J Bus Ethics (2014) 120:275–289
DOI 10.1007/s10551-013-1659-5

Methodological Issues in the Design of Online Surveys for Measuring Unethical Work Behavior: Recommendations on the Basis of a Split-Ballot Experiment

Kristel Wouters · Jeroen Maesschalck · Carel F. W. Peeters · Marijke Roosen

Received: 11 September 2012 / Accepted: 18 February 2013 / Published online: 7 March 2013

Abstract In recent years, there has been an increasing interest in unethical work behavior. Several types of survey instruments to collect information about unethical work behavior are available. Nevertheless, to date little attention has been paid to design issues of those surveys. There are, however, several important problems that may influence reliability and validity of questionnaire data on the topic, such as social desirability bias. This paper addresses two important issues in the design of online surveys on unethical work behavior: the response scale for questions regarding the frequency of certain types of unethical work behavior and the location of the background questions in an online survey. We present the results of an analysis of a double split-ballot experiment in a large sample ($n = 3{,}386$) on governmental integrity. We found that, when comparing response scales that have labels for all categories with response scales that only have anchors at the end, the latter provided answers with higher validity. The study did not provide support for the conventional practice of asking background questions at the end.

Keywords Unethical work behavior · Social desirability bias · Participation bias · Measurement error · Response scales · Anchors/labels · Background questions

K. Wouters (✉) · J. Maesschalck
Faculty of Law, Leuven Institute of Criminology (LINC), KU Leuven (University of Leuven), H. Hooverplein 10, 3000 Leuven, Belgium
e-mail: kristel.wouters@law.kuleuven.be

C. F. W. Peeters
Quality of Governance Research Group, Department of Governance Studies, VU University Amsterdam, Amsterdam, The Netherlands

M. Roosen
Research Group Prevention and Approach of Violence, Thomas More Kempen University College, Geel, Belgium

Introduction

In recent years, there has been an increasing interest in unethical work behavior (UWB) in the public sector (e.g., Maesschalck 2004; Lasthuizen et al. 2011; Kolthoff 2007b) as well as the private sector (e.g., Victor et al. 1993; Peterson 2002). This increasing interest in UWB, has heightened the need for valid measurements of UWB. On the one hand, this allows for evaluating ethics management policies which were introduced as a reaction to a number of scandals. On the other hand, valid measurements of UWB are crucial for testing theories on UWB. One of the most important problems in studying UWB, however, is the sensitivity of this research topic. This sensitivity could create survey errors, such as participation bias and social desirability bias. To date, not enough methodological research has been done on how to prevent these errors in surveys on UWB. This paper seeks to address two important methodological issues in designing surveys on UWB. The first issue concerns the appropriate response scale for questions regarding the frequency of certain types of UWB. The second concerns the location of background questions (e.g., age, education, unit in the organization, length of service in the organization, hierarchical position, etc.) in an online survey on UWB.

"Unethical work behavior" is an umbrella term that covers a wide variety of behaviors ranging from minor integrity violations (e.g., coming to work late, gossiping, minor effort) to manifestly criminal behavior (e.g., theft,

corruption, fraud). While a variety of definitions of unethical work behavior have been suggested in the literature, throughout this paper the term will refer to *every action by public servants that defies and violates (a) shared organizational norms and expectations, and/or (b) core societal values, mores and standards of proper conduct* (based on Vardi and Wiener 1996).[1]

The paper will start with a brief introduction into the problems that play in researching unethical work behavior online (i.e., web surveys). This includes a discussion of measuring unethical work behavior through a self-administered questionnaire and a discussion of the two most important methodological problems in measuring unethical work behavior (social desirability response bias and participation bias). The second paragraph presents the two methodological issues that will be addressed in this article and introduces our hypotheses. The third paragraph presents the results of a split-ballot experiment to test those hypotheses. The paper concludes with a discussion of the findings in light of the research on unethical work behavior.

Measuring Unethical Work Behavior

A variety of methods could be used to assess unethical work behavior. Each has its advantages and drawbacks. Researchers have measured unethical work behavior in a variety of ways, including surveys (e.g., Kolthoff 2007b; Lasthuizen 2008), interviews (e.g., Maesschalck 2004) and observations (e.g., Mars 1999; Loyens 2012).

The focus of this article will be on the quantitative measurement of UWB through self-administered questionnaires, which is probably the most common method in the study of this research topic (Frederickson and Walling 2001). A quantitative research design makes a sensitive topic as unethical work behavior more accessible since participants have a higher degree of anonymity in comparison to qualitative research methods. This is a very important aspect in research on unethical work behavior because participants who believe they may be identified may answer or behave in a socially desirable manner (Wildman 1977; Frick et al. 2001). Moreover, a quantitative design makes it possible to reach a larger pool of possible participants and thus to gather a larger amount of information to test hypotheses and refine theories.

Surveys among employees can, at least theoretically, provide more accurate information in comparison to other methods, because a number of limitations that account for dark number (i.e., unregistered and therefore unknown violations) are not present. There are three obvious ways of measuring unethical work behavior in survey reports: (1) victim-reports, in which information on the victimization of the respondent is asked; (2) proxy-reports, in which the respondent is asked to report on others' behavior; and (3) self-reports, in which the behavior of the respondent himself is inquired. In this contribution, the focus will be proxy-reporting and self-reporting. Each of these methods has its own limitations, leading to a certain degree of underestimation of the real prevalence of unethical work behavior. In proxy- and victim-reports, for example, there is a problem of underreporting because of invisibility and victimless misbehavior. While that problem disappears in self-reports, a serious problem still remains: the tendency of the respondent to answer survey questions dishonestly because of their sensitivity. However, this problem is not unique to self-reports; we believe all three ways of measuring unethical work behavior suffer to some extent from the problem "socially desirable responding" (see below).

Several types of survey instruments to collect information about unethical work behavior are available (e.g., Bennett and Robinson 2000; Kaptein and Avelino 2007; Kolthoff 2007a; Delbeke et al. 2008). Nevertheless, to date little attention has been paid to design issues of those surveys (e.g., which response scale to use for UWB, the formulation of the questions on UWB). There are, however, several important problems that may influence the validity of questionnaire data on the topic, such as social desirability bias and cognitive limitations of the respondents. In dealing with these difficulties, we can find some insights in the literature on sensitive topics in general (e.g., Lee 1999) and the more specific literature on measuring crime and delinquency (e.g., Hindelang et al. 1981; Sampford et al. 2006), on integrity testing for personnel selection (e.g., Wanek 1999) and in the broader methodological literature on survey design (e.g., Fowler 1995). However, much of the published research is not necessarily applicable for the measurement of such a sensitive subject as unethical work behavior and more research is needed to help solving methodological issues. This paper will focus on a relatively new way for administering a standardized questionnaire, which is the online survey. The lack of research on methodological challenges when measuring unethical work behavior is even more prevalent for this type of survey.

[1] Unethical work behavior is a topic receiving much attention in fields such as business ethics, administrative ethics (subfield of public administration), criminology, organizational psychology etc. The various angles from which unethical work behavior is studied give rise to a wide variation of concepts. In criminology, concepts such as employee deviance and organizational misbehavior are used, whereas in business and administrative ethics research is often focused on integrity violations or unethical behavior. When studying the operationalisations of these concepts, however, it becomes clear that they often mean the same. In this study the term "unethical work behavior" is chosen.

Methodological Problems and Unethical Work Behavior

At the most general level, survey methods suffer from the three problems that have been brought together under the umbrella concept "total survey error" (e.g., Junger-Tas and Marshall 1999; Weisberg 2005): (1) "coverage error" (the planned sample is not representative for the population), (2) "participation error" (nonresponse error) and (3) "measurement error" (distortion of responses). In this contribution we focus on those that follow from the sensitive nature of surveys on unethical work behavior. This sensitivity might particularly increase participation error (people might decide not to respond) and measurement error (people might decide to respond in a socially desirable way, particularly by underreporting deviant behavior, i.e., social desirability bias). As we hypothesize that the errors caused by the sensitive nature of the questions are systematic we will henceforth refer to them as "biases".

Sensitivity-Induced Participation Bias

When preparing a survey to measure unethical work behavior, one should particularly be concerned about problems caused by the sensitivity of the questions. Asking a respondent, for example, whether he has stolen something at work in the past year, is sensitive since admitting to it makes him vulnerable for sanctioning. This sanctioning could be formal (e.g., disciplinary measure) or informal (e.g., indignation of colleagues about the behavior).

As a consequence of the sensitive nature of questions probing for unethical work behavior, some respondents could decide not to respond (van der Heijden et al. 2000) leading to participation bias. Therefore, "sensitivity-induced participation bias" will be used as a criterion to evaluate the addressed methodological issues in this research. We will henceforth refer to this "sensitivity-induced participation bias" as "participation bias".

Participation bias could be present in a variety of situations (Heerwegh 2005). In this study we will make a distinction between three types of nonresponse. Participation bias could be present at the level of the questionnaire as a whole—in this article this will be studied by looking at the "unit nonresponse"—and at the level of the specific questions—which will be studied by looking at the "item nonresponse" and "partial nonresponse".

Unit nonresponse refers to the absence of all data for a sample unit (Heerwegh 2005). The most important reason for unit nonresponse is respondents deciding not to cooperate in the survey. Another reason could be that the respondent did not receive the survey (for example due to a wrong e-mail address). According to Groves and Cooper (1998), unit nonresponse constitutes a major threat for all types of survey research, because differences between respondents and non-respondents can introduce bias in the population parameter estimates.

Nonresponse can also occur at the level of the survey items. In this article we focus on two types of item-missing data: item nonresponse and partial nonresponse. First, item nonresponse could occur in a survey when a question is skipped by mistake or on purpose (Heerwegh 2005). If item-missing data do not occur at random, the results can be biased. In the case of research into unethical work behavior this bias could be the result of respondents deciding not to answer the sensitive questions of the survey. Second, partial nonresponse refers to the observation that from a distinct point forward, all data are missing. Respondents thus end their cooperation in the study. This type of nonresponse is also called "drop-out."

Social Desirability Bias

In addition to participation error, research into unethical work behavior also suffers from the problem of "measurement error." In the context of sensitive research, this particularly takes the form of socially desirable responding (Arnold and Feldman 1981; Randall and Gibson 1990): answering questions dishonestly by selecting the response choice that will place oneself in a favorable light (Edwards 1957). In the context of unethical work behavior, social desirability implies that respondents may not always answer truthfully about their misbehavior due to perceived social unacceptability (Tourangeau and Smith 1996; Sloan et al. 2004) and may tend to put themselves and their colleagues in a more positive light, in spite of their actual behavior. This will result in the systematic distortion called "social desirability bias" (SDB).

SDB is possibly the most important problem in research on unethical work behavior (Randall and Fernandes 1991; Crane 1999). Chung and Monroe (2003), however, argued that it is only recently that business ethics researchers pay more attention to SDB (e.g., Cohen et al. 1998, 2001; Schoderbek and Deshpande 1996; Randall and Fernandes 1992). The literature on the nature and the source of this bias in a business ethics decision setting is limited. Cohen et al. (1998, 2001) and Schoderbek and Deshpande (1996) found a gender effect on this bias and Randall and Fernandes (1992) studied the impact of social desirability on self-reported ethical conduct. Nonetheless, far too little attention has been paid to design issues of surveys and the relationship with SDB.

Ganster et al. (1983) consider social desirability (SD) problematic for three reasons. First, SD may provide a false or spurious correlation between independent and dependent variables, for example when SD is correlated with both the independent and dependent variables of interest. An

observed correlation between the independent and dependent variables might then be due to their shared variance in social desirability and not due to shared variance in the constructs that the measures purport to tap. Second, SD may moderate the relationship between independent and dependent variables. The distinguishing feature of this model is that there is an interaction effect between the independent variable and SD (Ganster et al. 1983, p. 324). Finally, social desirability bias may mask the relationship between two or more variables. This last aspect is most important in research on unethical work behavior. Because social desirability reduces the variance of unethical work behavior (respondents might underreport their behavior and hence be closer to a prevalence of 0 than reality) (Skipper and Hyman 1993), this might make it more difficult to find a relationship with another variable. As a conclusion, researchers must avoid the potential effects of social desirability response bias by, among other things, focusing on aspects of survey design. Therefore, social desirability response bias will be used as a criterion to evaluate the two addressed methodological issues in this research.

Arguably, social desirability can be considered to influence self-reporting as well as proxy-reporting (Nederhof 1985). As for self-reporting, respondents may be untruthful about their own behavior, for example out of fear of being identified, stigmatized or incriminated (Lee 1999). As for proxy-reporting, respondents could for example want to defend the image of their organization. Since public organizations have an exemplary function, public servants might feel that disclosure of unethical work behavior could have negative consequences for their organization and government in general since this might result in public distrust which in the long run could undermine a democratic society.

Coping with the Two Biases

There are two ways of coping with participation bias and social desirability bias.

A first way is to measure the bias. Participation bias could, for example, be measured by comparing responders with non-responders (e.g., Stoop 2005). SDB could be measured with SD-scales (McDonald 2000). In those scales, respondents are inquired about behavior that is positively evaluated, but that in reality is considered to be rare, or, vice versa, respondents are asked about behavior that is negatively evaluated, but occurs frequently in reality. In SD-scales researchers for example ask the respondent to rate items like "*I am sometimes irritated by people who ask favors of me*," which is negatively evaluated in society. The researcher, however, assumes that everybody is sometimes irritated by people who ask favors, and therefore respondents disagreeing with the item are answering in a socially desirable way. With the results on the SD-scale, the researcher can statistically control for this tendency of the respondent. However, a growing body of literature considers the measurement of bias, especially the use of SD-scales, to have limited value (e.g., Borkeneau and Ostendorf 1992; Beretvas et al. 2002; Holtgraves 2004).

The second way of dealing with bias is to prevent (or at least reduce) it. That will be the focus of this contribution. This is particularly done by adjusting characteristics of the survey. The literature abounds with hypotheses about how adjustments to characteristics might indeed reduce sensitivity-induced bias (e.g., Sudman and Bradburn 1982), but empirical tests of these hypotheses often lead to mixed results because of their different research contexts. In addition, their relevance for research into unethical work behavior is limited. Therefore, we will test some specific hypotheses in this article for research into UWB.

Response Scale and Question Location as Validity Issues

The Response Scale

In this section, we address the issue of response scale formats that can be used to measure the frequency of unethical work behavior. We studied the frequency of self-reported behavior and of unethical work behavior observed by proxies, in this case colleagues. Published research on UWB does not provide much guidance for making decisions about the scale format of UWB. Therefore, this paper will first look into the broader literature on scale characteristics to set up the hypotheses of this paper.

The two most important ways in which scale characteristics can be varied are the numbers of scale levels (which will not be addressed in this paper) and the way in which scale levels are defined and presented to the respondents (Finn 1972; Fowler 1995; Krosnick and Fabrigar 1997). The latter is of course dependent on the first and concerns the choice of defining every response category or only the anchors. The former are called all-category-defined scales (or a scale with labels), the latter response scales with anchors.

All-category-defined scales seem intuitively more accurate than scales where only the anchors are labeled. They give the impression that the researcher really knows what he is measuring. Moreover, the degree of abstract thinking required from respondents is, according to some, supposed to be lower for all-category-defined scales than for scales with only the end points defined (Borgers et al. 2003). When researchers choose to use an all-category-defined scale there are two options: (1) they can denominate the frequencies relatively exactly (e.g., once, twice,

three, or four times a week, weekly, daily) or (2) use more vague quantifiers (e.g., never–one time–several times–regularly–often) (Gaskell et al. 1994). The former response format could be considered very threatening since respondents have to denote relatively exactly how often they have seen their colleagues engaging in certain types of unethical work behavior (proxy-report) or have engaged themselves in unethical work behavior (self-report). The problem with the latter format is that the labels of the response scale should have the same meaning for each respondent and also in every context and this is not always the case (e.g., Bradburn and Miles 1979; Pepper 1981; Newstead 1988; Moxey and Sanford 1993; Wright et al. 1994). Several studies have revealed that the interpretation of vague quantifiers could be influenced by a number of factors, such as the expected frequency of the event, attitudes, experience and type of activity (Newstead 1988; Schwarz 1999). Gaskell et al. (1994), for example, found higher reported frequencies of feeling annoyed when using a scale with response alternatives that indicate high frequencies in contrast to a low frequency response scale. When opting for an all-category-defined scale it is thus important to choose the labels wisely because they are not neutral (Gaskell et al. 1994). In contrast to all-category-defined scales, response scales with only the end points labeled have the advantage that the researcher does not need to find adequate qualifiers and can choose for a response scale with more scale points. Yet, the researcher is forced to view his results only in relative terms (Wells and Smith 1960). Of course, for policy makers this is less interesting. Moreover, just like with vague quantifiers it is possible that not every participant places himself in the same way on the response scale (Schwarz 1999).

Relatively few studies have dealt with the difference between anchors and labels. Finn (1972) investigated whether the way in which response scales are presented affected the means of ratings. He found no difference in results as a consequence of variations in scale point labels. However, he only used formats with anchors and no format in which all points had a verbal label. Dixon et al. (1984) found that the difference between a format with anchors and an all-category-defined scale were not significant. Moreover, their research also showed that respondents had no format preference. However, Frisbie and Brandenburg (1979) found in their study on student evaluation that the scale format with anchors had higher mean rankings than the all-category-defined scale.

In this study, we will compare two response formats which could be used to measure UWB: a scale with numeric labels accompanied by anchors at the extremes (never 0–1–2–3–4 often) and an all-category-defined scale (never—one time—several times—regularly—often) (see Fig. 1). The former format originates from the "Integriteitmeter" of the Dutch bureau of Integrity in municipalities (Kolthoff 2007a). Much ethics research in the Netherlands and Belgium is (partially) based on this survey instrument, and thus an evaluation of this response scale is useful. The latter format is an alternative with labels. The only difference between the two formats is the fact that, in the all-category-defined scale, the qualifying adverbs are printed above the radio buttons, and in the format with anchors only numbers are printed above the bullets accompanied by labeled anchors at the end.

For the evaluation of the response scales in our study, we build on the above discussion of social desirability response bias (see "Measuring Unethical Work Behavior" section) and argue that, for items on unethical work behavior, the key question should be which response scale generates the highest mean score and variability. The underlying assumption is that, for sensitive items, a higher mean score and variance can be assumed to be more valid, as the incentives for socially desirable answers regarding questions that are potentially incriminating or intrusive will normally lead respondents to underreport the behavior. Social desirability would influence respondents to use only the extremes of the response scale. For unethical work behavior, the majority of respondents will by default only consider a small part of the response scale as good response options (whether this is congruent with the reality or not), more specifically only the left side of the response scale, which refers to "never" or "only a few times."

We expect the response scale with anchors to be less threatening to respondents because they only need to place

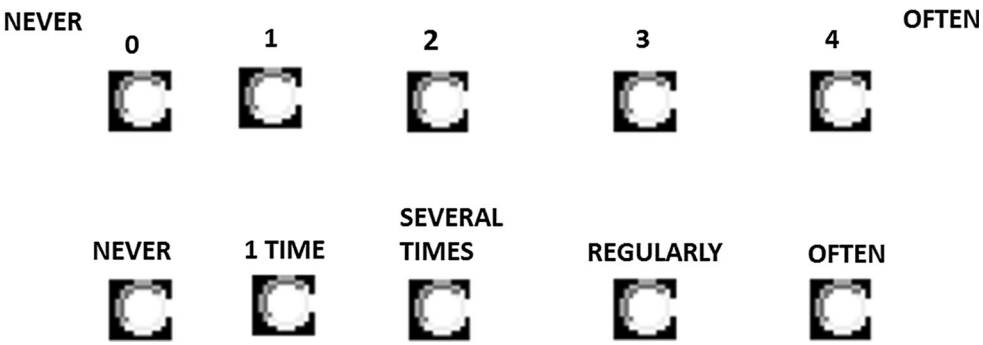

Fig. 1 The response scales in the study (*top* anchor format, *bottom* all-category-defined format)

themselves on a continuum and do not have to situate themselves relative to fairly specific labels. As such, they experience less pressure when answering the questions on unethical work behavior, which will result in a higher mean. We also expect this response format to have a higher variance since respondents will tend to use more categories of the response scale because of its lower threat. Thus, we formulate the following hypotheses:

Hypothesis 1 The response scale with anchors has a higher mean on the questions for unethical work behavior than the all-category-defined scale.

Hypothesis 2 The response scale with anchors has higher variance on the questions for unethical work behavior than the all-category-defined scale.

Location of the Background Questions

The issue whether background questions should be placed at the beginning of an online survey on unethical work behavior is of crucial importance.

The methodological literature typically recommends placing background questions at the end of the survey because of anonymity and commitment-issues (Babbie 1997; Billiet and Waege 2006; Dillman 2007). When anonymity is emphasized in the introduction of the survey (as is often done in surveys on sensitive issues such as unethical work behavior), it is not unlikely that some respondents could become suspicious if this reassurance of anonymity is immediately followed by background questions that might, in theory, allow them to be identified (e.g., if you are the only male staff member with a particular educational level in your unit). Placing background questions at the end of the survey should provide higher response rates and less socially desirable responses. Moreover, placing background questions at the beginning could result in motivational problems because filling out questions on personal information is not very interesting (Dillman 2007).

However, for online surveys one could argue in favor of placing these background questions at the beginning of the survey, because of possible drop-out of respondents during the survey. Answers to background questions are often crucial for data analysis since they are frequently used to describe the data and act as control variables in the analysis. When the background questions are placed at the end of the survey, the researchers are confronted with the risk of receiving no information on these questions for all the participants who dropped out during the survey. Moreover, we argue that in the case of research into unethical work behavior, the background questions are easier to respond to than the questions on unethical work behavior and on other, explanatory variables (for example ethics policy, ethical climate and employee satisfaction) and thus perhaps could reduce (instead of increase) motivational problems early in the survey.

However, before researchers can confidently place the background questions at the beginning of an online survey on unethical work behavior, they have to assess whether this will not have a serious effect on the return rate and the response behavior of the participants. Placing these background questions in the beginning of the survey could of course influence the perceived anonymity of the respondent and thus lead to more socially desirable responses or nonresponse.

In this study, we therefore compare both locations of the background questions: at the beginning of the survey and at the end by evaluating two aspects: participation bias and measurement bias. If the two conditions do not differ on these aspects, or the results are in favor of the location at the beginning, we can confidently place these questions in the beginning of the online survey on unethical work behavior.

A first issue in evaluating the best place for background questions (in the beginning or at the end), is the effect of the location on participation of respondents. To evaluate participation bias it is useful to look at three aspects: (1) unit nonresponse, (2) representativeness of the realized sample, and (3) item-missing data.

First, placing the background questions at the beginning should not result in a serious decline in unit nonresponse. Giles and Feild (1978) studied the effect of location, format and amount of background questions on return rates of a paper survey on job satisfaction. They found no effect of these three characteristics on the return rate. Roberson and Sundstrom (1990), also using a paper survey, however, found return rates to be higher when background items were placed at the end of the survey than at the beginning. In sum, the research is inconclusive and additional research is needed with a specific focus on unethical work behavior. In keeping with this research, the following hypothesis will be tested in this research:

Hypothesis 3a There is no difference in unit nonresponse of the survey between the survey with background questions located at the end and the survey with background questions located at the beginning.

Second, the location of the background questions could deter certain types of respondents to fill out the survey. Because no research could be found that surveyed this issue, we test the following proposition:

Hypothesis 3b There is no difference in the type of respondents that return the survey between the survey with background questions located at the end and the survey with background questions located at the beginning.

A third aspect which will be considered is nonresponse in the survey. We will focus on two types of item-missing data: item nonresponse and partial nonresponse. On the one hand, it is possible that respondents who are asked to fill out the background questions in the beginning of the survey answer fewer questions or dropout early during the survey because they might think their anonymity is not completely guaranteed. On the other hand, respondents might as well be more motivated to answer all the questions in the survey—and thus dropout less—once they have given their personal information. Frick et al. (2001) studied the influence of the location of personal information in online studies and found that asking respondents for personal information at the beginning of the questionnaire did not increase drop-out. On the contrary, drop-out occurred even less when this kind of information was asked at the beginning. Because the arguments are in favor as well as against locating the background questions in the beginning, the following hypotheses will be tested:

Hypothesis 4a There is no difference in drop-out between the survey with background questions located at the end and the survey with background questions located at the beginning.

Hypothesis 4b There is no difference in item nonresponse (specifically the number of questions that were filled out) between the survey with background questions located at the end and the survey with background questions located at the beginning.

In addition to participation bias, placing background questions at the beginning could also lead to measurement bias and particularly more socially desirable responding and thus more underreporting of unethical work behavior. Giles and Feild (1978) also studied the effect of location, format and amount of background questions on the item-response in a job satisfaction survey and found no impact of the amount and location of these questions. However, the type of format (specifically questionnaires with all categorical answers versus those with a combination of open questions and categories) in which the background questions were asked did cause response bias. Because explicit empirical studies on more sensitive topics is lacking, the work of Giles and Feild (1978) is the only available theoretical basis for hypothesizing that the location of the background questions does not influence respondents to underreport their responses:

Hypothesis 5 There is no difference in social desirability response bias between the survey with background questions located at the end and the survey with background questions located at the beginning.

Methods

Survey Implementation

Data collection in this study was done by means of an online survey that was distributed via e-mail in November 2007–February 2008 in nineteen entities of the Flemish government (a regional government of Belgium) who were willing to cooperate in the study. A self-selected sample was used: the organizations, informed and invited through a central government and an ethics committee, participated freely in this research. The survey was distributed among all public servants of the organizations. They received an e-mail request to fill out the survey containing a personalized URL-link with the subject line "survey integrity at work."

An online design was chosen for two reasons. First, it allows to reach a larger pool of possible participants and thus to gather a large amount of information (Dillman 2007). Second, an online survey is considered to create higher perceived anonymity than a paper survey (Frankel and Siang 1999; Ahern 2005). The latter aspect is very important in research on unethical work behavior because participants who perceive they may be identified will probably answer in a more socially desirable way (Frick et al. 2001). However, this argument may be overrated since merit of the online survey in obtaining more reliable information due to this higher perceived anonymity could possibly decrease over time, as more people will be familiar with the possibilities of computers to breach privacy (Peeters 2006).

From a total of 7,060 online surveys that were sent out, 7,017 reached the respondents. 3,386 respondents (48.3 %) opened the survey; 3,167 respondents (44.9 %) answered at least one question; 2,712 respondents (38.6 %) answered at least 50 % of all questions in the survey, and 2,558 respondents (36.4 %) completed the survey.

The typical response rate of surveys on this research topic is low. Randall and Gibson (1990) found in a meta-analysis that the response rates in business ethics research ranged from 10 % in a Business and Society review survey (1975) to 96 % in another study, with a mean response rate of 43 %. Specifically for the current study, we suspect that the relatively low response rate was on the one hand due to the sensitivity of the topic of unethical work behavior and on the other hand due to the length of the survey (see "Questionnaire" section). Therefore, the response rate of this study is considered acceptable. Readers are, however, cautioned when generalizing the findings beyond this sample.

In this article, we use the data of the respondents who opened the questionnaire and thus had some interest in filling out the survey. This refers to a realized sample of

3,386 respondents. Hence, respondents did not have to fill out questions to be included in the sample.

Questionnaire

The online survey "Integrity at the workplace" was developed by researchers of the Leuven Institute of Criminology and of the Public Management Institute, both at the KU Leuven in Belgium (Delbeke et al. 2008). The questionnaire itself consisted of seven parts and addressed several themes related to integrity such as ethical climate, ethical leadership, fairness, and integrity policy. The survey also contained three sets of items measuring different types of unethical work behavior. First, a list of thirty-two items (see Appendix) was drawn from the measurement instrument "Integriteitmeter" of the Bureau of Integrity in Dutch Municipalities (Kolthoff 2007a) that is based on the nine-fold classification of unethical work behavior of Huberts et al. (1999). The advantage of this classification is that it not only includes less serious (e.g., arriving too late at work, little effort, …) but also the more serious forms of unethical work behavior (e.g., corruption) which is a plus point in comparison to other typologies (Lasthuizen et al. 2011). Respondents were successively asked about the unethical work behavior of their colleagues in their workplace in the last year (UWB set 1, proxy-report). Second, a series of 7 items measuring unethical work behavior based on the ethical climate questionnaire (Maesschalck 2004) were included. For these items, respondents were asked to report on the behavior of their colleagues (UWB set 2, proxy-report) as well as their own behavior (UWB set 3, self-report) (see Appendix).[2] As mentioned in part 1.1, both self- and proxy-reporting have their advantages and disadvantages for measuring UWB. A number of closed-ended background questions addressed age, sex, level, and length of service in the public sector. Respondents were never forced to provide answers before allowed to move on.

The researchers acknowledge that sending a survey on UWB to all employees of an organization might of course look like the organization is organizing a witch hunt, which might result in people choosing not to participate. To avoid this as much as possible, the researchers have emphasized the anonymity of respondents in the invitation letter for the survey and in the section that contains the background questions. In the cover letter, the researchers explained several aspects of anonymity: (1) the data gathering and data analysis would be completely anonymous, (2) the researchers would not ask the name of the respondent, (3) the survey pages would be sent directly to the researchers, and (4) the results would only be published as a summary. In the background questions section, the researchers communicated that this personal information was necessary to achieve the research objectives and would certainly not be used to identify wrongdoers.

Split-Ballot Design

To answer the questions raised above, a split-ballot experiment was carried out using four versions of the survey instrument "Integrity at the workplace." We used a two-by-two design, as shown in Table 1. Sampling for these conditions was done at the level of the participating organizations. The workforce of each organization was randomly divided into four groups, representing one of the four experimental conditions.

Two effects (each with two conditions) were studied in the experiment: the effect of the response format on the response behavior and the effect of the position of the background questions in the survey-instrument (see Table 1). In two of the four versions a response scales with labels was used (surveys 2 and 4) and in the other two a scale with anchors (surveys 1 and 3). In addition, in two of the four versions the personal information was requested at the beginning of the survey (surveys 3 and 4) and in the other two at the end of the survey (surveys 1 and 2). Survey 1 and 2 started with some questions on the ethics policy in the respondents' organization.

Results

Anchors Versus Labels

In the first experiment, we compared the two response scales. The purpose of this study was two-fold. First, we wanted to verify whether the all-category-defined (verbal) format would yield significantly different results from the

Table 1 The four versions of the survey-instrument "integrity at the workplace"

	Location of the background questions	
	At the end	In the beginning
Response scale		
Anchors	Survey 1	Survey 3
Labels	Survey 2	Survey 4

[2] The authors note that the measurement of "unethical work behavior" in OMB set 2 and set 3 was not ideal. The items were formulated in a broad way (e.g., "violating laws, rules or procedures to help a friend"). The reason for this way of formulating the items was two-fold. A first reason was based on theoretical arguments. The objective was to evaluate the correlation between specific types of ethical climate and the associated types of unethical work behavior. A second reason was that formulating more specific items would lead to an increased length of the survey.

one with only the anchors defined (nonverbal). Second, we wanted to determine whether there was a difference in variance of the items with both formats.

In order to easily compare the results drawn from both formats and to avoid doing too many tests, we generated a composed variable that combines the answers to more specific questions on unethical work behavior. Specifically, we generated a summative index for the thirty-two items of unethical work behavior based on the typology of Huberts et al. (1999), which was a proxy measurement, (set 1 UWB), an index for the proxy results on the measurement based on the ethical climate typology (set 2 UWB) and one based on the self-report results based on that same typology (set 3 UWB) (see Appendix).

To test hypothesis 1, differences between the two formats in means were examined using a Mann–Whitney U test. A non-parametric technique was used because data violated the assumption of normality. The results of our experiment show that the group that answered the questions with the format with only the anchors defined has indeed a significantly higher mean rank than the group with the all-category-defined format ($p \leq 0.01$ for all UWB sets) (see Table 2).

Differences in variances (hypothesis 2) were examined using a Levene's test for equality of variances. The null hypothesis of equal variances could not be maintained for UWB set 1 ($F = 33.479$; $p = 0.000$) and UWB set 3 ($F = 5.300$; $p = 0.021$). However, for UWB set 2 the null hypothesis was maintained ($F = 1{,}665$; $p = 0.197$).

Thus, as hypothesized, the response scale with anchors provides a higher mean rank (hypothesis 1) for all UWB sets and a higher variance (hypothesis 2) on UWB set 1 and 3. As argued above, we assume that higher means and higher variance reflect less socially desirable responding. Hence, we can conclude that for the measurement of unethical work behavior, it might be better to use the response scale with numeric labels accompanied by anchors at the extremes than the all-category-defined format.

Table 2 Results for hypothesis 1

	N	Mean rank	Sum of ranks
Set 1 UWB (proxy-report)*			
Anchors	964	1,002.98	966,873.00
Labels	863	814.61	703,005.00
Total	1,827		
Set 2 UWB (proxy-report)*			
Anchors	1,131	1,176.85	1,331,019.00
Labels	1,050	998.53	1,048,452.00
Total	2,181		
Set 3 UWB (self-report)*			
Anchors	1,223	1,334.95	1,632,642.00
Labels	1,151	1,030.83	1,186,483.00
Total	2,374		

* Statistically significant at the 0.01 level

Location of the Background Questions

The second experiment examined whether respondents filled out the questionnaire differently when the background questions were placed at the beginning or at the end of the survey.

First, it is useful to check whether our assumption holds that placing the background questions at the beginning indeed generates a higher level of filled out background questions (see Table 3). In order to compare both situations an index was composed which measured the amount of background questions that were filled out. Not surprisingly, the results indicate that significantly more background questions were answered when these questions were assessed in the beginning of the survey (Mann–Whitney U test, $p = 0.000$). This difference could be ascribed to drop-out during the online survey. Another explanation might be that respondents who have just reported on their own and others' misbehavior might reconsider when asked to provide potentially identifiable info. Placing background questions at the end of a survey on UWB might heighten concerns about anonymity leading to more unwillingness to report anything that could possibly be used to link the respondent to his answers.

Although placing the background questions at the beginning of an online survey indeed generates more background information, it is also important to consider two other aspects: the influence on the return rate of the survey and the influence on the response behavior.

We examined whether there was a difference in survey unit response (at least one question filled out). There are two aspects that need to be considered. First, we assessed whether the location of the background questions influenced unit nonresponse (cf. hypothesis 3a). Second, we examined whether the location of the background questions influenced certain types of respondents to reconsider filling out the survey (cf. hypothesis 3b). After all, if some groups

Table 3 Respondents who answered the background questions in percentages

	At the beginning	At the end
Age	93.0	76.1
Gender	92.4	75.8
Level	92.4	75.7
Length of service	92.5	75.8
Organization	90.7	73.9

Table 4 Cross table for the response on the versions with personal information at the beginning and at the end

	Location of the background questions		
	At the end	In the beginning	Total
No return or answers			
Count	2,095	1,798	3,674
Expected count	2,052.9	1,840.1	
Std. residual	0.9	−1.0	
Respondents who filled out at least one question			
Count	1,628	1,539	3,167
Expected count	1,670.1	1,496.9	
Std. residual	−1.0	1.1	
Total of the surveys sent out			
Count	3,723	3,337	7,060

of respondents, as a consequence of the background questions being assessed in the beginning of the survey, systematically do not fill out the survey (for example the respondents from a lower hierarchical level who perhaps do not trust the promised anonymity as much as respondents who are more educated), the sample is not representative and thus the data are not reliable.

The unit response rate for the survey with the background questions in the beginning was 46.1 % and for the survey with the background questions at the end 43.7 %. This finding suggests that placing background questions in the beginning leads to a higher response rate and that starting with questions on the ethics policy in the respondents' organization, as was done in the survey with the background questions in the end, frightens respondents or leads to motivational problems. To test hypothesis 3a and explore if this difference was significant, a χ^2 was calculated (see Table 4). The results show that the two groups (background questions in beginning and background questions at the end) did have significantly different response rates ($\chi^2 = 4.067$a; $p = 0.044$). We can conclude that placing the questions about personal information in the beginning of the survey does not negatively influence the respondents' decision to fill out the questionnaire.

A next question is whether, as a consequence of the location of the background questions, some types of respondents reconsider filling out the survey. We examined whether some respondents reconsider filling out the survey while others do not, using a χ^2 test (hypothesis 3b). Because the different versions of the survey were randomly assigned to the respondents, we can assume that the background variables of the respondents would be equally distributed in the versions. When we tested this hypothesis and compared both types of surveys (personal information assessed in the beginning or at the end), we indeed found no significant differences between the groups in gender ($\chi^2 = 0.158$; $p = 0.691$) nor in age ($p = 0.348$), hierarchical level ($p = 0.663$), or length of service ($p = 0.360$).[3] Hence, we can conclude that the location of the background questions does not have an impact on which type of respondent agrees to cooperate. Of course, we only studied those who filled out the survey. We cannot rule out that the group of respondents differs from the group of non-respondents.

We did not only consider the unit nonresponse, but also the influence of the location of the background questions on item-response. There are two aspects which were examined in this study. A first aspect is the drop-out in the survey (cf. partial nonresponse). A second aspect is the number of questions filled out in general (cf. item nonresponse).

First, the drop-out in the survey was examined (hypothesis 4a). The survey consisted of seven parts, after which respondents needed to click the "further" button and data were sent to the database. Early drop-out during the survey was lower when respondents were asked for personal information at the beginning of the survey. When background questions were located at the beginning of the survey, 82.5 % of the respondents who answered at least one question of the survey also completed the survey. For the survey in which background information was assessed at the end of the survey, this percentage was slightly lower (78.8 %). To explore whether this difference was significant (with a significance level of 0.05) confidence intervals were estimated for the percentages of respondents who started and ended the survey. The results show that the difference was not significant.

Third, we analyzed item nonresponse, specifically whether respondents answered more (or less) questions depending on the location of the background questions (hypothesis 4b). Item nonresponse as a consequence of the location of the background questions, should of course be avoided. It is possible that respondents who receive a survey with the background questions in the beginning go through the survey without filling out many questions. If the assessment of personal information in the beginning of the survey strongly increases item nonresponse, this location obviously needs to be reconsidered. To study the relationship between item-response and the location of the background questions, a new variable was composed that counted the questions which were filled out by the respondent. We found, however, no significant difference between the two groups (personal information assessed in the beginning of the survey or at the end) (Mann–Whitney U test; $p = 0.466$). This suggests that the location of the background questions does not influence item nonresponse.

Finally, we also studied the effect of the location of the background questions on socially desirable responding (cf.

[3] To test possible differences, we used a χ^2 test for gender and Mann–Whitney U tests for age, level and length of service.

Table 5 Results for hypothesis 5

Location of the background questions	*N*	Mean rank	Sum of ranks	*p* value
Set 1 UWB (proxy-report)				
At the beginning	886	918.23	813,556.00	0.739
At the end	941	910.01	856,322.00	
Total	1,827			
Set 3 UWB (self-report)				
At the beginning	1,181	1,183.68	1,397,924.00	
At the end	1,193	1,191.28	1,421,201.00	0.784
Total	2,374			
Set 2 UWB (proxy-report)				
At the beginning	1,084	1,065.02	1,154,478.00	
At the end	1,097	1,116.68	1,224,993.00	0.050
Total	2,181			

hypothesis 5). In order to test hypothesis 5, we performed a Mann–Whitney *U* test with the composite scales of the three sets measuring unethical work behavior (see "Anchors Versus Labels" section) (see Table 5).

The results indicate that no differences were statistically significant. Therefore, we can assume that the early positioning of the background questions did not have any influence on the questions concerning integrity violations. This suggests that placing the background questions in the beginning of the survey does not result in more socially desirable responding. This finding corresponds with Giles and Feild (1978) findings, which showed that the location of the background items did not affect the responses themselves.

Possible Interaction–Effects

A next and final step was to test whether there was an interaction effect between the location of the background questions and the type of response scale. In order to test this, we generated three new composed variables of UWB. Since adding the frequency items leads to skewed dependent variables, dichotomization was preferred in this stage (Osgood et al. 2002). Each question was, therefore, recoded into 0 (original score 0) and 1 (original scores 1–4) and then summed into an index. Osgood et al. (2002) argued that this method was the most successful among several alternatives. We performed two types of analyses to analyze whether there was some evidence of interaction.

First, we compared all four types of surveys on the three summative indexes using a Bonferroni test. These tests showed that there were only significant differences between the surveys that differed in the response format (anchors and labels) and not the location of the background (see example for UWBset1 proxy, Table 6).

Second, we performed a regression analysis to examine a possible interaction effect between location of the background questions and response format (see Table 7). Multiple ordinary least squares (OLS) regression is probably the most widely used variant of regression analysis. One of the assumptions of ordinary least square estimates is that the errors are normally distributed. If this assumption is not met, a traditional multiple regression analysis will not be feasible. A plot of the observed residual distribution against the expected distribution shows the normality assumption is approximately met with the dichotomized indexes. We estimated a regression model with location of the background questions, type of response scale and the interaction between both as independent variables and the three composed variables of UWB as dependent variables. The results showed that in neither of the models the interaction between location of the background questions and response format was significant. Only the type of response format had a significant impact on the results.

Table 6 Results of the Bonferroni test

Survey (*I*)	Survey (*J*)	Mean difference (*I* − *J*)	SE	Sig.	95 % CI	
					Lower bound	Upper bound
1	2	4.06699*	0.44940	0.000	2.8801	5.2539
	3	0.15372	0.44444	1.000	−1.0201	1.3275
	4	4.27346*	0.47923	0.000	3.0077	5.5392
2	1	−4.06699*	0.44940	0.000	−5.2539	−2.8801
	3	−3.91328*	0.43738	0.000	−5.0685	−2.7581
	4	0.20647	0.47269	1.000	−1.0420	1.4549
3	1	−0.15372	0.44444	1.000	−1.3275	1.0201
	2	3.91328*	0.43738	0.000	2.7581	5.0685
	4	4.11975*	0.46798	0.000	2.8837	5.3557
4	1	−4.27346*	0.47923	0.000	−5.5392	−3.0077
	2	−0.20647	0.47269	1.000	−1.4549	1.0420
	3	−4.11975*	0.46798	0.000	−5.3557	−2.8837

* The mean difference is significant at the 0.05 level

Table 7 Results of the multiple regression analysis with UWB as dependent variable

Variable	UWB set 1	UWB set 2	UWB set 3
Intercept	10.829***	2.276***	1.917***
Location of background questions	0.011$^{(ns)}$	0.056$^{(ns)}$	0.005$^{(ns)}$
Type of response scale	−0.287***	−0.214***	−0.290***
Response scale* location	0.003$^{(ns)}$	−0.013$^{(ns)}$	0.017$^{(ns)}$
Adj. R^2	0.079***	0.049***	0.077***

(ns) Not significant

* Statistically significant at the 0.05 level

** Statistically significant at the 0.01 level

*** Statistically significant at the 0.001 level

Discussion and Conclusion

This paper focused on two methodological issues of survey design of questionnaires on unethical work behavior. The goal was to perform experiments to design a survey instrument on UWB that is well-adapted to its research context. A split-ballot experiment allowed us to answer the two questions raised above. First, which type of response scale (one with anchors or one with labels) is best used to measure unethical work behavior? Second, does placing the background questions at the outset of an online survey indeed lower response rates and/or generate different response behavior? Despite some restrictions, the research has produced several conclusions concerning the measurement of unethical work behavior in the public sector.

Before we can conclude, three limitations need to be considered. First, the survey included a self-reported set of UWB as well as two sets of proxy-reported UWB. We could not compare proxy and self-reporting. Further research might investigate whether the same results can be found in a survey only measuring proxy evaluations on UWB and a survey only measuring self-reports UWB. Second, the response rate in this study was not very high but typical for research into UWB. Therefore, readers are cautioned when generalizing the findings beyond this sample. Third, the research population could have had an influence on the results. In general, public servants are used to fill out surveys (occasionally with background questions in the beginning) and are often used to work with computers. More research is needed to assess whether these findings hold in other organizations and other cultural contexts. In addition, it would be interesting to test them for other sensitive and non-sensitive research topics.

The present study, however, makes several noteworthy contributions to the literature on measuring UWB. First, this study has shown that the items for the frequency of unethical work behavior are not equivalent if only the endpoints of the response scale are defined or all categories are defined. In general, the group which responded to the questions in the format with anchors had a significantly higher mean and variance than the group which responded to the format with labels. Based on these findings, we can conclude that the former format is probably better suited to measure the relative frequency of unethical work behavior than the latter, because they are probably less subjected to social desirability response bias. Of course, we could not assess whether respondents lied in filling out the survey. Socially desirable response bias cannot be prevented entirely given the nature of the topic. However, every honest answer is a decrease of the "dark number" of unethical work behavior and thus we can conclude that the response scale with anchors is more valid. However, more research is needed. A response scale with more points could be better (or worse) to measure the frequency of unethical work behavior and other labels could also have other results. Further research needs to be undertaken to investigate whether it might be better (or worse) to use exact frequencies or to use an open question to assess the frequency of unethical work behavior. In the present study, numbers were printed above the bullets of the non-verbal response scale. Omitting these numbers in future research could have different results because of possible interpretation difficulties.

Second, our findings did not provide support for the conventional practice of asking background questions at the end of a survey so as to avoid concerns about anonymity that might in turn influence the reliability and validity of the data (because of social desirability effects). Not surprisingly, the results indicated that more demographic questions were answered when they were placed at the beginning of the survey. More importantly, the results also indicated that this advantage does not come at a cost. The location of the background questions did not influence the return rate nor the response behavior of participants. These results indicate that, when background data are indispensable, placing the background questions at the beginning of a survey about unethical work behavior in a large sample is a risk worth taking. In spite of some restrictions, we can conclude that placing the background questions in the beginning of an online survey on unethical work behavior in the public sector does not systematically influence data. We reason that this might be due to the fact that the researchers provided information about the aim of the study, how the data would be used and how the anonymity of the respondents would be protected. We argued that giving information was necessary for two reasons. First, unethical work behavior is a sensitive topic and informing the participants about the aim of the study and the use of the data could help them better to understand why researchers ask them certain questions. Second,

anonymity had to be explicitly assured to the respondents because if they believed they could or would be identified they might not have agreed to participate in the study or perhaps would have answered in a more socially desirable way (Frick et al. 2001). Further work needs to be done to determine whether providing information decreases socially desirable responding.

Acknowledgments This research was supported by project grants from the "Steunpunt Bestuurlijke Organisatie Vlaanderen." The authors wish to thank Karlien Delbeke, Annelies De Schrijver, Arne Geeraerts, Annie Hondeghem, Kim Loyens and Stefaan Pleysier for their helpful comments and suggestions. The authors would also like to thank the Editor and the anonymous reviewer for constructive comments.

Appendix

List of Items of Unethical Work Behavior of UWB Set 1 (Proxy-Report)

1. Minimal effort by employees (laziness).
2. Gossiping.
3. Use of the internet, e-mail, or telephone above the permitted standard.
4. Favoritism by superiors.
5. Accepting small gifts from external parties.
6. Falsely reporting in sick.
7. Use of organizational resources for private purposes.
8. Careless handling of employees or external parties.
9. Neglecting core tasks or responsibilities in order to engage in more pleasant business.
10. Bullying (e.g., teasing, ignoring, or isolating).
11. Careless use of organizational properties.
12. Executives placing unaccepted pressure to influence things.
13. Careless handling of confidential information.
14. Disclosing confidential information to external parties.
15. Politicians placing unacceptable pressure to influence things.
16. Excessive use of alcohol while on duty.
17. Concealing information from the supervisory authorities.
18. Theft of organizational properties.
19. Favoring of friends or family outside the organization.
20. Setting a bad example in private time.
21. Deliberately delaying decision-making processes.
22. Incorrect handling of expense claims.
23. Not reporting illegal behavior.
24. Giving advice to externals in private time concerning the organizational specialism.
25. Discrimination based on sex, race or sexual orientation of colleagues.
26. Sideline activities or jobs that might pose a conflict of interest.
27. Unauthorized use of a colleague's password or access code.
28. Deliberately giving false information in reports and/or evidence.
29. Accepting bribes (money or favors) to do or neglect something while at work.
30. Accepting gifts of more serious value from external parties.
31. Sexual intimidation.
32. Procuring confidential information to third parties for remuneration.

List of Items of Unethical Work Behavior of UWB Set 2 (Proxy-Report) and UWB Set 3 (Self-Report)

1. Violating laws, rules or procedures because you do not agree with them by your own personal beliefs.
2. Ignoring important goals to work efficiently.
3. Violating laws, rules or procedures to protect your own interest.
4. Violating laws, rules or procedures to help a friend.
5. Violating laws, rules or procedures to protect colleagues from the same team or group.
6. Violating laws, rules or procedures to help a citizen in the course of your occupation.
7. Hiding unethical issues from people outside the organization to protect the image of the organization.

References

Ahern, N. R. (2005). Using the internet to conduct research. *Nurse Researcher, 13*, 55–70.

Arnold, H. J., & Feldman, D. C. (1981). Social desirability response bias in self-report choice situations. *Academy of Management Journal, 24*, 377–385.

Babbie, E. R. (1997). *The practice of social research*. Belmont: Wadsworth Publishing.

Bennett, R., & Robinson, S. L. (2000). Development of a measure of workplace deviance. *Journal of Applied Psychology, 85*, 349–360.

Beretvas, S. N., Meyers, J. L., & Leite, W. L. (2002). A reliability generalization study of the Marlowe–Crowne social desirability scale. *Educational and Psychological Measurement, 62*, 570–589.

Billiet, J., & Waege, H. (2006). *Een samenleving onderzocht. Methoden van sociaal-wetenschappelijk onderzoek*. Antwerpen: De Boeck.

Borgers, N., Hox, J., & Sikkel, D. (2003). Research quality in survey research with children and adolescents: The effect of labeled response options and vague quantifiers. *International Journal of Public Opinion Research, 15*, 83–94.

Borkeneau, P., & Ostendorf, F. (1992). Social desirability scales as moderator and suppressor variables. *European Journal of Personality, 6*, 199–214.

Bradburn, N. M., & Miles, C. (1979). Vague quantifiers. *Public Opninion Quarterly, 43*, 92–101.

Chung, J., & Monroe, G. S. (2003). Exploring social desirability bias. *Journal of Business Ethics, 44*, 291–302.

Cohen, J. F., Pant, L. W., & Sharp, D. J. (1998). The effect of gender and academic discipline diversity on the ethical evaluations, ethical intentions and ethical orientation of potential public accounting recruits. *Accounting Horizons, 13*, 250–270.

Cohen, J. F., Pant, L. W., & Sharp, D. J. (2001). An examination of differences in ethical decision-making between Canadian business students and accounting professionals. *Journal of Business Ethics, 3*, 319–336.

Crane, A. (1999). Are you ethical? Please tick yes□ or no□ on researching ethics in business organizations. *Journal of Business Ethics, 20*, 237–248.

Delbeke, K., Geeraerts, A., Maesschalck, J., & Hondeghem, A. (2008). *Naar een meetinstrument ter ondersteuning van het ambtelijk integriteitsbeleid. Handleiding survey 'Integriteit op het werk'*. Leuven: Bestuurlijke Organisatie Vlaanderen.

Dillman, D. A. (2007). *Mail and internet surveys. The tailored design method* (2nd ed.). New York: Wiley.

Dixon, P. N., Bobo, M., & Stevick, R. A. (1984). Response differences and preferences for all-category-defined and end-defined likert formats. *Educational and Psychological Measurement, 44*, 61–66.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. New York: Holt, Rinehart and Winston.

Finn, R. H. (1972). Effects of some variations in rating scale characteristics on the means and reliabilities of rating. *Educational and Psychological Measurement, 32*, 255–265.

Fowler, F. J. (1995). *Improving survey questions. Design and implementation*. London: Sage Publications.

Frankel, M. S., & Siang, S. (1999). *Ethical and legal aspects of human subjects research on the internet*. Washington, DC: AAAS Program on Scientific Freedom, Responsibility and Law.

Frederickson, H. G., & Walling, J. D. (2001). Research and knowledge in administrative ethics. In T. L. Cooper (Ed.), *Handbook of administrative ethics* (2nd ed., pp. 37–58). University Park: The Pennsylvania State University.

Frick, A., Bächtiger, M. T., & Reips, U.-D. (2001). Financial incentives, personal information and drop-out rate in online studies. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of internet science* (pp. 209–219). Lengerich: Pabst Science Publishers.

Frisbie, D. A., & Brandenburg, D. C. (1979). Equivalence of questionnaire items with varying response formats. *Journal of Educational Measurement, 16*, 43–48.

Ganster, D. C., Hennessey, H. W., & Luthans, F. (1983). Social desirability response effects: Three alternative models. *Academy of Management Journal, 26*, 321–331.

Gaskell, G. D., O'Muircheartaigh, C. A., & Wright, D. B. (1994). Survey questions about the frequency of vaguely defined events. The effects of response alternatives. *Public Opninion Quarterly, 58*, 241–254.

Giles, W. F., & Feild, H. S. (1978). Effects of amount, format, and location of demographic information on questionnaire return rate and response bias of sensitive and nonsensitive items. *Personnel Psychology, 31*, 549–559.

Groves, R. M., & Cooper, M. P. (1998). *Nonresponse in household interview survey*. New York: Wiley.

Heerwegh, D. (2005). Effects of personal salutations in e-mail invitations to participate in a web survey. *Public Opninion Quarterly, 69*, 588–598.

Hindelang, M. J., Hirschi, T., & Weis, G. (1981). *Measuring delinquency*. Thoasand Oaks: Sage Publications.

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin, 30*, 161–172.

Huberts, L. W. J. C., Pijl, D., & Steen, A. (1999). Integriteit en corruptie. In C. Fijnaut, E. Muller, & U. Rosenthal (Eds.), *Politie. Studies over haar werking en organisatie* (pp. 433–472). Alphen aan den Rijn: Samsom.

Junger-Tas, J., & Marshall, I. H. (1999). The self-report methodology in crime research. *Crime and Justice, 25*, 291–367.

Kaptein, M., & Avelino, S. (2007). Measuring corporate integrity: A survey-based approach. *Corporate Governance, 5*, 45–54.

Kolthoff, E. (2007a). *Ethics and new public management: Empirical research into the effects of businesslike government on ethics and integrity*. Den Haag: Boom Juridische uitgevers.

Kolthoff, E. (2007a). *Bing Integriteitmeter* (unpublished survey). Den Haag: Dutch Office of Local Government Ethics.

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 141–165). New York: Wiley.

Lasthuizen, K. (2008). *Leading to integrity. Empirical research into the effects of leadership on ethics and integrity*. Amsterdam: Vrije Universiteit Amsterdam.

Lasthuizen, K., Huberts, L. W. J. C., & Heres, L. (2011). How to measure integrity violations. Towards a validated typology of unethical behavior. *Public Management Review, 13*, 383–408.

Lee, R. M. (1999). *Doing research on sensitive topics*. London: Sage.

Loyens, K. (2012). *Integrity secured. Understanding ethical decision making among street-level bureaucrats in the Belgian Labor Inspection and Federal Police*. Leuven: KU Leuven.

Maesschalck, J. (2004). *Towards a public administration theory on public servant's ethics*. Leuven: Katholieke Universiteit Leuven.

Mars, G. (1999). *Cheats at work: An antrophology of workplace crime*. Aldershot: Ashgate.

McDonald, G. (2000). Cross-cultural methodological issues in ethical research. *Journal of Business Ethics, 27*, 89–104.

Moxey, L. M., & Sanford, A. J. (1993). *Communicating quantities: A psychological perspective*. Hove: Lawrence Erlbaum.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology, 15*, 263–280.

Newstead, S. E. (1988). Quantifiers as fuzzy concepts. In T. Zétényi (Ed.), *Fuzzy sets in psychology* (pp. 51–72). Amsterdam: Elsevier Science Publishers.

Osgood, D. W., McMorris, B., & Potenza, M. T. (2002). Analyzing multiple-item measures of crime and deviance I: Item response theory scaling. *Journal of Quantitative Criminology, 18*, 267–296.

Peeters, C. F. W. (2006). *Measuring politically sensitive behavior. Using probability theory in the form of randomized response to estimate prevalence and incidence of misbehavior in the public sphere: a test on integrity violations*. Amsterdam: Dynamics of Governance, Vrije Universiteit Amsterdam.

Pepper, S. (1981). Problems in the quantification of frequency expressions. In D. Fiske (Ed.), *New directions for methodology of social and behavioral science: Problems with language imprecision* (pp. 25–41). San Francisco: Jossey-Bass.

Peterson, D. K. (2002). The relationship between unethical behavior and the dimensions of the Ethical Climate Questionnaire. *Journal of Business Ethics, 41*(4), 313–326.

Randall, D. M., & Fernandes, M. F. (1991). The social desirability response bias in ethics research. *Journal of Business Ethics, 10*, 805–817.

Randall, D. M., & Fernandes, M. F. (1992). Social desirability bias in ethics research. *Business Ethics Quarterly, 2*, 183–205.

Randall, D. M., & Gibson, A. M. (1990). Methodology in business ethics research: A review and critical assessment. *Journal of Business Ethics, 9*, 457–471.

Roberson, M. T., & Sundstrom, E. (1990). Questionnaire design, return rates, and response favorableness in an employee attitude questionnaire. *Journal of Applied Psychology, 75*, 354–357.

Sampford, C., Shacklock, A., Connors, C., & Galtung, F. (2006). *Measuring corruption*. Hampshire: Ashgate.

Schoderbek, P. P., & Deshpande, S. P. (1996). Impression management, overclaiming, and perceived unethical conduct: The role of male and female managers. *Journal of Business Ethics, 15*, 409–414.

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

Skipper, R., & Hyman, M. R. (1993). On measuring ethical judgements. *Journal of Business Ethics, 12*(7), 535–545.

Sloan, J. J., Bodapati, M. R., & Tucker, T. A. (2004). Respondents misreporting of drug use in self-reports: Social desirability and other correlates. *Journal of Drug Issues, 34*, 269–292.

Stoop, I. A. L. (2005). *The hunt for the last respondent*. The Hague: Social and Cultural Planning Office of the Netherlands.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions. A practical guide to questionnaire design*. London: Jossey-Bass Publishers.

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format and question context. *Public Opninion Quarterly, 60*, 275–304.

van der Heijden, P. G. M., Van Gils, G., Bouts, J., & Hox, J. J. (2000). A comparison of randomized response, computer-assisted self-interview, and face-to-face direct questioning. *Sociological Methods and Research, 28*, 505–537.

Vardi, Y., & Wiener, Y. (1996). Misbehavior in organizations: A motivational framework. *Organization Science, 7*, 151–165.

Victor, B., Trevino, L. K., & Shapiro, D. L. (1993). Peer reporting of unethical behavior: The influence of justice evaluations and social context factors. *Journal of Business Ethics, 12*(4), 253–263.

Wanek, J. E. (1999). Integrity and honesty testing: What do we know? How do we use it? *International Journal of Selection & Assessment, 7*, 183–195.

Weisberg, H. F. (2005). *The total survey error approach: A guide to the new science of survey research*. London: University of Chicago Press.

Wells, W. D., & Smith, G. (1960). Four semantic rating scales compared. *Journal of Applied Psychology, 44*, 393–397.

Wildman, R. C. (1977). Effects of anonymity and social setting on survey responses. *Public Opninion Quarterly, 41*, 74–79.

Wright, D. B., Gaskell, G. D., & O'Muircheartaigh, C. A. (1994). How much is 'Quite a bit'? Mapping between numerical values and vague quantifiers. *Applied Cognitive Psychology, 8*, 479–496.