Journal of Management Inquiry
1–13

DOI: 10.1177/1056492620938139
journals.sagepub.com/home/jmi


Behind the Numbers: Questioning Questionnaires

Katja Einola[1] and Mats Alvesson[2]

Abstract

Is complex, ambiguous, and fluctuating social reality measurable? Sometimes yes, perhaps, but often not. At least not in the fairly straightforward way assumed by many researchers. This study is an ethnographic inquiry into data collection during a survey research project. Based on our observations of participants' spontaneous thoughts and confusions as they filled in questionnaires on "leadership" and "teamwork", we draw attention to hidden problems in much organizational research. Many respondents found measures ambiguous, irrelevant, or misleading. We (a) underline the inherently interpretative nature of research into complex organizational phenomena, (b) warn against lack of reflexivity and overreliance on existing survey instruments when we study complex social aspects of organizations, (c) identify five categories of possible problems, and (d) suggest paths towards better informed research that take context seriously.

Keywords

survey studies, questionnaires, critical study, quality of research, qualitative study, quantitative study, leadership, teams, research methods, ethnography

Introduction

This study came about unexpectedly. It is to some extent inspired by the influential *Laboratory Life* by Latour and Woolgar (1979), and other research that emerged as a side product of what at first seemed to be a well-planned and confident research process, but then turned out to be something much more uncertain, open-ended, and even counter-intuitive. We believe that what we found out deserves the attention of those who like us conduct research into complex, social phenomena such as teamwork and leadership. Let us start with a personal account reflecting the first author's experiences:

> Our research group aimed to conduct a survey study on leadership and project-based teamwork in the field of strategic advisory. The design of the questionnaire took a year to complete as we combed through all top journals for the best theories and constructs and tested them with our contacts in the field for a "fit". As part of our carefully crafted research design, our team conducted a pre-study. We then designed a research protocol that included data collection in three phases to reduce the possibility of biases. For the phase two, I would collect survey responses for our predictor variables and meet over hundred respondents in person in order to assure a high response rate and accuracy from a hard-to-reach, busy and rather limited pool of respondents. The decision was taken because this part of our data collection was the most intensive, time consuming and risky part of our entire research plan. All this careful preparation promised to lead to a successful completion of the project.
>
> Our approach allowed me to both collect survey responses and spend a significant amount of time at the consulting companies learning about these firms, the work conducted there and the people employed by them outside the survey-filling instances. More importantly for this study, however, it enabled me to observe people as they filled in questionnaires. Rather quickly, I noticed that many questions did not really work for several respondents. Questions were confusing or experienced as irrelevant or impossible to answer. Many basic concepts like "the leader", "the project team" and "cultural differences" were highly ambiguous, and some simple words like "well-being" did not seem to fit the respondents' reality even though the word itself was easily understood. These spontaneous remarks surfaced with regularity in all countries and companies I visited. Concerned with where our research was going, I started taking notes and paying close attention to the context our data collection took place in. This way, I gradually gained a much more nuanced understanding of the phenomena studied and respondents than in our rather limited pre-study that focused mainly on fine-tuning the questionnaire without challenging the underlying assumptions.

[1]Hanken School of Economics, Helsinki, Finland
[2]Lund University, Lund, Sweden

Corresponding Author:
Katja Einola, Hanken School of Economics, Arkadiankatu 22, Helsinki, 00100, Finland.
Email: katja.einola@hanken.fi

> Gradually, a picture of what was going on behind the numbers started to emerge. Over a time period of about 18 months, I managed to collect enough questionnaires for our project. At the same time, I also recorded observations and reflections of how respondents spontaneously reacted to our survey instrument, without eliciting questions purposefully. These reactions could include laughing or smiling, raising eyebrows, looking confused, appearing irritated, asking questions about the questions, sighing, or simply taking a relatively long time to answer. Clearly many found it difficult to match their experiences and knowledge with questions. The seemingly clear and what, beforehand, were thought of as easily answerable questions clashed with a social world not easily transferable to an "X" on a scale. These observations triggered an interest in what can happen when researchers try to "collect" and respondents try to "produce" truthful representations of phenomena.

In this study, we use naturally occurring data recorded in a fieldwork diary and show *how* respondents interpreted questions during a research project, not unlike many others. The interpretative nature of much research in social sciences is an important aspect we, researchers, easily forget when we focus all our attention on collecting, processing, and presenting *data* rather than shedding light on questions and the phenomenon that triggered our study in the first place. We thus report on a study of how questionnaires, at least sometimes, "really" work (see also Gobo & Mauceri, 2014), and discuss how questionnaire responses often are the result of a long chain of (mis)interpretation between the original author of a questionnaire and the final survey respondent. Our study does not consider strictly individual, psychological issues, for example, personality or values, where the subject matter is in a sense more "intact" or "contained", nor do we question all the questionnaires or usefulness of quantitative research in general. Rather, we address teamwork and leadership, the fields the original study was designed to contribute to, and investigate how people actually thought about, made sense of, and found questions (ir-)relevant and (non-) answerable in one particular project.

Our study contributes to the debate over the inherently contextual nature of research into complex social phenomena, which we suggest presents significant challenges for both quantitative and qualitative research, and that we need to pay much more careful attention to (e.g., Alvesson, 1996; Fischer, 2018; Schwarz, 1999). We explore how, at times, survey research fails to do justice to the interpretive, ambiguous, and contextual dimensions of the phenomena it sets out to rigorously study (cf. Potter & Wetherell, 1987).

Limits of the Method

"Measurement is at the heart of any field of scientific inquiry" (Scherbaum & Meade, 2009, p. 636). This idea is taken for granted by many if not most students of organizations. But is social reality always, or even normally, measurable? While deficiencies in research when it comes to the lack of adequate measurements of abstract constructs have been identified as "the greatest challenge to understanding the behavior of people in organizations" (Hinkin, 1998, p. 104; see also Schriesheim et al., 1993 and Schwarz, 1999), the remedy to this *adequacy problem* has largely focused on technical refinement of the tools and methods used, and guidance towards building better survey instruments and scales (Hinkin, 1998; Hogan & Nicholson, 1988).

Some attention has been paid to the context of data collection. For instance, Schwarz (1999) suggests that the researcher involvement in the questionnaire-filling process can be problematic if he or she reveals what the topic of interest is, and if the sequencing of questionnaire themes influences responses. Much of this quality improvement work, however, is silent about what is not readily observable and measurable: the *meaning* respondents attribute to the questions they answer. While quality assurance to assess scale and survey validity has become synonymous with using statistical tools and standards, little attention has been paid to the questions themselves and the way in which individuals interpret them (Gobo & Mauceri, 2014; Hardy & Ford, 2014). Researchers may often neglect whether targeted phenomena are clear and unambiguous enough to begin with so that they can be meaningfully accessed via surveys and scales borrowed from other studies.

Quantitative researchers, just like those of us doing qualitative work, should not underestimate the impact of language, multiple meanings, and interpretation on the quality of studies they conduct. Linguistic and contextual misunderstandings or respondent carelessness are difficult to prevent or detect by subsequent statistical analysis or sophisticated scale development, especially as researchers are not usually attentive to these issues. Our research communities have routinized ways of working and focusing on "data" that are only rarely questioned. Because respondents' interpretation of questions is difficult to observe and study, it is tempting to ignore the issue altogether. The use of a validated instrument is supposed to guarantee reliable data. The rhetorical appeal of certainty of numbers may well obscure the processes of construction and interpretation they are built upon (Amis & Silk, 2008; Poovey, 1998; Porter, 1996).

Whilst different wordings and scale alternatives have been shown to lead to radically different responses and results (Schwarz, 1999), there are few studies on the actual interpretation of questions by respondents in organizational survey research. Hardy and Ford's (2014) study on linguistic miscomprehension, and Galasinski and Kozlowska's (2010) research on survey respondent strategies to cope with inadequate or confusing survey items are exceptions. Both found severe issues with questions asked and scale adequacy due to matters of interpretation that further statistical refinement could not fix. At the heart of the problem is that researchers tend to solely focus on aspects they consider, falsely or

rightly, readily measurable. According to Hardy and Ford (2014), if the degree of variance observed linguistically would be observed numerically, it would be a serious cause for concern. But as it is today, this variance is often ignored, with very negative effects on the quality of our research output.

This is a situation akin to the metaphorical ostrich burying its head in the sand. Survey respondents routinely ignore our questions, fail to follow instructions, answer carelessly, adapt their reality to the survey, reinvent questions, and misunderstand words and sentences (Credé, 2010; Galasinski & Kozlowska, 2010; Gobo & Mauceri, 2014; Hardy & Ford, 2014). Often this is a matter of complex reality being impossible to reduce to single, seemingly clear statements, possible to answer through a Likert scale, for instance. According to Robbins (2002), rather than offering insight into what the respondents actually think or feel, data are at worst an "unknown mixture of politeness, boredom, and a desire to be seen in a good light" (p. 213).

What Does Data Tell Us?

A fundamental premise of any successful survey research is that questions asked can be used to accurately describe practices, conditions, experiences, personal characteristics, or opinions of respondents. Whether these questions are of simple, objective facts (e.g., height, salary, volume of sales), subjective states (e.g., job satisfaction, well-being), complex practices (e.g., decision-making, project management), or about more arbitrary concepts (e.g., "authenticity", "emotional intelligence"), matter enormously.

In quantitative organizational research, perhaps more so than in qualitative research, the *method itself* is a widely accepted guarantor of its own quality. Whether criteria such as reliability and internal, external, and construct validity are really met in the practice of conducting research, however, may remain debatable. Still, numbers and statistical analysis are trusted, and there is the seductive simplicity of clear roadmaps and criteria to follow. Yet, the very meaning of objectivity is in many ways a cultural construct embodying theoretical assumptions of a given scientific community and it can and should be critically scrutinized (Deetz, 1996; Latour & Woolgar, 1979/2013; Poovey, 1998; Porter, 1996). When techniques and procedures are key, reflections outside a narrow instrument-focused reasoning are sometimes absent. The superiority of quantification is taken for granted (e.g., Scherbaum & Meade, 2009), despite some recognition of how answers are an outcome of how questions are asked, how the context around questions looks like, and how scales are structured (Schwarz, 1999).

Data are often seen (or at least treated) as reliable indicators of reality, as robust building blocks for the construction of true knowledge. Ambitious researchers work with measurements they consider as established, adapt them to the specific research questions and context, do a pre-study to learn something about those studied, engage in a formal quality check to ensure face value of their survey instrument, and try to make adjustments before starting data collection. But when data are gathered, there may be little or no knowledge about what is behind respondents' numerical responses. One can only hope that the questionnaire respondents "got it", that is, understood and answered the questions in a standardized way and in line with research intentions, and that the measurements used were "correct".

Problematic Assumptions

Surveys rest on the assumption of an unbroken chain of comprehension between the mind of the researcher through the survey instrument and to the mind of the recipient—and back again (Hardy & Ford, 2014). Yet, questionnaire responding happens as in a black box—researchers have no access to what goes on in respondents' minds and have no control over the actual activity of questionnaire responding. Hardy and Ford (2014) provided strong evidence for three types of widespread miscomprehension when using many popular measurements and supposedly validated scales: (a) instructional (respondents do not follow instructions), (b) sentential (unintended interpretation of a sentence), and (c) lexical (erroneous or uncertain understanding of the meaning of a word in a given context). Galasinski and Kozłowska (2010) listened to people trying to fill in a questionnaire and found that they generally had great problems doing so, as the questionnaire did not allow them to express their experiences. They argue that questionnaire research "assumes that people have clear and well-formed opinions or views, know what they feel or believe, and are able to transform them into the categories offered by the instrument; second, the instrument is actually able to accurately capture all those views, opinions, or feelings in their complexity" (p. 271).

Researchers have often a limited understanding of the relationship between questionnaire-filling behavior and the phenomena they claim to say something about. Just working with and trying to minimize bias through, for instance, clearly expressed formulations or carefully crafted scales/response alternatives only mitigate some of the problems. The fundamental premise of survey research—that questions and constructs used reflect a real-life phenomenon as perceived by respondents—is merely an assumption that cannot be taken for granted. That respondents tend to respond in similar ways to similar-sounding questions—often viewed as a guarantee of construct and measure validity—cannot be considered to be so in absolute terms nor warrant the use of the construct in another setting at another point in time. We do not suggest that all questionnaires lack value, but that the problems we describe reach into "normal practice" much more deeply than is openly acknowledged, and that these cannot be explained away by simply suggesting that a certain

study where issues of interpretation surface, is an anomaly—a rare and flawed exception.

In research, it is important to create *shared meanings* rather than strict adherence to the correspondence theory of truth where the elusive "real world" meets what empirical data "yields".

First, the items on the instrument somehow need to reflect the meaning I ascribe to the construct that is my concern. Second, I need to try to ensure that the items on the instrument communicate my meaning relating to the construct to potential respondents. Whether the construct is real is irrelevant or at best ancillary to my concern about whether I can somehow create shared meaning with potential respondents. (Weber, 2004, p. 7)

This statement implies that for the construct to "count", there should be shared meanings between (a) the researcher who created and validated the instrument and the original respondents, (b) that original researcher and the one adopting the instrument, and (c) the second researcher and research subjects in another empirical setting, and (d) that second researcher and those peers evaluating whether the piece of research is worthy of publication. This cannot be taken for granted. Given the complexity of meaning, the contextual and performative nature of language and that the same use of words may conceal highly diverse meanings (Hardy & Ford, 2014), one could even say that getting all these shared meanings in place is a difficult if not an impossible enterprise (Cunliffe, 2011). We assume that broadly shared meanings are possible, but this is not a given but rather an accomplishment. Only focusing on the relation between the second researcher and his/her respondents (here, the option "c"), an important question emerges: Do we really know what meaning our respondents ascribe to the questions in the research instrument we expect them to fill, and does that meaning correspond with ours—or are we rather entering a grey area?

Filling in Questionnaires: A Case Study

The survey study we report on took place in Northern Europe and involved face-to-face encounters with 110 survey respondents in four countries and six companies engaged in consultancy work in strategic advisory in the banking industry. Because our findings apply equally across all countries and companies, we do not make comparisons *within* the sample. The unit of analysis was "projects" consultants had been working on in the past six months, and the total sample size was 434. The overall purpose of the research project was to study how individual, leadership, and team-based variables influenced junior consultants' learning and project effectiveness. The role of the researcher (the first author of this current article) was to be present and record mostly Likert scale survey answers on behalf of the respondents. She met respondents in one-on-one meetings that lasted between 30 and 60 minutes, and simply read out loud from a dummy questionnaire (also given to the respondents) item by item and then recorded the answers in writing. The interaction and physical presence of the researcher was seen as the only or at least the best way of getting an acceptable response rate from the very busy and hard-to-access people in the sample. Well-known problems with questionnaires—low response rates, careless responses, obvious misunderstandings, or outright confusions—could then be reduced.

Quality Assurance

The research team took ambitious measures to minimize errors and to strengthen the quality of the research. These included a careful survey design with extensive familiarization with the relevant theoretical literature informing hypothesis, constructs, and most suitable scales, preferably published and cited in quality academic journals in similar research. The team had also close personal connections to people in the industry who in many iteration rounds helped make sure the survey instrument reflected the reality of the field, and that its structure was appropriate when it came to, that is, answerability, clarity, and suitability. The survey protocol was conducted in three phases per interviewee to avoid common method bias: the *first survey* with questions about the interviewee was sent via an internet-based tool a few days before the meeting, the *second survey* with questions about project teams and leadership (the focus of this present study) was collected face-to-face, and the *third survey* with questions about project effectiveness and learning (dependent variables) was conducted over the phone about two weeks after the meeting. A careful pre-check and qualification of respondents to only include junior consultants were secured for a homogenous sample. Moreover, a pre-study was conducted in one firm. Some of the scales were reworded for better clarity by a native English speaker familiar with the industry, and some questions were omitted as they were deemed to be of low relevance by either the company representatives participating in the pre-study or by the research team.

With regard to the problems of cultural and linguistic variations in responses to questionnaires, the study focused on a very homogenous group in terms of age, profession, and education in an industry where firms are very similar in terms of corporate culture and English is the global *lingua franca*, making that problem very small compared to most studies aiming for a broader sample (e.g., managers in different companies or in different functions, or immigrant workers from various countries in different industries).

Troubling Observations

This rather unusual method for collecting survey data allowed us to study how people reacted to the questionnaire as it was being filled it in. Many questions caused ambiguity, mixed feelings, or misalignments of some sort. Over time,

clear patterns could be observed indicating that much of the data collected was a result of questions that were experienced as ambiguous, irrelevant, and not being answerable in a clear way beyond random or exceptional confusions. Answers to many questions could go in very different directions, depending on how they were interpreted. The researcher was able to partially capture, correct, and with time even anticipate doubts by being available to clarify respondents' concerns. This, of course required being constantly alert not to interfere too much in the process. How to be both near and far, inside and outside of the questionnaire-filling situation, was a source of constant concern and not an easy issue to reconcile in practice. Often, there were no "right" readings of a question as multiple versions made equal sense also to the researcher. In case of doubt, the instruction was for the respondent to just answer to best of his or her capabilities and then to elaborate further afterwards.

Respondents' reactions were often easy to observe. That some of the interviewees explicitly asked for clarification was a sign of honesty and candid interest in doing a good job at answering the survey, and that they were comfortable enough with the data collection situation to reveal potentially embarrassing confusion/ignorance. Some of the problems we discovered can be attributed in part to usual and known sources. For instance, the researcher needed to point out repeatedly that one question started with a negation (*No one* in the team undermined my efforts), which often made a radical difference in the score given as many respondents that failed to capture the scale had to be turned the other way around to answer "correctly". However, other problems that were far from marginal were not fully explained by otherwise well-documented biases and errors in this type of research.

The Questionnaire: Areas of Ambiguities Emerge

Even though the research protocol consisted of three separate survey instruments, the observations here concern only the second one used when the researcher met the respondents face-to-face and when most of the interaction took place.

Most respondents spontaneously asked questions about the survey or reacted to it with gestures or in other ways. We report on those themes that were brought up by at least five respondents and that stirred most discussions during the interviews. Interpretations that remained under the surface were not captured, and respondents were not actively encouraged to give subjective meanings to the questions because the purpose of the fieldwork was to simply collect questionnaire responses. There is no reason to expect that just because people did not signal any problems or doubts, questions were seen as answerable in a simple way or that there were no major problems with misunderstandings, unreflective box ticking, and careless or random responding. These went just unnoticed. The respondents were generally busy and probably not too inclined to prolong the questionnaire filling work by frequently asking questions or referring to complications.

While the note-taking of only spontaneously raised issues may signal a lack of rigor and perhaps be considered a shortcoming of our study, we want to highlight that our findings were unexpected and emerged gradually as a by-product of another study. Unexpected findings are crucial for innovation and for new insights to emerge (Kuhn, 1970; Latour & Woolgar, 1979/2013). Our aim here is to qualitatively illuminate how questionnaires may be interpreted in the field based on our ethnographic work, the purpose of which was not to count frequencies of responses. Hannah and Lautsch (2011), in fact, advice against counting when it is important for researchers to gain access to the perspectives of insiders and when researchers wish to pursue unexpected findings during an inductive data collection process.

To analyze our data, we grounded ourselves on standard procedures from the field of ethnography (Gilbert, 2001) that we adjusted to our research setting as the study started to take shape. We collected field notes during, before, and after the interviews in the companies we visited (six companies in total, four of which had multiple sites). We were not able to even consider audio-recording the interviews due to high security protocols that prevail in the banking sector that made the situation highly sensitive in many ways from the outset. Additionally, internet access was limited on site and visitors were confined in a special section of the building with no access to the office space the respondents worked in. The note taking that took place during the meetings was done in short-hand writing in a notebook. Here, the role of mental notes was important because the note-taking could not be allowed to interrupt the recording of the survey responses. These mental notes were added to the notebook after the respondent had left. All these notes were then cleaned, organized, and rewritten in the evening of the day of the interview. After each site visit, patterns were searched for across the interviews.

To cut up and mark the data, different color highlights were used to code the most striking findings indicating "unexpected interpretations and reactions" (the most critical information from the point of view of the quality of the survey project) and "other" that included general observations that helped to better understand the actual research context. This included everything from information obtained from administrative and other staff, managers, cafeteria talks, dynamics observed in the coffee room, the type of buildings and decor, safety protocols and dress codes for staff, crowded late night moods in the office, and so on. Researcher concerns and feelings about the research project as it progressed were left in her personal diary. Subsequent analysis focused solely on the "unexpected interpretations". The data was regrouped in emerging patterns and reconstructed under categories. These evolved over time as we balanced between

clarity, completeness, and parsimony. For instance, the initial category "floating and ambiguous object of response" was separated in two: (a) floating concept and (b) ambiguous wording. This separation was deemed necessary because the former indicated interpretations of variables under study (teamwork, project, and leadership), and the latter different meanings respondents attributed to other words in the questionnaire (diversity, conflict/disagreement).

Further we discuss five areas of ambiguities that emerged of our analysis: (a) confusing unit of analysis, (b) ambiguous wording, (c) poor contextual fit, (d) low practical relevance of a measure, and (e) floating concept.

Confusing unit of analysis. Our participants worked mainly on two types of projects: (a) sales pitches (projects to get a contract) and (b) delivery projects (projects to fulfil a contract). "Sales pitch projects", a term suggested by participants in the pre-study, appeared to be a relevant and manageable category—avoiding problems of comparing extremely varied types of projects and facilitating comparisons between similar projects. However, even though the researcher prompted the respondents to consider sales pitches only, she caught them on many occasions responding on delivery projects, that is, projects following from bids the company had actually won and executed as contracts. Many of the pitches became full projects with time and most respondents worked on many projects at any given time so the confusion was understandable. However, mixing leadership and team experiences in fast-paced, standardized sales pitches with much more complex and lengthy delivery projects could muddle comparisons. It became unclear what was actually being studied, not only the core "anchoring unit", for example, the same and comparable type of projects, but also everything else being addressed. This is not a trivial issue, as all the follow-up issues on, for example, diversity and project leadership become very different if the respondents focus on sales pitches or large and more complex projects.

Some respondents indicated that we had chosen a less relevant category and openly suggested that if we were interested in learning about consultants' project work, we should be asking about the more time-consuming delivery projects or *both* pitches and delivery projects, not sales pitches only.

If you want to know about our project work. . . you should ask about live projects we deliver. . . more complicated. . . more people involved. . . pitches are done by juniors mainly, in a rather standard way.

In any case, sales pitches were extremely varied; from last minute ad hoc informal customer meetings with only days, sometimes hours of preparation time to very large, openly competitive requests for proposals that the firm was well aware of ahead of time and that took months to complete. Had we chosen the appropriate unit of analysis when we focused on sales pitches? Perhaps not[1].

Ambiguous wording. Some apparently simple words, like "diversity" seemed to turn into ambiguous phenomena in respondents' minds.

> Q: How diverse was the team in terms of nationality?
> A: We were two Swedes, one of us with Spanish parents, one Finnish Swede and one Norwegian. Was this team then diverse—I mean we were all Scandinavians and spoke English and Swedish, sort of?
> Q: How diverse was the team in terms of country/regional experience?
> A: I really do not know that well where people in this project had been before, but most of us here have been abroad as a trainee or as an exchange student.
> Q: How diverse was the team in terms of industry experience?
> A: What does "industry experience" mean? Do you mean customer segments we cater to here at my company (retail, raw materials, etc.) or me having worked in other industries than consulting?

The "objective measure" of the degree of diversity (whatever that is) may trigger very different agree/do not agree numeric positions on a given scale. People with "two Swedes, one of us with Spanish parents, one Finnish Swede and one Norwegian but all being Scandinavian and speaking Swedish" in a team may be as likely to see this situation as low or as high on diversity, depending on their personal views or just associations to the signifier "diversity". Researchers have no insight about what a specific numerical response says.

Another example was the set of questions about conflict/disagreements/differences of opinion that respondents found ambiguous and hard to make sense of, many visibly hesitated and backtracked previous answers to change them.

> Q1: How often did people in your team disagree about opinions regarding the work being done?
> Q2: How frequently were there conflicts about ideas in your team?
> A: What is the difference between a conflict and a difference of opinion?

In the rather hierarchical work and high-paced environment, severe conflicts paralyzing work are extremely rare, yet some respondents clearly considered "conflict" as a sort of a heated discussion necessary for delivering a good job, while others clearly did not see how "conflicts", interpreted here more as "fighting", were relevant in their place of work. Here, the researcher that avoided making any other comments than "conflict" should be considered in rather neutral terms, as sometimes leading to good and sometimes to bad results. How the respondents made sense of these questions in the end, is totally unclear. It seems likely that some people

had similar experiences, but answered the questions in very different ways—depending on how they interpreted the signifiers "disagree" or "conflict".

Poor contextual fit. With the questions on the adequacy of workload, respondents often smiled or laughed because the term "fair and reasonable workload" in consultancy work is different from that in many other fields. Thus, the reference point involved with this question was ambiguous, making the question sound odd to the respondents.

> Q: The amount of work that I was expected to do was fair and reasonable.
> A: Do you mean as in a normal job or in consulting. . . my job here?

The concept of a "reasonable workload" relevant in many other industries and in Nordic countries in general, is not applicable in many professional services firms following the Anglo-Saxon tradition in which juniors are expected to work long hours. It is difficult to know if the respondents referred to the norm in their industry (i.e., where a long workweek is standard) rather than what is perceived as "reasonable" in other jobs in the country in question. "Fair and reasonable" may thus be addressed in many and fluctuating ways.

Another major problem is how constructs that appear to be relevant in some situations may be irrelevant or problematic in others. In the present study, the words "leader concern for employee well-being" often provoked smiles or laughs and did not seem to be a relevant concept in this context, but rather an alien category.

> Q: The project leader showed concern for my well-being.
> A: . . . I work about 80 hours per week. This is expected and I do not want my senior to fuss over my well-being either. I do not see how this question reflects my workplace.
> A: The senior may be sympathetic if I spend a red-eye night, but still. . . the job has to be done.

The problem here is that the question assumes that showing "concern" is good/wanted while unwanted concern is not really to be considered. Many saw "concern" as "fuss". In the absence of a response with an alternative, signaling, for example, the irrelevance or negative assessment of "concern", the question cannot really be answered in a way indicating how respondents see the issue. The expression "concern with well-being" is loaded with diverse meanings making it difficult for the respondent to produce a simple X-answer. Work–life balance may be a popular topic in some streams of organizational literature, but there are industries where there may be different ideals or logics that the employees have accepted and internalized.

> Q: I was able to keep up with all my personal responsibilities outside of work.
> A: When I started this job, I knew this is what it was going to be like the first years: long hours. So, I am here at work a lot with my colleagues and I have accepted that I no longer have much of a social life. It is my choice.
> A: My friends in normal jobs do not understand this line of work anyway. So personal responsibilities are not an issue at this time.

In consulting, one typically works extra hard until a certain seniority level is reached, after which the workweek becomes shorter along with a higher pay and attractive fringe benefits, a situation many consider worth making some temporary sacrifices for. Most respondents were also males between 22 and 32 years old, and many interpreted a "personal responsibility" to be a child or an elderly parent to attend to, a situation that only rarely reflected the reality of their life at the time of interview.

Low practical relevance of a measure. Even some apparently very straightforward questions were a source of ambiguity. For instance, it was difficult to count how many projects the respondents had at any given time since some were more active than others, and along the way, depending on prioritization, some tasks on less important projects were put on hold if a priority project needed more manpower.

> Q: How many projects did you pursue at the same time with this one?
> A: Hard to say. . .. There were more and less active projects at the time. . .
> A: The sales pitch was on and off. . . and while I worked on this one, some of other projects were on and off. . . let me think. . .
> A: I cannot say exactly. . . how many projects! But I remember I missed the Christmas party so I must have worked about 90 hours that week.

Overall, the relevance of asking this question relating to one's workload appears to be very limited in this context. In particular, the assumption that people' s personal commitment to a specific project had any particular significance (managing multiple demands from the seniors and the overall hours worked during that time seemed to be more critical) was misleading and the idea of trying to get measures on the number of simultaneous projects appeared almost pointless. More basic is perhaps the almost obvious insight that what appears to be a simple measurable phenomenon—number of projects—may not be possible to answer, as its nature is, in fact, fluctuating.

Floating object of response. With questions about the "project team" it was often unclear who was supposed to be considered

as a team member. Team sizes and complexities also varied enormously, also *within* a project. Sometimes remote members were involved or outside experts were consulted (are they part of the team?), sometimes the team was very big involving many firms and professional groups, and fluctuated in size over time. Sometimes most of the work was done between two juniors sitting next to each other with hardly any involvement from the seniors (or outsiders) until the end.

Who should I think of as forming part of the team—everyone who took part in it in some capacity or only me and Harry, and perhaps Will who stepped in every now and then, who put most of the pitch together?

Questions about the "leader" were, according to the research design (as well as most leadership studies), thought to be important. Yet, who the "leader" was (and if there was any leader to begin with) seemed to be difficult to point out in project-based work where the juniors had many hierarchical seniors over them at any given time but with varying levels of involvement depending on the project and phase in the project. Typically, seniors get more involved towards the end, and the more important the deal is, and the less trustworthy the juniors are perceived to be, the more active they are in leading the project. There is also an enormous variation between the juniors. A first-year analyst is just learning the job while a sixth-year associate is often able to act as a project manager, a de facto informal leader.

Leader. . .do you mean the account director, vice-president or the managing director. . .?

Who do you mean as the leader. . .? Here I as a senior junior took care of the project as none of the seniors was available and they trust me anyhow. So, should I rate myself?

"The leader" is not an unproblematic concept. Apart from all the ambiguities around almost any statement of "leadership", for example, how to capture this mystical quality, including how to distinguish it from, for example, management (Alvesson et al., 2017; Carroll et al., 2008), we have the fundamental problem in a project-based setting to identify who is supposed to be *the* leader—and even a member due to the fluid nature of the teams. If it is an experienced junior or a more senior person that one meets fairly infrequently, the questions may gain different meanings and lead to different responses, that is, an employee is likely to have a close relationship and get more support from the former than the latter. So, measures of leaders and leadership may in the statistics seem to refer to the same phenomenon, but may in reality represent an unrecognized and confusing mix of stated perceptions of everything from junior persons to the vice president to the managing director. Again, the researchers have no idea of what is actually being studied and what the X's on questionnaire responses mean. It is likely that the quantitative indicators point in all directions in terms of relationships and qualities, without the researcher having any idea of this.

Similarly, the only question from Gardner's (2012) performance pressure scale that was deemed to be highly relevant according to the pre-study (this project was very important for my organization) was ambiguous to some, even though account executives tend to communicate clearly which projects are "must-wins". Those studied typically work in groups nested in larger firms. Sometimes a project can be important for the smaller unit or even an individual manager, but much less so for the parent organization—or vice versa. It can even happen that a junior is pulled in two directions by two seniors, both acting as if *their* project is the most important one in one same unit. It is not clear what "the organization" is when it comes to individual employees at the bottom of the corporate hierarchy. The idea of "the organization" assumes a sometimes false unity of a variety of people, groups, and considerations.

Discussion

Many researchers rather unreflectively believe that "questionnaires have long demonstrated their usefulness, validity, and reliability in the measurement of leadership" (Kroeck et al., 2004, p. 85) as well as of many other fields. Scherbaum and Meade (2009) claim that "measurements can provide relevant, consistent, and accurate information" (p. 637). We would say that while this *can* be true, many studies have made these assumptions rather overconfidently and without really demonstrating that these claims *are* true. Some taken-for-granted habits and beliefs, or what Lance (2011) and Vandenberg (2006) refer to as "methodological myths and urban legends" deeply anchored in our research communities that influence the way we conduct research, are badly in need of a reality check. Many believe that surveys are "becoming rather overused" (Antonakis, 2017, p.13) and that areas like leadership are "over-reliant on questionnaire studies" (Bryman, 2011, p. 26).

No statistical or methodological sophistication is able to deal with the type of problems we have outlined. Just because numbers add up and correlations are produced, does not automatically mean that knowledge obtained stands on a firm footing. Reviewing the literature on leadership and innovation brings Hughes and colleagues (2018) to the conclusion that "we play 'fast and loose' with construct definitions and the procedures we follow when translating these definitions into measurement scales" (p. 563). Many seemingly convincing results about "positive" leadership leading to "positive" outcomes may simply reflect positive views the employees have of the leader rather than actual behavior and/or the use of items combining behaviors and evaluations of behavior in the same instrument (Fischer, 2018).

A careful inquiry, like the present one uncovering *how* people interpret survey questions against their personal experience and organizational context(s), raises serious doubts about claims for solid methodological rigor and objectivity of questionnaire studies. Some firm believers in the superiority and accuracy of questionnaires may see the basic problems

identified in our study as an indicator of poor measurements and an invalid use of scales, and suggest that the use of better measurements more carefully qualified for a contextual fit would fix or reduce the problem (Hogan & Nicholson, 1988; Schwarz, 1999). Although we admit that this *may* be the case, the research project discussed in this study was hardly less well prepared or relied on weaker measures than most other similar studies. The careful, resource-intensive method that was used for data collection, and the alertness of the researcher responsible for data collection to "quality issues" made it possible to unravel otherwise difficult to detect problems usually hidden from survey researchers.

We need more critical reflection on how to conduct survey-based research and present knowledge claims. Cronbach (1989, p. 147), after a long career in this area, writes that "the literature on construct validation wavers across the range from utopian doctrine to vapid permissiveness". Still today, the issue of what constitutes a "validated measure" is debated. Hence, we concur with Granlund and Lukka (2017, p. 65) in that "the question of reliable and valid measures is always, in principle, an open one". What is "valid" and "reliable" is, in practice, a social construction, the result of a tacit negotiation between the members of a given community of researchers with little inclination to change from within (Vandenberg, 2006).

Common practice in many research areas seems to assume that questionnaires work simply because others have used them and the research has been published in highly ranked journals. This is illustrated by two recent examples from a leading journal (picked at random, as copies of the journal were on the second author's desk). Busch and colleagues (2017) used items like "my peer-mentor (boss) helps to make my work easier" and refer to the questionnaire and another researcher having "successfully used it", and then sent it to a sample of low-skilled immigrant workers from a range of countries. Van Gils and colleagues (2018) used a four-item "performance quality scale" that asked "leaders" to assess whether a follower "delivers work of high quality", and a "respectful leadership scale", used by other researchers, where followers are asked to respond to items like "my leader takes me and my work seriously". Respondents were then asked to put an X on a scale varying from 1 completely disagree to 5 completely agree. The questionnaire was sent to 214 "leaders" and 214 "followers" in 10 German organizations. We use citation marks here as it is doubtful whether people in the sample really saw themselves as, or are best described as "leaders" and "followers". Just like people in our study, many others may find these labels ambiguous and problematic (cf. DeRue & Ashford, 2010; Learmonth & Morrell, 2017). It is likely that a careful investigation, like the one we report on in our study, would reveal similar problems in these two studies. Perhaps, by comparison, the Busch et al. (2017) and Van Gils et al. (2018) studies composed of much more diverse and heterogeneous samples than ours, could face even much greater problems than our study of junior professionals in a consultancy, representing a Western, well-educated, homogenous group with excellent skills in English.

Like our study indicates, researchers may be inclined to ask questions that appear to be easy but are then hardly relevant and/or answerable and the meanings of reported answers may be almost impossible to decipher. Describing one's "leader", for instance, where there is none (or there are many that could fit the label), assuming someone's attitude is fixed when it really is ambivalent, using apparently simple but in the context confusing words, asking someone to give an answer about something he or she is rather ignorant about, asking about one's feelings about something when there are next to none to consider, or giving a single response in terms of a number to something that is inherently fluctuating, complicated and ambiguous, may be totally noninformative. We could add to this the important issue of random or careless sampling (Credé, 2010). One may assume that including a number of questions that respondents feel difficult to interpret or irrelevant to them in a survey will also increase the number of random responses even to relevant or answerable questions due to decreased motivation.

Implications

Based on this study, but also in consideration of broader methodological reflections, we suggest five types of implications mitigating fundamental problems with using questionnaires to capture complex social phenomena, such as leadership or teamwork. While we do not claim that all of the mentioned problems are always present, many questionnaire studies are likely to be impacted by these or other similar hidden issues that are yet to be identified. We also acknowledge that many of the problems we discuss here are not unique to questionnaires, but also concern many interview-based studies (Alvesson, 2011; Potter & Wetherell, 1987; Silverman, 2006).

1. *The research community needs to become more aware of and open to issues related to interpretation, language, and communication when conducting or assessing the quality of a survey study.* The idea of so much of social reality being readily measurable (or even straightforwardly reported in interview statements) needs to be critically addressed. How people fill in questionnaire items may be at times arbitrary and stand in a very ambiguous relation to the phenomenon (believed to be) studied. Ambitious efforts to minimize bias may reduce basic problems but do not necessarily eliminate them and can perhaps even create new ones. Questionnaires should most likely be used much less frequently than is currently the case and survey researchers be more open about using alternative methods in their research. In

addition, researchers should become more modest about and cautious with the knowledge claims they make, especially when it comes to supposed objectivity and methodological rigor. In particular, researchers need to carefully think through how they move from produced data to making knowledge claims, and consider (and even reconsider) what may be fundamental language and interpretation problems hidden behind seemingly robust data.

2. *Questionnaire researchers, journal editors and reviewers should be more careful and suspicious about using published measures in management research.* Designing new questionnaires is tricky and time consuming, so it is tempting to use and re-use existing ones for practical and legitimation reasons (Scherbaum & Meade, 2009). Moreover, the use of established questionnaires is hoped to allow for comparisons across many studies. All this, of course, is potentially helpful for advancing our fields of study, but often the temptation should be resisted. Research protocols should be designed, and if necessary, adjusted first and foremost to serve the purpose of finding answers to interesting and relevant research questions, rather than for the convenience of the research team.

Many published measures have serious foundational flaws, and often poorly reflect the real-life phenomenon they are meant to mirror (Fischer, 2018; Ford & Scandura, 2007; Hardy & Ford, 2014; Hinkin, 1998; Schwarz, 1999). One may assume that a measurement that has been used before probably will function for a new sample, but the opposite assumption that it will probably *not* may be equally or more valid. Our study reveals that many of the questionnaire items did not work particularly well in the specific context (and presumably not in many others as well). This had much more to do with fundamental problems related to multiple meanings and questions being at odds with the specifics of the studied work context than with more widely acknowledged issues such as careless answering and acquiescence bias (Credé, 2010; Kam & Meyer, 2015).

Differences between cultures are sometimes mentioned as a hinder for reuse of questionnaires, but almost all use of "validated" instruments face problems unless developed for and then only used on a very distinct and homogenous group during a limited time period. There may be far-reaching cultural differences between not only countries but also across time, generations, nations, social classes, ethnic groups, regions, industries, professions and organizations. Using a questionnaire for a broad sample of, for example, "managers", may mean that a number of unknown and diverse industrial, professional and organizational situations and cultures are targeted. "Managers" are far from a homogenous group. The only way to uncover possible diverse associations and meanings is to carefully explore the specific situation of the category addressed. This would probably mean that the repeated use of questionnaires would be much more limited, unless very simple and straightforward questions are asked—as opposed to questions built around ambiguous constructs, like in our research context "leadership" and "teams" turned out to be. Given the work needed to carefully develop and validate a questionnaire for a new group one could speculate whether taking this seriously would mean a drastic reduction in the use of questionnaires, currently often used because of convenience, persistent habits in our research communities that have gone unquestioned for too long, and low costs.

3. *Management scholars conducting questionnaire studies should strive to become much more knowledgeable about their empirical domain than they often do.* As our case demonstrates, contacts with people in the targeted domain and a brief "pre-study" may not lead to any real knowledge. In-depth pre-knowledge is important to be able to ask the relevant questions and interpret the answers in an informed way. As we demonstrated earlier, more often than not there are organizational and occupational specificities, work conditions, norms, group associations and idiosyncratic uses of words making the use of standard measures problematic. Related to this, the increasing sophistication in software and other tools used in modeling and data analysis is not a substitute for in-depth knowledge of the empirical field targeted.
4. *Researchers should build in ambiguity sensitizing devices in questionnaires*. Questionnaire studies typically demand a seemingly distinct and unambiguous answer, for example, a number on a one to seven-point Likert scale. Designers also demand that all items should be answered. Our case shows that questions can be answered in different ways or do not make much sense for the respondents. Realizing that this may not make life easier for researchers (or respondents), one could use questionnaires so that uncertainties and un-answerable questions are highlighted rather than denied. Gobo and Mauceri (2014) and Galasinski and Kozlovska (2010) strongly suggest interactional modes of conducting surveys. One could ask respondents to answer only those questions they find answerable, that is, "please answer only the questions that you understand, find relevant and capture your experience. Skip the rest." Or add a column, with an instruction, "please put an 'X' here if you find the question vague, irrelevant or for other reasons difficult to answer", and then offer respondents space to elaborate further in their own words. Alternatively, they could answer in different ways depending on the various meanings, that is, put

different X's on the same scale. All this would amount to an inbuilt validity check. If, for instance, 20% of respondents do not answer the question, put an X in a specific "difficult to answer" box or fill in more than one X, the researcher may decide to not use the item or consider the response in a different way, for instance, by engaging in the largely forgotten technique of deviant case analysis (Gobo & Mauceri, 2014). Or simply see the non- or multiple meaning responses as clues for further study, perhaps a qualitative inquiry. Here we also see potential for using digital technology to make questionnaires more interactive and "intelligent".

5. *Researchers should conduct serious multi-method studies more frequently.* Researchers conducting questionnaire-based studies should supplement them with a strong qualitative component more often. In our case, listening to respondents' thoughts about questionnaire items was highly enlightening. Applying mixed methods in a way that steers away from narrow methodological choices and allows for both quantitative and qualitative approaches to be used simultaneously and fully to better understand a phenomenon is one so far little applied way forward (cf. Tashakkori & Teddlie, 1998). The qualitative part should be a significant element in the research design and be reported in tandem with the quantitative work.

In the study we discussed here, the interaction between the researcher and the respondents consisted mainly of questionnaire-filling support. This is time-consuming. One option is to be present at, for instance, 25% of all the questionnaire filling events, and supplement listening to and recording spontaneous comments with a follow-up interview. Respondents could be asked to think aloud (i.e., express their spontaneous thoughts) when filling in the questionnaire. It would be valuable to get additional in-depth information alongside filling in the numeric answers and then carefully analyze data from both sources on equal terms. This provides insights about the limitations of studied variables and may provide the researcher with inputs to revisions and expansions of the hypothesis and models they work with. This would offer significantly more reliable data. The quantitative data used should then be only on items where few or no respondents expressed difficulties in answering questions. Others should be removed from the study—even if this would mean less material for analysis.

For those issues where the questionnaire works badly, qualitative material could be used more fully. In the case presented here, we probably learned much more about the phenomena through the ad hoc comments on the questionnaire and informally making observations, interacting in situ with the respondents, their managers, and support staff in general, than through the actual responses to the questionnaire. A key insight is to pay much more attention to what goes on outside the questionnaire respondents' box filling behavior. Of course, incorporating these richer sets of data complicates the research work. But as there is no point in pretending that responses that are to a large extent arbitrary and misleading offer a basis for measurement and aggregation, it is necessary to acknowledge that there are strong limits to the easy quantification of ambiguous phenomena.

Reflections and Conclusions

It is important to study how science is done and to be critical of the process in order to strengthen its quality like Latour and Woolgar (1979/2013) highlight in their *Laboratory Life*. Similarly, social research worth doing calls for much more ambitious efforts to investigate what goes on *behind the numbers*. Too often, basic assumptions and quality standards are accepted as given, rather than scrutinized (Alvesson & Sandberg, 2013; Vandenberg, 2006; Welch & Piekkari, 2017). Our research communities tend to downplay the inherently interpretative nature of our social world—a condition that applies not only to qualitative but also to much quantitative research.

What would we do differently in our next project? This is not a question we have a simple answer or a "recipe" for. Hindsight-wisdom is also always problematic—and some of the answers are better to be left to private discussions. But thinking of these experiences as a learning opportunity and what we would like to share with PhD students (or other interested researchers), our lessons learned would include: (a) Conducting a much more thorough pre-study than we did, spending time in the field and getting to know the work environment, people who work in that industry and the business itself. (b) Informing ourselves better by reading much more secondary data. There is plenty of published research and books on global consulting industries, including ethnographic and auto-biographic work by those who have worked in the field. (c) Designing the research in such a way that time would be allocated for the collection of both qualitative and quantitative data when meeting with the respondents. (d) Mobilize considerable support (by supervisors, colleagues) during the work with the "collection" of data. Fieldwork in this type of research requires experience and skills that are only learned by doing. Moreover, data collection is a crucial aspect of any research project like this, not a mechanical activity that, unlike apple picking, can readily be done by someone new to the field without proper training and experience.

The methodological divide between quantitative and qualitative researchers is lamentable. Dichotomizing research and researchers this way and assigning them to camps suspicious of each other is not helpful if we want to pay serious attention to asking good questions and

understanding complex phenomena. We know many researchers who feel the same way. There is nothing stopping us as far as we can see to start learning more about different methods, joining forces and conducting better research.

While many survey researchers may realize that a significant part of social sciences builds on interpretation, this awareness tends to remain hidden when we confidently collect, analyze, and report findings based on *data*, leaving the questions we ask in the shadows. Just like our respondents in the survey study that triggered this article who questioned the questionnaire, we should, perhaps more often, question our questions and measurement tools. It is far too easy to take much for granted and assume that we know what questions are relevant and answerable to our respondents. We need to be more sensitive to our research context and phenomenon we want to shed light on, even if this means complicating data collection, adjusting our research design and prolonging our publication cycles. Acknowledging privately and during method seminars that interpretation is "there" does not help if we collectively maintain a discourse of false objectivity and report on our research as if we did not need to consider this in our research practice.

Acknowledgments

The authors wish to thank Knut and Alice Wallenberg's Foundation, as well as Jan Wallander and Tom Hedelius' Research Foundation for support received.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Mats Alvesson https://orcid.org/0000-0001-8709-4684

Note

1. The type of project to be included in the study was determined by their supposed comparability, letting the ambition to make comparisons guide the design. Had we been more interested in the perspectives of our respondents, then we had structured our research around their work life in which many kinds of projects, both large and small, delivery and sales, were common. These insights were, however, gained only after the research was too advanced to make changes.

References

Alvesson, M., (1996). Leadership studies: From procedure and abstraction to reflexivity and situation. *Leadership Quarterly*, *7*(4), 455–485.

Alvesson, M. (2011). *Interpreting interviews*. London: Sage.

Alvesson, M., Blom, M., & Sveningsson, S. (2017). *Reflexive leadership*. London: Sage.

Alvesson, M., & Sandberg, J. (2013). *Constructing research questions: Doing interesting research through problematization*. London: Sage.

Amis, J. M., & Silk, M. L. (2008). The philosophy and politics of quality in qualitative organizational research. *Organizational Research Methods*, *11*(3), 456–480.

Antonakis, J. (2017). On doing better science: From thrill of discovery to policy implications. *Leadership Quarterly*, *28*(1), 5–21.

Bryman, A. (2011). Research methods in the study of leadership. In A. Bryman, D. Collinson, K. Grint, B. Jackson, & M. Uhl-Bien (Eds.), *The SAGE handbook of leadership* (pp. 15–28). London: Sage.

Busch, C., Koch, T., Clasen, J., Winkler, E., & Vowinkel, J. (2017). Evaluation of an organizational health intervention for low-skilled workers and immigrants. *Human Relations*, *70*(8), 994–1016.

Carroll, B., Levy, L., & Richmond, D. (2008). Leadership as practice: Challenging the competency paradigm. *Leadership*, *4*(4), 363–379.

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, *70*(4), 596–612.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy. Proceedings of symposium in honor of Lloyd G. Humphreys* (pp. 147–171). University of Illinois Press.

Cunliffe, A. L. (2011). Crafting qualitative research: Morgan and Smircich 30 years on. *Organizational Research Methods*, *14*(4), 647–673.

Deetz, S. (1996). Describing differences in approaches to organization science: Rethinking Burrell and Morgan and their legacy. *Organization Science*, *7*(2), 191–207.

DeRue, D. S., & Ashford, S. J. (2010). Who will lead and who will follow? A social process of leadership identity construction in organizations. *Academy of Management Review*, *35*(4), 627–647.

Fischer, T. (2018). *Leadership: Processes and ambiguities* (PhD thesis). University of Lausanne Press.

Ford, L. R., & Scandura, T. A. (2007). *Item generation: A review of commonly used measures and recommendations for future practice*. Paper presented at the Annual Meeting of the Southern Management Association, Nashville.

Galasinski, D, & Kozłowska, O. (2010). Questionnaires and lived experience: Strategies of coping with the quantitative frame. *Qualitative Inquiry*, *16*(4), 271–284.

Gardner, H. K. (2012). Performance pressure as a double-edged sword: Enhancing team motivation but undermining the use of team knowledge. *Administrative Science Quarterly*, *57*(1), 1–46.

Gilbert, N. (2001). Ethnography. In N. Gilbert (Ed.), *Researching social life* (pp. 145–163). Sage.

Gobo, G., & Mauceri, S. (2014). *Constructing survey data: An interactional approach*. Sage.

Granlund, M., & Lukka, K. (2017). Investigating highly established research paradigms: Reviving contextuality in contingency theory

based management accounting research. *Critical Perspectives on Accounting*, *45*, 63–80.

Hannah, D. R., & Lautsch, B. A. (2011). Counting in qualitative research: Why to conduct it, when to avoid it, and when to closet it. *Journal of Management Inquiry*, *20*(1), 14–22.

Hardy, B., & Ford, L. R. (2014). It's not me, it's you: Miscomprehension in surveys. *Organizational Research Methods*, *17*(2), 138–162.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, *1*(1), 104–121.

Hogan, R., & Nicholson, R. (1988). The meaning of personality test scores. *American Psychologist*, *43*(8), 621–626.

Hughes, D. J., Lee, A., Tian, A., & Newman, A. (2018). Leadership, creativity, and innovation: A critical review and practical recommendations. *The Leadership Quarterly*, *29*(5), 549–569.

Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods*, *18*(3), 512–541.

Kroeck, G., Lowe, K., & Brown, K. (2004). The assessment of leadership. In A. Antonakis, A. T. Cianciolo, & R. Sternberg (Eds.), *The nature of leadership* (pp. 71–98). Sage.

Kuhn, T. (1970). *The structure of scientific revolutions*. University of Chicago Press.

Lance, C. E. (2011). More statistical and methodological myths and urban legends. *Organizational Research Methods*, *14*(2), 279–286.

Latour, B., & Woolgar, S. (1979/2013). *Laboratory life: The construction of scientific facts*. Princeton University Press.

Learmonth, M., & Morrill, K. (2017). Is critical leadership studies "critical"? *Leadership*, *13*(3), 257–271.

Poovey, M. (1998). *A history of the modern fact: Problems of knowledge in the sciences of wealth and society*. University of Chicago Press.

Porter, T. M. (1996). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.

Potter, J., & Wetherell, M. (1987). *Discourse and social psychology: Beyond attitudes and behaviour*. Sage.

Robbins, C. (2002). *Real world research: A resource for social scientists and practitioner-researchers*. Blackwell.

Scherbaum, C., & Meade, A. (2009). Measurement in the organizational sciences: conceptual and technological advances. In D. Buchanan & A. Bryman (Eds.), *The SAGE handbook of organizational research methods* (pp. 636–653). Sage.

Schriesheim, C., Powers, K. J., Scandura, T., Gardiner, C. C., & Lankau, M. J. (1993). Improving construct measurement in management research: Comments and a quantitative approach for assessing the theoretical content adequacy of paper-and-pencil survey-tape instruments. *Journal of Management*, *19*(2), 385–417.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*(2), 93–105.

Silverman, D. (2006). *Interpreting qualitative data*. Sage.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. Sage.

Van Gils, S., Van Quaquebeke, N., Borkowski, J., & Van Knippenberg, D. L. (2018). Respectful leadership: Reducing performance challenges posed by leader role incongruence and gender dissimilarity. *Human Relations*, *71*(12), 1590–1610.

Vandenberg, R. J. (2006). Statistical and methodological myths and urban legends: Where, pray tell, did they get this idea? *Organizational Research Methods*, *9*(2), 194–201.

Weber, R. (2004). The rhetoric of positivism versus interpretivism: A personal view. *MIS Quarterly*, *28*(1), iii–xii.

Welch, C., & Piekkari, R. (2017). How should we (not) judge the "quality" of qualitative research? A re-assessment of current evaluative criteria in International Business. *Journal of World Business*, *52*(5), 714–725.