Method Effects in Psychological Assessment

Karl Schweizer[1,2]

[1] Institute of Psychology, Goethe University Frankfurt, Frankfurt a. M., Germany
[2] Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou, China

Abstract:
Method effects are described as systematic variation observed in measurement that originates from the method of measurement instead of from the attribute, which the scale or measurement procedure is expected to capture. Method effects are major sources of impairment of the quality of measurement. Because of a method effect a scale or measurement procedures does not or only partly measure what is expected to measure. Method effects and statistical methods for the identification and control of method effects are discussed. Special emphasis is given to the item-position, speed and wording effects.

Keywords:
Method effect, measurement, item-position effect, speed effect, wording effect

Method effects

A method effect is an effect that is like an experimental effect due to a source. Whereas in the experimental effect the source is the experimental manipulation, the method effect is due to the method. It is the method employed in measurement. Therefore, method effects are closely linked to measurement (Maul, 2013). Because of this close relationship it is convenient to outline the meaning of measurement before considering method effects in more detail.

In following basic ideas by Suppes and Zinnes (1963) we describe measurement in psychology as the mapping of a human attribute to a numeric scale. Measurement is expected to yield scores that reflect the variation of the attribute in the population. The distribution of the scores achieved when measuring an attribute in a sample should reflect the distribution of the attribute in the population in the sense of a linear relationship. The mapping is accomplished by the method of measurement. This extends to observations by observers, the application of questionnaires and tests, the use of apparatuses providing reaction times, recordings of EEG and others.

We follow Sechrest et al. (2000, p. 64) who describe a method effect as variation that does not originate from the attribute to be measured but is characteristic of the method of measurement. It is additional systematic variation that is part of the observed variation. This means that variation due to the method of measurement is a recurring effect that shows the property of replicability. This property distinguishes method effects from random influences. This view of the method effect is in line with the idea underlying Kowles‘ (1986) statement that „measuring changes the measure“. According to this statement the input to the method of measurement is the attribute whereas the output that is the result of the process of mapping of this attribute to the scale additionally shows characteristics of the process.

The process of mapping the attribute to the scale has been found to be open to a larger number of distorting influences. In the early studies on method effects the use of questionnaires for data collection turned out to be a major source of such an effect in personality research (Campbell & Fiske, 1959). Furthermore, considered in more detail, questionnaires exhibit characteristic features (e.g., item formats, types of response options, presentation modes) that influence the outcome of responding to the items in an idiosyncratic way. Moreover, observers were identified as source of systematic variation in measurement (Byrne, 2016).

Some method effects are direct method effects whereas others are indirect ones. We consider effects as direct method effects if the process of mapping the attribute to the scale does not involve the stimulation of reflective cognitive processes. The item-position effect and other serial effects are examples (Schweizer, 2012; Schweizer & Troche, 2016). The item-position effect unfolds when completing one item after another item of the scale. Another example of a direct effect is the effect of the time limit in testing that is also known as speededness (Oshima, 1994). This effect is due to the termination of the opportunity to go on with completing further items by introducing a time limit in testing. If there is no indication of the closeness to end of the time span allowed for completing the items, the process of completing items is simply terminated. The consequence is that there are not-reached items and an incomplete dataset.

In contrast, indirect method effects include the evaluation of input and the integration of the consequence of the evaluation into the response as part of the process of mapping this attribute to the scale. An example of an indirect method effect is the acquiescence effect (Hinz, Michalski, Schwarz, & Herzberg, 2007; Rammstedt & Farmer, 2013). In the case of this effect the response to the item does not only depend on the outcome of the processing the item but also on the tendency to establish agreement with the narrative of the item. Another example of an indirect method effect is social desirability (Grimm, 2010; Rauch, Schweizer, & Moosbrugger, 2007). If the evaluation of the input reveals that the response to the item may lead to undesirable consequences, the tendency to avoid them by responding in an agreeable way is stimulated. As is obvious in both examples, although the effect is created during the process of mapping the attribute to the scale, additional processing is characteristic of such indirect method effects.

The method effect as systematic variation observable in addition to the variation due to the attribute to be measured has given rise to concerns regarding the validity of measurement. In their seminal paper on convergent and discriminant validity Campbell and Fiske (1959) demonstrated that method variation can impair estimates of convergent validity. Shared method variation amounts to an overestimation of convergent validity when computing correlations between scale scores. In the same way shared method variation leads to the overestimation of reliability. Furthermore, the case of unique method variation is to be considered. Whereas in the case of shared method variation all scores show the same specific method effect, in the case of unique method variation different method effects characterize the various observed scores. In this case the method variation decreases the relative amount of variation, which different scale scores have in common. For example, a specific cognitive process is assessed by measures yielding reaction times in one case and accuracies in the other one. In this situation the investigation of the convergent validity is likely to underestimate the relationship because of the unique method variation originating from reaction times on one hand and accuracies on the other hand.

The negative consequences for the validity of measurement have effectuated that the identification and control of method effects have been given high priority in science. Campbell and Fiske (1959) proposed the multitrait-multimethod design for the identification of method effects. But, after the proposal of this design, it took almost 20 years until the first confirmatory factor model became available that enabled the separation of trait and method influences in data showing the multitrait-multimethod design (Kenny, 1976). Since that time the investigation of multitrait-multimethod data has been advanced. Recently Byrne (2016) distinguished three major multitrait-multimethod models (the general confirmatory factor analytic model, the correlated uniqueness confirmatory factor analytic model, the correlated trait-correlated method minus one (CT-C(M-1) model). These models include trait and method factors. In the case of the correlated uniqueness confirmatory factor analytic model the method factors are unique factors. The factor loadings on these factors enable the estimation of the variance components due to traits and methods. All these models decompose the systematic variance into two parts so that purified representations of the traits become available.

The multitrait-multimethod design in combination with the confirmatory factor models for investigating multitrait-multimethod data have been important steps in advancing the validity of measurement. They constitute an approach that is preferably selected if there is reason for suspecting that a scale shows a specific method effect. This approach requires data collection according to an extensive design. The design must include several trait scales and several methods of measurement. Furthermore, the systematic combination of the trait scales and the methods is required. What, however, appears to be to some degree an open question is how to create the method variation that is appropriate for revealing a method effect. For example, does the combination of self rating, teacher rating and parent rating lead to the same method variation as the combination of self rating, sibling rating and peer rating. This means that the selection of measurement methods may create some vagueness regarding what is actually captured. Nevertheless, this approach offers the opportunity to make a purified representation of the trait that is to be measured available.

What is not possible by means of the multitrait-multimethod approach is the ex post investigation of data regarding the presence of a method effect. Such an effect may occur despite provisions that exclude method effects. There is virtually always the possibility that the circumstances of data collection enable method effects to impair the validity of the collected data. For example, the researcher selects a time limit for testing that can be expected to enable the members of the prospective sample to complete all items; but the actual sample includes participants who are too slow for the time limit as, for example, elderly people (Salthouse, 2000) so that the data show some degree of speededness. Although this effect may have been correctly considered in planning of the study, there are speeded data. The multitrait-multimethod approach is not appropriate for the ex-post identification of speededness unless this possibility has been anticipated by additionally considering different time limits. This situation requires an approach that enables the detection of speededness without the variation of the time limit. An approach that meets this requirement is speed-effect analysis (Schweizer, Reiß, Ren, Wang, & Troche, 2019). Speed-effect analysis proceeds from assumptions on how speededness unfolds for its detection. Furthermore, speed-effect analysis includes a representation of latent processing speed by a latent variable that enables the decomposition of the latent variance into components associated with speed and the genuine source of the scale.

The contributions to the special issue describe and apply methods for investigating method effects that qualify for ex post investigations. They focus on the item-position effect, the speed effect and the wording effect. The item-position effect has been reported in the 50th for the first time (Campbell & Mohr, 1950; Mollenkopf, 1950). Although there has been research into the item-position effect over the years, a broad interest in this effect has evolved just recently (e.g., Birney, Beckmann, Beckmann, & Double, 2017; Debeer & Janssen, 2013; Embretson, 1991; Hartig & Buchholz, 2012; Kubinger, 2008; Lozano, 2015; Schweizer, Schreiner, & Gold, 2009; Verguts & De Boeck, 2000; Zeller, Reiss, & Schweizer, 2017). A reason for the before limited effort in investigating the item-position effect has probably been the opposition to the assumption that the responses to the items of a scale are independent of each other. The independence assumption excludes such an effect. The recently soaring interest in the item-position effect can be ascribed to the availability of statistical models with an enhanced capability for investigating the substructure of data.

The speed effect that is also referred to as speededness (Oshima, 1994) is another effect that is likely to influence the validity of data. This effect is observed if participants have not enough time for completing all items. The time limit leads to not-reached items. Because of individual differences in the speed of completing the items, some participants may reach all items whereas others may end up with different subsets of completed items. It is an old issue of psychological assessment (Kelley, 1927). The first attempt to control the influence of speededness on the outcome seemed to have been the distinction of power and speeded testing (Gulliksen, 1950). Since virtually all performance measures are applied with a time limit in testing, it is rather likely that at least a few participants reach not all items in data collection. The not-reached items impair the validity of data (Lu & Sireci, 2007).

The wording effect appears to originate from the advice to vary the wording of the items (e.g., Nunnally, 1978). This advice was given to test constructors because it was feared that otherwise participants would cease in reading the items carefully and resort to a schematic response style instead. The consequence of varying the item wording was an increased degree of heterogeneity in the responses. Investigations of the homogeneity of scales including equal numbers of positively and negatively worded items revealed a decrease of homogeneity due to the change of the wording (e.g., DiStefano & Motl, 2006; Vautier, Steyer, Jmel, & Raufaste, 2005). The detection of the change in item wording as source of the decreased homogeneity has provided the inspiration for the denotation as item wording effect. This effect is also open to ex post investigations by advanced structural equation models.

References

Birney, D. P., Beckmann, J. F., Beckmann, N., & Double, K. S. (2017). Beyond the intellect: complexity and learning trajectories in Raven's Progressive Matrices depend on self-regulatory processes and conative dispositions. *Intelligence*, *61*, 63-67. doi: 10.1016/j.intell.2017.01.005

Byrne, B. M. (2016). Using multitrait-multimethod analysis in testing for evidence of construct validity. In K. Schweizer & C. Distefano (Eds.), *Principles and methods of test construction* (pp. 288-307). Göttingen: Hogrefe Publishing.

Campbell, D. T., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Campbell, D. T., & Mohr, P. J. (1950). The effect of ordinal position upon responses to items in a check list. *Journal of Applied Psychology, 34*, 62-67.

Deeber, D**.**, & Janssen, R. (2013). Modeling item-position **e**ffects within an IRT framework**.** *Journal of Educational Measurement, 50,* 164-185. doi: 10.1111/jedm.12009

DiStefano, C., & Motl, R. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling, 13,* 440-464.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*, 495–515. doi: 10.1007/BF02294487

Grimm, P. (2010). Social desirability bias. In J. N. Sheth & N. K. Malhotra (Eds.), *Wiley International Encyclopedia of Marketing*. John Wiley & Sons Ltd.

Gulliksen, H. (1950). Speed versus power tests. In H. Gulliksen (Ed.), *Theory of mental tests* (pp. 230–244). New York: John Wiley & Sons.

Hartig, J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modelling*, *54*, 418–431.

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

Kelley, T. (1927). *Interpretation of educational measurements*. Yonkers, N. J.: World Book co.

Kenny, D.A. (1976). An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology*, *12*, 247-252.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology*, *55*(2), 312–320. doi:10.1037/0022-3514.55.2.312

Kubinger, K. D. (2008). On the revival of the Rasch model-based LLTM: From constructing tests using item generating rules to measuring item administration effects. *Psychology Science Quarterly*, *50*, 311–327.

Lozano, J. H. (2015). Are impulsivity and intelligence truly related constructs? Evidence based on the fixed-links model. *Personality and Individual Differences*, *85*, 192–198. doi: 10.1016/j.paid.2015.04.049

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement, 26*, 29-37. doi 10.1111/j.1745-3992.2007.00106.x

Maul, A. (2013). Method effects and the meaning of measurement. *Frontiers in Psychology*. doi: 10.3389/fpsyg.2013.00169

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166. doi: 10.1037/0033-2909.105.1.156

Mollenkopf, W. G. (1950). An experimental study of the effects on item-analysis data of changing item placement and test time limit. *Psychometrika*, *15*, 291–315. doi:10.1007/BF02289044

Nunnally, J. M. (1978). *Psychometric theory*. New York: McGraw-Hill.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31*, 200–219. doi: 10.1111/j.1745-3984.1994.tb00443.x

Rammstedt, B., & Farmer, R. F. (2013). The impact of acquiescence on the evaluation of personality structure. *Psychological Assessment, 25*, 1137-1145. Doi: 10.1037/a0033323

Rauch, W. A., Schweizer, K., & Moosbrugger, H. (2007). Method effects due to social desirability as a parsimonious explanation of the deviation from unidimensionality in LOT-R scores. *Personality and Individual Differences, 42*, 1597-1607.

Salthouse, T. A. (2000). Aging and measures of processing speed. *Biological Psychology, 54*, 35-54. doi: 10.1016/S0301-0511(00)00052-1

Schweizer, K. (2012). The position effect in reasoning items considered from the CFA perspective. *International Journal of Educational and Psychological Assessment, 11*, 44-58.

Schweizer, K., & Troche, S. (2018). Is the factor observed in investigations of the item-position effect actually the difficulty factor? *Educational and Psychological Measurement, 78*, 46-69. doi: 10.1177/0013164416670711

Schweizer, K., Reiß, S., Ren, X., Wang, T., & Troche, S. (2019). Speed effect analysis using the CFA framework. *Frontiers in Psychology* (Section Quantitative Psychology and Measurement), *10*, ArtID 239. doi: 10.3389/fpsyg.2019.00239

Schweizer, K., Scheiner, M., & Gold, A. (2009). The confirmatory investigation of APM items with loadings as a function of the position and easiness of items: a two-dimensional model of APM. *Psychology Science Quarterly, 51*, 47-64.

Sechrest, L., Davis, M. F., Stickle, T. R., & McKnight, P. E. (2000). Understanding ‚method‘ varance. In L. Bickman (Ed.), *Research design: Donald Campbell's legacy* (pp. 63 – 87). Thousand Oaks, CA: Sage Publications.

Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush, & E. H. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 1 – 76). New York: Wiley.

Vautier, S., Steyer, R., Jmel, S. & Raufaste, E. (2005). Imperfect or perfect dynamic bipolarity? The case of antonymous affective judgements. *Structural Equation Modeling, 12*, 391-410.

Verguts, T., & De Boeck, P. (2000). A Rasch model for detecting learning while solving an intelligence test. *Applied Psychological Measurement*, *24*, 151–162. doi: 10.1177/01466210022031589

Zeller, F., Reiss, S., & Schweizer, K. (2017). Is the item-position effect in achievement measures induced by increasing item Difficulty? *Structural Equation Modelling, 24,* 745-754. doi: 10.1080/10705511.2017.1306706

Corresponding author:
Karl Schweizer
[1] Institute of Psychology
Goethe University Frankfurt
Frankfurt, Germany
[2] Department of Psychology and Behavioral Sciences, Zhejiang
University, Hangzhou, China
K.Schweizer@psych.uni-frankfurt.de
ORCID iD: 0000 0002 3143 2100