Review Article

A Primer and Pragmatic Guide to Developing Survey Questionnaires for Population Health Research

Nashit Chowdhury[1,2], Irfan Hyder[1,2], Mohammad M. H. Raihan[1,2], Tanvir C. Turin[1,2]

[1]Department of Family Medicine, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada; [2]Department of Community Health Science, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

Abstract

Surveys are indispensable in population and public-health research for measuring knowledge, perceptions, and behaviours not captured in administrative data, especially in cardiovascular (CVS) research (e.g., diet, activity, medication adherence). However, poorly constructed questionnaires introduce systematic measurement errors and biases. This paper synthesizes best practices for designing reliable and valid self-report instruments, emphasizing a respondent-centred, systematic approach. We present evidence-based strategies integrating (1) item generation through operationalizing theoretical constructs into measurable indicators, (2) tailored item formulation addressing the four cognitive stages of survey response (comprehension, retrieval, judgment, and response formatting) to minimize measurement error, (3) questionnaire composition strategies for logical structure, flow, and managing contextual and order effects, and (4) systematic questionnaire testing and refinement through pre-testing (qualitative item refinement, individual item quality assessment, and holistic questionnaire appraisal), pilot testing with psychometric validation (reliability, validity, and item discrimination), and evidence-based item reduction and finalization. Practical tools (e.g., BRUSO checklist, cognitive interviewing, factor analysis) are highlighted throughout. Applying these strategies yields questionnaires that are more interpretable, less biased, and better suited to intended analyses. Key operational recommendations include defining the constructs and using appropriate question items and types to measure them, prioritizing brevity and clarity, preferring closed items with exhaustive response categories, allowing for the capture of unforeseen responses, conducting pre-testing and pilot testing, and thoroughly evaluating the questionnaire.

High-quality questionnaire development is a resource-intensive, iterative process that combines qualitative insight and quantitative rigour. Adoption of the outlined evidence-based steps substantially reduces measurement error and enhances the utility of survey data for CVS and broader public-health research.

Key words: questionnaire development; survey methods; cognitive interviewing; psychometrics; cardiovascular research; measurement error; tailored-design method.

[JNHFB 2026; 15 (1) : 12-19] doi.org/10.61819/jnhfb.v15i1.19

Introduction

Population and public health research increasingly employ survey methods to explore community and patient population knowledge, perceptions, and related behaviors regarding various health issues. Surveys serve as vital instruments for understanding population needs, tracking health trends, and informing evidence-based interventions[1]. They enable researchers to capture social, behavioral, and environmental determinants of health that are not readily available in administrative or clinical data. For example, in cardiovascular (CVS) health research, survey has been particularly effective in examining dietary habits, physical activity, medication adherence, and risk perceptions, illustrating its immense contribution to advancing disease prevention and health promotion efforts[2].

Survey research constitutes a foundational methodological approach within scientific inquiry, holding significant merit for rigorous design and analysis[3]. An essential element of this approach is development of high quality questionnaires to collect reliable (consistent) and unbiased data[3]. The process of developing a questionnaire to achieve these standards is inherently challenging and labour-intensive. It requires a comprehensive, critical, and systematic approach to ensure minimal bias and optimal validation[1]. This process necessitates adopting a tailored-design method, where researchers must get into a "state of mind" to understand what the questions and questionnaire will look like from the respondent's perspective, helping them easily understand and navigate the survey instrument[4].

Corresponding Author
Dr Tanvir C Turin MBBS, PhD
Department of Family Medicine, Room G012, Health Sciences Center,
3330 Hospital Drive Northwest, Calgary, Alberta T2N 4N1, Canada.
Email: turin.chowdhury@ucalgary.ca

A poorly constructed questionnaire, including ambiguously phrased questions, poorly designed response formats, and overall flaws such as excessive length, poor logical flow, or lack of thematic consistency, is a major source of significant measurement errors and biases in a survey[5]. These flaws deviate the results or inferences from the truth, and persist as pervasive problem in questionnaire design[6]. Subsequently, they undermine the validity and utility of the resulting data, requiring diligent attention to detail that elevates questionnaire design from a simple exercise to a mixture of art and science[5]. This article aims to describe the systematic and evidence-based processes and methodological considerations involved in developing a rigorous survey instrument, contributing to ongoing efforts to improve the quality and reliability of survey research in population and public health sciences.

[I] Item or Individual Question Generation: Operationalizing Concepts into Measures

The initial phase of survey questionnaire development represents the critical juncture where theoretical concepts are operationalized into empirical measures. Rigorous adherence to systematic procedures in this phase is paramount. This includes conducting thorough literature review, developing a comprehensive research protocol, and having the protocol expert/peer-reviewed for mitigating specification errors, which arise when key theoretical concepts or constructs are not properly reflected in the survey variables[7]. For example, the failure to include questions about smoking in a survey measuring perception of myocardial infarction (MI) risk among individuals exemplifies a specification error.

A. Defining Research Objectives and Variables of Interest

A clearly articulated research objective and a distinct, well-defined target population are essential prerequisites for developing a rigorous survey[3]. Refining these initial research objectives requires the specification of the topic, the intended respondents, and the primary research questions to be addressed. Once these foundations are established, researchers must carefully identify and define the main variables of interest, which are central to the study's objective. A variable refers to any measurable attribute or characteristic that can differ across individuals or groups. Many of these variables represent broader, abstract ideas known as constructs[7]. These constructs, for example, stress, social support, health literacy, attitudes, or beliefs, are fundamentally unobservable, latent variables. Because such latent constructs cannot be directly observed or measured, they must be translated into measurable forms through indicators (questions or items)[7,8].

The variables of interest possess dual, yet complementary, definitions: a conceptual definition (nominal) ensuring the meaning within the study context and a construct definition (operational) guiding the measurement[7,8]. For example, when asking about smoking in a survey, one must first define that smoking in this context means consuming tobacco products (conceptual). Then, the operational definition would specify how to measure it, such as "Have you smoked at least 100 cigarettes in your lifetime?" or "How many cigarettes do you currently smoke per day?"

B. Item Formulation

Overarching Principles

Once the objectives and constructs are defined, survey questions or items are formulated. A comprehensive approach is to generate all potential questions pertinent to the objective and variables of interest, thereby maximizing the ability to tap into all important domains or themes[3]. Items can be generated through several structured methods such as thorough literature reviews, discussion through focus groups or interviews, often combining these approaches with input from potential respondents or subject matter experts[9].

Individual item construction should adhere to certain methodological constraints linked directly to the construct. These include ensuring that each question stem focus on a single construct. Items must be short, clear, and precise. However, many constructs (such as happiness or intelligence) are unobservable and complex, complex concepts should often be measured using multiple items or scales rather than a single question. When formulating items, researchers should carefully consider whether pre-existing measures or validated scales can be utilized, but only if they are well-written and appropriately suit the survey's specific needs[4]. Once maximum possible number of items generated, they can be organized into groups/categories that share similar themes[3].

Tailored-design Item Formulation

The construction of individual survey items is a critical source of measurement errors[7]. The goal of this phase is to structure questions that minimize misperception of individual items and their response options, inapplicable or biased responses by reducing ambiguous wording, confusing response formats, and jargon or language barriers[7]. A tailored-design approach emphasizes that item formulation must account for how respondents understand, interpret, and mentally process questions to ensure accurate and meaningful responses[4,7]. The process of generating a survey response typically involves four stages: (1) Comprehension, (2) Retrieval, (3) Judgment and (4) Response[10]. Being aware of potential measurement errors/biases that can occur in different stages (Table 1) can help better formulation of the items.

Table 1: Stages of the survey response process and common biases at each stage

Cognitive Stage	Associated Biases/Errors
Stage 1: Comprehension (Interpreting the question and relating it to relevant concepts)	• Misunderstanding due to complex wording • Inconsistent interpretation of jargon/technical terms • Confusion from double-barreled questions • Cognitive overload from negative wording (especially double negatives) • Response bias from leading/loaded questions • Varied interpretation of vague terms
Stage 2: Retrieval (Searching memory for specific information, using recall cues)	• Recall failure or inaccuracy • Memory decay over long time periods • Telescoping (misplacing events in time)
Stage 3: Judgment (Assessing information's relevance and integrating it)	• Overestimation in hypothetical scenarios • Social desirability bias • Acquiescence bias • Framing effects
Stage 4: Response (Formatting and selecting the final answer)	• Forced choice error (guessing when uncertain) • Imprecise categorization due to insufficient response options • Primacy/Recency bias • Inability to express unanticipated answers • Unclear scale meanings

Stage 1: Comprehension - Ensuring Question Clarity Rigorous item formulation begins with ensuring clarity and brevity, making the question easily comprehensible and unambiguous to the target population[11].

Brevity and Simplicity: To mitigate respondent burden and potential cognitive fatigue, questions should be precise and brief[7]. Question stems should generally target fewer than 20 words and employ clear, succinct, nonjudgmental language appropriate for the respondents' educational level[11]. Replace complex words with simpler alternatives, particularly scrutinizing words longer than six to seven letters[4].

Jargon and Terminology: Technical jargon, uncommon words, complex terminology, and abbreviations must be strictly avoided, as these increase the risk of inconsistent interpretation and potential alienation of respondents[11].

Avoiding Double/Multi-Barreled Questions: Combining two or more issues, respectively, in a single question confuses the respondent and also prevents the investigator from discerning which component the respondent actually answered[12].

Avoiding Negative Wording: The use of negative questions is discouraged, particularly those with double negatives as they confuse respondents, potentially leading to frustration and non-response[12]. Question like "Do you disagree that you don't exercise regularly?" is cognitively challenging and can cause respondents to make erratic response[13].

Avoiding Leading and Biased Questions: Leading or loaded questions contain emotionally charged words or are phrased to direct the respondent toward a specific answer, thereby creating an unnecessary bias that increases response errors[7] Questions containing absolute terms (e.g., "always", "none", or "never") should be avoided[3].

Mitigating Vagueness: Error is more likely when the question is vague, which makes it difficult for respondents to understand what was asked of them, potentially leading to misperception of the question. Vague questions such as "Do you have a healthy diet?", where "healthy" is a vague term and is vulnerable to varied subjective interpretation can have detrimental effects on both the validity and reliability of the data[14].

Ensuring Technical Accuracy: Questions must use correct units of measurement and specialized terms appropriate for the population, consulting with content experts when necessary to ensure precision.

Stage 2: Retrieval - Facilitating Memory Access
Measurement error often stems not from ill-intent, but from the inherent difficulty respondents face in recalling past memory and behaviors[6].

Type of Information Being Sought: Recognize that factual or demographic questions are generally easier for respondents to answer than attitude or opinion questions, which require more cognitive processing and are more susceptible to context effects[6]. Questions about behaviors and events are particularly prone to recall problems, as memory fades over time and people may not categorize events by exact timeframes. Improve recall accuracy by asking about recent or memorable events and providing clear definitions or examples.

Enhancing Retrieval: To maximize the quality of recall, questions should facilitate memory retrieval by posing a task or providing simple illustrative context that aligns closely with how information is organized in memory[15]. For example, a more effective way to rephrase a question like, "How many times did you feel chest pain in last month?" can be "In the last month, on how many separate occasions did you experience chest pain lasting 5 minutes or longer?"

Time Frame Specification: Questions requiring recall should specify clear, manageable time frames to reduce

memory burden and improve accuracy. For instance, asking about "the past week" typically yields more reliable responses than asking about "the past year"[15]. Otherwise, respondents may experience telescoping—misplacing events in time by either recalling past events as more recent (forward telescoping) or more distant (backward telescoping).

Stage 3: Judgement - Supporting Response Formation
Respondents need to assess the relevance of retrieved information and determine how to integrate it into their answer. Measurement errors can occur when questions require complex processing or ask respondents to make judgments beyond their actual experience.

Hypothetical vs. Behavioural Questions: Questions that rely on hypothetical situations (e.g., "If your doctor advised you to, would you be willing to quit smoking?") risk overstating outcomes by asking people to imagine scenarios they are not actually in[14]. The investigator must determine whether the purpose of the question is to collect data regarding a belief or an actual behaviour, and design the question accordingly[15].

Avoiding Framing Effects: The framing effect occurs when responses differ depending on whether a question is presented positively or negatively, even though the underlying information is identical. For example, asking "How confident are you that regular exercise helps prevent heart disease?" may yield higher agreement than "How confident are you that lack of exercise causes heart disease?", despite both assessing the same belief.

Minimizing Social Desirability Bias: Questions on sensitive topics should be worded to reduce pressure to provide socially desirable answers. Using normalizing language (e.g., "Many patients find it difficult to take medications as prescribed...") can encourage more honest responses and reduce judgment-related error.

Minimizing Acquiescence Bias: Respondents may exhibit a tendency to agree with statements regardless of content (acquiescence bias). Agree/Disagree, Yes/No, True/False, type questions are vulnerable to this type of bias. To support respondents to adequately think and then answer the item need to be formulated thoughtfully. For example, instead of a statement phrased as "I am very concerned about my heart health" (Strongly agree to Strongly disagree), consider "How concerned are you about your heart health?" (Extremely concerned to Extremely unconcerned)

Stage 4: Response – Formatting and Recording Answers
The format and characteristics of the response options profoundly influence measurement precision, especially regarding the choice between open-ended and closed-ended items, and the treatment of uncertainty.

Open vs. Closed Questions: The determination of question type—open-ended (free text) or closed-ended (structured options)—should be a direct result of how variables were conceptualized and constructed[12,16] Generally, the preferred option for efficient analysis is the closed-ended question, but this requires response categories to be exhaustively and mutually exclusively defined[7].

Handling Indeterminacy and Novelty: Including options such as "I don't know," "uncertain," or "no opinion" is crucial as they acknowledge respondent uncertainty or indecisiveness[3]. Forcing a choice when a respondent genuinely lacks knowledge can lead to mere guessing[14]. Likewise, providing an "other" response option allows for unanticipated answers, potentially identifying new issues, and facilitating respondent expression[3].

Order Effects (Primacy/Recency): In visually administered modes (like paper or web surveys), there is a tendency for respondents to favor options listed first (primacy bias). In telephone or other verbally administered surveys, respondents often favour options presented last (recency bias). These biases happen due to differences in attention and memory. Either way, respondents tend to forget or overlook items in the middle of the list because of increased cognitive load and task burden[15].

Response Format and Labelling: The specific response format (e.g., binary, nominal, ordinal, Likert, etc.) must be intentionally chosen based on the question stem, as it determines the subsequent type of statistical analysis[3]. For rating items, survey designers may want to label all response options (e.g., instead of rating medication adherence 1-5, explicitly label all five options as 1: No Adherence, 2: Low Adherence, 3: Moderate Adherence, 4: High Adherence, 5: Full Adherence) to help each choice seem equally viable and clarify its meaning, thereby improving item quality[13]. Insufficient categories can force imprecise choices, necessitating a scale with sufficient discriminating power[15].

Response Options for Rating scales: The optimal number of response options for rating items typically ranges between four and seven. For unipolar continua (like frequency of a behavior), five options may be optimal, while for bipolar continua (like measuring positive or negative attitudes), seven options may be ideal[13]. The categorization for Likert scales generally varies between 3 and 11 choices, with 5 and 7 classifications displaying better statistical properties for discriminating between responses[8].

[II] Questionnaire Composition and Contextual Effects
The next phase of questionnaire development transitions the focus from individual items to the structural, formatting, and flow characteristics of the questionnaire as a whole. This step is crucial because the architecture of the questionnaire, its composition and sequence, may introduce sources of substantial measurement error that are independent of individual item-related flaws[6].

A. Structure, Layout, and Flow

Survey respondents perceive the questionnaire not as a mere collection of discrete items but as a complete entity searching for an overall logical arc or narrative.

Logical Organization: Questions must be clearly numbered, organized into coherent topics, and arranged in a logical order to assist the respondent's thought processes and facilitate the flow of the questionnaire[3]. Mixing topics haphazardly can create unnecessary confusion and increase respondent burden[7].

Visual and Aesthetic Consistency: Maintaining a professional and consistent visual layout throughout the instrument is critical, as it efficiently directs the respondent's attention to vital information[13]. Font style and size should be easy to read and appropriate.

Satisficing and Fatigue: When questionnaires are too long, respondents may experience fatigue, leading them to provide "uniform and inaccurate answers", often exhibiting behaviors known as satisficing (putting forth suboptimal cognitive effort) bias[7,15].

Item Branching and Filtering Logic: Failure to use logical skip patterns or inclusion of non-applicable questions alienates respondents, breaks rapport, and increases the likelihood of item non-response or survey termination[15,17].

B. Contextual and Order Effects

In addition to easy and smooth navigation, the sequence and surrounding context of questions can also influence how people respond. These psychological factors are known as contextual and order effects[14].

Sponsorship Effects: The sponsor of the survey and the introduction establish an interpretive frame that influences how respondents process information, potentially reinforcing social desirability and acquiescence bias[18]. The sponsor should convey legitimacy to encourage participation and trust. Neutral framing should be the default to prevent suggesting a specific interpretation.

Assimilation Effects: This occur when the target question is unclear/unfamiliar to them, and the respondents take cues from nearby questions, leading to responses that converge or become consistent with the earlier cues[14]. For example, if a survey first asks detailed questions about chest pain and then later asks a general question about "cardiac symptoms," respondents might think mainly about chest pain and overlook other important symptoms such as shortness of breath or fatigue.

Contrast Effects: Contrast effects occur when a specific question precedes a general question, leading respondents to mentally exclude the specific domain from the general question, thus causing divergent responses[14]. For example, if asked "How satisfied are you with your blood pressure control?" immediately before "How satisfied are you with your overall cardiovascular health?", respondents may exclude blood pressure from their overall health assessment.

Priming Effects: The simple presence of a preceding question may activate certain concepts or memories and subsequently influence answers to the latter questions[17]. For example, if a question about respondents' weight follows a question about dietary habits, responses are likely to differ compared to when the weight question comes first, because thinking about their weight first may prime respondents to reconsider or modify how they report their dietary habits.

Strategic Placement of Items: The organization of items should prioritize respondent engagement and cognitive capacity[13]. Simple, interesting, or attention-grabbing questions should be placed first to ease the respondent into completion[12]. The most important items should be placed earlier in the survey to ensure they are answered before respondent focus or effort declines[13]. Sensitive items, such as detailed demographic questions, should generally be reserved for the end of the questionnaire[3].

[III] Questionnaire Testing and Refinement

This testing and refinement process encompasses pre-testing, pilot testing with psychometric validation, and evidence-based item reduction to produce a finalized instrument[3,14,19]. Pre-testing is a systematic process of reviewing and revising items to confirm that respondents interpret them as intended[8]. Pilot testing represents the empirical phase where the refined questionnaire is administered to a small representative sample to evaluate its practical functionality, identify operational issues, and generate data for formal psychometric validation. The qualitative and quantitative evidence from pre-testing and pilot testing, respectively, informs the design of a refined final instrument.

A. Pre-testing: Item Refinement and Quality Appraisal

1. Qualitative Item Refinement

Before subjecting the questionnaire to empirical field testing, items are scrutinised through qualitative methods to ensure content fidelity, operational clarity, and avoidance of redundancy.

Expert Consensus and Review: Experts evaluate the item pool to identify inappropriate or redundant items, determining if the key underlying constructs are being sufficiently addressed. Three types of experts should ideally be consulted: content experts (specialists in the substantive topic area), questionnaire design experts (survey methodologists), and analysis experts (statisticians who can ensure the data collected will support intended analyses)[20]. Experts may use ranking or rating scales to judge the suitability of individual items, often providing binary responses (include/exclude). This expert review

process helps reduce the initial item pool by eliminating clearly problematic items before data collection.

Cognitive Interviewing (CI): Techniques such as cognitive interviewing are vital for identifying flawed items. CI involves administering draft questions to respondents while concurrently collecting verbal information about respondent's interpretations and thought processes (through thinking aloud technique), enabling researchers to determine whether the question extracts the information the author intended[19]. This formative testing helps identify items that respondents find difficult, confusing, or irrelevant, leading to item deletion or rephrasing.

2. Individual Item Quality Assessment

BRUSO checklist: The BRUSO checklist provides a framework for writing effective survey questions by ensuring they adhere to five key principles[21]. Each question should be evaluated to ensure it is, i) Brief, ii) Relevant for the respondent and to the study objective(s). iii) Unambiguous terminology or wording, iv) Specific, narrowly focused and v) Objective, devoid of bias such as non-leading or loaded (Table 2).

Table 2: An example of application of BRUSO checklist for cardiovascular (CVS) health related questions

Criterion	Poor Example	Effective Example
Brief	"To the best of your recollection, could you please indicate the frequency with which you engage in cardiovascular exercise for a duration of at least 30 minutes?" (excessively wordy and complex language)	"How many times a week do you exercise for 30+ minutes?"
Relevant	"What type of TV do you have?" (Not relevant to the research topic asking about CVS health)	Do not include this item unless it is clearly relevant to the research (e.g., assessing sedentary behaviors).
Unambiguous	"Do you eat heart-healthy?" (Uses a subjective and vague term)	"In a typical week, on how many days do you eat vegetables?"
Specific	"Do you think a low-fat diet and regular exercise can reduce your risk of heart attack?" (a double-barreled question)	Break this into two separate questions: "How much can a low-fat diet reduce your risk of a heart attack?" "How important is getting 150 minutes of exercise per week for preventing heart disease?"
Objective	"Given that high blood pressure is a silent killer, how important is it to get it checked regularly?" (Uses leading and emotionally loaded language)	"How important is it to have your blood pressure checked once a year?"

3. Holistic Questionnaire Appraisal

Applying checklists to assess questionnaires: Apart from BRUSO for individual items, there are checklists available to evaluate both individual items and/or questionnaire as a whole. For example, the "survey checklist manifesto"13 functions as a quantitative scale to help researchers evaluate the quality of existing surveys or guide the creation of new items. The tool comprises 17 scorable items organized across three domains: the formulation of items (e.g., avoiding agree-disagree formats), the crafting of response options (e.g., ensuring all options are labeled), and the overall formatting and organization of the survey (e.g., placing important items earlier). This process serves as a final check to mitigate sources of measurement error before the instrument is piloted.

For a CVS or other health-related questionnaire Clinical Sensibility Testing (CST), checklist can be used to assess the comprehensiveness, clarity, and overall face validity of the questionnaire[3] This checklist involves administering a 1-page assessment sheet to respondents, posing items as questions with either Likert scale or nominal response formats (e.g., asking how many items were inappropriate or redundant)

B. Pilot Testing and Psychometric Validation

Pilot testing represents the empirical phase where the refined questionnaire is administered to a small representative sample (typically requiring a minimum of five respondents per candidate item) to evaluate its practical functionality, identify operational issues, and generate data for formal psychometric analysis and evidence-based item reduction[20,22].

Pilot Testing: Pilot testing involves running an empirical examination of the penultimate draft on a small sample of participants drawn from the general population of interest[22]. Key elements to examine include completion rates and time required, patterns of item nonresponse, response distributions (ceiling or floor effects, unexpected patterns), and functionality of skip logic and branching[19,20]. Data from the pilot study provides the empirical evidence necessary for statistical item reduction and psychometric validation.

Formal Psychometric Validation: Psychometrics provides the statistical assurance of quality, focusing on reliability and validity measures.

Reliability Assessment: Reliability refers to the consistency or stability of an instrument in its measurements[8]. The specific reliability assessment (test–retest, interrater, internal consistency) must be chosen based on the survey's objective and the format of the responses[3]. Test–retest reliability assesses whether the same question yields consistent results over time, requiring the administration of the final questionnaire to selected respondents on two occasions at least 2–6 weeks apart[1,3]. Internal consistency assesses the extent to which items within a domain are related to the same underlying concept and typically measured as Cronbach's α during pilot testing[8]. Low Cronbach's alpha (α) values may indicate problematic items requiring deletion, while excessively high values may suggest item redundancy.

Validity Assessment: Validity confirms that the instrument accurately measures what it purports to measure[22]. In

addition to face and content validity assessed through expert review and cognitive interviewing (as described in Section A), construct validity and criterion validity can be evaluated statistically from pilot testing data. Construct validity, which evaluates whether the instrument measures the theoretical construct it was designed to measure, can be assessed using factor analysis (exploratory or confirmatory) to determine if items cluster into expected dimensions[8]. Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) examine how items cluster together into underlying factors or constructs. Items demonstrating weak correlation with the hypothesized construct (low factor loadings) or cross-loading onto multiple factors become candidates for removal. If statistical analysis reveals multidimensionality in an item set meant to measure a single construct, the measure lacks internal validity. Advanced models from modern test theory, such as Item Response Theory (IRT) and Structural Equation Modeling (SEM), offer powerful tools for tackling these dimensionality challenges[11,14]. Construct validity also includes convergent validity (demonstrating correlation with related measures) and discriminant validity (showing weak correlation with unrelated constructs). Further, criterion validity assesses how well the instrument correlates with a gold standard or external criterion, including concurrent validity (correlation with established measures administered simultaneously) and predictive validity (ability to predict future outcomes, such as whether a cardiovascular risk questionnaire predicts subsequent cardiac events)[8].

Assessment of Item Discrimination: Floor and ceiling effects occur when responses cluster overwhelmingly at the minimum (floor) or maximum (ceiling) of the scale[23]. These clustering patterns, identified through examining response distributions in pilot data, indicate that an item fails to capture sufficient variance in the population and is therefore poorly discriminating. Items exhibiting these effects are strong candidates for removal or modification of the response format.

C. Item Reduction and Finalization

Following pre-testing and pilot testing, the process of item reduction synthesizes qualitative and quantitative evidence to make final decisions about which items to retain, revise, or eliminate, ultimately producing a parsimonious yet valid instrument. This critical phase balances the need for comprehensive construct coverage against the imperative to minimize respondent burden, with most research objectives adequately addressed by 25 or fewer items[24].

Item reduction decisions should integrate qualitative findings (expert consensus, cognitive interviewing results, conceptual overlap assessments) with quantitative indicators (factor loadings, inter-item correlations, internal consistency measures, floor/ceiling effects, and response patterns). Priority should be given to retaining items that demonstrate strong construct alignment, adequate response variance, expert-judged essentiality, and minimal respondent burden while maintaining measurement quality. The finalization process concludes with a refined instrument that has been systematically reduced based on empirical evidence while preserving the scientific integrity required to validly measure the intended constructs. This final questionnaire is now ready for full-scale deployment in the target population.

Conclusion

Developing a reliable and valid questionnaire requires a holistic tailored design approach that considers all aspects of the respondent's experience and how these elements interact to influence their ability to provide accurate and meaningful answers. Such questionnaire construction is a complex process that demands intellectual rigour, substantial effort, and strict adherence to systematic, iterative procedures. This process transcends mere item drafting; it necessitates methodical transitions from defining abstract theoretical concepts to generating precise, empirically measurable constructs. The commitment to continuous drafting, testing, and revising is the most likely strategy to ensure the final tool is maximally answerable and achieves its research objective. By following the systematic steps outlined here—from meticulous item generation and formulation, through strategic questionnaire composition and managing contextual/order effects, to systematic testing and refinement via pre-testing, pilot testing, and evidence-based item reduction, researchers can significantly minimize errors and bias, thereby ensuring the resultant data are high-quality, reliable, and valid, ultimately advancing knowledge in their field.

Conflict of interest: None.
Funding: The authors have no relevant financial or non-financial interests to disclose.

Declarations: We (the authors) take full responsibility for the content of this paper. We acknowledge the use of AI (editgpt) for assistance with English language editing. We used prompts to improve the structure of sentences that we deemed could be further improved. AI was prompted to improve clarity by improving grammar and choice of vocabularies used in the text. All suggestions were critically reviewed and revised to ensure the integrity of our own expressions. The authors have no conflicts of interest, relevant funding, employment, financial or non-financial interests to declare.

References

1. Beatty P. Developing Measures of Health Behavior and Health Service Utilization. In: Johnson TP (ed) Handbook of Health Survey Methods. Hoboken, NJ: John Wiley & Sons, Inc., 2015, pp. 171–190.
2. Ren J, Guo XL, Lu ZL, Zhang JY, Tang JL, Chen X, et al. Ideal cardiovascular health status and its association with socioeconomic

factors in Chinese adults in Shandong, China. BMC Public Health. 2016;16(1):942.

3. Burns KEA, Duffett M, Kho ME, Meade MO, Adhikari NKJ, Sinuff T, et al. A guide for the design and conduct of self-administered surveys of clinicians. CMAJ. 2008;179 (3):245-252.
4. Dillman D, Smyth J, Christian L. The fundamentals of writing questions. Internet Phone Mail Mix-Mode Surv Tailored Des Method 4th Ed Hoboken N J Wiley.
5. Biemer PP. Total Survey Error: Design, Implementation, and Evaluation. Public Opin Q 2010; 74: 817–848.
6. Bautista R, Bilgen I, Truesdale D. Design and evaluation of survey questions. SAGE Publications Ltd, 2020.
7. Gideon L. The Art of Question Phrasing. In: Gideon L (ed) Handbook of Survey Methodology for the Social Sciences. New York: Springer, 2012, pp. 91–107.
8. Kishore K, Jaswal V, Kulkarni V, De D. Practical Guidelines to Develop and Evaluate a Questionnaire. Indian Dermatol Online J. 2021 Mar 2;12(2):266-275.
9. Walters WH. Survey design, sampling, and significance testing: Key issues. J Acad Librariansh 2021; 47: 102344.
10. Tourangeau R, Rips LJ, Rasinski K. The Psychology of Survey Response. Cambridge: Cambridge University Press. 05 June 2012. doi: 10.1017/CBO9780511819322
11. Neuman WL. Survey Research. In: Neuman WL (ed) Social Research Methods: Qualitative and Quantitative Approaches. Harlow: Pearson Education Limited, 2014, pp. 315–366.
12. Slattery EL, Voelker CC, Nussenbaum B, Rich JT, Paniello RC, Neely JG. A practical guide to surveys and questionnaires. Otolaryngol Head Neck Surg. 2011;144(6):831-837.
13. Gehlbach H, Artino Jr AR. The survey checklist (manifesto). Acad Med 2018; 93: 360–366.
14. Yang S-M, Hinkle JC. Issues in Survey Design: Using Surveys of Victimization and Fear of Crime as Examples. In: Gideon L (ed) Handbook of Survey Methodology for the Social Sciences. New York, NY: Springer, pp. 443–462.
15. Choi B, Pak A. A catalog of biases in questionnaire research. Prev Chronic Dis Health Res Pract Policy; 2.
16. Neuman WL. Qualitative and Quantitative Measurement. In: Neuman WL (ed) Social Research Methods: Qualitative and Quantitative Approaches. Harlow: Pearson Education Limited, 2014, pp. 201–244.
17. Strack F. "Order Effects" in Survey Research: Activation and Information Functions of Preceding Questions. In: Schwarz N, Sudman S (eds) Context Effects in Social and Psychological Research. New York, NY: Springer New York, pp. 23–34.
18. Shine B, Dulisse B. Does Paying More Mean Getting a Better Product: Comparison of Modes of Survey Administration. In: Handbook of Survey Methodology for the Social Sciences. 2012, pp. 361–373.
19. Willis G. Pretesting of Health Survey Questionnaires: Cognitive Interviewing, Usability Testing, and Behavior Coding. In: Johnson TP (ed) Handbook of Health Survey Methods. Hoboken, NJ: John Wiley & Sons, Inc., 2015, pp. 217–241.
20. Dillman D, Smyth J, Christian L. Ordering questions and testing for question order effects. Internet Phone Mail Mix-Mode Surv Tailored Des Method Wiley 2014; 228–257.
21. Peterson RA. Constructing effective questionnaires. SAGE Publications, https://cir.nii.ac.jp/crid/1970023484883700765 (2000).
22. Kelley K, Clark B, Brown V, Sitzia J. Good practice in the conduct and reporting of survey research. Int J Qual Health Care. 2003;15(3):261-266.
23. Rockwood T. Assessing physical health. Health Surv Methods 2014; 105–141.
24. Sharma H. How short or long should be a questionnaire for any research? Researchers dilemma in deciding the appropriate questionnaire length. Saudi J Anaesth 2022; 16: 65–68.