Article

Educational and Psychological
Measurement
2016, Vol. 76(3) 454–469
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164415593602
epm.sagepub.com
SAGE

Can Reliability of Multiple Component Measuring Instruments Depend on Response Option Presentation Mode?

Natalja Menold[1] and Tenko Raykov[2]

Abstract

This article examines the possible dependency of composite reliability on presentation format of the elements of a multi-item measuring instrument. Using empirical data and a recent method for interval estimation of group differences in reliability, we demonstrate that the reliability of an instrument need not be the same when polarity of the response options for its individual components differs across administrations of the instrument. Implications for empirical educational, behavioral, and social research are discussed.

Keywords

confidence interval, composite reliability, group difference, multicomponent measuring instrument, rating scale polarity, response option

Measuring instruments consisting of multiple components, such as tests, inventories, testlets, scales, self-reports, questionnaires, surveys, and so on (referred to as ‘‘instruments’’ for short below), are very often used in the educational, behavioral, social, marketing, business, and biomedical sciences. Such instruments are highly popular in these and related disciplines in part due to their theoretically and empirically appealing property of providing interrelated converging pieces of information about

[1] Leibniz Institute for the Social Sciences, Mannheim, Germany
[2] Michigan State University, East Lansing, MI, USA

Corresponding Author:
Natalja Menold, Leibniz Institute for the Social Sciences, B2, 1, Mannheim, Germany.
Email: natalja.menold@gesis.org

latent constructs of main substantive interest (e.g., Raykov & Marcoulides, 2011). Reliability of these instruments is a main index of the quality of measurement accomplished with them. Owing to the fact that the reliability coefficient is defined as the ratio of true to observed variance, as pointed out in numerous sources, this coefficient is dependent on both the instrument and the population on which it is used (e.g., McDonald, 1999).

This dual-dependency is widely known among educational and psychological researchers and has received ample discussion in the substantive and methodological literature over the past several decades (e.g., Crocker & Algina, 1986). While appreciating this population dependency, however, it would be incorrect to imply that there are no other factors that could contribute to the value of the reliability coefficient for a given multicomponent measuring instrument. The realization of this additional potential relationship has begun to emerge particularly in recent decades, following an increasingly widespread trend in Internet-based data collection. One of the important features of such empirical studies is the possibility of presenting the same instrument with the same components (questions or items) in different response option formats to distinct groups of subjects from the same studied population.

The present article aims to address this potentially consequential issue for empirical educational and behavioral research, as well as to contribute to the extant literature on effects of item format on reliability of multi-item measuring instruments. In the remainder, we will be interested in the impact that different response options associated with the individual instrument components may have on the overall reliability coefficient. To accomplish our goals, we use a recent method for evaluation of group differences in composite reliability (Raykov & Marcoulides, 2015) on data from an Internet-based study and demonstrate that notable measurement consistency discrepancies could result when the polarity of item response options is altered in an instrument.

Presentation Format of Item Response Options and Measurement Quality: What Do We Already Know From Prior Research?

As is well known, different response options can be associated with the components of a multi-item measuring instrument, such as for instance dichotomous, nominal, or ordered categories. An item with ordered response options is often referred to as rating scale. A rating scale represents typically a continuum pertaining to a given item or question in an instrument, with individual response options that can extend from one extreme to the other, for example, ranging from ''strongly agree'' to ''strongly disagree.'' As has been found in prior research, rating scale formats can affect overall instrument reliability as a result of differences in (a) the interpretation of individual question categories by the examined subjects and/or (b) the specific conditions under which instrument administration occurs (e.g., Krosnick & Presser, 2009; Menold, Kaczmirek, Lenzner, & Neusar, 2014; Saris & Gallhofer, 2007). These and related effects of differences in item presentation format on the functioning of a multicomponent instrument have been known for a considerable period of time.

The most researched issues in this context have been the effects of the number of categories on different aspects of measurement with rating scales. These aspects were discriminability of ratings (Garner, 1960), results of factor analysis (Schutz & Rucker, 1975), equidistance of rating scale categories (Wakita, Ueshima, & Noguchi, 2012), as well as reliability and validity of measurement (Matell & Jacoby, 1971; Parker, Vannest, & Davis, 2013; Preston & Colman, 2000). With respect to reliability, some studies have found that the number of categories in rating scales can affect it (Lozano, García-Cueto, & Muñiz, 2008; Parker et al., 2013; Preston & Colman, 2000). However, results are also available that do not suggest a relationship between number of categories and reliability (e.g., Leung, 2011; Matell & Jacoby, 1971; Wakita et al., 2012). Besides the number of categories, other aspects of rating scales have been found to be associated with effects on reliability, such as (a) the labeling used for categories (e.g., Alwin & Krosnick, 1991; Krosnick & Berent, 1993; Menold et al., 2014; Saris & Gallhofer, 2007; Weng, 2004) and (b) the use of a middle category (Alwin & Krosnick, 1991). All these aspects of rating scales need therefore to be considered by empirical educational and behavioral scientists when deciding on appropriate rating scale formats with the aim to increase measurement quality, and specifically instrument reliability.

Against the backdrop of this extensive body of research, however, an important further characteristic of rating scales, scale polarity, has received only limited attention. Research in this area has typically differentiated between unipolar and bipolar rating scales. As discussed in detail for instance in Schaeffer and Presser (2003), bipolar scales can be seen as including two opposite rating dimensions, such as disagreement and agreement, liking and disliking, ease or difficulty. Bipolar scales combine in this way a negative and a positive rating dimension (with response options ranging, e.g., from -5 to $+5$), and usually include a neutral or zero point in the middle. Unlike bipolar scales, unipolar scales consist of one rating dimension, for example, agreement, liking, goodness, or importance, with response options typically ranging from zero to a higher degree (e.g., symbolized by 5, 7, or 10) and the middle option(s) expressing a moderate degree.

Commonly used agree–disagree Likert-type items or questions have been thereby typically considered to represent bipolar scales (e.g., Krebs, 2012). Other examples of bipolar scales are those with rating options ranging from ‘‘very difficult’’ to ‘‘very easy,’’ or from ‘‘very good’’ to ‘‘very bad,’’ and so on. Conversely, an agreement scale with response options extending from ‘‘do not agree at all’’ to ‘‘fully agree’’ or ‘‘does not apply at all to me’’ to ‘‘entirely applies to me’’ is usually treated as a unipolar scale.

A typical finding with respect to bipolarity is that use of negative numbers as response options can lead to response bias (e.g., Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark, 1991; see Schaeffer & Presser, 2003, for a review). This bias consists in the fact that subjects tend to avoid the negative part of the associated item scale and produce more positive results (e.g., higher means) as compared with items using only positive numbers as response options. In general, verbal and numeric labels may have independent effects on the responses selected by examined persons (O'Muircheartaigh, Gaskell, & Wright, 1995). Fully verbalized rating scales have been found thereby to be associated with higher reliability (e.g., Menold et al., 2014). With this in mind, it becomes important to shed light on the question as to whether verbal polarity can affect reliability of measures. This polarity is present in an item when numeric labels are not used to denote pertinent response options, but instead the latter are verbally expressed. In this manner, they convey the implied polarity through suitable words rather than numbers.

Although researchers in the social and behavioral disciplines are frequently faced with the need to make a decision to use either unipolar or bipolar rating scales, only a few studies have addressed the possible effect of verbal rating scale polarity on instrument reliability. Recently, Krebs (2012) compared Cronbach's coefficient alpha between unipolar and bipolar scales expressed with verbal versus numeric polarity when measuring attitudes toward foreigners in Germany. She found higher alpha coefficients in case of bipolar scales, irrespective of whether polarity was expressed verbally or with positive and negative numbers. However, coefficient alpha cannot be considered in general an accurate reliability measure, since it can be strictly interpreted as a reliability coefficient only if certain measurement model features hold, namely, that the items evaluate the same latent dimension with the same units of measurement and no error correlations (e.g., Raykov, 2012). Thus, no direct interpretation of Krebs' (2012) results is possible with respect to instrument reliability itself, due to the fact that these features have not been thereby taken into account. Hence, it is important to investigate the impact of verbal polarity of item response options on instrument reliability in the case of congeneric measures that is the most general and widely used setting of items evaluating a single underlying latent dimension (e.g., Jöreskog, 1971).

It is this research issue that the present article is concerned with. The remaining discussion is specifically focused on the possible dependence of a multi-item instrument's reliability on the response option presentation format of its individual components. We argue below that even in the popular case of homogeneous instruments, that is, with unidimensional items, different reliabilities may result in distinct groups from the same population when an instrument is administered under different polarity of its item response options. To achieve our aims, we use an empirical data set obtained from two random samples of German adults collected in an Internet-based survey and the popular latent variable modeling (LVM) methodology (e.g., Muthén, 2002). We find this possibility for reliability differences in relation to presentation format of response options to be worth articulating, as it extends in an important way the understanding of behavioral measurement in general and of composite reliability in particular. Specifically, as we discuss next, in addition to the instrument and population of interest reliability may also be influenced by the specific way in which the instrument components (e.g., questions or items) and especially their response options are presented to the studied subjects.

Composite Reliability Differences Across Item Response Presentation Modes

Group differences in reliability of measuring instruments, which are of special interest in this article, have received considerable attention over the past several decades by methodologists as well as substantive scholars. Beginning perhaps with the work by Feldt (1969) that was focused exclusively on group differences in coefficient alpha, educational and behavioral scientists have been well aware of the possible discrepancies in consistency of measurement when a multi-item instrument is presented to different populations. Thereby, what may be seen as an earlier implication from that and related research is the expectation that as long as the studied population and individual components are fixed, the instrument could be assumed to be functioning in essentially the same way.

Recent developments in behavioral and social measurement have provided methodological means to address this consequential assumption and in particular some important aspects of it. In the context of the present article, one may well argue that the above-mentioned implication would entail that there would be no subgroup differences in composite reliability when the instrument is administered using different item presentation formats in a given population, in particular with respect to polarity in the response options available for its individual components. To examine this conjecture, use can be made of a method in Raykov and Marcoulides (2015), who outline a generally applicable, finite mixture modeling based procedure for studying scale reliability differences across the mixture components. A special case of their approach is also applicable when the number of mixture components is known beforehand as is subjects' membership in them. This special case can be directly used to address the query of relevance in the current article, viz., whether there may be differences in composite reliability when the same instrument is presented with different polarity of its items' response options to different groups that are randomly sampled from the same population of interest.

The essence of the method in Raykov and Marcoulides (2015) for the case of say $g = 2$ groups and known subject group membership, which is of importance in the rest of this article, is based on an application of the LVM methodology to point and interval estimation of group differences in composite reliability. In the unidimensional case of concern here, the underlying single-factor model of relevance is

$$y = \mu + \beta\eta + \varepsilon, \tag{1}$$

where y is the $p \times 1$ vector of instrument elements (questions, items, components; $p > 1$), η is the construct evaluated by them, β is the vector of component loadings on the construct, and ε is the $p \times 1$ vector of unique factors comprising observed measure specificity and "pure" measurement error (e.g., Raykov & Marcoulides, 2011).

In this setting, the group difference in sum score reliability has been shown there to equal

$$\begin{aligned}\Delta\rho_{u,w} = {} & \phi_u\left(1+b_2+\cdots+b_p\right)^2/[\phi_u\left(1+b_2+\cdots+b_p\right)^2+\nu_{u,1}+\cdots+\nu_{u,p}] \\ & -\phi_w\left(1+b_2+\cdots+b_p\right)^2/[\phi_w\left(1+b_2+\cdots+b_p\right)^2+\nu_{w,1}+\cdots+\nu_{w,p}],\end{aligned} \tag{2}$$

where the groups are formally indexed by *u* and *w*, the *b*s are the loadings of the individual components on the assumed single construct they are evaluating, ϕ_u and ϕ_w denote its variances in the two respective groups, and the *v*s are the associated error term variances. Point and interval estimation of the reliability group difference is possible then within the framework of the popular LVM methodology and is discussed in detail in Raykov and Marcoulides (2015). We present next an application of that general method in the case of two groups that is of relevance in the remainder and suffices for the aims of the present article.

An Empirical Study of Composite Reliability Differences and Item Response Option Polarity

In this section, to accomplish the goals of the article, we use empirical data that were collected within a probabilistic online assess panel (OAP) on German-speaking adults living in the Federal Republic of Germany and aged 18 years or older. In that study, users of Internet for non-work-related purposes were recruited between February and August 2011, by a telephone survey using dual frame (cell phone and land line numbers). For further details on this sample, reference is made to Struminskaya, Kaczmirek, Schaurer, and Bandilla (2014). The data from this OAP used in the current section were collected in July and August 2012. In the remainder of the section, we use a measuring instrument consisting of five items on Powerful Others Control Orientation from a short version of a questionnaire on locus of control (Bornmann & Daniel, 2000). The latent construct evaluated by Powerful Others Control Orientation is defined as a general expectation that important life occurrences in one's life are determined by powerful other persons. The text of the items (translated into English) is provided in Table 1.

This five-item instrument was presented to two randomly drawn groups from the original OAP sample, which differed in the polarity of the available response options on all items. The first group, consisting of 268 adults, used a verbal bipolar scale for their responses on each of the items, with 7 options ranging from ''completely disagree'' to ''completely agree.'' The specific response options (translated into English) on each of the questions are presented in Table 2. The second group comprised 269 adults and used instead a unipolar version with 7 possible responses on each of the questions, which ranged from ''does not apply at all to me'' to ''fully applies to me.'' The text of the item response options available on each question to this group is provided in Table 2. The subjects in both groups were asked to evaluate each item using the corresponding item response options provided to either group.

Table 1. Items of the Powerful Others Control Orientation Instrument (Translated Into English).

Item 1: I have the feeling that a lot of what occurs in my life depends on other people.
Item 2: Other people often hinder the realizations of my plans.
Item 3: I have only limited possibility to carry through my interests against the will of other people.
Item 4: My satisfaction with life depends strongly on the behavior of other people.
Item 5: In order to give my plans a chance, I develop them in accordance with the wishes of other people.

Table 2. Response Options Available to the Two Groups, for Each of the 5 Items of the Used Instrument (Translated Into English).

Group 1 (bipolar format)
- Completely disagree
- Disagree
- Disagree to some extent
- Partly agree/partly disagree
- Agree to some extent
- Agree almost fully
- Fully agree

Group 2 (unipolar format)
- Does not apply at all to me
- Does not apply to me
- It applies to me to a minor degree
- It applies moderately to me
- It applies to me to some notable degree
- It applies to me
- It entirely applies to me

Employing on this data set the method in Raykov and Marcoulides (2015) for the case of $g = 2$ groups, we begin by fitting the single-factor model simultaneously in both groups that we denote by M1.[1] (For completeness of this article, we provide in the appendix the Mplus input file used for its analytic purposes that is a minor modification of the corresponding code in that original source; see below; Muthén & Muthén, 2014.) Given that each item has 7 possible response options in each group, we use thereby the robust maximum likelihood method for model fitting and parameter estimation (e.g., Raykov & Marcoulides, 2011). This model is associated with fit indices that cannot be considered tenable however: $\chi^2 = 56.152$, degrees of freedom (*df*) = 14, root mean square error of approximation (RMSEA) = .106, with a 90% confidence interval [.078, .136]. Examining its modification indices suggests considering inclusion of the error covariance between Questions 2 and 3. Closer inspection of their text (see Table 2) indicates that this would be a substantively sound and justified decision. Indeed, both questions effectively ask about the degree

Table 3. Parameter Estimates, Standard Errors, *t* Values, *p* Values, and Confidence Intervals (Further Below) Associated With Model M3 (Software Output Format).

```
                     Estimate      S.E.     Est./S.E.    Two-Tailed P-Value

Group BIPOLAR
F            BY
  Q1                   1.000      0.000      999.000            999.000
  Q2                   0.968      0.093       10.382              0.000
  Q3                   0.713      0.113        6.307              0.000
  Q4                   0.991      0.113        8.780              0.000
  Q5                   0.674      0.115        5.847              0.000
Q3           WITH
  Q2                   0.223      0.099        2.246              0.025
Means
  F                    0.000      0.000      999.000            999.000
Intercepts
  Q1                   3.803      0.068       56.046              0.000
  Q2                   3.187      0.071       44.610              0.000
  Q3                   2.810      0.062       45.248              0.000
  Q4                   3.236      0.082       39.657              0.000
  Q5                   3.118      0.071       44.079              0.000
Variances
  F                    0.930      0.154        6.024              0.000
Residual Variances
  Q1                   0.669      0.114        5.865              0.000
  Q2                   0.882      0.162        5.461              0.000
  Q3                   1.013      0.131        7.712              0.000
  Q4                   1.669      0.200        8.340              0.000
  Q5                   1.287      0.123       10.505              0.000
Group UNIPOLAR
F            BY
  Q1                   1.000      0.000      999.000            999.000
  Q2                   0.999      0.080       12.494              0.000
  Q3                   0.813      0.083        9.813              0.000
  Q4                   1.015      0.117        8.674              0.000
  Q5                   1.047      0.133        7.881              0.000
Q3           WITH
  Q2                   0.339      0.085        3.962              0.000
Means
  F                   −0.121      0.088       −1.370              0.171
Intercepts
  Q1                   3.803      0.068       56.046              0.000
  Q2                   3.187      0.071       44.610              0.000
  Q3                   2.810      0.062       45.248              0.000
  Q4                   3.236      0.082       39.657              0.000
  Q5                   3.118      0.071       44.079              0.000
Variances
  F                    0.879      0.147        5.997              0.000
Residual Variances
  Q1                   0.728      0.088        8.264              0.000
  Q2                   0.834      0.110        7.600              0.000
  Q3                   0.791      0.110        7.167              0.000
  Q4                   1.215      0.140        8.669              0.000
  Q5                   0.825      0.183        4.502              0.000
New/Additional Parameters
  REL_G1               0.746      0.033       22.610              0.000
  REL_G2               0.805      0.024       33.033              0.000
  DELTA                0.058      0.041        1.421              0.155
```

(continued)

Table 3. (continued)

```
Confidence Intervals of Model Results

                  Lower   Lower   Lower             Upper   Upper   Upper
                  0.5%    2.5%     5%    Estimate    5%     2.5%    0.5%

Group BIPOLAR
F          BY
  Q1              1.000   1.000   1.000    1.000    1.000   1.000   1.000
  Q2              0.728   0.785   0.814    0.968    1.121   1.150   1.208
  Q3              0.422   0.491   0.527    0.713    0.899   0.934   1.004
  Q4              0.700   0.769   0.805    0.991    1.176   1.212   1.281
  Q5              0.377   0.448   0.484    0.674    0.863   0.899   0.970
Q3         WITH
  Q2             −0.033   0.028   0.060    0.223    0.386   0.417   0.478
Means
  F               0.000   0.000   0.000    0.000    0.000   0.000   0.000
Intercepts
  Q1              3.629   3.670   3.692    3.803    3.915   3.936   3.978
  Q2              3.003   3.047   3.070    3.187    3.305   3.327   3.371
  Q3              2.650   2.688   2.708    2.810    2.912   2.932   2.970
  Q4              3.026   3.076   3.102    3.236    3.370   3.396   3.446
  Q5              2.936   2.980   3.002    3.118    3.235   3.257   3.301
Variances
  F               0.532   0.627   0.676    0.930    1.184   1.232   1.328
Residual Variances
  Q1              0.375   0.445   0.481    0.669    0.856   0.892   0.962
  Q2              0.466   0.566   0.616    0.882    1.148   1.199   1.298
  Q3              0.675   0.756   0.797    1.013    1.229   1.271   1.351
  Q4              1.153   1.277   1.340    1.669    1.998   2.061   2.184
  Q5              0.972   1.047   1.086    1.287    1.489   1.527   1.603
Group UNIPOLAR
F          BY
  Q1              1.000   1.000   1.000    1.000    1.000   1.000   1.000
  Q2              0.793   0.843   0.868    0.999    1.131   1.156   1.205
  Q3              0.600   0.651   0.677    0.813    0.950   0.976   1.027
  Q4              0.714   0.786   0.822    1.015    1.207   1.244   1.316
  Q5              0.705   0.787   0.828    1.047    1.265   1.307   1.389
Q3         WITH
  Q2              0.118   0.171   0.198    0.339    0.479   0.506   0.559
Means
  F              −0.348  −0.293  −0.266   −0.121    0.024   0.052   0.106
Intercepts
  Q1              3.629   3.670   3.692    3.803    3.915   3.936   3.978
  Q2              3.003   3.047   3.070    3.187    3.305   3.327   3.371
  Q3              2.650   2.688   2.708    2.810    2.912   2.932   2.970
  Q4              3.026   3.076   3.102    3.236    3.370   3.396   3.446
  Q5              2.936   2.980   3.002    3.118    3.235   3.257   3.301
Variances
  F               0.501   0.591   0.638    0.879    1.120   1.166   1.256
Residual Variances
  Q1              0.501   0.555   0.583    0.728    0.873   0.901   0.955
  Q2              0.551   0.619   0.653    0.834    1.014   1.049   1.116
  Q3              0.507   0.575   0.609    0.791    0.972   1.007   1.075
  Q4              0.854   0.940   0.985    1.215    1.446   1.490   1.576
  Q5              0.353   0.466   0.523    0.825    1.126   1.184   1.297
```

(continued)

Table 3. (continued)

Confidence Intervals of Model Results

	Lower 0.5%	Lower 2.5%	Lower 5%	Estimate	Upper 5%	Upper 2.5%	Upper 0.5%
New/Additional Parameters							
REL_G1	0.661	0.682	0.692	0.746	0.801	0.811	0.831
REL_G2	0.742	0.757	0.765	0.805	0.845	0.852	0.867
DELTA	−0.047	−0.022	−0.009	0.058	0.126	0.139	0.164

Note. Q1 through Q5 = items of used instrument; F = latent construct evaluated by instrument; S.E. = standard error; Est./S.E. = *t*-value for the test of the null hypothesis of pertinent parameter being 0 in the population; New/additional parameters = reliability coefficient in each group and their difference, DELTA, which are added in model M2 without affecting its fit or parameter estimates and S.E.s, leading to the present model M3.

to which one's implementation of his/her own plans is made more difficult by other people, thus making it rather likely that the common factor shared by these and the remaining 3 items would not be completely capable of explaining the interrelationships between Questions 2 and 3. (This observation in effect makes in each group their error covariance an a priori relevant, omitted model parameter.)

With this in mind, in the next model version, denoted M2, we add the error covariance for these two items in both groups, which leads to the following goodness of fit indices (90% confidence interval of RMSEA stated after its point estimate): χ^2 = 30.526, *df* = 12, RMSEA = .076 [.043, .110]. The corrected difference in chi-square test (e.g., Muthén & Muthén, 2014) is thereby found to be significant: corrected χ^2 = 19.557, $df = 2, p < .001$. (For completeness of this article, an *R* function for conducting this test is also provided in the appendix.) This result suggests that the error covariance for Questions 2 and 3 is not zero in both groups.

We consider the goodness of fit indices of the last model, M2, as tenable, that is, as indicative of a model that is plausible as a means of data description and explanation. In a final fitted version of the model, denoted M3, we point and interval estimate the reliability coefficients of the 5-item instrument in each of the two groups as well as their difference. This model, M3, is associated with the same fit indices, parameter estimates, and standard errors as M2, since it does not impose any additional restrictions in the latter model that are consequential for its fit to the data (see the appendix for its Mplus source code). The parameter estimates in model M3, with standard errors and 95% confidence intervals, are presented in Table 3.

As seen from Table 3, the overall instrument reliability is higher in the group with unipolar response presentation format than in the group with bipolar format. Thereby, the 95% confidence interval for their group difference in reliability is [−.022, .139]. That is, any value between −.02 and .14 (rounded off) is equally plausible as any other value in that interval for the difference in instrument reliability in the studied population of German-speaking adults. Despite the fact that 0 is covered by this interval, and in line with the aims of this article that are unrelated to hypothesis testing to begin with, we should stress that this particular value is equally plausible as a population value of the group reliability difference as any other number between the endpoints of this interval, and hence as any number between say .13 and .14. That is, the latter considerable reliability difference (in the low to mid-teens) across the two groups with different item response formats—bipolar versus unipolar—is just as plausible as any other difference within the above confidence interval that includes also such differences with notable magnitude.[2]

We can therefore interpret the control orientation study used in this section as an empirical example where polarity discrepancies in item response option format may produce notable differences in instrument reliability even when both the components of a given instrument and the studied population are the same.

Conclusion

This article was concerned with the possibility that reliability of a multicomponent measuring instrument may also depend on the presentation mode of its items, questions, elements, or components, and specifically on the format in which their response options are offered to the studied subjects. Using an empirical data set, we examined the discrepancy in composite reliability as a function of the polarity of its items and in particular response options. A main aim of our discussion was to show that even if the same measuring instrument is presented to samples from the same population, its reliability need not be the same but may in fact be related to the verbal polarity of the individual questions or items that the instrument consists of, and especially to the form in which their response options are presented to the examined persons.

Our article does not aim to suggest that in any educational or behavioral study the reliability of a given scale will depend on rating scale polarity. Rather, our goal was merely to show, by using an empirical example, that even in case of the same population being measured with the same set of items reliability of a multi-item instrument could be related to the way in which its individual components are presented, and specifically to the verbal polarity of their response options.

The preceding discussion in this article suggests that additional attention needs to be routinely paid not only to the number of categories and their verbalization but also to the polarity of rating scale response options that is verbally expressed. In particular, well-established measuring instruments, which may have been already used for a while in some substantive fields in the educational, behavioral, and social disciplines, need not be associated with the same reliability even in the same studied population unless the format of presentation of the response options on their individual items is preserved. That is, what may at first glance seem like a harmless change in the item response options could in fact have a more profound effect on the quality of measurement of the underlying latent construct, in particular, potentially altering its degree of measurement consistency. This warning may be more serious than it may appear at first, since as is well known reliability and validity of measurement are not unrelated. In fact, an effect on reliability of an instrument could have an impact also on its validity and thus potentially lead to loss in validity. This possibility is real in empirical research since as is well known (e.g., McDonald, 1999) reliability is in general an upper bound of validity. Hence, any instrument modification, such as changing polarity of original response options on some or all of its components, could lead to changes in the reliability and validity of the resulting modified instrument. If this modification is to be pursued for substantive reasons, however, new studies with representative samples from the population in question need to be conducted before one could claim in a more trustworthy way what the reliability and validity of the instrument altered in this way would be, even if (a) none of the original items is dropped or new items added to it, (b) the same number of response options is retained on all items, and (c) the same population is still of interest to be studied with that instrument. The present article has also exemplified how one could readily examine reliability differences across instrument modifications of various nature in empirical educational and behavioral research (for a more generally applicable method, see Raykov & Marcoulides, 2015).

Appendix

Mplus Source Code for Fitting Models M1 Through M3

```
TITLE:              EVALUATION OF GROUP DIFFERENCES IN COMPOSITE
                    RELIABILITY DEPENDING ON POLARITY.
DATA:               FILE = <name of multi-group raw data file> ;
VARIABLE:           NAMES ARE G0 Q1-Q5 G;
                    USEV = Q1-G;
                    GROUPING = G(1=BIPOLAR, 2=UNIPOLAR);
ANALYSIS:           ESTIMATOR = MLR;
MODEL:              F BY Q1@1
                    Q2-Q5 (B12-B15);
                    F(FI1);
                    Q1-Q5(V11-V15);
                    Q3 WITH Q2 (PSI1_32); ! may not need this parameter
MODEL UNIPOLAR:
                    F BY Q1@1
                    Q2-Q5(B22-B25);
                    F(FI2);
                    Q1-Q5(V21-V25);
                    Q3 WITH Q2 (PSI2_32); ! may not need it in general
MODEL CONSTRAINT:
                    NEW(REL_G1, REL_G2, DELTA);
                    REL_G1 = FI1*(1+b12+b13+b14+b15)**2
                            /(FI1*(1+b12+b13+b14+b15)**2
                                +v11+v12+v13+v14+v15+2*PSI1_32);
                                ! this is the composite reliability in
                                the bipolar group
                    REL_G2 = FI2*(1+b22+b23+b24+b25)**2
                           /(FI2*(1+b22+b23+b24+b25)**2
                              +v21+v22+v23+v24+v25+2*PSI2_32);
                              ! this is the composite reliability in
                              the unipolar group
                    DELTA = REL_G2-REL_G1; ! this is their difference
OUTPUT:             CINTERVAL ; !requests confidence intervals for all
                    !parameters
```

Note. Drop the entire Model Constraint section when fitting model M1, as well as the error covariance in both groups. Add then the 2 lines for the latter covariance, declaring it a model parameter (with the keyword ''WITH''), when fitting model M2. To fit model M3, use the entire above source code. (For a brief introduction to the syntax of Mplus, see, e.g., Raykov & Marcoulides, 2006.)

R-Function for Conducting the Corrected Chi-Square Difference Test on Nested Models

```
corr.chisq.diff.test = function(d0, d1, c0, c1, t0, t1){
  t = (t0 - t1)*(d0-d1)/(d0*c0 - d1*c1)
p=1-pchisq(t,d0-d1) # d – df's, c – scaling correction factors,
# t – (corrected) chi-square values; all output by Mplus
c(t,d0-d1, p) # 0 - index of nested model, 1 – index of full model
}
```

Note. At the R prompt, paste this function and call it subsequently providing correspondingly the degrees of freedom (the d's), scaling correction factors (the c's), and chi-square values (the t's) for the restricted and the relaxed model in a pair of nested models fitted to a given data set. Text within a row, which is preceded by '#', is only an annotating comment and hence can be dropped thereby. (For a nontechnical introduction to the freely available software R, see, e.g., Raykov & Marcoulides, 2012; for formal details on the corrected chi-square difference test, see www.statmodel.com.)

Authors' Note

This research was in part conducted while T. Raykov was visiting the Leibniz Institute for the Social Sciences, Mannheim, Germany. We are indebted to G. A. Marcoulides for helpful discussions on reliability and group differences examination, to B. Rammstedt for valuable support, and to the editor and two anonymous referees for critical comments on an earlier version that have contributed significantly to its improvement.

Declaration of Conflicting Interests

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

Funding

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

Notes

1. We fitted this model without the constraint of measurement invariance imposed on the factor loadings and mean intercepts (e.g., Raykov, Marcoulides, & Millsap, 2013). The reason for our decision is the fact that the goal of this article is to alert empirical scientists of the possibility that instrument reliability may depend on the polarity of item response options. Given that in a typical applied setting where a measuring instrument is used one has access to a single group of subjects evaluated with it, the issue of measurement invariance is nonexisting then. Our aim in this note is also to provide an instance of possible reliability differences if one were to decide for using a different response option format relative to a given or earlier established one, in which case the concern is again with single group instrument reliability when the measurement invariance issue is similarly void. Last but not least, imposing the measurement invariance constraints in the fitted models in this section does not change the interpretation of its results, and particularly that of the group reliability difference confidence interval as consisting of equally plausible population values including such markedly away from 0. (The resulting 95% CI is then [−.024, .134], with effectively the same group reliability difference estimate and associated standard error in practical terms; see below in main text.)
2. We emphasize that our article is not concerned with hypothesis testing, since the latter is not of relevance for accomplishing its aims. Rather, as seen also from its title, the article is interested merely in answering the question whether it is possible that item answer polarity could affect instrument reliability in empirical educational and psychological research. For this reason, the confidence interval of the group difference in reliability found in this empirical example does need to be only interpreted as presenting an interval of equally plausible values for the population reliability discrepancy (at the confidence level chosen, viz. 95%). Therefore, it is not correct to imply from this interval that there are no reliability differences in the population at large, since any nonzero value that is covered by the confidence interval is just as plausible as 0 (at that confidence level). (Note that the use of a confidence interval for testing a null hypothesis is only then meaningful when this hypothesis testing needs to be conducted, which is not the case in this section or elsewhere in the article.) With this in mind, the used empirical example in fact achieves the aims of the article in that it allows one to answer affirmatively the question representing its title, which is the only concern of this article.

References

Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociological Methods & Research*, *20*, 139-181. doi:10.1177/0049124191020001005

Bornmann, L., & Daniel, H.-D. (2000). Reliabilität und Konstruktvalidität des Kurzfragebogens für Kontrollüberzeugungen (FKK). Überprüfung der Testgütekriterien im Rahmen einer Mehrthemenbefragung unter Studierenden [Reliability and construct validity of a short questionnaire of control beliefs. Examining test criteria within multiple construct survey of students]. *Empirische Pädagogik*, *14*, 391-407.

Crocker, L., & Algina, J. (1986). *Classical and modern test theory*. Baton Rouge, FL: Harcourt Brace Jovanovich.

Feldt, L. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient is the same for two tests. *Psychometrika*, *30*, 357-370.

Garner, W. R. (1960). Rating scales, discriminability, and information transmission. *Psychological Review*, *67*, 343-352. doi:10.1037/h0043047

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

Krebs, D. (2012). The impact of response format on attitude measurement. In S. Salzborn, E. Davidov, & J. Reinecke (Eds.), *Methods, theories, and empirical applications in the social sciences* (pp. 105-113). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Krosnick, J. A., & Berent, M. K. (1993). Comparisons of party identification and policy preferences: The impact of survey question format. *American Journal of Political Science*, *37*, 941-964.

Krosnick, J. A., & Presser, S. (2009). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (pp. 263-313). Bingley, England: Emerald.

Leung, S.-O. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert scales. *Journal of Social Service Research*, *37*, 412-421. doi: 10.1080/01488376.2011.580697

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *4*(2), 73-79. doi: 10.1027/1614-2241.4.2.73

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. *Educational and Psychological Measurement*, *31*, 657-674. doi:10.1177/001316447103100307

Menold, N., Kaczmirek, L., Lenzner, T., & Neusar, A. (2014). How do respondents attend to verbal labels in rating scales? *Field Methods*, *26*, 21-39. doi:10.1177/1525822X13508270

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

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, *29*, 81-117.

Muthén, L. K., & Muthén, B. O. (2014). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

O'Muircheartaigh, C. A., Gaskell, G., & Wright, D. B. (1995). Weighing anchors: Verbal and numeric labels for response scales. *Journal of Official Statistics*, *11*, 295-307.

Parker, R. I., Vannest, K. J., & Davis, J. L. (2013). Reliability of multi-category rating scales. *Journal of School Psychology*, *51*, 217-229. doi:10.1016/j.jsp.2012.12.003

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*, 1-15. doi:10.1016/S0001-6918(99)00050-5

Raykov, T. (2012). Scale construction and development using structural equation modeling. In R. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 472-492). New York, NY: Guilford.

Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling*. Mahwah, NJ: Erlbaum.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Taylor & Francis.

Raykov, T., & Marcoulides, G. A. (2012). *Basic statistics. An introduction with R*. New York, NY: Rowman & Littlefield.

Raykov, T., & Marcoulides, G. A. (2015). Scale reliability evaluation with heterogeneous populations. *Educational and Psychological Measurement*. Advance online publication. doi:10.1177/0013164414558587

Raykov, T., Marcoulides, G. A., & Millsap, R. E. (2013). Examining factorial invariance: A multiple testing procedure. *Educational and Psychological Measurement*, *73*, 713-727.

Saris, W. E., & Gallhofer, I. N. (2007). Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey Research Methods*, *1*, 31-46.

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology*, *29*, 65-88. doi:10.1146/annurev.soc.29.110702.110112

Schutz, H. G., & Rucker, M. H. (1975). A comparison of variable configurations across scale lengths: An empirical study. *Educational and Psychological Measurement*, *35*, 319-324. doi:10.1177/001316447503500210

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, *55*, 570-582.

Struminskaya, B., Kaczmirek, L., Schaurer, I., & Bandilla, W. (2014). Assessing representativeness of a probability-based online panel in Germany. In M. Callegaro, R. Baker, J. Bethlehem, A. Göritz, J. A. Krosnick, & P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective* (pp. 61-84). New York, NY: Wiley.

Wakita, T., Ueshima, N., & Noguchi, H. (2012). Psychological distance between categories in the Likert scale: Comparing different numbers of options. *Educational and Psychological Measurement*, *72*, 533-546. doi:10.1177/0013164411431162

Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, *64*, 956-997. doi:10.1177/0013164404268674