes on each response, you will probably come to loathe this process and never want to

An example of a more ambiguous quote coded as 'unclassified/other':
Here is another code we put in the 'unclassified/other' category but which another coder might have assigned the 'Guilt more likely' code:

P10 "If you are on trial for robbery and if the judge is determining whether x will be detained before trial there is a high probability that X will get out of prison"

While this participant might be implying the 'Guilt more likely' inference—that because the judge detained the defendant, they are more likely to be guilty—it wasn't quite explicit enough for us, we couldn't say we were 'confident'. This is the kind of response that will typically end up being discussed in the disagreement resolution phase.

Figure 10. *An example of a more ambiguous quote coded as 'unclassified/other'.*

do it again. Remember that the final set of codes will be determined by both you and the second coder, and so many 'incorrect' decisions[5] on your or their part will be picked up this way. In the end, some small number of responses will be in unclassified/other that should perhaps have been in one of the main codes, but these should be minimal and will make little impact on your overall frequencies. An example of a more ambiguous 'unclassified/other' coded response can be seen in Figure 10.

4.2.3. Mutual exclusivity

Your main set of codes might be mutually exclusive, or not. By this, we mean that each response might be assigned only one code, precluding any other, or it could be assigned multiple codes. This is not a choice you make arbitrarily, but is determined by the nature of the phenomenon you are coding. In the detainment study example, a person could not simultaneously infer 'Guilt more likely' and state 'No information' as these thought processes are contradictory. However, in other studies, such as Dewitt, Adler, et al. (2023), participants could be assigned all of the main codes. Your study may even have a mixture, with some codes mutually exclusive with each other, and others not. The 'unclassified/other' code, of course, is always mutually exclusive with the main codes—a response is only placed there if it is not assigned one of the main codes. Furthermore, all responses should either get a main code or go in unclassified/other. No response should be completely uncoded.

4.2.4. Coding blind to condition and quantitative data

If you plan to make claims about coding frequencies with respect to conditions, then both coders must analyze the data blind to the condition. We achieve this blindness by, in a simple spreadsheet (1) creating a column of random numbers (which could just be participant number if random with respect to condition) and ordering participants accordingly, then (2) either using the 'hide' function on columns you wish to hide or cutting and pasting all qualitative data and random numbers to a new spreadsheet for the analysis. This ensures that you are not privy to participants' conditions or other important information during coding. Sometimes, qualitative data can reveal clues about the condition, such as references to condition-specific aspects of a vignette. This cannot always be avoided, and in such cases, you will not be able to make strong claims about code frequencies across conditions. You can still present such frequencies, but should clearly state that the analysis was not fully blind to condition, adding an extra layer of tentativeness.

While there might be instances where looking at quantitative data during qualitative coding is beneficial, we believe that, by default, you should code blind to this too (and clearly state this in your paper). First, like with condition, you may also want to make statistical correlation claims between qualitative and quantitative data, for example, that participants who gave a particular type of qualitative

[5] In the sense that you would change your mind if given a long time to reflect.

An example of distinguishing between hypothesis-driven and exploratory coding:
In the detainment paper, by the time we got to the qualitative coding, we 'expected' higher levels of 'Guilt more likely' in the 'Unexplained-Detained' condition than in the 'Explained-Detained' condition. However, as this 'expectation' emerged at some point during the research process, we did not attempt to present this as a true hypothesis. If we conducted a follow up to this study, we might be prepared to make a true prediction of this sort, now that we have a firm idea of the kind of responses to expect.

Figure 11. *An example of distinguishing between hypothesis-driven and exploratory coding.*

An example of deciding on the level of independence needed for the second coder:
In the detainment paper, the second coder was the student who's third year dissertation the project was. While this was acceptable for publication as preliminary and exploratory work as a conference paper, if we had aimed to publish in a full journal or to make strong claims we would have needed to recruit a new, fully independent second coder.

Figure 12. *An example of deciding on the level of independence needed for the second coder.*

response had higher scores on some quantitative measure (e.g., Dewitt, Adler, et al., 2023). Such a claim can only be made if this was conducted blind.

4.2.5. Exploratory vs hypothesis-driven

The vast majority of the time, this kind of qualitative analysis and any associated comparisons of coding frequencies between conditions should be exploratory. We agree with the Scheel et al. (2021) criticism that our discipline is often too eager to test hypotheses when we are still actually in exploratory mode, and also with papers like Glass (2010) that research questions should be primary drivers rather than hypotheses. However, if you do have a strong hypothesis with respect to, for example, coding frequencies across conditions, there is no reason it cannot be preregistered like any quantitative hypothesis. However, if you intend to make such strong claims with your qualitative data, you should anticipate greater scrutiny of your analytical approach.

A hypothesis-confirmatory qualitative analysis would typically really need to follow an exploratory study, as even with a strong theoretical foundation, it is difficult to predict what participants will say. However, if you have conducted an initial exploratory study, it might be reasonable to predict in a follow-up study that certain responses will occur more frequently if you manipulate certain variables or change the framing (e.g., see Figure 11).

4.3. Stage two—Second coder recruitment, communication, and coding

While recruiting and working with a second coder may seem straightforward, there are many pitfalls to avoid. The purpose of this process is to validate your coding scheme and provide greater confidence to your readers that the coding scheme and associated code frequencies are not merely a fiction of the first coder's mind. To achieve this, it is essential to minimize the influence of the first coder's desired outcomes on the second coder, thereby maximizing their independence (see Figure 12 for some thoughts on this).

This independence can be challenging to achieve because the first coder will often recruit the second coder and has various kinds of 'power' over them[6]. For instance, second coders are typically added as authors on papers, so they may feel pressure to please the first coder (usually the first author) to secure their authorship. Additionally, more senior authors often recruit students, creating a seniority disparity.

[6]This is our experience as UK-based researchers, however international colleagues have made us aware that this process may well be mediated by a lab manager. This is good for maximizing independence, but not available to all researchers.

Lastly, the first author is usually the authority on the paper and coding scheme, having made it, and possesses the overall vision for the paper and analysis. To address these issues, an environment must be established where the second coder is not only given permission but is actively encouraged to hold the first coder accountable. We recommend including a statement like the following when sending coding instructions to the second coder.

> Thank you for agreeing to be a second coder. It is your job to provide a validation check on the coding scheme I have come up with and to code the data using that scheme to the best of your ability. It is your job to hold me to account and not just to sign off on what I have done. During this process, please have permission to argue, debate and be as disagreeable as you need to be. My aim is to produce a coding scheme and set of coded responses which as much as possible reflects the truth in the data rather than what I would prefer to be true.

When choosing a second coder, you face a tension. Ideally, your second coder would have no prior involvement with the paper and should not be too familiar with you either. However, this must be balanced with the need for the coder to comprehend the study, especially if it is complicated. It also must be balanced with pragmatic constraints of who is available. Independence becomes more crucial if you intend to make strong claims regarding coding frequencies. If your qualitative component is purely exploratory and tentative, the independence requirement can be somewhat relaxed (e.g., see Figure 12). However, if you are aiming to make strong claims about coding frequencies across conditions to support a theoretical position, it is essential to find a second coder unaware of those aims and to maintain that unawareness throughout the process. This might require looking beyond your immediate lab, especially if you have presented early ideas locally. In terms of competency, we recommend recruiting someone within the same field who ideally has some previous qualitative experience to understand the requirements. If they have no experience with this or similar processes, having them read this paper would be a good idea. Above all, you should be transparent in your paper about the second coder's background and any previous involvement in the study.

4.3.1. Second coder instructions

You will need to send the second coder written instructions as well as a prepared spreadsheet for them to conduct their coding independently. There are two options we have used here for communicating this to the second coder:

1) **Text only:** if independence is the primary concern, and the coding scheme/code definitions are relatively straightforward, we recommend just sending the instructions along with the second coder spreadsheet by email or other text channel.
2) **Text + calibration meeting:** if communication clarity is the primary concern, for example, if the coding scheme is complicated or the definitions are quite nuanced, you could have a calibration session with your second coder where you go through a few responses together to agree on what they would be. This meeting should cover the minimum number of responses possible because these responses should not be included in any final coding frequencies.

Whichever of these two routes you choose should be clearly stated in the paper, and for route two, the number of responses used for calibration (and therefore removed) should be reported. If the calibration meeting leads to any changes or clarifications in the definitions of the codes, the second coder's instructions should be edited to reflect this. We have several examples of second coder instructions and accompanying spreadsheets ready for the second coder (e.g., with condition / quantitative data hidden) on the online repository: https://osf.io/cyjhd/ (also see Figure 13 for an example of a single code). When sending these, the same balance of understanding vs independence continues. The second coder will need some basic background on the study (and potentially the phenomenon under study, depending on their expertise) to do their job effectively, but you also do not want to overly influence them, and certainly not to give away what you expect to find (e.g., a large number of participants saying X, or more participants saying X in one condition than another). Again, it is important to make sure that all

An example of a code definition and example code sent to the second coder for the running 'legal detainment' example:
Here is the code definition and one example code sent to the second coder for the 'Guilt more likely' code:

Guilt more likely (Definition: the participant states that because of the detainment decision the defendant is more likely to be guilty). For example:

P135 "If the judge decided to detain him, that means he thought he was likely to flee, or not show up for the trial. That might tip the scales in favour of him being guilty."

Figure 13. *An example of a code definition and example code sent to the second coder for the running 'legal detainment' example.*

information you send to your second coder is clearly available to your readers in an online repository so they can judge for themselves if the process was robust.

After providing some background, your instructions should include your 'codebook', containing your coding scheme. This codebook should contain definitions for each code (typically a few sentences) along with one to two exemplar quotes that you think perfectly capture the essence of each code. Importantly, not all responses need to be as perfectly articulated as these exemplars: participants often vary in their clarity of expression (more on this soon). As mentioned previously, creating these instructions may prompt you to rethink your coding scheme, which is normal and should be embraced. It is much better to get it all straight now than face a large number of disagreements in stage three.

Second coders, especially inexperienced ones, sometimes struggle with how much to 'read between the lines'. Generally, it is better to be conservative but otherwise to just use their best judgment. We instruct them to apply the code if they, personally, are 'confident' (see Figure 9) that it is correct. If they feel like they are 'guessing', they should just place the response in 'unclassified/other'. However, this is not a surface-level form of content analysis, where the person has to literally say an exact phrase or word. Both coders use their natural human intelligence (and professional domain knowledge) to decide whether they think the person is trying to express the cognitive process of interest, and the agreement between them provides some (not absolute) validation of the final set of coded responses.

There is an issue of communication in all of this. The first coder has an idea of what their code means in their head and attempts to communicate this, but this may not always be entirely effective. The second coder should be encouraged to ask questions if they are completely unclear about a code's meaning, but should otherwise do their best autonomously to maximize their independence. If you are going down route one (text only), you should avoid having repeated back-and-forth conversations with your second coder about whether particular responses fit a code; they must code independently to the best of their ability. If this does not seem to be working, it might be best to go down route two and have a calibration meeting.

4.4. Stage three—Disagreement resolution

After the coding is complete, the first and second coders should meet to discuss responses they disagree on and compile a set of 'final' agreed coded responses. Before beginning the discussion, it is beneficial to remind the second coder that they have your permission to push back and hold you accountable.

During the discussion, it seems to be a good rule that if the two coders cannot agree within a short time (e.g., one minute), then the response should just be put into 'unclassified/other'. This mirrors the personal '10 second' version of this rule as a lone coder if one is struggling to decide. This achieves the two goals of reducing the chance that the first coder dominates the second and also reduces the chance of a disagreeable dynamic emerging. If both coders remember that it is not 'bad' for a response to go into

An example of a code definition and example code sent to the second coder for the running 'legal detainment' example:
Several statistical metrics are commonly employed to measure the level of agreement among coders in many domains, and could be used for inter-coder agreement:

1. **Percentage Agreement**: Calculates the proportion of coding decisions in agreement among coders, providing a simple but limited measure.
2. **Cohen's Kappa**: A robust statistic for two coders using nominal categories, considering agreement beyond chance.
3. **Fleiss' Kappa**: Extends Cohen's kappa for studies with three or more coding categories and multiple coders.
4. **Scott's Pi**: Measures agreement beyond chance for two or more coders, particularly suitable for imbalanced coding distributions.
5. **Krippendorff's Alpha**: Versatile for various levels of measurement and accommodating missing data and binary/multi-category coding.
6. **Intraclass Correlation Coefficient (ICC)**: Applied to continuous or ordinal data with multiple coders, commonly used in psychometrics and behavioural research.

The choice of the appropriate measure depends on factors such as research context, number of coders, complexity of coding categories, and data characteristics.

Figure 14. *Examples of statistics that can be used for calculating inter-coder agreement.*

'unclassified/other' because you do not really want to pollute your main coding frequencies with more ambiguous responses, then this should ensure a smoother process. While the time recommendation here might seem short, there may be many responses to deal with, and you will see the need to move relatively swiftly once you begin this process. Typically, this meeting will take 30 minutes to an hour, but it could take a lot longer if you exhaustively debate every disagreement.

In the final spreadsheet you upload to your online repository, you should show both coders' initial codes and the final agreed code after discussion. In the online repository associated with this paper (https://osf.io/cyjhd/), we provide several example spreadsheets, including one for the running 'detainment' example and another for a different study (Dewitt, Adler, et al., 2023). We have also included a template spreadsheet which is set up to display which codes between the first and second coder require discussion, accompanied by a 'read me' document with some tips on how to use it. Also see Figure 14 for some possible metrics for inter-coder agreement.

This transparency allows an inquisitive reader to see what transpired during the disagreement resolution discussion. For example, in Dewitt, Adler, et al. (2023), there were 28 disagreements for one code, all of which stemmed from the second coder assigning the code while the first coder had put them in unclassified/other. This might suggest that the second coder was being less conservative, with a lower confidence threshold for assigning the code than the first coder (i.e., 'over-coding'). The spreadsheet shows that nine out of the 28 times, the final agreed code reflected the second coder's initial coding, while in 19 cases, it reflected the first coder's. Although this slightly favors the first coder, it does not seem overly concerning, and a reader could look to see if this pattern was there across multiple codes. It is good practice and a sign of transparency to make a note of such things for inquisitive readers, for example, by using the comment function in an appropriate part of the spreadsheet.

One approach to minimize the first coder's dominance during disagreement resolution (and save time at this phase) is, before the discussion, to just assign any response that either coder placed in 'unclassified/other' to that category in the final agreed set of codes. We used this method in Dewitt, Adler, et al. (2023), resolving 31 of 91 disagreements this way, significantly reducing the proportion resolved through potentially biased discussions. While this need not be a universal rule, as there are cases where a coder genuinely errs by categorizing responses as 'unclassified/other', this highly conservative approach can make this stage more robust, especially if low inter-coder agreement is achieved at the end of stage two.

An opportunity for the reader to get practise as a second coder:
To get a feel for all this, we have provided materials in the online repository so that you can get experience acting as a second coder (it's not really possible to give first coder experience, you can only really do that with your own study, but this should still help make that feel less alien too). In the repository, navigate to the 'Legal detainment' folder, which contains all the materials associated with the running example. You will find a spreadsheet file called 'Legal detainment blank' which is set up for a second coder. Columns have been created for the five main codes + unclassified/other, participant order has been randomised, and condition / quantitative data has been hidden. Before you attempt to do the coding, look at the 'Code definitions for second coder' document. Then, have a go at coding as many responses as you need to get a feel for the process. Once you have done this, you can either compare your responses with those in 'Legal detainment first coder only' (the first author's codes) or with both the first and second coder in 'Legal detainment initial final codes' and see which one of them you agreed with more.

Figure 15. *An opportunity for the reader to get practice as a second coder (https://osf.io/cyjhd/).*

4.4.1. When things go wrong

What if you encounter very low inter-coder agreement, discover that your second coder has completely misunderstood your code definitions, or realize during the discussion that you need to change your coding scheme (e.g., the second coder thinks you missed an important code, or two of your codes are actually the same thing)?

While we have never faced major issues of this sort, we have heard from other researchers who have. Following a structured process like the one outlined here should minimize the chance of these issues occurring (if you want to get some practice now, check out Figure 15). Small changes can often be fixed with full transparency in the paper and the online repository. For example, if during the discussion process, the first coder realizes that the second coder's interpretation of a particular code is actually more accurate than the first coder's, the spreadsheet might show a large proportion of discrepancies resolved in favor of the second coder's initial code. This might be followed by a change to the coding definition as it is written in the paper, which is not a problem.

However, if your process has gone completely awry and your second coder is no longer the independent checker you need, you may need to start over. In this case, consider the process so far as a 'revision' of your coding scheme or a 'dry run'. This is perfectly acceptable as long as it is explained openly. Essentially, you just unintentionally hired someone to help refine your coding scheme before engaging in the second coder process, so the person you thought was your second coder has now become part of your first coder team. You now need to hire a completely new second coder to start again. Clearly detail this in your paper, explaining that you hired a second coder, realized your coding scheme was unclear (report the inter-coder agreement), revised the scheme, and undertook the process again. This is not shameful and does not need to be 'dressed up' as anything else; it is just part of the real, messy process of research.

5. Writing up your qualitative section

For someone who is used to writing quantitative papers, transitioning to a mixed methods paper of this sort introduces a few new sections to master: the description of your coding process and the detailing of your qualitative results. The description of your coding process could either be a part of your method section (e.g., within 'Data Analysis') or could be described at the start of your qualitative results section, or a mixture, as appropriate to guide your reader. The key here is to describe as openly as possible the whole process you went through as was detailed in the previous sections i.e., which author created the coding scheme, how the second coder was recruited, who they are, how that process was managed (ideally with reference to materials in an online repository), whether coding was blind to condition / quantitative response, how discrepancies were resolved, and if something went wrong, what was done to deal with it.

Table 1. *The percentage of participants assigned each code by condition in Dewitt, Glatzel, and Lagnado (2023). The most frequent code assigned other than 'Unclassified/Other' is made bold for each condition. Initial inter-coder agreement is shown for each code in the final column.*

	Unexplained		Explained		Grand Total	
	Not detained	Detained	Not detained	Detained		
n	93	91	84	95	363	ICA
Guilt more likely	–	**54.9**	1.2	8.4	16.3	97.7%
Guilt less likely	**39.8**	–	3.6	–	11.0	97.2%
Confound	–	–	8.3	13.7	6.1	98.3%
No Information / Evidence	23.7	13.2	**36.9**	**25.3**	24.5	99.5%
Innocent Until Proven Guilty	1.1	3.3	1.2	4.2	2.5	99.7%
Unclassified / Other	35.5	28.6	48.8	48.4	39.7	97.5%

For the results section, we typically divide it into 'Quantitative' and 'Qualitative' subheadings, usually presenting the quantitative results first. While there may be instances where the two analyses are intertwined (e.g., see how we handled this kind of issue in the 'legal detainment' example paper: Dewitt, Glatzel, & Lagnado, 2023), this format is more easily interpreted by an audience accustomed to quantitative-only analyses. The qualitative section generally begins with a section including a table or diagram with the final agreed coding frequencies and any statistical analyses conducted on those frequencies. For example, Table 1 shows the table of coding frequencies we provided in the running 'legal detainment' example (Dewitt, Glatzel, & Lagnado, 2023). We have also here added the initial inter-coder agreement for each code (far right column), as we have recently found this a natural place to include this.

In Table 1 we can tentatively observe (presuming that the reader has some faith in the coding process that produced these percentages) supportive evidence for the overall research claim that participants infer guilt from detainment when it is unexplained (54.9% coded that way in 'Unexplained-Detained') but not when explained (only 8.4% coded as 'Guilt more likely' in 'Explained-Detained', with the modal response at 25.3% being 'No Information / Evidence'). In this paper, conducting statistical analyses on these frequencies did not seem appropriate, but we did use binary logistic regression (as codes are usually binary 0/1 data) at certain points in Dewitt, Adler, et al. (2023) to support specific claims, which were still presented as fully exploratory.

The second part of your qualitative results section should contain definitions and exemplar quotes for each of the codes in the table to help your reader interpret these coding frequencies more deeply. This should be similar to the codebook you sent to your second coder as their instructions. There is little else to say about this, so we point you to a range of papers that serve as examples of how to structure and phrase things (Adler & Dewitt, 2024; Dewitt, Adler, et al., 2023; Dewitt, Glatzel, & Lagnado, 2023; Liefgreen et al., 2020). We will continue to add new papers to the online repository as further work is published. We should also note that this process has been in development over several years, and writing this paper has brought further clarity, so more recent papers are likely to follow this process more closely than our earlier work.

For the second section, where you include definitions and quotes, there is one important thing to note. It is inevitable that the author will use their 'best' quotes here, those that most clearly match the code definition. These are sometimes called exemplars or 'gem' quotes. It should be generally understood that if the reviewer or reader examines all the responses coded under a particular code in the online repository, the vast majority will not be as 'clean' as those provided in the paper. This should not be considered cherry-picking but should be accepted by the reader, who should not be misled into thinking that all responses are as clear as those in the paper. It would be helpful for the author to support this by using the language of 'exemplars', such as: 'Here are some exemplar quotes which capture the core

idea of the code most clearly'. This language makes it clear that you are not trying to hide anything or mislead your reader.

6. A checklist for reviewers

In this section, we provide a checklist for reviewers to guide them in gaining a holistic view of whether an analysis of this sort has been undertaken in good faith. We try to keep the below brief, and the related sections of the paper cover more ways in which one could get a feel for the analysis. However, there is no single way to conduct research, and authors may follow the spirit of this guide while taking a different approach appropriate to their area of study. Therefore, few of the following should be considered absolute requirements; instead, they combine to create a picture of openness and transparency. Importantly, a robust process should be demanded in proportion to the strength of the claims made by the authors. If the authors present their qualitative analysis as completely exploratory with a 'take it or leave it' attitude, less scrutiny of the process that produced those codes is appropriate. However, if they use their coding frequencies to make strong claims, such as about frequencies across conditions or as strong evidence for a theory, a highly robust process and high levels of openness and transparency with materials detailing the process in an online repository should be expected. We provide this list as a fillable spreadsheet to support reviewers in the online repository: https://osf.io/cyjhd/

6.1. Stage one—First coder creation of coding scheme

There are no strict 'rules' for the creation of the coding scheme, and this may even involve computational assistance such as an LLM in some form. Like scientific theories, ideas can come from anywhere; however, it is a good sign of openness if the first author describes this process thoroughly.

- If the authors are making claims about coded response frequencies with respect to condition (e.g., more people saying X in a given condition), or a quantitative response (e.g., higher average response among those coded as Y), was coding conducted blind to condition / quantitative response?

6.2. Stage two—Second coder recruitment, communication, and coding

Please read the stage two section for further details, but here are some key indicators to look out for:

- Have the authors explained who the second coder was (they should ideally be named if not an author on the paper) and how they were recruited?
- Is there a potential power discrepancy or a shared desire for a particular outcome that could affect independence?
- Have the authors clearly explained the communication process with the second coder?
- Have they uploaded the instructions they sent to the second coder?

6.3. Stage three—Disagreement resolution and final codes

Please read the stage three section for further details, but here are some key indicators to look out for:

- Have the authors described the process for disagreement resolution?
- Have the authors uploaded a spreadsheet showing the qualitative data for all participants, including initial codes for both coders and final agreed codes?

Data availability statement. All materials associated with the guide, including example papers and coding spreadsheets, example second coder instructions, practise materials, assistive checklists for reviewers and the data and materials for the small study mentioned can be found at the following link: https://osf.io/cyjhd/.

References

Adler, N., & Dewitt, S. H. (2024). Re-Examining Base-rate neglect: The effect of context. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

Alaswad, A., Kalganova, T., & Awad, W. S. (2023). Using ChatGPT and other LLMs in professional environments. *Information Sciences Letters*, *12*(9), 2097–2108.

Bengtsson, M. (2016). How to plan and perform a qualitative study using content analysis. *NursingPlus open*, *2*, 8–14.

Bhatia, S., Walasek, L., Slovic, P., & Kunreuther, H. (2021). The more who die, the less we care: Evidence from natural language analysis of online news articles and social media posts. *Risk Analysis*, *41*(1), 179–203.

Bijker, R., Merkouris, S. S., Dowling, N. A., & Rodda, S. N. (2024). ChatGPT for automated qualitative research: Content analysis. *Journal of Medical Internet Research*, *26*, e59050.

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, F. Demircan, C., Eckstein, M, K., Éltető, N., Griffiths, T, L., Haridi, S., Jagadish, A. K., Alexander Kipnis, L, J-A., Kumar, S., Ludwig, T., Mathony, M., Mattar, M., Modirshanechi, A. et al. (2024). Centaur: A foundation model of human cognition. *Preprint*, arXiv:2410.20268.

Binz, M., Alaniz, S., Roskies, A., Aczel, B., Bergstrom, C. T., Allen, C., et al. (2025). How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences*, *122*(5), e2401227121.

Bishop, D. (2019). Rein in the four horsemen of irreproducibility. *Nature*, *568*(7753), 435.

Cokely, E. T., & Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision making*, *4*(1), 20–33.

Decorte, T., Malm, A., Sznitman, S. R., Hakkarainen, P., Barratt, M. J., Potter, G. R., . . . Frank, V. A. (2019). The challenges and benefits of Analyzing feedback comments in surveys: Lessons from a cross-National Online Survey of small-scale cannabis growers. *Methodological Innovations*, *12*(1), 205979911982560.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., et al. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), 688–701.

Dewitt, S., Hsu, A., Lagnado, D., Connor-Desai, S., & Fenton, N. (2019, July). Nested sets and natural frequencies. In COGSCI'19: Creativity+ cognition+ computation (Vol. 41, pp. 1633–1639). Cognitive Science Society.

Dewitt, S. H., Adler, N., Li, C., Stoilova, E., Fenton, N. E., & Lagnado, D. A. (2023). Categorical updating in a Bayesian propensity problem. *Cognitive Science*, *47*(7), e13313.

Dewitt, S. H., Glatzel, S., & Lagnado, D. (2023). How does knowledge of detainment affect juror reasoning? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

Eccles, D. W., & Arsal, G. (2017). The think aloud method: What is it and how do I use it? *Qualitative Research in Sport, Exercise and Health*, *6778*, 1–18.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215.

Feuerriegel, S., Maarouf, A., Bär, D., Geissler, D., Schweisthal, J., Pröllochs, N., et al. (2025). Using natural language processing to analyse text data in behavioural science. *Nature Reviews Psychology*, 1–16.

Feynman, R. P., Leighton, R., & Hutchings, E. (2007). *Adventures of a curious character*. Vintage.

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, *137*(2), 316–344.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704.

Glass, D. J. (2010). A critique of the hypothesis, and a defense of the question, as a framework for experimentation. *Clinical Chemistry*, *56*(7), 1080–1085.

Glöckner, A., & Betsch, T. (2011). The empirical content of theories in judgment and decision making: Shortcomings and remedies. *Judgment and Decision making*, *6*(8), 711–721.

Glockner, A., Gollwitzer, M., Hahn, L., Lange, J., Sassenberg, K., & Unkelbach, C. (2024). Quality, replicability, and transparency in research in social psychology. *Social Psychology*, *55*(3), 134–147.

Grice, H. P. (1975). Logic and conversation. In Speech acts (pp. 41–58). Brill.

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, *316*(5827), 998–1002.

Haidt, J., & Jussim, L. (2016). Psychological science and viewpoint diversity. *APS Observer*, *29*(2).

Hall, K. H. (2002). Reviewing intuitive decision-making and uncertainty: The implications for medical education. *Medical Education*, *36*(3), 216–224.

Hallsworth, M. (2023). A manifesto for applying behavioural science. *Nature Human Behaviour*, *7*(3), 310–322.

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: What natural frequencies are and what they are not. *Cognition*, *84*, 343–352.

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9), 1277–1288.

Hubbs, D. L., & Brand, C. F. (2005). The paper mirror: Understanding reflective journaling. *The Journal of Experimental Education*, *28*(1), 60–71.

Jamieson, M. K., Govaart, G. H., & Pownall, M. (2023). Reflexivity in quantitative research: A rationale and beginner's guide. *Social and Personality Psychology Compass*, *17*(4), e12735.

Johansson, P., Hall, L., & Sikström, S. (2008). From change blindness to choice blindness. *Psychologia*, *51*(2), 142–155.

Jussim, L., Crawford, J. T., Anglin, S. M., & Stevens, S. T. (2015). Ideological bias in social psychological research. *Social psychology and politics*, 107–126.
Kuusela, H., & Paul, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *The American Journal of Psychology*, *113*(3), 387–404.
Lejarraga, T., & Hertwig, R. (2021). How experimental methods shaped views on human competence and rationality. *Psychological Bulletin*, *147*(6), 535–564.
Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255–275.
Liefgreen, A., Pilditch, T., & Lagnado, D. (2020). Strategies for selecting and evaluating information. *Cognitive Psychology*, *123*, 101332.
Macchi, L. (2000). Partitive formulation of information in probabilistic problems: Beyond heuristics and frequency format explanations. *Organizational Behavior and Human Decision Processes*, *82*(2), 217–236.
Madsen, J. K., de-Wit, L., Ayton, P., Brick, C., de-Moliere, L., & Groom, C. J. (2024). Behavioral science should start by assuming people are reasonable. *Trends in Cognitive Sciences*, *28*(7), 583–585.
Mayser, S., & von Wangenheim, F. (2013). Perceived fairness of differential customer treatment: Consumers' understanding of distributive justice really matters. *Journal of Service Research*, *16*(1), 99–113 Processes, 82(2), 217–236.
Nicmanis, M. (2024). Reflexive content analysis: An approach to qualitative data analysis, reduction, and description. *International Journal of Qualitative Methods*, *23*, 16094069241236603.
Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231.
O'Grady, C. (2021). Fraudulent data raise questions about superstar honesty researcher. Science. https://www.science.org/content/article/fraudulent-data-set-raise-questions-about-superstar-honesty-researcher.
Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*, 1596–1618.
Olmos-Vega, F. M., Stalmeijer, R. E., Varpio, L., & Kahlke, R. (2023). A practical guide to reflexivity in qualitative research: AMEE guide no. 149. *Medical Teacher*, *45*(3), 241–251.
Ranyard, R., & Svenson, O. (2011). Verbal data and decision process analysis. In M. Schulte-Mecklenbeck, A. Kühberger, & A. &. R. Ranyard (Eds.), *A handbook of process tracing methods for decision research: A critical review and user's guide* (pp. 115–130). Taylor & Francis.
Reips, U.-D. (2002). Standards for internet-based experimenting. *Experimental Psychology*, *49*(4), 243–256.
Reips, U.-D. (2021). Web-based research in psychology: A review. *Zeitschrift für Psychologie*, *229*(4), 198–213.
Rodger, A., Sanna, G. A., Cheung, V., Raihani, N., & Lagnado, D. (2025). Negative anecdotes reduce policy support: Evidence from three experimental studies on communicating policy (in) effectiveness. *OSF*. https://doi.org/10.31219/osf.io/e2kxc_v1.
Rouder, J., Saucier, O., Kinder, R., & Jans, M. (2021). What to do with all those open-ended responses? Data visualization techniques for survey researchers. *Survey Practice* August. https://doi.org/10.29115/SP-2021-0008.
Saldaña, J. (2021). *The coding manual for qualitative researchers*. SAGE Publications Ltd.
Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, *16*(4), 744–755.
Schulte-Mecklenbeck, M., Kühberger, A., & Johnson, J. G. (2019). *A handbook of process tracing methods*. Routledge.
Schulte-Mecklenbeck, M., Kühberger, A., & Ranyard, R. (2011). The role of process data in the development and testing of process models of judgment and decision making. *Judgment and Decision making*, *6*(8), 733–739.
Sela, H. K. U., Ab, P., & Pau, I. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *The American Journal of Psychology*, *113*, 3.
Singer, E., & Couper, M. P. (2017). Some methodological uses of responses to open questions and other verbatim comments in quantitative surveys. *Methods, Data, Analyses*, *11*(2), 114–134.
Toomela, A. (2008). Variables in psychology: A critique of quantitative psychology. *Integrative Psychological and Behavioral Science*, *42*(3), 245–265.
Weiss, D. J., & Shanteau, J. (2021). The futility of decision making research. *Studies in History and Philosophy of Science*, *90*, 10–14.
Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R. C., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832.
Wilson, T. D. (1994). The proper protocol: Validity and completeness of verbal reports. *Psychological Science*, *5*(5), 249–252.

Cite this article: Dewitt, S. H., Liefgreen, A., Adler, N., and Strittmatter, L. E. (2025). 'Please explain your response': A guide to uncovering cognitive processes from open-text box data using pragmatic and reflexive content analysis. *Judgment and Decision Making*, e39. https://doi.org/10.1017/jdm.2025.10010