Survey Research Methods (2020)
Vol. 14, No. 3, pp. 253-266
doi:10.18148/srm/2020.v14i3.7613

© European Survey Research Association
ISSN 1864-3361
http://www.surveymethods.org

# Designing better questions for complex concepts with reflective indicators

Willem Saris
Sociometric Research Foundation
Barcelona, Spain

Irmtraud Gallhofer
Sociometric Research Foundation
Barcelona, Spain

There are many concepts in the social sciences that are measured using multiple indicators. Such concepts have been called by Blalock (1968) concepts-by-postulation because one needs a theoretical argument to define them. Within the set of concepts-by-postulation, a distinction has been made between concepts with reflective indicators and concepts with formative indicators. This distinction refers to the assumption whether the latent concepts determine the observed indicators (reflective) or that the indicators together determine the latent concept of interest (formative). Blalock complained that developing measurement procedures for complex concepts, researchers think mainly about questions not about concepts that these questions measure. In this way the questions used contain unique components which reduce the quality of the composite score based on these questions as measure for the complex concept of interest. Saris and Gallhofer have shown how alternative formulated questions can be developed to measure so called concepts-by-intuition. In this paper we will show that the same procedure can be used to avoid unique components in the measurement of complex concepts with reflective indicators and that in this way the quality of the composite score for complex concepts can considerably be increased.

*Keywords:* design of survey questions; reflective indicators; quality of composite scores

## 1 Introduction

The effects that the wording of survey questions can have on their responses have been studied in depth by Alwin and Krosnick (1991), Andrews (1984), Költringer (1993), Molenaar (1986), Saris and Gallhofer (2007), Scherpenzeel and Saris (1997), Schuman and Presser (1981), Sudman and Bradburn (1983). In contrast, very little attention has been given to the problem of translating concepts into questions (De Groot & Medendorp, 1986; Hox, 1997). Saris and Gallhofer (2007, 2014) have tried to fill this gap in the literature making use of the valuable work in this context of Blalock (1961, 1968, 1990), who following Northrop (1947) distinguished between concepts-by-intuition and concepts-by-postulation.

Regarding the differentiation between concepts-by-intuition and concepts-by-postulation, (Blalock, 1990, p. 34) asserts the following:

> Concepts-by-postulation receive their meaning from the deductive theory in which they are embedded. Ideally, such concepts would be taken either as primitive or undefined or as defined by postulation strictly in terms of other concepts that were already understood. Thus, having defined mass and distance, a physicist defines density as mass divided by volume (distance cube). The second kind of concepts distinguished by Northrop are concepts-by-intuition, or concepts that are more or less immediately perceived by our sensory organs (or their extensions) without recourse to a deductively formulated theory. The color "blue" as perceived by our eyes, would be an example of a concept-by-intuition, whereas "blue" as a wavelength of light would be the corresponding concept-by-postulation.

The distinction he makes between the two concepts follows the logic that concepts-by-intuition are simple concepts, the meaning of which is immediately obvious, while concepts-by-postulation are less obvious concepts that require explicit definitions. Note that not all simple observations represent concepts-by-intuition. For example reading the temperature for a thermometer is a typical example of a concept-by-postulation because it requires a theoretical argument why the height of mercury in a tube indicates the concept "temperature in physics". On the other hand a simple question, "How warm do you feel?" is a measure for a concept-by-intuition "human temperature". Note that this difference does not say anything about the quality of the mea-

Contact information: Willem Saris, SRF, Calle Josep Pla 27 9-4, 08019 Barcelona, Spain (E-mail: w.saris@telefonica.net)

sures.

Examples of concepts-by-intuition include judgments, feelings, evaluations, norms, and behaviors. Most of the time, it is quite obvious that a text presents a feeling (x likes y), a norm (people should behave in a certain way), or behavior (x does y). Examples of concepts-by-postulation might include "ethnocentrism", different forms of "racism", and "attitudes toward different objects". One item on its own in a survey cannot present an attitude or racism. For such concepts, more items are necessary and therefore, these concepts need to be defined and thus are concepts-by-postulation.

One of the major problems in the operationalization process of concepts-by-postulation is, as Blalock suggested, that the researchers are not thinking in terms of concepts-by-intuition, but only in terms of questions. They operationalize concepts-by-postulation without a clear awareness of the basic concepts-by-intuition being represented by the questions.

This observation leads us to suggest a two steps procedure: first of all a study of the definition of concepts-by-postulation by concepts-by-intuition and secondly the specification of questions for concepts-by-intuition. In this paper, we want to show that in this way one will get better measures for concepts-by-postulation. Therefore, we will concentrate on the definition of concepts-by-postulation through concepts-by-intuition. Next we will show that concepts-by-postulation based on alternative forms of concepts-by-intuition provide better measures for the concepts-by-postulation then the use of questions for different concepts-by-intuition. Then we will discuss how alternative questions for concepts-by-intuition can be formulated and how possible memory effects can be reduced.

We start with a simple example: the measurement of "job satisfaction". We define this concept as the feeling a person has about his/her job. We believe that though this feeling exists in people's minds, it is not possible to observe it directly. Therefore this unobserved variable is called a *latent variable*. We give this latent variable the name "job satisfaction" and a short form "JS".

## 1.1 Job satisfaction as a concept-by-intuition

Measuring job satisfaction can appear to be a simple task if one thinks of it as a concept-by-intuition that can be measured with a direct question: *How satisfied or dissatisfied are you with your job?*

1. Very satisfied
2. Satisfied
3. Dissatisfied
4. Very dissatisfied

Indeed, many past studies (Blauner, 1966; Robinson, Athanadiou, & Head, 1969) as well as more recent ones (ESS Round 6: European Social Survey Round 6 Data, 2012) have relied on this direct question, or a variation of it. Such an operationalization assumes that people can express their job satisfaction in the answer to such a simple question. However, we must accept that errors will be made in the process, whether due to mistakes in respondents' answers or in interviewers' recordings of them.

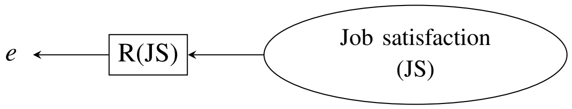

*Figure 1.* A measurement model for a direct measure of job satisfaction

In Figure 1 we present this process through a path model. This model suggests that people express their job satisfaction directly in their response with the exception of some errors. The variable of interest is job satisfaction. This *latent* or *unobserved* variable is presented in the circle. The responses to the direct question presented can be observed directly. Such variables are usually presented in squares while the random errors, inherent in the registration of any response, are normally denoted by *e*. This model suggests that the verbal report of the question is determined by the unobserved variable job satisfaction and errors. As shown in the model, the response to the JS question is denoted as R(JS). We will use this notation throughout this paper.

This approach to measure job satisfaction with a direct question presupposes that the meaning of job satisfaction is obvious to everyone and that people share a common interpretation of it. In other words, it assumes that when asked about their job satisfaction, all respondents are answering the same question.

The approach discussed here, assuming that the concept of interest is a concept-by-intuition that can be measured by a direct question, can be applied to many concepts, such as "political interest", "left-right orientation", "trust in the government", and many other attitudes.

However, it has also been criticized as being over-simplistic. For example, with respect to the direct measure of job satisfaction, some argue that asking people about their degree of job satisfaction is naïve because such a question requires a frank and simple answer with respect to what may be a complex and vague concept (Blauner, 1966; Wilensky, 1964). These researchers deny that job satisfaction can be seen as a concept-by-intuition. Others have said that such a direct question leads to too many errors and offers too low reliability (Robinson et al., 1969). Let us therefore look at the alternatives. We will first discuss the complexity problem and then follow with the reliability issue.

## 1.2 Job satisfaction as concept-by-postulation

Many scholars have suggested that one's feelings about one's job are based on one's satisfaction with its different aspects. Clark (1998) mentions that the following aspects are highlighted in the literature: salary and working hours,

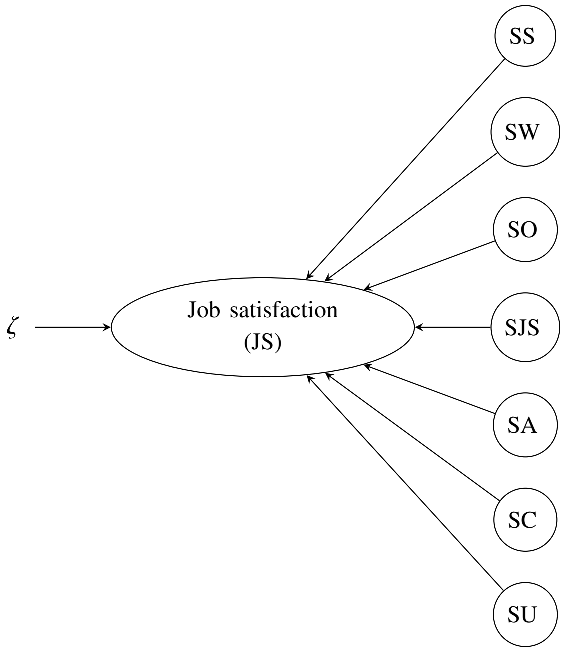

*Figure 2.* The operationalization of job satisfaction by a set of formative indicators where SS=satisfaction with the salary, SW=satisfaction with the working hours, SO=satisfaction with opportunities for advancement, SJS= satisfaction with job security, SA=satisfaction with autonomy, SC=satisfaction with contacts and SU= satisfaction with usefulness of the job while $\zeta$ is the disturbance term in this measurement model.

opportunities for advancement, job security, autonomy in the work, social contacts and usefulness of the job for society. This operationalization suggests that job satisfaction is affected by satisfactions with these different aspects of the job. This is different from the situation we depicted above. In the previous section, we suggested that an opinion of job satisfaction determines the response, which is the measure for job satisfaction. Here, we are suggesting that it is the level of satisfaction with the different aspects of a job that determine or form a person's job satisfaction. Therefore, the measures of these aspects are called *formative indicators* for the concept-by-postulation Job satisfaction. This leads to a different model as shown in Figure 2.

So far, we have only defined the concept-by-postulation through other concepts which are causes of job satisfaction. We have done this in order to go from the concept-by-postulation to the concepts-by-intuition. If this theory is correct, then we can ask respondents about their satisfaction with these different aspects and therefore, obtain information about their job satisfaction. For example, we can ask:

*How satisfied or dissatisfied are you with the following aspects of your job? Give your judgment in a number from 0 to 10, where 0 means completely dissatisfied and 10 completely satisfied*

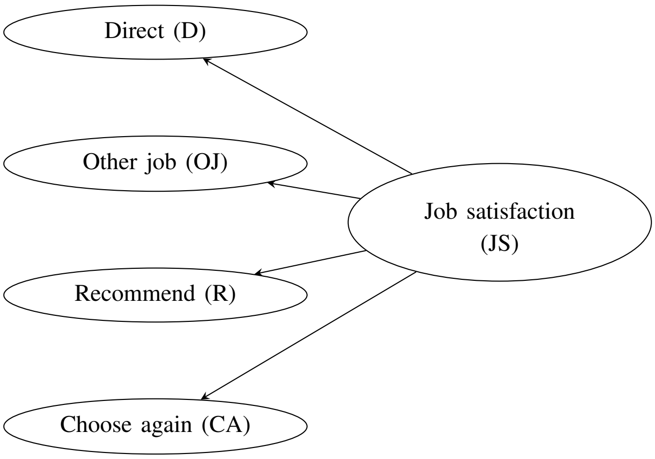

*Figure 3.* The measurement model for a concept-by-postulation with reflective indicators

- Your salary (SS)
- Working hours (SW)
- Opportunities for advancement (SO)
- Job security (SJS)
- Autonomy in the work (SA)
- Social contacts (SC)
- Usefulness of the job for society (SU)

This procedure is used often for the operationalization of complex concepts or, as we call them, concepts-by-postulation. The idea is to determine different causes of the concept-by-postulation and ask questions about these aspects. These so-called "indicators" can be concepts-by-intuition that can be directly converted into questions as shown for job satisfaction. However, it may be that these aspects are still too complex themselves and need to be decomposed even further before one ends up with concepts-by-intuition.

Although this procedure for the operationalization of concepts-by-postulation with formative indicators seems very logical, it also has some very serious limitations:

1. All important aspects need to be included. If not, the measurement is incomplete and therefore will be *invalid.*

2. As mentioned before the model in Figure 2 is assumed to be a causal model for which the effect parameters have to be estimated. However, the problem is that the dependent variable JS is a latent variable. Such a model is not identified, which means that the effect parameters can only be estimated if one adds to the model extra dependent variables. But with the chosen variables the causal effects will vary and so the relationships between the latent variable of interest and the composite score obtained for this variable will vary too.

3. The effects of the satisfactions of each cause can vary

for different groups. Scherpenzeel and Saris (1996) have shown that this is the case for the different causes of life satisfaction.

4. Ignoring these differences would mean that the *researcher* determines what job satisfaction is. This definition might be quite different from the latent variable that exists in the mind of the respondents.

5. It is not necessarily true that all aspects affect the latent variable of interest. It has been found that for some people, the latent variable (life satisfaction) determines the satisfaction with its supposed causes (Scherpenzeel & Saris, 1996).

This overview shows that this approach of measuring concepts-by-postulation using formative indicators encounters serious problems. Given these problems this approach has heavily been criticized (Aguirre-Urreta, Rönkkö, & Marakas, 2016; Borsboom, Mellenbergh, & Van Heerden, 2003; Edwards, 2011; Lee & Cadogan, 2013), even so much that it was suggested by Hardin and Marcoulides (2011) "to consider temporary suspending the use of formative measurement till these problems have been solved". We agree with this point of view and will concentrate mainly on the concepts-by-postulation with reflective indicators.

## 1.3 Operationalization using reflective indicators

Given these problems in operationalization and the possible unreliability of the direct questions, we will restrict the discussion to the concepts-by-postulation with reflective indicators and illustrate this alternative, once again with the example of job satisfaction. This approach assumes that an individual's job satisfaction has effects on other opinions. In several studies, Kalleberg (1974, 1975, 1977) has suggested, in addition to using the direct question, to use an indicator that we shall call "other job". The idea is that an individual who is very satisfied with his job will want to continue in the same job, whereas someone who is dissatisfied will prefer the possibility of another job. Another indicator he has suggested is denoted as "recommendation". Here, the assumption is that satisfied people will recommend their jobs to friends, while dissatisfied people will not. A third one is called "choose again", which is based on the idea that someone who is satisfied would choose this job again if he had the opportunity to do so, whereas a dissatisfied person would choose a different job. As we can see, all these cases are based on the assumption that job satisfaction determines the other opinions. In other words, such indicators "reflect" an individual's feeling of job satisfaction. For this reason, we shall call them "reflective indicators". In this case, the concept-by-postulation (job satisfaction) affects the different indicators. This is illustrated in Figure 3.

If the different indicators are seen as concepts-by-intuition, then one can develop direct questions for each of them. This possibility has indeed been used in several studies by Kallenberg and others. For example, one could use the following questions to measure the concepts-by-intuition that are used as reflective indictors for job satisfaction: *Would you say that you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree with the following statements?*

- Overall I am satisfied with my job (D)
- I would like to have a different job (OJ)
- I would recommend my job to a friend (R)
- I would choose my job again if I had the opportunity (CA)

Note that the responses to these questions are expected to be a consequence of the opinions about these concepts-by-intuition. This leads us to extend the model in Figure 4 by including these effects as well as the possibility of errors.

The reason for which researchers suggest using not only one question, such as the direct question, is that they expect that this question alone will contain too many errors. The idea is therefore that the combination of responses to several questions, which are all observable indicators of the concept of interest, will provide a more reliable measure of that concept. This explains why researchers normally use a weighted or unweighted sum of the observed scores on the different indicators as the measure for job satisfaction. We present the final model of this measurement process in Figure 5.

It will be clear that the same process can be applied and has been applied on many other concepts-by-postulation. This is therefore also an illustration of a general approach. One can look for different reflective indicators for a specific concept of interest. If these indicators represent concepts-by-intuition, the concepts can be directly transformed into questions. After collecting responses to these questions, the researcher can combine the scores and obtain a composite score for the concept-by-postulation being measured. Of course, the composite score is only as good as the theory used in the model and the size of the measurement errors in the observed variables.

## 1.4 The problems of these concepts-by-postulation

The direct question most likely is only measuring how a person feels about his job. No other perceptions will influence this response. However, when asked about "whether he/she would like to have another job, his opinion is not only affected by his satisfaction with his own job, but also by the satisfaction he/she could have in other jobs. The variable "Other Job" is a concept-by-postulation with two formative indicators where one is Job satisfaction. Similarly, with respect to the question "Recommend to a friend", the respondent will not only reflect on his/her personal job satisfaction, but also on the capacities of his/her friend as well as his/her friend's own job satisfaction. Finally for the concept "Choose the job again" also different possible options next to his present job will be compared.

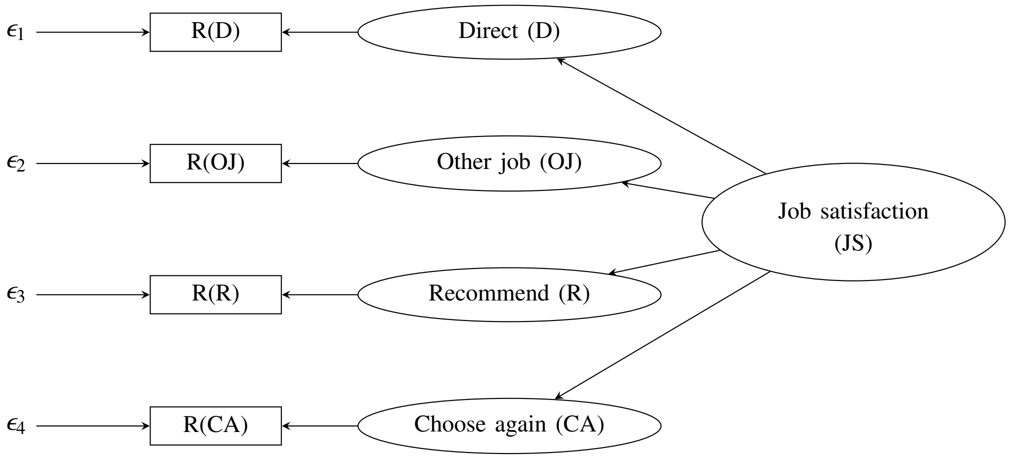

*Figure 4*. The measurement model for "job satisfaction" using concepts-by-intuition as reflective indicators

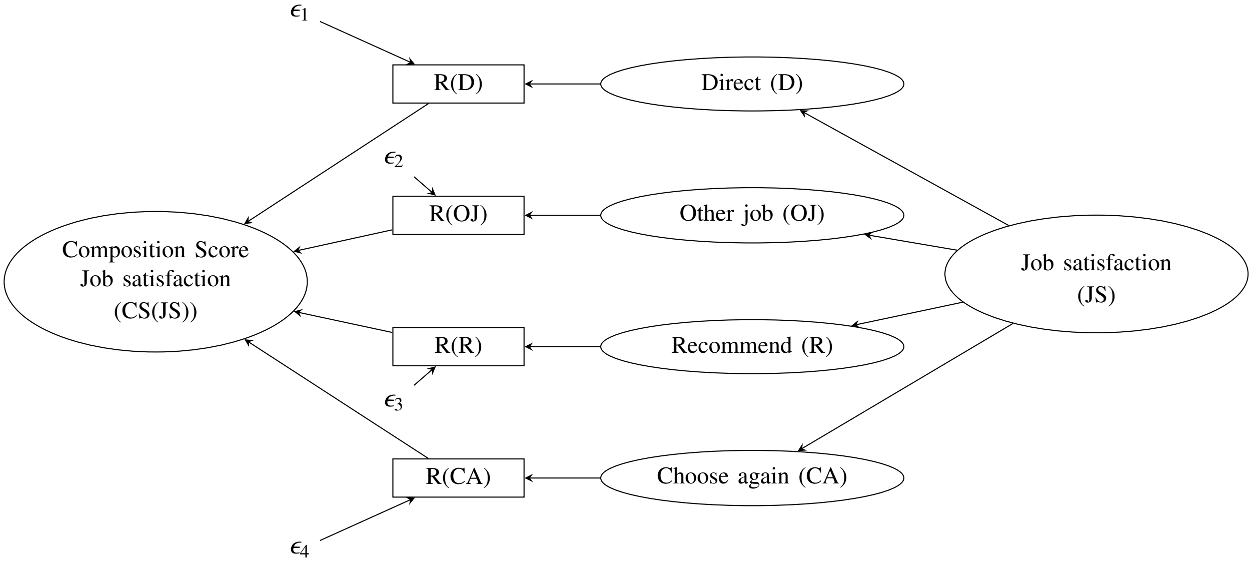

*Figure 5*. The model for determineing the composite score of "job satisfaction" using reflective indicators

This means that while the answer on the direct Job satisfaction question will only be determined by the latent variable Job satisfaction, the other variables will not only be determined by Job satisfaction, but also by systematic effects of specific other variables. While one wants to measure job satisfaction, it turns out that some of these reflective indicators are not just influenced by the variable of interest, but also by other variables. These variables often are denoted for simplicity as unique components of the different concepts. Because these variables are quite different from each other they are supposed to be uncorrelated. Only the question called "Direct" is not effected by a unique component because it is a direct measure of job satisfaction and the response variable contains only random measurement errors.

One might expect that when calculating the composite score, the number of random errors in this measure for job satisfaction will be smaller than for the single question R(D) because the random errors cancel each other out. But the unique components remain present in the composite score whenever the number of indicators is small, as is the case in most survey research.

Saris (1981) has studied this problem and found that the indicators "recommend to a friend" and "choose again" have an overlap of only 70% after correction for random measurement errors. This suggests that the unique components in these questions are quite large. Given this situation one may wonder if the composite score for such a concept-by-postulation with reflective indicators is a better measure of the concept of interest than a direct question. We will show that this is indeed not the case. This suggests that one can doubt the quality of the many measures developed in this way, analyzed with factor analysis and for which composite

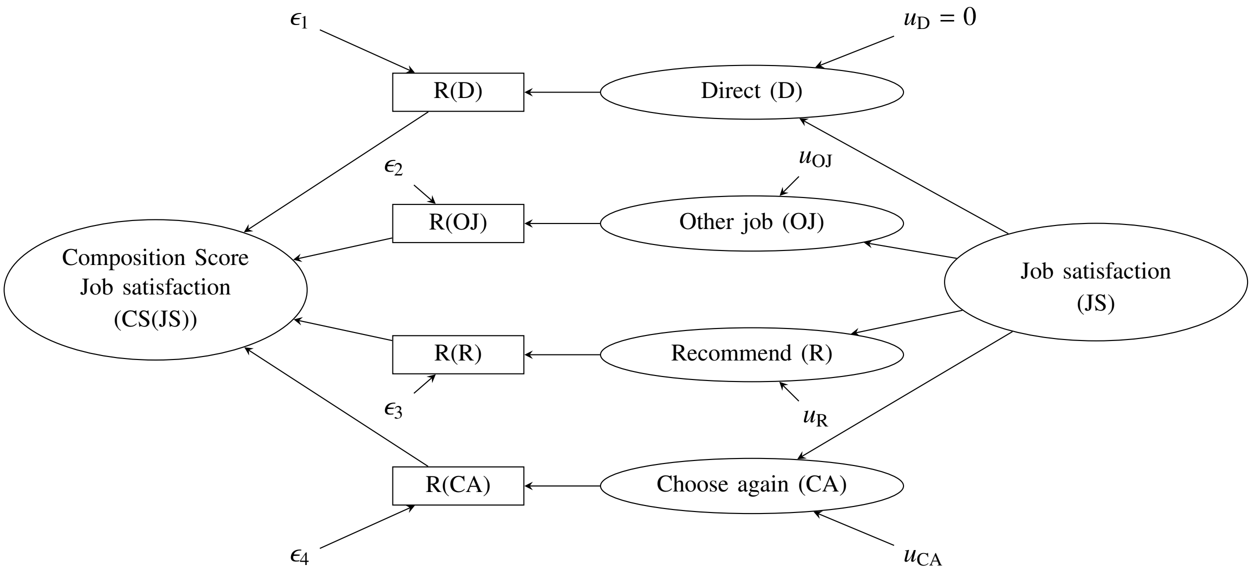

*Figure 6*. The complete model for measurement of "job satisfaction" using reflective indicators

scores are computed.

The logical solution to this problem would be to avoid using different questions and instead use the same direct question repeatedly in order to increase the reliability of the composite scores. While this simple procedure can be used in physics, it leads to problems in the social sciences because of memory effects. Therefore, this solution requires special attention.

Saris and Gallhofer (2014) suggest that researchers should allow for sufficient time between the first presentation of questions and the second presentation of the same questions to avoid memory effects. In order to reduce the time between the repetitions, it is possible to vary the response options. In that way the content of the question remains the same, while the respondent cannot rely on the memory of the first answer in responding to the second question.

In this paper we will first show that the quality of a composite score based on questions that measure only the same latent variable is better than the quality of a single question but also that the composite score of questions which don't measure only the same latent variable have a lower quality than the composite score of questions that only measure the one latent variable of interest.

Next we like to suggest other possibilities than just repeating the same question to increase the quality of the composite scores. The new possibilities make use of differently formulated questions which also measure only the same concept of interest without unique components. The way these questions can be formulated is based on the procedure of Saris and Gallhofer (2014) for designing valid questions for concepts-by-intuition. Then we will show that this procedure can be used to generate alternative valid questions for the same concepts-by-intuition. In the next section we will present 4 different approaches to create questions for concepts-by-postulation with reflective indicators. Then we will suggest briefly two ways to reduce memory and method effects. Finally we will make an overview of all possible ways questions for concepts-by-postulation with reflective indicators can be formulated together with the approaches to reduce the memory and method effects and we will evaluate these different procedures.

## 2 Effect of unique components on the quality of measures

In order to show that using questions with unique components to measure a concept-by-postulation will decrease the quality of the composite score we use again as an example the two questions about job satisfaction for illustrative purposes:

1. How satisfied are you with your job?
2. Would you like to have another job?

The first question is a direct question for the concept-by-intuition job satisfaction. This measure is only determined by the opinion about job satisfaction (JS) except for random errors.

This means for the observed variable $Y_1$ that:

$$Y_1 = \mathrm{JS} + \epsilon_1 \quad . \tag{1}$$

The variables are expressed in deviation from their means and the covariance of JS and $\epsilon_1$ is zero. The *reliability* of the measure $Y_1$ for JS has been defined (Lord & Novick, 1968) as:

$$\text{Reliability} = \frac{\sigma^2_{\mathrm{JS}}}{\sigma^2_{Y_1}} \quad . \tag{2}$$

If in equation 1 the variables are standardized (in italics) we get

$$Y_1 = \frac{Y_1}{\sigma_{Y_1}} = \left(\frac{\sigma_{\mathrm{JS}}}{\sigma_{Y_1}}\right)\mathrm{JS} + \left(\frac{1}{\sigma_{Y_1}}\right)\epsilon_1 \quad , \tag{3}$$

or

$$Y_1 = \lambda_1 \mathrm{JS} + e_1 \quad , \tag{4}$$

where $\lambda_1 = \frac{\sigma_{\mathrm{JS}}}{\sigma_{Y_1}}$ and $e_1 = \frac{1}{\sigma_{Y_1}}\epsilon_1$. From equations 4 and 5 follows that

$$\text{Reliability} = \frac{\sigma^2_{\mathrm{JS}}}{\sigma^2_{Y_1}} = \lambda_1^2 \quad . \tag{5}$$

So the reliability can be computed in two ways depending whether unstandardized or standardized variables are used. In the latter case the reliability is the squared effect of the variable of interest on the observed score for this variable.

The response to the second question ($Y_2$) represents a concept that we have called Other Job (OJ) which will be affected by JS but also by the opinion about a possible other job, a unique component ($u$). The standardized coefficient for the effect of JS on the latent variable OJ is denoted by $\rho_2$. The coefficient $\rho_2^2$ can be interpreted as the *validity* of OJ as a measure for JS. We can formulate:

$$\mathrm{OJ} = \rho_2 \mathrm{JS} + u \quad , \tag{6}$$

where both variables are standardized and Cov(JS$u$). As before in equation (4) we can write for the standardized measure $Y_2$ of OJ

$$Y_2 = \lambda_2 \mathrm{OJ} + e_2 \quad . \tag{7}$$

We can also derive by substituting equation (6) in (7) that

$$Y_2 = \lambda_2 \rho_2 \mathrm{JS} + u + e_2 \quad . \tag{8}$$

Note that now the effect of JS on $Y_2$ is the product of the *reliability* and the *validity* coefficient. This means that $(\lambda_2\rho_2)^2$ is now the strength of the relationship of JS with the variable $Y_2$, therefore, we will call the strength of this relationship the *quality* of the measure $Y_2$ for JS.

$Y_1$ is a direct measure of it concept-by-intuition JS and therefore there is no unique component and therefore the quality of Y1 for JS was equal to the reliability of $Y_1$.

As usual the composite score is computed as the new measure for the concept-by-postulation (JS) with equal or unequal weights ($w_i$). Lawley and Maxwell (1971) have suggested procedures to determine weights which maximize the relationship between the composite score and the latent variable of interest.

$$\mathrm{CS} = w_1 Y_1 + w_2 Y_2 \quad . \tag{9}$$

The model for the specification of this measurement process is presented in Figure 7.

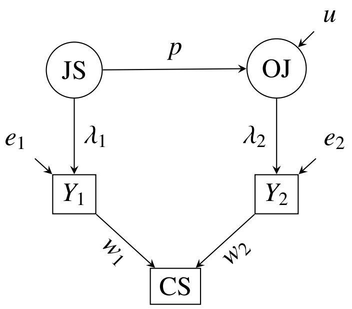

*Figure 7.* A simple measurement model using two questions for Job satisfaction where JS is: How satisfied are with your job? and OJ is: Would you like to have another job?

In order to standardize CS this variable is divided by its standard deviation $\sigma_{\mathrm{CS}}$ which gives

$$CS = \frac{\mathrm{CS}}{\sigma_{\mathrm{CS}}} = \frac{1}{\sigma_{\mathrm{CS}}}(w_1 Y_1 + w_2 Y_2) \quad . \tag{10}$$

Substitution of $Y_1$ and $Y_2$ by equalities (4) and (8) we get

$$CS = \frac{1}{\sigma_{\mathrm{CS}}}(w_1\lambda_1\mathrm{JS} + w_1 e_1 + w_2\lambda_2\rho_2\mathrm{JS} + w_2\lambda_1 u + w_2 e_2) \quad , \tag{11}$$

and by reordering we get

$$CS = \frac{1}{\sigma_{\mathrm{CS}}}(w_1\lambda_1 + w_2\lambda_2\rho_2)\,\mathrm{JS} + \frac{1}{\sigma_{\mathrm{CS}}}(w_1 e_1 + w_2\lambda_2 u + w_2 e_2) \quad . \tag{12}$$

The last term represents the errors in the composite score CS. The effect of JS on CS is called the quality coefficient of CS and is denoted by $q_{\mathrm{CS}}$. The quality[1] of the composite score is equal to $q^2_{\mathrm{CS}}$, with

$$q^2_{\mathrm{CS}} = \left[\frac{1}{\sigma_{\mathrm{CS}}}(w_1\lambda_1 + w_2\lambda_2\rho_2)\right]^2 \quad . \tag{13}$$

This result can also be written as

$$q^2_{\mathrm{CS}} = \left[\frac{1}{\sigma_{\mathrm{CS}}}w_1\lambda_1 + \frac{1}{\sigma_{\mathrm{CS}}}w_2\lambda_2\rho_2\right]^2 \quad . \tag{14}$$

And this shows that the quality of CS can be derived using path analysis as the squared correlation $\rho^2_{\mathrm{CS,JS}}$ which is the squared sum of the indirect effects of JS on CS through the observed variables.

Based on this model for a latent variable JS and two reflective indicators $Y_1$ and $Y_2$ we want to proof two theorems about the quality of composite scores of concepts. In order

[1] If we would assume that $\rho_2 = 1$, then $\rho_2\lambda_2 = \lambda_2$ and equation (14) becomes $\left[\frac{1}{\sigma_{\mathrm{CS}}}(w_1\lambda_1 + w_2\lambda_2)\right]^2$ which is the definition of the reliability coefficient $\omega$ of McDonald (1999) if the weights are equal to 1. This shows that our definition is more general. Our approach has earlier been specified by Heise and Bohrnstedt (1970).

to give these proofs we assume that the latent variables are standardized as well as the two observed variables $Y_1$ and $Y_2$ and that the quality coefficients for both indicators for the concepts they measure are equal. In the Figure 7 that would mean that $\lambda_1 = \lambda_2 = \lambda$.

If the composite score, CS, is computed as the unweighted sum $w_1 = w_2 = 1$ of the two observed variables, it can be shown that the variance of the composite score is $\sigma^2_{\mathrm{CS}} = 2 + 2\lambda^2\rho_2$. The standard deviation of the composite score is denoted by $\sigma_{\mathrm{CS}}$.

Using these assumptions we can derive the following results.

**Theorem 1.** *The quality of a single item without an unique component is smaller than the quality of a composite score of two questions with no unique components which means that* $\rho_2 = 1$.

So we have to prove that

$$\lambda^2 < \left[\frac{1}{\sigma_{\mathrm{CS}}}(w_1\lambda + w_2\lambda\rho_2)\right]^2 \quad .$$

If we assume that the weights are equal to 1 and $\rho_2 = 1$ it follows that

$$\lambda^2 < \frac{1}{\sigma^2_{\mathrm{CS}}} 4\lambda^2 \quad .$$

Multiplying both sides with $\sigma^2_{\mathrm{CS}}$ we get

$$\sigma^2_{\mathrm{CS}}\lambda^2 < 4\lambda^2 \quad ,$$

or

$$\left(2 + 2\lambda^2\right)\lambda^2 < 4\lambda^2 \quad .$$

So

$$2\lambda^2 + 2\lambda^4 < 4\lambda^2 \quad ,$$

and

$$2\lambda^4 < 2\lambda^2.$$

Because $\lambda < 1$ the left side is always smaller than the right side which proves theorem 1.

**Theorem 2.** *The quality of the composite score of two questions which don't measure only the same latent variable of interest is always lower than the quality of the composite score of two questions that measure only the same latent variable.*

Now we have to prove that

$$\left[\frac{1}{\sigma_{\mathrm{CS}}}(w_1\lambda + w_2\lambda\rho_2)\right]^2 < \left[\frac{1}{\sigma_{\mathrm{CS}}}(w_1\lambda + w_2\lambda)\right]^2 \quad .$$

Assuming again equal weights and equal quality of the questions it follows that

$$\left[\frac{1}{\sigma_{\mathrm{CS}}}\lambda\rho_2\right]^2 < \left[\frac{1}{\sigma_{\mathrm{CS}}}\lambda\right]^2 \quad .$$

This is only not true if $\rho_2 = 1$ which proves theorem 2.

These results indicate that the composite score of two questions[2] that measure the same concept without unique components is a better measure for the concept of interest than a single question without a unique component and also better than the composite score of two questions of which at least one contains a unique component[3].

The last result of course depends on the size of $\rho_2$. It will be clear that the difference in quality will be small if the coefficient $\rho_2$ is very close to 1. If the correlation is perfect and the quality of both questions is the same ($\lambda = 0.7$) the quality (equal to the reliability in this case) of the composite score will be 0.658. If the correlation between the two latent variables is 0.7, as Saris (1981) has found for two of these variables, then the quality of the composite score is much lower 0.527. If the correlation between the two indicators is 0.5 the quality of the composite score goes down to 0.445. A third indicator does not improve the situation very much because then the quality of the composite score is only minimally better ( 0.464).

Given this result, it makes sense to look for approaches to formulate different questions for a concept-by-postulation that as much as possible measure the same latent variable so that the composite score is indeed a better measure of the variable of interest. However, before doing this, we have to introduce briefly how different valid questions for concepts-by-intuition, i.e. without unique components, can be formulated.

## 3 Design of valid questions for concepts-by-intuition

In order to be sure that the questions measure the concept of interest that one wants to measure, Saris and Gallhofer (2014) have developed an approach called "the three steps procedure" that guarantees that the questions really measure the concept of interest. There is a nearly endless list of possible concepts in the social sciences that one may want to measure. Of course, one cannot specify how to formulate questions for all these concepts, but, if one can reduce the number of concepts by classifying them into a limited number of classes of basic concepts, then this problem may be

[2] One may wonder what happens if one uses several questions which have unique components. In that case the proofs become more complicated because the variance of the composite score varies, but numerically it can be checked that also with three (see above) or four indicators the reliability of the composite score is still smaller than the reliability of the composite score of two questions which measure the same concept without unique components.

[3] "Better" means that the relationship between the concept of interest and the measure is stronger. Some researchers call the strength of the relationship the reliability. We prefer to speak of quality because it is a combination of reliability and validity of the items.

solved. Saris and Gallhofer (2014) worked out this possibility. Based on an analysis of the literature they suggested a list of basic concepts including evaluations, feelings, preferences, norms, values, behaviour etc. Here we illustrate the use of these basic concepts for just one basic concept, a Feeling. There are many concepts mentioned in the literature that belong to this basic class, for example: Social trust, Political trust, Satisfaction with anything, Happiness etc. The basic idea was that the concepts that belong to the same basic concept could be formulated in the same way. So, the first step in the 3 steps procedure is to determine which basic concept one likes to measure.

Although the aim is to specify questions, it is more efficient to start with the formulation of assertions for the basic concepts because there are many different ways to formulate questions, while linguistic research (e.g. Koning & van der Voort, 1997) has shown that there are only three basic structures of assertions.

Therefore Saris and Gallhofer (2014) specified the basic forms of the assertions that are typical for the different concepts. The second step in the procedure is to choose one of the possible forms. For our chosen basic concept, a Feeling, especially for one's job, there are indeed three different ways to formulate assertions representing it.

First of all, the form **x I f**. We can fill in for **x** "my job", **I** stands for "is" and the feeling **f** is "satisfying". The sentence then is: *My job is satisfying*. The second form is **x F y**. This form can be used as follows: **x** is "I", **F** is "am satisfied with" and **y** is " my job". So the sentence becomes: *I am satisfied with my job*. There is a third form, namely, **x P** $\mathbf{y_f}$ where **x** is "my job" **P** is a neutral verb "gives me" and $\mathbf{y_f}$ is a substantive with a feeling connotation like "satisfaction". So the sentence becomes: *My job gives me satisfaction*.

There is no doubt using this approach that these assertions represent feelings about one's job. In the same way assertions for other objects than jobs can be formulated. For more examples and details we refer to Saris and Gallhofer (2014). The third step in the procedure is to change the assertion in a question. These assertions can be changed in questions or as we called "requests for answers". We use this term because the assertions can't only be transformed in sentences in an interrogative form but also in imperative and declarative forms. As all forms used, aim at getting an answer from a respondent we prefer to use the term "request for an answer" instead of questions. Here we will discuss only requests for answers in the interrogative form. For the other forms, we refer to Saris and Gallhofer (2014).

In this case the request for an answer can be created from an assertion by the inversion of the (auxiliary) verb with the subject component. The construction of direct requests by the inversion of the verb and subject component is quite common in many languages but also other forms can be used[4]. Using the same examples, the requests for an answer can be formulated as follows:

- Is your job satisfying?
- Are you satisfied with your job?
- Does the job give you satisfaction?

Note that we have developed three alternative requests for an answer which all three are request for an answer for the concept-by-intuition job satisfaction which means that all three measures are only effected by the opinion of "job satisfaction" without unique components. These questions are attractive candidates to be reflective indicators for the latent variable "job satisfaction" because their composite score will not contain unique components and will have a higher quality than the indicators with unique components we have presented earlier.

In order to avoid specifying leading questions or awkward sentences with too many adjectives it is advisable to substitute them with a so called WH word like "how" or "how much", as in the example below:

- How satisfying or dissatisfying is your job?
- How satisfied are you with your job?
- How much satisfaction does your job give you?

All these requests start with a WH word. This approach can be applied to all assertions.

The use of WH words, which in English refer to words such as "who", "which", "what" but also "how", "to what extent", "to what degree" allows capturing the gradation in the answers. Question designers have to pay attention that the concept they are interested in remains the same. For more details we refer again to Saris and Gallhofer (2014).

To make the requests ready for use in a questionnaire several extra parts have to be added, at least the response procedure has to be specified that the respondent has to use. For other components and the evaluation of the quality of questions we refer to the survey literature (e.g. Revilla, Zavala-Rojas, & Saris, 2016).

## 4 Designing valid questions for concepts-by-postulation

Earlier we have argued that the safest way to develop reflective indicators for a concept is to use requests that measure as much as possible the same concept without unique components. Below we will suggest several approaches to realize this aim.

*The first approach* to guarantee that one measures the same concept is to use the same question repeatedly at different places in a survey. For example one can use the following

[4] In French it is also possible to place the question formula "Est-ce que" in front of a declarative sentence to indicate the interrogative form. Spanish, for instance, constitutes an exception since one does not have to use the inversion, as rising intonation of the declarative form is already enough. Interrogatives are indicated by two question marks, one in front of the clause (¿) and the other at the end of the clause (?).

question in the beginning of the questionnaire and again later in the survey:

*How satisfying or dissatisfying is your job?*
1 Very dissatisfying
2
3
4
5 Very satisfying

In this case there is no doubt about the fact that the question measures the same concept but the problem of memory effect may occur if the time lag between the two requests is too short (Van Meurs & Saris, 1990).

*The second approach* is that we make use of the alternative grammatical structures that are possible for requests of feelings. This leads to the alternatives mentioned already above which are now completed with the same response scales:

*How satisfying or dissatisfying is your job?*
1 Very dissatisfying
2
3
4
5 Very satisfying

*How satisfied or dissatisfied are you with your job?*
1 Very dissatisfied
2
3
4
5 Very satisfied

*How much satisfaction or dissatisfaction does your job give you?*
1 Very much dissatisfaction
2
3
4
5 Very much satisfaction

One question could be used in the beginning of the survey and the other(s) later. In this case the questions still measure the same concept but due to the different formulation the memory effect may be smaller.

*A third approach* is that one uses one of the possible forms for measurement of job satisfaction mentioned above. For example, the first form and then one formulates an alternative for the question by substituting one or more important terms by synonyms. We start with the request for an answer:

*How satisfying or dissatisfying is your job?*
*My job is*
1 very dissatisfying
2
3
4
5 very satisfying

Next we can create alternatives by substituting the word "job" by the synonym "work" and or substituting the words "satisfying and dissatisfying" by "fulfilling and unfulfilling". In this way we get 3 different alternatives:

*How satisfying or dissatisfying is your work?*
1 Very dissatisfying
2
3
4
5 Very satisfying

*How fulfilling or unfulfilling is your job?*
1 Very unfulfilling
2
3
4
5 Very fulfilling

*How fulfilling or unfulfilling is your work?*
1 Very unfulfilling
2
3
4
5 Very fulfilling

One of these questions is asked in the beginning of the survey and the other later. In this case the memory problem is probably less, but now arises the problem whether the synonyms are indeed synonyms. The change of job into work probably will not create too much of a problem but whether "satisfying" is the same as "fulfilling" is already more difficult to determine without further research.

A *fourth approach* is to be a bit less strict in changing the words in the requests. One may consider that people who are satisfied with their job also like their job. That means that one can for example use the following requests:

*How much does your job satisfy or dissatisfy you?*
*My job is*
1 very dissatisfying
2
3
4
5 very satisfying

*How much do you like or dislike your job*
1 I very much dislike my job
2
3
4
5 I very much like my job

In this case one may think that the similarity is still quite good but certainly not perfect, because being satisfied with a job is not the same as liking a job. However this question will contain much less unique components than the question about another job. On the other hand, the memory effect even will be smaller than in case of the use of synonyms. This can be expected because the questions are more different in this case. Whether questions really measure the same or not can be tested using the congeneric test model (Jöreskog, 1971).

## 5 Reduction of the memory and method effect

The different requests mentioned in section 4 were developed to measure as much as possible the same concept and to reduce the memory effect. However, there are alternative possibilities to reduce the memory effect. The first is to vary the response scale while keeping the request the same. Saris (1981) has shown by an example using the congeneric test model of Jöreskog (1971) that such requests still measure the same concept because the correlation between the latent concepts, after correction for random errors, is not different from 1.

To give an example: one request could ask for answers in categories as before and the other in numbers:

*How much does your job satisfy or dissatisfy you?*
1 Very dissatisfying
2
3
4
5 Very satisfying

*Indicate by a number between 0 and 100, where 0 means completely dissatisfying and 100 means completely satisfying, how much your job does satisfy you?*

Number:

A second way to reduce the memory effect is to increase the time lag between the repeated questions so much that the memory effect is reduced to zero. Such a study for questions concerning political issues has been done by Van Meurs and Saris (1990). They showed that in a survey, where the specific repeated question was followed by a number of similar questions about the same topic, after 20 minutes the memory effect was reduced to zero for people who have no extreme opinions. Persons with extreme and strong opinions will always give the same answer but that is not a memory effect.

For other topics similar research has to be done because the needed time lag between the requests that are repeated may differ from concept to concept. Fortunately several studies for different questions are done and will be published soon (Rettig, Höhne, & Blom, 2019; Revilla & Höhne, 2019; Schwarz, Revilla, & Weber, 2019). The time between the repetitions can of course also be reduced by using different formulations of the request and the response scale together as well.

## 6 An evaluation of the possibilities

We have presented here four different approaches to develop items for a concept with reflective indicators which differ with respect to the quality criterion in as far as they measure the same concept. Besides we have suggested two procedures to reduce the memory effect in case of repeating approximately the same questions. In the table below we present our hypotheses of the positive and negative qualities of the different combinations of the approaches.

The first row for each approach indicates by “the number of + signs” the similarity of the questions and the second row represents with “the number of – signs” the negative effect of the memory effect given the time lag between the questions.

This table suggests the following. If the first approach is used, i.e., the same questions are repeated, there is no problem with respect to the equality of the measured concept but there is a problem with memory effect unless the time lag between the repetitions of the questions is long enough.

Going from the first approach to the second approach, grammatical variants of the same request still measure the same but the memory effect is reduced by the use of different forms of the requests, different response scales and possibly other aspects.

With the third approach, using synonyms for some words of the request, the equality of the questions is slightly reduced but this approach will even more reduce the memory effect. As a consequence repetition in shorter interviews is possible.

In the fourth approach, where no synonyms are used but still terms which deviate not very much, there will still remain quite some similarity between the concepts measured, although not as much as in the other approaches. One has to test whether one really still measures the same concepts with these requests. The memory effect is close to zero even in short interviews.

Table 1
*The expected quality of the different approaches to measure reflective indicators for a concept of interest such as Job satisfaction.*

| | same scale<br>short survey | same scale<br>long survey | different scale<br>short survey | different scale<br>long enough survey |
|---|---|---|---|---|
| First approach | + + + | + + + | + + + | + + + |
| Same request twice | – – – | – – | – – | 0 |
| Second approach | + + + | + + + | + + + | + + + |
| Grammatical Variants | – – | – | – | 0 |
| Third approach | + + | + + | + + | + + |
| Use of synonyms | – | 0 | 0 | 0 |
| Fourth approach | + | + | + | + |
| Nearly synonyms | 0 | 0 | 0 | 0 |

## 7 Conclusions

Blalock (1968) following Northrop (1947) suggested that concepts-by-postulation should be based on theoretical arguments. We argued that concepts that are measured using several indicators are "concepts-by-postulation". Blalock (1968) made a distinction between concepts with reflective and formative indicators. In this paper we concentrated on concepts with reflective indicators. For these concepts it is essential that the concept measured by the different indicators is the same or nearly the same because only under that condition one can be sure that the combination of the indicators will measure the concept of interest better than a single indicator.

We have provided procedures to develop alternative formulations of questions for the same concept. These formulations could be created using different grammatical questions for the same concept or by applying the same grammatical forms but using synonyms for the most typical words in the questions. This approach was based on the work of Saris and Gallhofer (2014) concerning the formulation of valid questions for different basic concepts in the social sciences.

Furthermore, we have suggested that one can also change the response scale which may not change the concept measured as was shown by Saris (1981).

Combining these characteristics we have created a table summarizing this information and indicating how similar the measures will be and under what conditions the repeated observation can be made without memory effects (Table 1).

Based on the expectations presented in the table, we would suggest to use one of the first three approaches for designing the requests for an answer for reflective indicators. Research is needed in nearly all these cases to determine how long the time lag has to be between the repetitions. However it will be clear that the more similar the formulation of the requests, the longer the time lag has to be between the two requests asked in the survey.

If the procedures suggested above for the construction of a measurement instrument for a concept with reflective indicators is used, one can expect that the error terms in the factor model will only contain random errors and no unique components. In this case one can use a weighted average to estimate a composite score on the scores obtained for the indicators and one can be sure that this weighted average will be a better measure of that concept than any of the indicators for the concepts individually.

## Acknowledgement

We are very grateful for the useful comments of Melanie Revilla and Wiebke Weber on earlier versions of this paper. We are also very grateful for the detailed and very useful comments of the anonymous reviewers.

## References

Aguirre-Urreta, M. I., Rönkkö, M., & Marakas, G. M. (2016). Omission of causal indicators: Consequences and implications for measurement. *Measurement: Interdisciplinary Research and Perspectives*, *14*(4), 170–175.

Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement. The influence of question and respondent attributes. *Sociological Methods and Research*, *20*(1), 139–181.

Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly*, *48*(2), 409–442.

Blalock, H. M. J. (1961). Theory, measurement, and replication in the social sciences. *American Journal of Sociology*, *66*(4), 342–347.

Blalock, H. M. J. (1968). The measurement problem: A gap between the languages of theory and research. In H. M. J. Blaclock & A. B. Blalock (Eds.), *Methodology in Social Sciences* (pp. 5–27). London: Sage.

Blalock, H. M. J. (1990). Auxiliary measurement theories revisited. In J. J. Hox & d. J. Jong-Gierveld (Eds.), *Operationalization and Research Strategy* (pp. 33–49). Amsterdam: Swets and Zeitlinger.

Blauner, R. (1966). Work satisfaction and industrial trends in modern society. In R. Bendix & S. Lipset (Eds.), *Class, Structure and Power* (pp. 473–487). New York: The Free Press.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 201–219.

Clark, A. F. (1998). Measurement of job satisfaction; What makes a good job? Evidence from OECD countries. In *OECD Labour Market and Social Popicy Occasional papers* (pp. 11–30).

De Groot, A. D., & Medendorp, F. L. (1986). *Term, begrip, theorie: inleiding tot signifische begripsanalyse*. Meppel: Boom.

Edwards, J. R. (2011). The fallacy of formative measurement. *Organizational Research Methods*, *14*(2), 370–388.

ESS Round 6: European Social Survey Round 6 Data. (2012). Data file edition 2.4. NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC. doi:10.21338/NSD-ESS6-2012

Hardin, A., & Marcoulides, G. A. (2011). A commentary on the use of formative measurement. *Educational and Psychological Measurement*, *71*(5), 753–764.

Heise, D. R., & Bohrnstedt, G. W. (1970). Validity, invalidity, and reliability. In F. Borgatta & G. W. Bohrnstedt (Eds.), *Sociological methodology* (pp. 104–129).

Hox, J. J. (1997). From theoretical concept to survey questions. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey Measurement and Process Quality* (pp. 47–70). New York: Wiley.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*(2), 109–133.

Kalleberg, A. L. (1974). A causal approach to the measurement of job statisfaction. *Social Science Research*, *3*(4), 299–322.

Kalleberg, A. L. (1975). *Work values, job rewards and job satisfaction: A theory of the quality of work experience* (Doctoral dissertation, University of Wisconsin).

Kalleberg, A. L. (1977). Work values and job rewards: A theory of job satisfaction. *American Sociological Review*, *42*, 124–143.

Költringer, R. (1993). *Gültigkeit von Umfragedaten*. Wien: Bohlau.

Koning, P. L., & van der Voort, P. J. (1997). *Sentence Analysis*. Groningen: Wolters–Noordhoff.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method*. London: Butterworth.

Lee, N., & Cadogan, J. W. (2013). Problems with formative and higher-order reflective variables. *Journal of Business Research*, *66*(2), 242–247.

Lord, F. N., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Molenaar, N. J. (1986). *Formuleringseffecten in Survey-Interviews* (Doctoral dissertation, Free University, Amsterdam).

Northrop, F. S. C. (1947). *The logic of sciences and the humanities*. New York: World Publishing Company.

Rettig, T., Höhne, J. K., & Blom, A. (2019). Recalling survey answers: A comparison across question types and different levels of online panel experience. Under review.

Revilla, M., & Höhne, J. K. (2019). Repeatedly measuring political interest: Can we reducerespondents' recall ability and memory effects in surveys using memory interference tasks? Under reviewing.

Revilla, M., Zavala-Rojas, D., & Saris, W. E. (2016). Creating a Good Question: How to Use Cumulative Experience. In *The SAGE-Handbook of Survey Methodology* (pp. 236–254). Sage Publications Ltd.

Robinson, J., Athanadiou, R., & Head, K. B. (1969). *Measurement of occupation attitudes and Occupational characteristics*. Ann Arbor: Institute of Social Science.

Saris, W. E. (1981). Different questions, different variables. In C. Fornell (Ed.), *A second generation of multivariate analysis* (Vol. 2, pp. 78–96). New York: Praeger.

Saris, W. E., & Gallhofer, I. N. (2007). *Design, evaluation and analysis of questionnaires for survey Research*. Hoboken: Wiley.

Saris, W. E., & Gallhofer, I. N. (2014). *Design, evaluation and analysis of questionnaires for survey Research* (2nd ed.). Hoboken: Wiley.

Scherpenzeel, A. C., & Saris, W. E. (1996). Causal direction in a model of life satisfaction: The top-down/bottom-up controversy. *Social Indicators Research*, *38*(2), 161–180.

Scherpenzeel, A. C., & Saris, W. E. (1997). The validity and reliability of survey questions: A meta-analysis of MTMM studies. *Sociological Methods & Research*, *25*(3), 341–383.

Schuman, H., & Presser, S. (1981). *Questions and Answers in Attitude Survey: Experiments on Question Form, Wording and Context*. New York: Academic Press.

Schwarz, H., Revilla, M., & Weber, W. (2019). Memory Effects in Repeated Survey Questions – Reviving the Empirical Investigation of the Independent Measurements Assumption. Under review.

Sudman, S., & Bradburn, N. M. (1983). *Asking Questions: A Practical Guide to Questionnaire Design*. San Francisco: Jossey Bass.

Van Meurs, A., & Saris, W. E. (1990). Memory effects in MTMM studies. In A. Van Meurs & W. E. Saris (Eds.), *Evaluations of measurement instruments by meta-analysis of Multi-trait Multi-method studies* (pp. 134–146). Amsterdam, North Holland.

Wilensky, H. L. (1964). Varieties in work experience. In H. Borow (Ed.), *Man in the world at work* (pp. 125–154). Boston: Houghton-Mifflin.