Question Wording and Item Formulation

In: J. Matthes, R. Potter, & C. S. Davis (Eds.), *International Encyclopedia of Communication Research Methods*. Wiley-Blackwell.

Malte Elson
Educational Psychology Research Group
Ruhr University Bochum
malte.elson@rub.de

Abstract

In most empirical research with human subjects, at least some data are collected by recording self-reported responses. However, the overlap between exhibiting a behavior and reporting about the same is less than perfect. Responding to self-report instruments is itself a behavior governed by a number of factors, such as item wording, that can bias data and inferences. Thus, the methodology underlying the formulation of questions is essential to many topics in communication research. Researchers can adopt strategies to maximize the diagnostic value of their items and scales, such as avoiding item order effects, selecting an appropriate response format, reducing ambiguity in the wording, and controlling for careless responding.

Keywords: Scale development, item phrasing, survey methods, questionnaires, self-report

Self-report data are ubiquitous in communication research, the general social sciences, and beyond. In most empirical research with human subjects, at least some data are collected not through direct observation of a behavior or other variable of interest, but by prompting study subjects with questions or items (these terms are used interchangeably here) in a given format and recording their immediate response to them. Individual interviews and focus groups, computerized test batteries and street intercept surveys, standardized screenings and assessment centers, clinical interviews and witness reports - all rely on the principle of gaining insight into a human characteristic, their comprehension of a phenomenon, or relationship between variables by querying the source in one way or another. Thus, self-report data and their foundation - the methodology underlying the construction of questions and items to which subjects respond - are essential to many topics

in communication research.

Responses to items are, broadly speaking, recorded for three different purposes: They can be used as *predictor* variables, i.e. observables that are thought to deterministically affect another variable. They can be used as *outcome* variables, i.e. observables that are thought to be influenced or induced by another variable or experimental procedure. Finally, they can be used as *covariates*, i.e. observables that are thought to play a meaningful role in the relationship between predictor and outcome variables. *Manipulation checks* constitute a special case of self-report items as they usually inform the researcher about the success of an experimental procedure, but are rarely used to predict outcomes themselves. In communication research, it is not uncommon to find all types are used simultaneously in an empirical study.

There are two dominating paradigms for the design, analysis, and scoring of psychometric tests. In Classical Test Theory, the psychometric quality of a test or scale is commonly judged by proof for its *objectivity*, *reliability*, and *validity*. Accordingly, each individual item that a scale comprises should be considered a diagnostic instrument on its own, whose usefulness accordingly can be evaluated on these three criteria. The more recent psychometric Item Response Theory (IRT) models the relationship between the response a test item and an overall measure (the instrument) of the characteristic that item was designed to measure. Self-reported responses are a mathematical function of two components or parameters: person characteristics (e.g. abilities or traits) and item characteristics (such as the wording).

Notably, there are many types or variations of such queries of respondents, yet they all share a critical quality which may limit the usefulness of data they yield: responding to self-report instruments is itself a behavior governed by a number of factors, of which only one is the variable (or construct) that the given items are supposed to model (the "true value"). A number of known and unknown influences ("error"), such as the wording and order of items, can bias or pollute response behavior if not controlled for.

Plainly speaking, the overlap between exhibiting a behavior and reporting about the same is less than perfect, although a substantial proportion of communication research relies on using the latter as a proxy for the former. It is this disparity that researchers aim to minimize with the methodology of question wording and item formulation. There are several strategies researchers can adopt to maximize the diagnostic value of their items and scales, of which some will be briefly discussed in the following.

Order v. Variety

One principal decision of every researcher collecting self-report data is the positioning of items, i.e. the order in which they are presented. An item order effect occurs whenever the response to one item influences the responses to a succeeding item or set of items. Item order effects can occur locally, i.e. in a topical item sequence or group, and globally, affecting subjects' overall response behavior.

Generally, self-report instruments should unfold much like an actual conversation between the researcher and the respondent. If possible, items should have a natural, coherent order in which one item (or set of items) logically leads to the next. Questions about the same subject, experience, or entity should be grouped together, as repeated shifting between topics and returning to questions already answered can be confusing and frustrating to study participants. Ideally, items grouped by topic should also exhibit a consistent response format. However, continuously responding in a highly repetitive format (or to repetitively worded items) can be exhausting, and cause subjects to pay less attention to the exact wording of questions (increasing probability of *careless responses*, see below). Therefore, response accuracy can often be increased by occasionally alternating the task or type of item subjects are asked to respond to (e.g. present an open-ended question after a series of closed-ended ones). Item order effects deserve particular consideration in longitudinal research, as a change to the item order between measurement intervals might cause differential responses.

In order to initiate and establish rapport with study participants, researchers might place simple opening questions that are easy to answer and do not touch on complicated or sensitive subjects at the beginning of the instrument. This can be a useful device to increase their involvement in the research and motivate them to complete the self-report instrument diligently and with necessary care. Consequently, items (and topics) should be ordered from easy to difficult throughout the whole survey.

Occasionally, communication researchers prompt subjects with a general question about an experience or event and one or multiple specific questions (for example, overall satisfaction and satisfaction with particulars). The order of these items might bias participants' response behavior. One potential effect of placing the general question before the particulars is that respondents might only consider a limited number of aspects relevant to form an overall evaluation. Thus, prompting the particular items first can be a useful device when a large number of factors can determine the general response.

That being said, researchers should be aware of two important effects that might occur as a result of specific items preceding a general item: assimilation, where the responses to the general item becomes more similar to that of the particulars, and contrast, where there are greater differences between the responses to the general item and the particulars (Schwarz, Strack, & Mai, 1991). Under item assimilation, subjects form their general response as a summary of the listed particulars (discounting other particulars that might exist, but were not included by the researcher). Conversely, an item contrast is present when respondents consider the general question as a residual category (consisting only of particulars *other than* those listed before).

Contingencies under which item assimilation or contrast occur are not sufficiently established. In principal, a higher number of particulars might increase probability of assimilation, while a rather low number of particulars (one or two) can elicit a contrast effect (as a summative response to the general question would be highly redundant with the particulars). However, there is some debate about the replicability of item order effects, and whether their magnitude has any practical significance (Schimmack & Oishi, 2005). Researchers are advised to prompt subjects with specific instructions how to form their

general response (e.g. *"Taking these aspects together..."* or *"Aside from these aspects..."*).

Closed-Ended v. Open-Ended

One basic characteristic of all items is the degree by which subjects' response is restricted to a given format. There are merits and disadvantages to each alternative that warrant careful consideration.

For a *Closed-Ended Question*, subjects are presented with a fixed set of response options to select from (e.g. multiple choice or Likert-type items). This allows indicating a response promptly and without having to conceive a formulation for an answer, which would potentially increase the ambiguity of responses. Thus, closed-ended questions are most useful when there is a limited number of possible responses to an item, the simplest form being a dichotomy (*"Do you own a hat?"*). They can also increase the accessibility of responses subjects might either not have considered on their own, or that they would have some trouble remembering (*"What is your favorite type of hat? Beret, Tricorne, Ushanka..."*). For researchers, closed-ended questions have utility because responses are collected swiftly and do not have to be subsequently transcribed and coded. Indeed, they can be sorted, categorized, grouped, and quantified effortlessly for further statistical analyses.

While this makes them arguably efficient, closed-ended questions come at a serious disadvantage: In their response behavior, subjects are restricted to a set of alternatives predefined by the researcher, and thus they might not be given the opportunity to provide their actual or preferred response (e.g., when *"Panama Hat"* is missing from the options). Some participants will approximate (i.e. give the answer that seems "closest", e.g. *"Fedora"*), some will not respond at all (reducing the amount of useful data) - but in the worst case, the researcher unknowingly collects a false response (i.e. an answer that is not close at all to the preferred response, e.g. *"Toque"*), yielding false data that might prompt false inferences.

Open-Ended Questions are a viable solution to many of the previously raised concerns. Here, subjects are not restricted to a selection of answers, but given the opportunity to respond in their own words. This is useful when some variability in responses across a sample is expected (*"What do you like most about hats?"*). It allows subjects to emphasize relevant parts of a response, to clarify remarks, and to qualify their understanding of the questions being asked. Responses to open-ended questions usually turn out a lot more diverse than anticipated, and thus enable exploratory research and generation of new hypotheses.

Naturally, preparing open responses for analyses is laborious, and usually requires iterative coder training to reduce, interpret, and classify data into a meaningful category system. This may introduce error in the processed data and reduce instrumental objectivity and reliability. Open-ended questions might have limited utility when they require pristine memory of a situation or experience (*"What did you feel when you wore a hat for the first time?"*). Sometimes, where response options could provide some context, respondents will answer open-ended questions in ways not intended by the researcher, or be overwhelmed by the amount of possible answers that need consideration (*"What is the best opportunity to*

wear hats?"). Further, participants might refuse to verbalize responses (e.g. to sensitive or embarrassing questions), or simply fail to articulate themselves sufficiently, only providing partially usable data.

Some remedy for the idiosyncratic limitations of closed-ended and open-ended questions can be found in questions with an *Open Response Option*. Here, similar to a closed-ended question, subjects are presented with predefined options, but they can choose to add their own responses in one or multiple open text fields (often labelled as *"Other"*). This can be a compromise between the two approaches, and it is most feasible when there are options known to be common or likely (which subjects can then choose from), but there could be unknown, less common responses. Essentially, this reduces the amount of coding necessary after data collection if it was an open-ended question without limiting the alternatives to the imagination of the researcher.

Precision v. Parsimony

The inverse relationship between the dogmas of "keeping it short" and "keeping it elaborate" can best be solved by the following rule of thumb: One should always try to phrase items as accurately as necessary and as briefly as possible. Parsimony is an important characteristic that can be applied to almost every part of the test material, from initial instructions to response options in closed-ended questions. It should not only govern the length and number of response options to choose from, but also the amount of information conveyed by each item.

At the same time, researchers need to exert precision and avoid ambiguity: Items and response options need to be sufficiently comprehensible so that the interpretation of what might be asked is not left to the subject. Depending on the individual that is being asked, a question as simple as *"How many hats do you own?"*, for example, might have a lot of ambiguity (*"Does it include caps? Bonnets? Other headgear, e.g. my trusty ICA2012 visor?"*). Ambiguous items run the risk of eliciting a large variability in responses, effectively decreasing the test's reliability.

Other item characteristics further increase ambiguity, or inadvertently bias responses in a direction. Items should be exhaustive, yet disjunctive: Researchers must be careful to avoid *double-barreled questions*, i.e. including two separate entities in the same item while only allowing for one answer (*"Do you ever wear top hats and crocs?"*), as this renders meaningful inferences virtually impossible. Here, a useful heuristic is to look out for the grammatical conjunction *"and"* (note that the same applies to double-barreled response options in closed-ended questions). *Either-or questions* that consist of two complementing options can also increase data noise (*"Do you think hats are great, or do you lack a fashion sense?"*), unless the researcher intentionally restricts alternatives, or only two possible responses truly exist. As a general rule of thumb, each relevant bit of information should be inquired separately with its own item.

Negatively worded questions can confuse respondents and are prone to yield incorrect answers, increasingly so when they are part of a large question battery. Confusion can

also be created when questions indicate a vague amount, frequency, or other quantification of an event or experience (*"Do you wear hats often?"*) that should rather be part of the response (instead ask *"At which frequency do you wear hats?"*). Subjects can have different understandings of what exactly such a quantifier means, but are usually not given an opportunity to qualify their response. Researchers who insist on using *fuzzy quantifiers* should consider providing a reference point or criterion to compare it to. Another error that might be easily committed is formulating *leading* or *suggestive questions* (*"Do you agree that everyone should wear hats?"*), which can either prompt subjects to respond uniformly and/or in a way that conforms to hypotheses, rendering the item useless.

There are two types of items that can yield data which might look regular "on the surface" (and even metrically), but usually have little informational value and rarely allow meaningful inferences: These are items the response to which requires knowledge participants do not possess, or questions about hypothetical situations (*"Would you wear multiple hats if you could?"*). In both cases, subjects might refuse to provide a response, or if they do, there is a certain risk it will be uninformed. Researchers need to make sure that all information necessary to give an informed response are available, including knowledge of characteristics that might be relevant in a given hypothetical situation.

Finally, the evidential value of items asking respondents to indicate their agreement with statements (usually on a Likert-type scale) can suffer from multiple limitations, two of which are emphasized here: Researchers can easily commit the inferential fallacy of treating the indicated disagreement with a given statement equal to the agreement with its semantic opposite. For example, respondents' recorded (dis-)agreement with the statement *"I like hats"* has limited informational value to the question of how much people like hats since it is not possible to indicate one dislikes them – the scale is essentially truncated. Instead of agreement to a fixed characteristic, researchers should consider phrasing items that can be responded to on a bipolar scale with semantic opposites of the characteristic in each extreme point (e.g., *"like"* and *"dislike"*), a so-called *semantic differential*.

Another source of inaccuracy in questions of agreement is *acquiescence bias* (or *yea-sayer effect*), the tendency of survey respondents to express agreement with items or to indicate a positive connotation. A good example for this bias is the finding that, more often than not, positive and negative affect towards a stimulus are not (strongly) negatively correlated, although an underlying bipolarity of the two constructs is assumed. Item formulation rules can help to minimize this bias: Again, the use of semantic differentials can be useful, but researchers will often find this impractical or limiting to their research questions. Quantification strategies for acquiescence bias have been suggested (e.g., Hinz, Michalski, Schwarz, & Herzberg, 2007) to statistically account for this phenomenon in a given sample.

Fairness, Sensitiveness, and Carelessness

When working on the wording for a new scale, or even a single question, researchers should be attentive to *item bias* (or *item fairness*). In abstract terms within the framework

of item response theory, item bias is present when the item characteristic curves of two groups do not coincide (Mellenbergh, 1989), or, more practically, when its characteristics (such as the wording) cause systematically differential responses in individuals of the same actual opinion, ability, or trait, but with different ethnicity, culture, sex, age, or socioeconomic status. Wording in particular can contribute to item bias when it contains language that subsamples are differentially familiar with, phrases that are differentially common in areas of a country, or terms that have differential meanings and connotations within subcultures or socioeconomic classes of a population. An example of this would be to compare populations differentially competent at understanding figures of speech by using items that contain idioms (*"Are you good at keeping something under your hat?"*).

Naturally, the potential of an item or scale being unfair or biased grows in increments of populations to which it is applied. Cross-cultural studies are prone to this problem by nature, and various measures to develop *culture fair items* have been suggested, all with specific merits and disadvantages. Test materials used in the Program of International Student Assessment (PISA), a worldwide study comparing adolescents' scholastic performance, for example, are developed by a consortium of researchers from multiple participating countries. This process ensures cultural differences are taken into account, but comes at the cost of being relatively laborious and resource-intensive. Intelligence researchers have taken a different approach, as they frequently rescind the use of verbal items in favor of completely nonverbal tests. Which of these (and other) approaches is feasible in communication research depends on multiple idiosyncratic and domain-specific considerations to be taken into account.

Sometimes, communication researchers study aspects of human nature by prompting respondents with delicate or awkward questions (e.g., drug habits, sexual preferences). Such *sensitive items* can elicit a number of biases in subjects' self-disclosure. The most common problem researchers will face is non-response, i.e. subjects refusing to provide sensitive information. Aside from the decreased amount of data collected, caution is advised when drawing inferences since item non-response can induce *selection bias* in the sample as the population of subjects willingly responding to sensitive questions might be qualitatively different from those who do not (this might be indicated, for example, by a non-normal distribution of responses to those items).

An arguably more severe bias in sensitive items is *social desirability*. Socially desirable responses occur mostly for two reasons: Either because subjects comply with the researchers' predictions (a form of *demand effects*), or because the (untruthful) response will cast a more positive light on them. While particularly common when subjects are asked to admitting to embarrassing or illegal behaviors, social desirability bias is ubiquitous in self-report (Fisher & Katz, 2000). Sensitive questions can also affect responses to subsequent, non-sensitive items by diminishing subjects' compliance or inclination to self-disclose to the researcher. Therefore, researchers should mind the order in which they prompt such items, and need to build a rapport with the respondents to minimize response bias (see above). Some researchers control for it with specifically designed social desirability scales; this, however, constitutes a tautological fallacy since response behavior to such scales is, of course, also shaped by social desirability.

In contrast to participants caring too much about what others will think of their responses, there are also those that care too little: Any set of self-report data will contain a certain proportion of *careless responses*. This occurs when subjects are inattentive, exhausted, or unmotivated to participate in a research project, for example when participation is mandatory for course credit. Research settings in which subjects are largely unmonitored, such as anonymous web-based surveys, may be prone in particular to careless responding. Researchers can decrease this problem by designing their survey in a way that is varied, interesting, and not too exhausting (e.g., using diversified phrases, limiting the number of items per page, avoid overstraining subjects' patience and goodwill). Several tools are at researchers' disposal to identify careless responses (Meade & Craig, 2012), such as special items designed to detect inattentiveness, response time thresholds, response consistency indices, and self-reported diligence and seriousness checks (Aust, Diedenhofen, Ullrich, & Musch, 2013).

References

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*(2), 527–535. doi:10.3758/s13428-012-0265-2

Fisher, R. J. & Katz, J. E. (2000). Social-desirability bias and the validity of self-reported values. *Psychology and Marketing*, *17*(2), 105–120. doi:10.1002/(SICI)1520-6793(200002)17:2<105::AID-MAR3>3.0.CO;2-9

Hinz, A., Michalski, D., Schwarz, R., & Herzberg, P. Y. (2007). The acquiescence effect in responding to a questionnaire. *Psycho-Social Medicine*, *4*, Doc07.

Meade, A. W. & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. doi:10.1037/a0028085

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127–143. doi:10.1016/0883-0355(89)90002-5

Schimmack, U. & Oishi, S. (2005). The influence of chronically and temporarily accessible information on life satisfaction judgments. *Journal of Personality and Social Psychology*, *89*(3), 395–406. doi:10.1037/0022-3514.89.3.395

Schwarz, N., Strack, F., & Mai, H.-P. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly*, *55*(1), 3–23. doi:10.1086/269239

Further Reading

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science, 2*(4), 396–403. doi:10.1111/j.1745-6916.2007.00051.x

Chesney, T., & Penny, K. (2013). The impact of repeated lying on survey results. *SAGE Open, 3*(1), 1–9. doi:10.1177/2158244012472345

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

Hendrick, T. A. M., Fischer, A. R. H., Tobi, H., & Frewer, L. J. (2013). Self-reported attitude scales: Current practice in adequate assessment of reliability, validity, and dimensionality. *Journal of Applied Social Psychology, 43*(7), 1538–1552. doi:10.1111/jasp.12147

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93–105. doi:10.1037/0003-066X.54.2.93

Stone, A. A., Bachrach, C. A., Jobe, J. B., Kurtzman, H. S., & Cain, V. S. (Eds.). (1999). *The science of self-report: Implications for research and practice.* New York, NY: Psychology Press.

Zijlstra, W. P., van der Ark, L. A., & Sijtsma, K. (2010). Outliers in questionnaire data: Can they be detected and should they be removed? *Journal of Educational and Behavioral Statistics, 36*(2), 186–212. doi:10.3102/1076998610366263

Author Biography

Malte Elson is a postdoctoral researcher in the educational psychology research group at Ruhr University Bochum (Germany) where he studies human learning. His research interests are the use of technology for education and learning, as well as social science methods and methodology. He is a dedicated proponent of open science practices in psychology and communication research, and is currently conducting research on academic peer review.