Reversed Items in Likert Scales: Filtering Out Invalid Responders

Article in Journal of Psychological and Educational Research · May 2017

2 authors:

Krisztian Jozsa
University of Szeged
202 PUBLICATIONS **1,188** CITATIONS

SEE PROFILE

George Arthur Morgan
Colorado State University
165 PUBLICATIONS **5,556** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Understanding research methods and statistics View project

Developing Diagnostic Assessments View project

JOURNAL OF PSYCHOLOGICAL
AND EDUCATIONAL RESEARCH

REVERSED ITEMS IN LIKERT SCALES: FILTERING OUT INVALID RESPONDERS

Krisztián Józsa[*]
University of Szeged, Hungary

George A. Morgan
Colorado State University, USA

Abstract
Likert type scales are commonly used in social sciences. Most of the Likert scales include both positively- and negatively worded items. However, the use of negatively worded (reversed) items is supported by some researchers but not others. This study analyzes the reversed items in educational settings. The school age, self-rating version of the Dimensions of Mastery Questionnaire (DMQ 17) was used. The sample consisted of 7261 Hungarian students, age 10 to 16. An iteration method was developed and used to filter our presumably invalid responders. The analysis is based on the empirical inconsistency between the reversed and the positively worded items. The iteration eliminated step-by-step the possibly invalid questionnaires. The reliabilities of the scales were increased with the iteration process. After eliminating about 20% of the sample, the reliabilities were somewhat higher with all scales having acceptable alphas. If one would like to use this iteration method for eliminating the invalid responders, he or she needs to oversample the accessible population. Based on this results we eliminated the reversed items form the new DMQ persistence and pleasure scales.

Keywords: reliability; Likert-scale; reversed items; invalid responses; mastery
           motivation

Introduction

Likert-scale questionnaires frequently have items that are worded negatively but later recoded so they can be combined with positively-worded

Correspondence concerning this paper should be addressed to:

[*] PhD, Institute of Education, University of Szeged. Address: Petofi street, No. 30-34, 6722, Szeged, Hungary. Email: *jozsa@edpsy.u-szeged.hu*

items to form a summated scale. These negative items are intended to encourage the respondents to read all items carefully rather than use a set pattern of responding. However, the use of negative items in questionnaire studies poses problems because some respondents don't read well or carefully. Instead, these respondents answer the negatively-worded items as if they were positively-worded or at least not consistent with the average of the positive items in the same scale or subscale. This inconsistent responding lowers the validity of the scale and also the internal consistency reliability (Cronbach's alpha) of the scale.

Likert scales

Rating scales are well known in the social sciences (Dillman, Smyth, & Christian, 2014; Gliner, Morgan, & Leech, 2017; Nunnally, 1978). One of the most frequently used rating techniques was developed more than 80 years ago by Likert (1932). He initially developed this method as a way of measuring attitudes about particular groups, institutions, or concepts. Researchers often develop their own scales for measuring attitudes or values, but there are also a number of standardized scales to measure certain kinds of attitudes, motivation. This type of rating scale was named after the creator, Rensis Likert, as a Likert-type scale, often called just Likert scale.

The term Likert scale is used in two ways: (1) for the summated scale; and (2) for the individual items or rating scales from which the summated scale is computed. Likert items are statements about a particular topic, and the participants are asked to indicate whether they strongly agree, agree, are undecided, disagree, or strongly disagree. The summated Likert scale is constructed by developing a number of statements about the topic, usually some of which are clearly favorable and some of which are unfavorable. These statements are intended to provide a representative sample of all possible opinions or attitudes about the subject. These statements are then presented to a group of participants who are asked to rate each statement from strongly disagree to strongly agree. To compute the summated scale score, each type of answer is given a numerical value or weighting, usually 1 for strongly disagree up to 5 for strongly agree. Some studies use another range of numbers, e.g., a 7-point scale, or an even-point scale (Gliner, Morgan, & Leech, 2017).

_____

Typically, there are several reversed, negatively word items in each Likert scale (Gliner, Morgan, & Leech, 2017; Hartley, 2013). However, some studies use equal number of reversed and non-reversed items (e.g., Baumgartner & Steenkamp, 2001). Some other researchers are against to use reversed items (e.g., DeVellis, 2003). An item can be reversed in different ways. For example, the item can be a negated statement which could include a negative word (e.g., not), or include an antonym (e.g., give up easily). When computing the summated scale, the negatively worded items need to be reversed in terms of the weighting. For example, in case of a 5-point scale strongly disagree is given a weight of 5 and strongly agree is given a weight of 1.

Data collected with summated rating attitude scales, like all the other data collection tools, need to be investigated for reliability. Internal consistency would be indicated if the various individual items correlate with each other, indicating that they belong together in assessing this attitude. Validity is seeing if this summated scale can differentiate between groups thought to differ on this attitude, or by correlations with other measures that are assumed to be related to this attitude. The construction of summated scales (for attitude or personality measurement) is discussed in depth by Spector (1992).

Some studies have focused on the research methodological aspects of reversed items. However, there are not many in educational settings (e.g., Barnette, 2001). Reversed items sometimes have lower item-total correlations, and lower model fit (e.g., Feifei & Tanner, 2013; Józsa & Molnár, 2013). The scales can have higher reliabilities after eliminating the reversed items (e.g., Barnette, 2000; Józsa, Wang, Barrett, & Morgan, 2014). Weijters, Baumgartner, and Schillewaert (2013) proposed an integrative model of three important sources of reversed item problems. They mention (1) acquiescence (preference for choice of a number from one side of the scale), (2) careless responding (random or nonrandom response, which is not related to the content), and (3) confirmation bias (tendency to activate beliefs that are consistent with the way in which the item is stated).

On the other hand, there are important advantages of including reversed Likert items in scales. Reversed items can improve scale validity. These items work as cognitive "speed bumps" and can cause a slower, more careful reading. Reversed items implicitly correct for acquiescence (Weijters & Baumgartner,

2012). These authors suggest to use reversed items in the scales. However, they suggest to use them with caution.

Definition and Importance of Mastery Motivation

The National Academy of Science report From Neurons to Neighborhoods (Shonkoff & Phillips, 2000) identified mastery motivation (the intrinsic drive to explore and master one's environment) as a key developmental concept, which should be included as part of a child's evaluation. Morgan, Harmon, and Maslin-Cole (1990) proposed that mastery motivation is a multifaceted, intrinsic psychological force that stimulates an individual to attempt to master a skill or task that is at least moderately challenging for him or her. Mastery motivation has two major aspects: instrumental and expressive (Barrett & Morgan, 1995). The instrumental aspect motivates a person to attempt, in a focused and persistent manner, to solve a problem or master a skill or task, which is at least moderately challenging for him or her (Morgan, Harmon, & Maslin-Cole, 1990). The expressive aspect of mastery motivation produces affective reactions while the person is working at such a task or just after completing it. This affect may or may not be overtly expressed and may assume different forms in different children as they develop.

The Development of Dimensions of Mastery Questionnaire

The Dimensions of Mastery Questionnaire (DMQ) assesses several aspects of perceptions of children's mastery related behaviors. This is a Likert type questionnaire. The DMQ was developed over the last 30 years, and is one of several measurement techniques, including challenging structured tasks and semi-structured play, developed to assess mastery motivation (Busch-Rossnagel & Morgan, 2013; Morgan, Józsa, & Liao, 2017).

When the development of this mastery motivation questionnaire began in the early 1980s, there were no parental report questionnaires designed to assess the motivation of infants and preschool children. Infant temperament questionnaires did assess perceptions of persistence, but none of them provided adequate coverage of the motivational aspects of toddlers' or preschoolers' attempted problem solving and mastery play. To our knowledge, the DMQ still is the only parental report measure of young children's mastery motivation.

Results of early versions supported the usefulness of the questionnaire, but we felt that the psychometric properties and age appropriateness of the questionnaire could be improved without losing the strengths. Revisions and expansions of the Dimensions of Mastery Questionnaire included the domains of persistence at gross motor and social tasks; the social mastery motivation scale were revised and split into two scales: social persistence with peers and social persistence with adults. In addition, scales of the expressive aspects of mastery pleasure, and negative reactions to failure in mastery situations were added.

Early versions were designed to assess parent or teacher ratings of toddlers and preschool children. Current versions of the DMQ were developed for school age children. The school-age versions had forms for the child to rate him or herself and a form for an adult (parent or teacher) to rate the child. All these age versions of the DMQ had 14 common items that were thought to be appropriate across ages. The remaining 31 items varied somewhat by age version but paralleled the items in the preschool version (Busch-Rossnagel & Morgan, 2013).

More than 15,000 children from 6-month to 19-year of age were rated with the DMQ 17. These included more than 1000 atypically developing children with a variety of delays and more than 500 children at risk due to low SES. Geographically and linguistically, the children were very diverse. Participants included English speakers from the United States, Canada, the UK, and Australia. Chinese speakers were from mainland China and Taiwan. In Hungary, more than 10,000 mostly typically developing school-age children rated themselves and/or were rated by their parents and teachers. In addition, a Spanish version of the DMQ 17 was used by Spanish speakers in the US, and translations into native languages also have been used to assess children from at least the Netherlands, Israel, and Korea. A number of journal articles, dissertations, and presentations have included the DMQ 17. Józsa (2007) published a book in Hungarian on his large sample studies of mastery motivation, cognitive skills, IQ, and school achievement. Overviews of DMQ 17 research on the Hungarian, English, and Chinese samples were published by Józsa and Molnar (2013), Morgan, Wang, Liao, and Xu (2013), and Józsa, Wang, Barrett, and Morgan (2014), Józsa and Morgan (2014). These papers summarized evidence for reliability and validity, relationships to other variables, and also compared the three cultures at similar ages and across ages.

However, a major issue was that the reverse coded items clearly caused problems for 10-20% of the raters, who did not seem to rate them accurately. This accuracy problem was inferred based on the assumption that rater's scores on the negatively worded item in each scale should (after it was recoded) be similar to the average of the positively worded items. If the discrepancy was large, the rater must not have been reading carefully (perhaps reading too fast), or have developed a response bias to use one end of the scale, or have been confused because of low reading ability.

To deal with this problem, Morgan, Busch-Rossnagel, Barrett, and Wang (2009) suggested a formula for deciding which questionnaires seemed to be invalid because of inaccurate reading of the negative items. However, decisions about the cut-point and how many questionnaires to delete were arbitrary. Thus, the present study developed and tested a computerized, iterative method for assessing the effect of deleting such inaccurate questionnaires based on the changes in the scale alphas. The iterative process began by filtering out the questionnaires with the biggest inconsistency with the positively-worded items; then the program moved to filter/delete slightly less inconsistent questionnaires and so forth.

Objectives

The goal of this study is to improve the reliability of Likert scales. We developed and tested a statistical method to increase the scale reliability. The research question was: Could a computerized, iteration method be used to effectively filter out presumably invalid questionnaires, and how would the Cronbach alpha reliability indices change after this filtration?

Method

Participants

The questionnaire was administered to 7261 10-16-year-old students. They studied in grades 4, 6, 8, and 10; 49% of the sample was male (Table 1). The sample was representative of Hungarian children according to gender, geographical distribution and parents' highest level of education.

_____

Table 1. Size and age of sample by school grade

Sample	Grade				Total
	4	6	8	10	
N	2448	1435	1389	1989	7261
Age Mean	10.85	12.92	14.91	16.76	13.61
Age SD	0.46	0.43	0.38	0.69	2.44
Boys (%)	49	50	49	50	49

Instrument

The Dimensions of Mastery Questionnaire DMQ (also known as DMQ version 17) is a self (or adult) rating of a child's motivation to master tasks or solve problems. In this study, five DMQ scales were used: four instrumental or persistence scales, and an expressive or affective scale. The instrumental scales are behavioral manifestations of persistence, which was a principle measure of mastery motivation in previous studies (e.g., Józsa, 2007; Józsa & Molnar, 2013; Morgan, MacTurk, & Hrncir, 1995). These scales include 1) cognitive persistence, 2) gross motor persistence, 3) social persistence with adults, and 4) social persistence with peers. For the expressive scale, labeled mastery pleasure, the items reflect positive affect during persistent mastery attempts or immediately after success.

Morgan et al. (2013) presented evidence that each of the four DMQ 17 instrumental/ persistence scales and mastery pleasure scale had acceptable to good internal consistency (alphas > .74) for both English and Chinese parent versions and the English version by teachers. Alphas for the child self-ratings were somewhat lower (.67 - .85) on these five scales. Some of the English-speaking children were 5-7 years old, probably too young to fully understand these self-ratings of their motivation, even when the items were read to them and the tester used visual aids.

There were also good Cronbach alphas for the Hungarian samples (Józsa, 2007; Józsa & Molnár, 2013), on the four instrumental/ persistence scales and the mastery pleasure scale for teachers and parents. Reliabilities of Hungarian teacher ratings were somewhat higher than those of parents. No significant age differences in alphas were found for either the teacher or the parent samples. However, reliability for student self-ratings were somewhat higher for older school-age groups than younger school-age groups. Development of reading comprehension undoubtedly influences the computed

reliability of the questionnaire, and it could be the reason for the increase in reliability indices with age. Total persistence, had an alpha of .92. For the Hungarian sample, Cronbach alphas ranged from .67, to .84 (median .74) with alpha of .88 for total persistence.

Józsa and Molnár (2013) also reported test-retest reliabilities, ranging from .61 to .94, for 98 Hungarian teachers, parents, and students on the four instrumental and two expressive scales. The median correlations for these scales were .83, .80, and .74 for teacher, parents, and students, respectively. These test-retest correlations were highest for cognitive/object and gross motor persistence, somewhat lower for the social mastery scales and mastery pleasure, and lowest for negative reactions to failure.

In this study the DMQ was an example of a summated scale, which included a few negatively-worded statements or items. Each scale contained one negatively-worded statement and 5 to 8 positively-worded statements. Children were assessed on 5-point Likert-type scales of how typical each of 35 behaviors is for the child (*see* Table 2).

Table 2. The Dimensions of Mastery Questionnaire (DMQ) Scales

DMQ Scales	*N* of Items	*M*	*SD*
Object Oriented Persistence	9	3.48	.59
Gross Motor Persistence	8	3.75	.81
Social Persistence			
with Adults	6	3.40	.71
with Children	6	3.83	.62
Mastery Pleasure	6	3.97	.70

The Hungarian version of the Dimensions of Mastery Motivation (H-DMQ) was used in this study. The H-DMQ was administered to school classes of Hungarian children by their teachers.

Two sample items for each of the five DMQ scales are shown in Table 3. For each scale, one of the several positively worded items is shown, and the one negatively item is also shown.

Table 3. Samples of Positive and Negative Worded Items

Object-Oriented (Cognitive) Persistence

9. If a task is hard to do, I stop trying after a short time. (R)

23. I work for a long time trying to do something hard.

Gross Motor Persistence

3. I give up easily if I <u>cannot</u> do physical skills well. (R)

12. I try to do well in physical activities even when they are hard for me.

Social Persistence with Adults

22. I try very hard to get adults to understand things.

33. I give up quickly, when I play with adults. (R)

Social Persistence with Children

32. I try to get included when other children are playing.

39. I avoid getting involved with other children. (R)

Mastery Pleasure

9. I do not smile after I make something happen. (R)

18. I get excited when I figure something out.

Note: An R notes that this item is reverse-scored

Design

This study was a cross-sectional data collection all over Hungary. We used the Hungarian Educational Authority's school database for random sampling. The questionnaires were filled out in a classroom setting part of a school class. The data collection was managed by the school teachers. Detailed instructions were sent to the teachers before the data collection.

Procedure

We designed SPSS syntax for making the iterative filtration. This syntax could handle simultaneous changes in multiple parameters. In addition to executing the statistical computations, it also saved the results of various filtrations into a database and displayed the Cronbach alphas.

The filtrations were conducted one dimension or subscale at a time. The respondents who rated both positive and negative items (before they were recoded) in essentially the same way as either high or low were filtered out being considered invalid. The filtration was conducted using steps of 0.2 difference in the mean scores of the positive items. For example, the first respondents to be excluded were the ones who had a mean rating of 5.0 on the positive and who rated the negative item in that scale as a 4 or 5. Then we moved to exclude those who had a positive item mean of 4.8 with a 4 or 5 on

the negative statement, and so forth down by 0.2 mean points to a positive item mean of 3.0. The analyses were conducted symmetrically; i.e., we used a similar procedure with those who rated both the negative item low and had a low mean on the positive statements. Throughout these steps we computed values of scale reliability (alpha) for the respondents who remained in the sample after filtering, and we examined the alpha's of the respondents who had been excluded.

Children who rated all items in a scale, both the positive items and the negative one, as 5 or 1, were excluded on the first step and their alpha was not calculated because there was no variability and the alpha would have been artificially inflated. These deleted questionnaires were less than 1% of the original sample.

Results

At each step of iteration, Cronbach alphas were computed for each scale separately for the respondents who were deleted and for the remaining respondents in the original sample. As shown in Figure 1 and 3, the alphas for the deleted (filtered out) respondents (those answering the negatively-worded items inaccurately) were quite low. The other three scales also showed unacceptable alphas. In Figure 1 for mastery pleasure, the alphas were very low for all iterations when the means of the positive items varied from 4.6 to 3.0. For social persistence with adults (Figure 3), the pattern was different, with the alphas gradually decreasing from .80 when the mean of the positive items was 4.8 to near zero when the mean of the positive items approached 3.0. This indicates that there was very low internal consistency reliability for the respondents who were inaccurately reading (answering) the negative items. However, for the first several iterations (positive item means near 5), there was little variation in either the positive or negative items, making the alphas artificially high as shown in Figures 1 and 3.

The change in alphas for the remaining participants (after those who misread were excluded) varied across the five dimensions. For mastery pleasure (Figure 2), alphas increased slightly from .71 to a maximum of .78 when the mean of positive items was 3.6 and 14% of original sample was filtered out; then the alpha remained essentially the same as the mean of positive items was decreased to 3.0. For persistence with adults (Figure 4) the

_____

alpha increased gradually from .64 to .76 when the positive item mean was 3.0 and 81% of the original sample remained. Table 4 shows how much the maximum alphas differed from the original alphas for each scale, and the percentage of respondents who would be deleted when the alpha reached its maximum. Note that although the increase in alphas was relatively small (.02-.12), when the reversed item was included, the two scales with alphas less than .70 increased to acceptable levels ($\geq$ .70). Notice also that the percentage of respondents filtered out varied a lot from 7% to 22%.
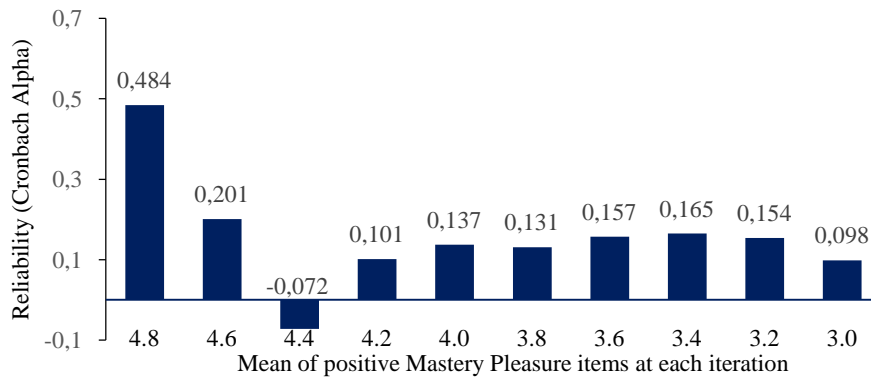

Figure 1. Participants filtered out at each iteration and reliabilities of the filtered out mastery pleasure scale
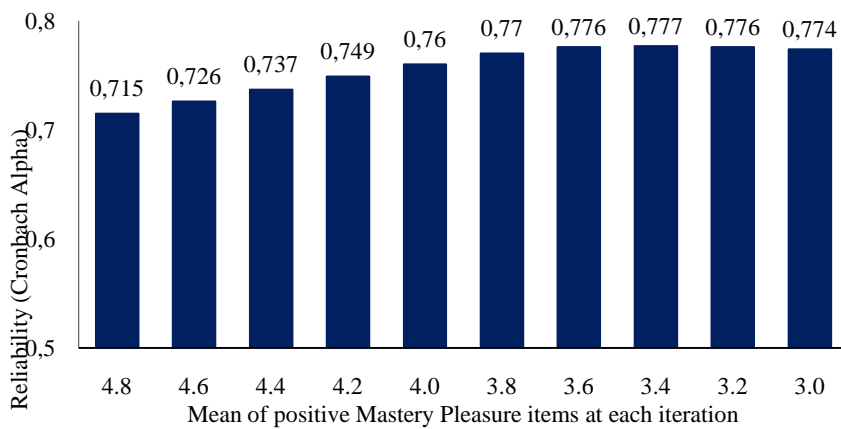

Figure 2. Questionnaires remaining after those filtered out were deleted and reliabilities of the questionnaires of the remaining mastery pleasure scale questionnaires at each iteration
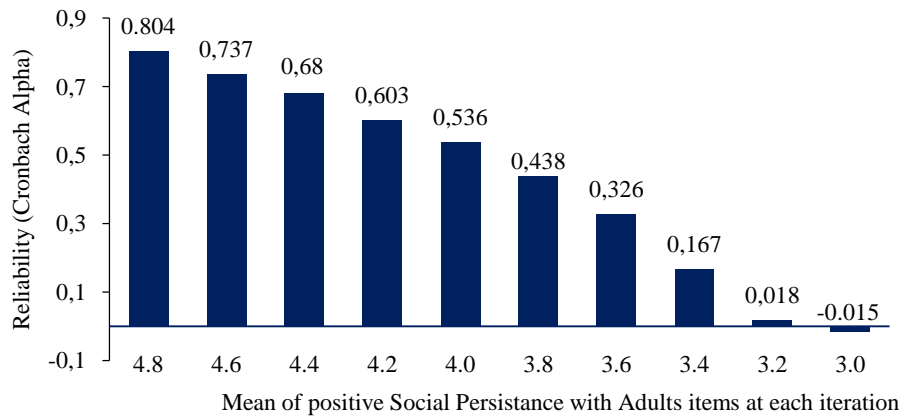
17

Figure 3. Participants filtered out at each iteration and reliabilities of the filtered out social persistence with adults scale
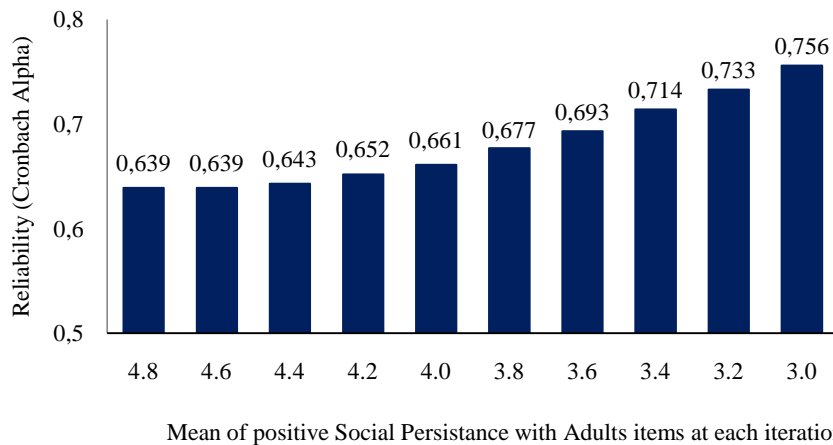

Figure 4. Questionnaires remaining after those filtered out were deleted and reliabilities of the remaining social persistence with adults scale questionnaires at each iteration

We also calculated the reliabilities just from the positive items. We did it both before the filtering iterations and after the last step of the iterations. Before the iterations, all of the scales had the same or higher alphas without the reversed items. After filtering out the presumably invalid responders, the alphas

were the highest, if we filtered out the invalid responders and then use just the positive items (*see* columns 4 of Table 4).

Table 4. Reliability (Cronbach-α) for the DMQ Scales Before and After Filtering out Invalid Questionnaires and the Percentage Deleted at the Maximum Alpha

DMQ scales	Including reversed item		Not including reversed item		Percentage filtered out
	Before filter	Max after filter	Before filter	Max after filter	
Object oriented persistence	.70	.74	.73	.75	22%
Gross motor persistence	.81	.83	.81	.83	7%
Social persistence					
with adults	.64	.76	.71	.78	19%
with children	.63	.72	.68	.73	21%
Mastery pleasure	.71	.78	.73	.80	14%

Initially, we had five separate iterations for the five DMQ scales. Next step, we combined the five iterations and made just one iteration for all the five scales. In this process, we eliminated a student from the sample if he or she was eliminated in at least four of the five separated iterations. That means, we used the same sample for all of the DMQ scales in this type of iteration. We computed the alphas after eliminating 10% and then 20% of the sample (Table 5). All of the alphas increased after eliminating 10% of the presumably invalid responders. The alphas are somewhat higher after eliminating 20% of the students, and the alphas are higher if we omit the reversed items from the scales.

Table 5. Reliability (Cronbach-α) for the DMQ Scales after filtering out 10% and 20% of presumably invalid responders

DMQ scales	Before filter	Including reversed item		Not including reversed item	
		10% filtered out	20% filtered out	10% filtered out	20% filtered out
Object oriented persistence	.70	.72	.74	.74	.76
Gross motor persistence	.81	.81	.83	.81	.83
Social persistence					
with adults	.64	.68	.74	.76	.76
with children	.63	.66	.71	.72	.73
Mastery pleasure	.71	.72	.79	.80	.81

19

Discussion

  The authors and others who have used the DMQ (and other widely used summated scales with negative items) have long noted that some respondents either (a) do not read well so are confused by the negatively-worded items or (b) read too quickly so misread the negative items, or (c) have such a strong response bias that they will not rate themselves (or their child) low on any items.

  It was surprising to us that the alphas were relatively high for the deleted questionnaires on the first several iterations and then dropped dramatically. This small percentage of respondents was very clearly not reading and rating validly, yet they had relatively good alphas. On closer examination we realized that most were consistently rating both the positive and the negative items high (i.e., 4or 5), but a few were doing the opposite, consistently rating both types as low (i.e., 1 or 2). After reverse coding the negative items, the interitem correlations were high as were the alphas, apparently because there was little variability on the object persistence scale.

  It is disturbing that such high percentages (up to 22% on the object persistence scale) seem to be answering invalidly. It seems important to ask why and what to do to remedy this. The cut point for exclusion of questionnaires needs to balance maximizing alpha and validity with minimizing the percentage of deleted questionnaires. A problem is that different scales required that different percentages of respondents be deleted in order to maximize the alpha. Assuming that a researcher wanted to delete whole participants rather than only selected scales, the balance is more difficult. Perhaps one could consider deleting only enough participants to make the lowest alphas be acceptable, perhaps .70 or above. That would probably reduce the percentage of deleted questionnaires to something like 10% of the total. If one were to delete a questionnaire as invalid based on procedures like those described in this paper, it would be wise to oversample by perhaps as much as 20%.

  A less iterative method similar to the one used here has been used with some success with Chinese and American DMQ data, but this method also should be tried with questionnaires other than the DMQ as suggested by Morgan, Busch-Rossnagel, Barrett, and Wang (2009). These analyses could contribute to defining specific criteria for improving the reliability and validity of Likert-scale studies. The influence of using this method also will be

examined in the future to see if the relationships with other variables; (e.g., IQ, cognitive skills, and school achievement), will increase the validity of the questionnaire when apparently invalid responders are deleted.

One of the reasons for these invalid responses could be reading comprehension problems. Based on the PISA (Program for International Student Assessment) result ca. 20% of the Hungarian students have very serious reading problems (OECD, 2016; Ostorics, Szalay, Szepesi, & Vadász, 2016). These children are functionally illiterate. This percentage of children with serious reading comprehension problems is in the line with our results, where we suggested deleting about 20% of the subjects. This reading comprehension issue could be very important in educational research settings, but deleting students with reading problems would make the sample less representative of the student population.

The Hungarian data was collected in many classrooms by the teachers; it is possible that some teachers did not administer the questionnaire correctly so that students did it carelessly or too quickly. There are known advantages, in terms of higher response rates, to questionnaire administered to "captive" audiences, but it may well be that there is a concurrent increase in invalid ratings.

Computer based data collection techniques can help the children who have reading problems. The computer can read aloud the statements. The visualization of the numbers is also useful; e.g., in the case of a 5-point Likert scale, 5 circles (from a little, light colored circle to a big, dark colored one). A short, ca. one minute long video explained the meaning of the circles to the students. In the case of younger children and poor readers, these computer-based techniques can increase the reliability of the Likert scales (Józsa, 2014; Józsa, Hricsovinyi, & Szenczi, 2015).

Limitations

This study was conducted in Hungary, and we used students self-rating questionnaires. Thus, our results are limited to Hungarian school-age children. We analyzed the DMQ's scales and tried to generalize the results to all Likert type scales. It would be beneficial to replicate the study in other cultures and also with other Likert type questionnaires. Also, future research should analyze other raters; e.g., parents' and teachers' ratings.

_____

Conclusions

Including negatively-worded items in the Likert scales can cause some problems. On the other hand, there are some advantages of using them. Carefully constructed negatively-worded items could be useful if one wanted to identify and then eliminate subjects who respond invalidly. After that, one can decide to summate just the positive items, or include the reversed items also in the summated variable.

The questionnaire items, especially negative worded items, may be confusing to raters, especially younger children and poor readers. We made considerable effort with the several revisions of the DMQ to make negative items be clear, including underling the word <u>not</u> in the few items that used it, so that it would not be missed easily. In earlier versions of the DMQ we had more negatively worded items. For the current study, we decided to retain one such negative item in each scale, in part to be able to check the validity of the responses.

After this study, we deleted the reversed items from the DMQ persistence and mastery pleasure scales. In the most recent version of the DMQ, we do not include any reversed items in these scales. However, we still have the negative reaction to failure items, which can serve the purpose of encouraging more careful reading (Józsa & Morgan, 2015).

References

Barnette, J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement, 60*(3), 361-370.

_____

Barnette, J. (2001). Likert survey primacy effect in the absence or presence of negatively-worded items. *Research in the Schools*, *8*(1), 77-82.

Barrett, K. C., & Morgan, G. A. (1995). Continuities and discontinuities in mastery motivation in infancy and toddlerhood: A conceptualization and review. In R. H. MacTurk, & G. A. Morgan (Eds.), *Mastery motivation: Origins, conceptualizations, and applications* (pp. 67-93). Norwood, NJ: Ablex.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*, 143-156.

Busch-Rossnagel, N. A., & Morgan, G. A. (2013). Introduction to the mastery motivation and self-regulation section. In K. C. Barrett, N. A. Fox, G. A. Morgan, D. J. Fidler, & L. A. Daunhauer (Eds.), *Handbook of self-regulatory processes in development: New directions and international perspectives* (pp. 247-264). New York, NY: Psychology Press.

DeVellis, R. F. (2003). *Scale Development: Theory and Applications.* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, Phone, Mail and Mixed-Mode Surveys: The Tailored Design Method*. 4th edition. John Wiley: Hoboken, NJ.

Feifei, Y., &Tanner L. W. (2013). Psychological Sense of School Membership Scale: Method Effects Associated With Negatively Worded Items. *Journal of Psychoeducational Assessment*, *32*(3), 202-215.

Gliner, J. G., Morgan, G. A., & Leech, N. L. (2017). *Research methods in applied settings: An integrated approach to design and analysis*. New York: Routledge/Taylor & Francis Group.

Hartley, J. (2013). Some thoughts on Likert-type scales. *International Journal of Clinical and Health Psychology*, *13*, 83-86.

Józsa, K., Hricsovinyi, J., & Szenczi, B. (2015). Számítógép-alapúElsajátításimotivációkérdőívekvaliditása és reliabilitása [Validity and Reliability of the computer-based Dimensions of Mastery Questionnaire]. In B. Csapó, & A. Zsolnai (Eds.), *Online diagnosztikusmérésekaziskolakezdőszakaszában* [Online diagnostics assessment in elementary school] (pp. 123-146). Budapest: Oktatáskutató és FejlesztőIntézet.

23

Józsa, K. (2007). *Azelsajátításimotivació* [Mastery motivation]. Budapest, Hungary: MűszakiKiadó.

Józsa, K. (2014). Developing new scales for assessing English and German language mastery motivation. In J. Horvath, & P. Medgyes (Eds.), *Studies in honour of Marianne Nikolov* (pp. 37-50). Pécs: Lingua Franca Csoport.

Józsa, K., & Molnár, E. D. (2013). The relationship between mastery motivation, self-regulated learning and school success: A Hungarian and European perspective. In K. C. Barrett, N. A. Fox, G. A. Morgan, D. J. Fidler, & L. A. Daunhauer (Eds.), *Handbook on self-regulatory processes in development: New directions and international perspectives* (pp. 265-304). New York, NY: Psychology Press.

Józsa, K., & Morgan, G. A. (2014). Developmental changes in cognitive persistence and academic achievement between grade 4 and grade 8. *European Journal of Psychology of Education*, *29*(3), 521-535.

Józsa, K., & Morgan, G. A. (2015). An Improved Measure of Mastery Motivation: Reliability and Validity of the Dimensions of Mastery Questionnaire (DMQ 18) for Preschool Children. *Hungarian Educational Research Journal*, *5*(4), 1-22.

Józsa, K., Wang, J., Barrett, K. C., & Morgan, G. A. (2014). Age and cultural differences in mastery motivation in American, Chinese, and Hungarian school-age children. *Child Development Research. 2014, 16 pp.* doi:10.1155/2014/803061

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, No. 140. New York.

Morgan, G. A., Busch-Rossnagel, N. A., Barrett, K. C., & Wang, J. (2009). *The Dimensions of Mastery Questionnaire (DMQ): A manual about its development, psychometrics and use*. Colorado State University, Fort Collins.

Morgan, G. A., Harmon, R. J., & Maslin-Cole, C. A. (1990). Mastery motivation: Definition and measurement. *Early Education and Development*, *1*, 318-339.

Morgan, G. A., Józsa, K., & Liao, H.-F. (2017). Introduction to the Special Issue on Mastery Motivation: Measures and Results across Cultures and Ages. *Hungarian Educational Research Journal*, *7*(2), (in press).

_____

Morgan, G. A., MacTurk, R. H., & Hrncir, E. J. (1995). Mastery motivation: Overview, definitions and conceptual issues. In R. H. MacTurk, & G. A. Morgan (Eds.), *Mastery motivation: origins, conceptualizations and applications* (pp. 1-18). Norwood, NJ: Ablex.

Morgan, G. A., Wang, J., Liao, H.-F, & Xu, Q. (2013). Using the Dimensions of Mastery Questionnaire to assess mastery motivation of English - and Chinese-speaking children: Psychometrics and implications for self-regulation. In K. C. Barrett, N. A. Fox, G. A. Morgan, D. J. Fidler, & L. A. Daunhauer (Eds.), *Handbook of self-regulatory processes in development: New directions and international perspectives* (pp. 305-335). New York, NY: Psychology Press.

Nunnally, J. C. (1978). *Psychometric Theory*. 2nd edition. New York, NY: McGraw-Hill.

OECD (2016). *PISA 2015. Results in focus*. Paris: OECD Publishing.

Ostorics, L., Szalay, B., Szepesi, I., & Vadász, Cs. (2016). *PISA 2015 összefoglaló jelentés* [PISA 2015 Technical Report]. Budapest: Oktatási Hivatal.

Shonkoff, J. P., & Phillips, D. A. (2000). *From neurons to neighborhoods: The science of early childhood development*. Washington DC: National Academy Press.

Spector, P. E. (1992). *Summated Rating Scale Construction*. Newbury Park, CA: SAGE Publications.

Weijters, B., & Baumgartner, H. (2012). Misresponse to Reversed and Negated Items in Surveys: A Review. *Journal of Marketing Research*, *49*(5), 737-747.

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods*, *18*(3), 320-334.