Questionnaire

Investigating subjective exhibition experiences via appropriate questions

Eva Specker and Helmut Leder

Introduction

Questionnaires are a classic method in psychology, museum studies, and other (related) disciplines. The major benefit of questionnaires is that they allow for a direct assessment of subjective experiences, which can be harder to assess with other methods. For example, if we want to know if a museum visitor likes an artwork, asking them if they like the artwork seems a straightforward and valid measurement (assuming the viewer does not lie in their response). This doesn't mean that other methods are not available. For example, we could also assess viewing time (Leder, Mitrovic, and Goller 2016) or bigger pupil size (Kuchinke et al. 2009), which are also associated with liking. However, many other factors (apart from liking) may influence viewing time, meaning this can only be an indirect measure. This is true for many physiological measures. Moreover, collecting data about people's experiences by employing questionnaires also allows for more differentiated information than any other measure. For example, we can assess positive and negative emotional responses by measuring facial muscles, smiles and frowns (Gerger et al. 2011). A questionnaire however enables us to distinguish and map many dozens of different emotional responses (Schindler et al. 2017). As far as we are interested in assessing aesthetic experiences, questionnaires generally allow the most direct assessment of these experiences. When thinking specifically about the implementation in an exhibition setting, questionnaires also have the benefit of being cost-effective (no expensive apparatus is required), time-effective, and relatively easy to implement.

However, in many ways, questionnaires can seem deceptively simple (e.g. Simms 2008). As most of us use language every day, making a questionnaire can appear as simple as typing a few questions on a page. In this chapter, we explain that it is worth spending time on developing the questionnaires used in research to ensure measurement validity. We will focus on quantitative measurement. Though questionnaires are flexible and can include open questions that will allow for qual-

itative analyses, this will not be our focus. In addition, questionnaires can be used in the form of tests (where we can answer questions correctly or incorrectly), for example when testing art knowledge (e.g. Specker 2021, Specker, Cotter & Kim 2023, Specker, Forster, et al. 2020). Though this is an option frequently used, this chapter focuses on questionnaires assessing the subjective experience (in the broadest sense) of exhibitions.

For the purpose of this chapter, we mainly focus on item development (i.e. developing your own questions) rather than other aspects of scale/questionnaire development (see e.g. Gehlbach and Brinkworth 2011 for a more extensive overview). This focus was chosen as we aim to enable researchers to develop items/questions that fit specific needs for their individual study rather than enable readers to get in-depth knowledge about scale/questionnaire development. What is worth noting here is that for single questions (e.g. 'How much do you like this artwork?') statistical validity evidence cannot be given and thus assessment of single questions relies mainly on face validity (Does the question seem to measure what we aim to measure?). Nonetheless, the validity of such single-item measurements can be improved by how these items are phrased, hence our focus on item development in the current chapter.

Aim of the method

In most cases, questionnaires are used as a measure of the dependent variable. What is crucial here is that in (quasi-)experimental design, we have an independent variable (or a 'cause') and a dependent variable (or 'the effect'). As noted above, this chapter focuses on questionnaires assessing subjective experience (in the broadest sense) as a dependent variable in a museum study. But questionnaires can also be used to measure other dependent variables as well as independent variables. In all cases, what is crucial is that the measurement is valid, which means that we measure what we want to measure. This is often easier said than done, and testing the validity of a questionnaire generally takes a lot of time and effort (e.g. Specker 2021, Specker, Cotter, and Kim 2023, Specker, Forster et al. 2020).

Step-by-step guideline

To use questionnaires, the first step would be to consider if your theoretical construct can be measured by a questionnaire and if there already exists a questionnaire that you could use (step 1 below). If this is the case, then steps 2–4 can be followed to construct the questionnaire.

Step 1: First you need to define the theoretical construct that you want to measure (Gehlbach and Brinkworth 2011). What is it that you want to measure? Liking? Beauty? Stress? After this has been decided, you can then consult the literature to see if there are any validated scales. Using a validated scale will ensure that your measurement is valid and make your work more comparable to other research. If no validated scale exists, you can assess how others have measured this in the past. In sum: why reinvent the wheel (Clark and Watson 1995)? That said, you may not find anything suitable, for example, because you want to ask something about the specific exhibition or museum that you are testing in.

Therefore, we will now turn our focus to item development. Some advice may seem intuitive in principle but tends to be hard in practice. In general, these practices have been empirically studied and are based on best-practice advice (e.g. Gehlbach and Brinkworth 2011, Simms 2008, for short overviews). Furthermore, we focus here on a description in easy-to-follow, jargon-free language with examples focused towards museum and exhibition studies.

Step 2: After you have decided what you want to measure, you have to make decisions on how to measure it. When writing questions yourself, the first thing to decide is the format of your questions. Should they be in the form of declarative statements, i.e. will you ask 'How much do you like this artwork?' or will you ask 'To what extent do you agree with the following statement: I like this artwork'? Please be aware that in principle, it is advised to avoid agree/disagree response formats (e.g. Krosnick 1999). This is because this is a cognitively demanding task that therefore can increases both error and effort on the side of the participant (Gehlbach and Brinkworth 2011), especially in the cases of item reversal or linguistic negation (i.e. double-negatives; Swain, Weathers and Niedrich 2008). For the same reason, it is advisable to stick to one question format if you plan to ask multiple questions about the same construct. If you switch, i.e. you use a question for artwork 1 and a declarative statement for artwork 2, this will be confusing to the participant of your study. It will be easier (and quicker) for participants to fill out your scale (accurately) when the structure of the questions is consistent.

For this reason, it is also advisable to keep the answering pattern the same. For example, say you use a 7-point scale, then it is much clearer if 1 always means e.g. 'not at all' and 7 always means 'very much'. Sometimes, changes in the format are chosen in order to "make the participant pay attention" or "keep participants honest", i.e. to avoid having participants respond to all questions the same (Gehlbach and Brinkworth 2011). But, in practice, changing scales, such as reversed scored items, reduces reliability (e.g. Benson and Hocevar 1985) and causes issues of "misresponse" (Swain, Weathers, and Niedrich 2008). Note that this applies to both how the items are formulated as well as the actual way the response is formulated. Specifically, items should be formulated in a way that a high rating, for instance, ('very much')

always indicates a positive response (e.g. liking an exhibition) rather than that the meaning switches.

Step 3: When starting to formulate your questions the main goal should be to be clear in how you phrase them (Clark & Watson 1995, Gehlbach & Brinkworth 2011, Simms 2008). This is easier said than done. We advise to

1. deal with only one central thought in each item
2. be precise
3. be brief
4. avoid awkward wording or dangling constructs
5. avoid irrelevant information
6. present items in positive language and avoid double negatives
7. avoid items like 'all' and 'none'
8. avoid items like 'frequently' and 'sometimes'

Why? Many of these rules adhere to how humans comprehend language. For example, double negation is hard to understand (e.g. Benson and Hocevar 1985). In other words, unclear questions make it harder for participants to accurately respond to your questions and for you as a researcher to interpret the resulting data. This is also the main aspect that point 7 and 8 are focused on. For example, 'I was mainly interested in the paintings'. This is a classic example of a question that is double-barreled: a question that assesses more than one thing (i.e. does not follow recommendation 1 above). Another example would be if we ask the participant to respond to the statement 'my museum experience was informative and pleasurable'. The participant may be at a loss as to what to answer when the experience was only informative or only pleasurable but not both.

What is relevant is: What do you want to measure? (see also step 1). If you are interested in knowing if participants are relatively more interested in paintings or sculptures, then it may make sense to ask this directly, "Where you more interested in paintings or in sculptures?" (like Reitstätter et al. 2020) with, for instance, scale points of 1 "more paintings" to 7 "more sculptures". If you want to know how interested participants were in paintings and sculptures individually, then it makes sense to split this into two questions and leave the "mainly" out. What we aim to illustrate here is that improving a question can be done in different ways and that this should be informed by step 1, the question what is it that I want to measure?

Furthermore, though we should aim to be brief we need to be precise. To provide a positive example (also from Reitstätter et al. 2020): "With how many people (excluding you) have you visited the museum today?" Here the inclusion of "excluding you" is essential, as otherwise the question is unclear. This then leads to issues for

the researcher as they will not know in the end if 'two' means the participant visited the exhibition with one or two other people.

Step 4: In a final step, consider your response options. Generally, there are three types of options: dichotomous (e.g. yes/no, true/false), categorical (e.g. nationality) or continuous (e.g. a 7-point scale). Categorical responses can be hard to analyze as a dependent variable and thus are generally used for questions that are focused on descriptive aspects of the sample (e.g. nationality). When using more than two answer options, it is important that response categories are evenly distributed. Not doing this, for example, in a 4-point scale where 1 is 'disagree', 2 is 'agree', 3 is 'strongly agree' and 4 'extremely agree', means that you are able to measure agreement well (with high measurement precision), but you are not able to measure disagreement well. Of course, the aim is to have as complete a measurement as possible, so this is suboptimal (Wenig 2004). As a heuristic, at least 5- or 7-point scales are recommended to be able to treat the data as continuous (Wenig 2004). That said, more scale points are not necessarily better. In fact, it may actually reduce the validity as participants are unable to make the subtler distinctions that are required (Clark and Watson 1995, Symonds 1924, Wenig 2004). Finally, sometimes an even number of scale points is preferred to avoid the potentially tricky interpretation of the midpoint value. However, this may force respondents to give answers that do not reflect their true opinion/feelings, which may be problematic (Clark and Watson 1995).

Regarding equipment, there are several ways to implement questionnaires. You can use pen and paper (e.g. Pelowski et al. 2022), tablets (e.g. Reitstätter et al. 2020) or participants can use their own phone (e.g. Specker et al. 2020). In the latter case, participants can scan a QR-code that will then open the questionnaire/link on their phone. A downside of pen and paper is the need to manually enter the responses in, for example, an Excel file in order to analyze the data. This can be rather time-consuming and may also lead to errors in the data by simple human mistakes. That said, it is easy to use and does not require technical infrastructure (ability to recharge the devices, software licenses and so on) nor any programming knowledge that the others may require.

Regarding human resources, in principle, all steps can be done by one single researcher who designs the study, selects or creates the used questions, collects the data, and then analyzes them. What would be required is to have expertise in experimental design for quantitative studies as well as statistical data analysis. Depending on the expertise and experience of the researcher(s) with quantitative methods, and the relative complexity of the study design, the time it will cost to complete each step will vary. In addition, if a validated scale can be found in step 1, then step 2–4 do not have to be completed, which would save time.

The duration of data collection would likely not depend on the researcher but on other factors such as the sample size aimed for (how many people you want to

test) as well as how many people would be able to participate. For example, when testing in a large museum like the Albertina in Vienna, you may be able to test a 100 people in only a few days (e.g. Specker et al. 2020), whereas in smaller museums with fewer visitors per day, you may need a longer time period. Depending on how long the questionnaire would be, participation can generally be relatively short. One thing to consider here is whether you're testing regular museum visitors or whether you're bringing people to the museum. In the first case, you need to remember that participants are there to visit the museum rather than to participate in your study, so the shorter you can keep the questionnaire the better. In the second case, you can generally ask more of your participants, as they came to the museum specifically to participate in your study.

Case study

An example of a study that used questionnaires in an exhibition setting is Specker, et al. (2020). In this study, we were interested in the curatorial narrative – i.e. the embedding of artworks or an entire exhibition inside a wider context of meaning and significance. In the study, half of our participants attended the Monet retrospective at the Albertina Museum in Vienna and the other half visited the *Monet to Picasso* permanent exhibition. In both exhibitions, people were asked to look at *The Water Lily Pond* (1917–1919) by Claude Monet. In the retrospective, this painting was hung in the room which marked the stylistic change in Monet's work towards more abstract(ed) painting. In the permanent exhibition, this work was hung in a room with other impressionist artworks. While there was a stylistic deviance in the first setting of the retrospective, this was not the case for the permanent exhibition. Previous research (e.g. Stamkou, van Kleef, and Homan 2018) had shown that viewers judge artists as more influential if their work is presented in a context of stylistic deviance, and we found the same in our exhibition study. That is, Monet was perceived as more influential in the retrospective exhibition than in the permanent exhibition. In this case, perceived influence as well as deviance was measured by a questionnaire. As we based our study on Stamkou et al. (2018), we could use the same questions and only had to adapt their questions to our exhibition context. For example, instead of asking: "What do you think of this artist?", we asked specifically: "What do you think of Monet as a painter?" Followed by the same four statements that Stamkou et al. (2018) used, but again changing "artist" to "Monet": e.g. "I think that [Monet] will continue to make a great contribution to art even after many generations of painters".

This study led to several insights: Firstly, this study supports a long-held assumption in curatorial practice – that the decision of the curator when composing an overall narrative for an exhibition (specifically, the ordering in which pieces are seen) does, in fact, lead to measurable differences. Secondly, these curatorial

narratives can shape our view of artists – even for very well-known and famous artists such as Monet. It seems even likely that the effects may be bigger for lesser-known artists. Overall, studies like this could provide a basis for an evidence-based approach to curation wherein curators can use empirical findings to achieve or modify curatorial goals and shape visitor experience.

Method reflection

As mentioned in the introduction and discussed above, questionnaires have the benefit that they allow us to assess subjective experiences of exhibitions in a direct way, are cost- and time-effective, flexible in the types of questions that can be asked, and relatively easy to implement (Simms 2008). When focusing on quantitative measurement, one benefit is that questionnaires allow for statistical analysis and relatively straightforward interpretations – an average score is a concise summary of the responses of various respondents. It would be much harder to create such a short summary of phenomenological interviews or other qualitative responses.

That said, these methods (both questionnaires as well as qualitative methods) rely on introspection. The underlying assumption is that participants have insight into their subjective experiences and can report on them accurately by responding to the questions asked. Nonetheless, one benefit of using more quantitative-oriented questionnaires is that it's generally easier for participants to respond to 'How beautiful do you think this artwork is?' on a designated scale e.g. from 1 ('not at all beautiful') to 7 ('very beautiful') rather than as an open question which would require participants being able to verbalize their thoughts which – depending on what and who is being asked – may be hard for them. However, a downside is that respondents are not flexible in their answers. We will only get answers to the questions we have asked, which inherently leads to a limited focus. In addition, as the questions are short and answer options are restricted, we will not get as rich a dataset as we could obtain with qualitative methods. For example, we may find out if things are liked, but not necessarily why they are liked, at least not on an individual level.

Furthermore, questionnaires (or other methods where participants are directly asked) can lose validity due to different kinds of response biases that may occur. For example, people may be motivated to not be completely honest if topics are e.g. socially sensitive – asking someone 'Are you racist?' or even 'Are you interested in art?' (for example, when this is asked in an art-related setting such as a museum) may not be the best method. That said, McKibben and Silvia (2017) report to have found no evidence for social desirability in responses to creativity and arts scales.

Finally, after having discussed all these aspects of questionnaires in art research, we want to stress that the benefit of using quantitative questionnaires is the highest when researchers have clear ideas regarding hypotheses that they want to test. And

despite the temptation to add various components, that it is always important to decide which statistical analyses they want to do, in order to test these in confirmatory research. With these two demands in mind, the reader can start to make good questions for a meaningful, scientific exhibition analysis.

References

Benson, Jeri and Dennis Hocevar. 1985. The Impact of Item Phrasing on the Validity of Attitude Scales for Elementary School Children. In *Journal of Educational Measurement* 22 (3): 231–40. https://sci-hub.st/10.2307/1435036 (05.08.2024).

Clark, Lee Anna and David Watson. 1995. Constructing Validity: Basic Issues in Objective Scale Development. *Psychological Assessment* 7 (3): 309–319. http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/04-questionnaire/04-Clark-1995.pdf.

Gehlbach, Hunter and Maureen E. Brinkworth. 2011. Measure Twice, Cut down Error: A Process for Enhancing the Validity of Survey Scales. *Review of General Psychology* 15 (4): 380–387. https://doi.org/10.1037/a0025704.

Gerger, Gernot, Helmut Leder, Pablo P. L. Tinio, and Annekathrin Schacht. 2011. Faces versus Patterns: Exploring Aesthetic Reactions Using Facial EMG. In *Psychology of Aesthetics, Creativity, and the Arts* 5 (3): 241–250. https://doi.org/10.1037/a0024154.

Kuchinke, Lars, Sabrina Trapp, Arthur M. Jacobs, and Helmut Leder. 2009. Pupillary Responses in Art Appreciation: Effects of Aesthetic Emotions. In *Psychology of Aesthetics, Creativity, and the Arts* 3 (3): 156–163. https://doi.org/10.1037/A0014464.

Leder, Helmut, Aleksandra Mitrovic, and Jürgen Goller. 2016. How Beauty Determines Gaze! Facial Attractiveness and Gaze Duration in Images of Real World Scenes. In *I-Perception* 7 (4). https://doi.org/10.1177/2041669516664355.

McKibben, William Bradley, and Paul J. Silvia. 2017. Evaluating the Distorting Effects of Inattentive Responding and Social Desirability on Self-Report Scales in Creativity and the Arts. *The Journal of Creative Behavior* 51 (1): 57–69. https://doi.org/10.1002/jocb.86.

Pelowski, M., Eva Specker, Jane Boddy, Beatrice Immelmann, Felix Haiduk, Giovanni Spezie, Paula Ibáñez de Aldecoa, Hillary Jean-Joseph, Helmut Leder, and Patrick S. Markey. 2022. Together in the Dark? Investigating the Understanding and Feeling of Intended Emotions Between Viewers and Professional Artists at the Venice Biennale. In *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/ACA0000436.

Reitstätter, Luise, Hanna Brinkmann, Thiago Santini, Eva Specker, Zoya Dare, Flora Bakondi, Anna Miscená, Enkelejda Kasneci, Helmut Leder, and Raphael Rosen-

berg. 2020. The Display Makes a Difference: A Mobile Eye Tracking Study on the Perception of Art before and after a Museum's Rearrangement. In *Journal of Eye Movement Research* 13 (2). https://doi.org/10.16910/jemr.13.2.6.

Schindler, Ines, Georg Hosoya, Winfried Menninghaus, Ursula Beermann, Valentin Wagner, Michael Eid, and Klaus R. Scherer. 2017. Measuring Aesthetic Emotions: A Review of the Literature and a New Assessment Tool. In *PLOS ONE* 12 (6): e0178899. https://doi.org/10.1371/JOURNAL.PONE.0178899.

Simms, Leonard J. 2008. Classical and Modern Methods of Psychological Scale Construction. In *Social and Personality Psychology Compass* 2 (1): 414–33. https://doi.org/10.1111/j.1751-9004.2007.00044.x.

Specker, Eva. 2021. Further Validating the VAIAK: Defining a Psychometric Model, Configural Measurement Invariance, Reliability, and Practical Guidelines. In *Psychology of Aesthetics, Creativity, and the Arts*. https://doi.org/10.1037/ACA0000427.

Specker, Eva, Katherine N. Cotter, and Kyung Yong Kim. 2023. The next Step for the VAIAK: An Item-Focused Analysis. In *Psychology of Aesthetics, Creativity, and the Arts*, https://doi.org/10.1037/ACA0000559.

Specker, Eva, Michael Forster, Hanna Brinkmann, Jane Boddy, Matthew Pelowski, Raphael Rosenberg, and Helmut Leder. 2020. The Vienna Art Interest and Art Knowledge Questionnaire (VAIAK): A Unified and Validated Measure of Art Interest and Art Knowledge. *Psychology of Aesthetics, Creativity, and the Arts* 14 (2): 172–185. https://doi.org/10.1037/aca0000205.

Specker, Eva, Eftychia Stamkou, Matthew Pelowski, and Helmut Leder. 2020. Radically Revolutionary or Pretty Flowers? An Experimental Museum Study of the Impact of Curatorial Narrative Highlighting Artistic Deviance on the Visitor's Assessment of Artist Influence. In *Psychology of Aesthetics Creativity and the Arts*. https://doi.org/https://doi.org/10.1037/aca0000320.

Stamkou, Eftychia, Gerben A. van Kleef, and Astrid C. Homan. 2018. The Art of Influence: When and Why Deviant Artists Gain Impact. In *Journal of Personality and Social Psychology* 115 (2): 276–303. https://doi.org/10.1037/pspi0000131.

Swain, Scott D., Danny Weathers, and Ronald W. Niedrich. 2008. Assessing Three Sources of Misresponse to Reversed Likert Items. In *Journal of Marketing Research* 45 (1): 116–131. https://doi.org/10.1509/JMKR.45.1.116.