METODE & FORSKNINGSDESIGN
METHODS & RESEARCH DESIGN
https://journals.aau.dk/index.php/mf

Nr. 5 • 2023

A Scoping Review of Rankings, Smileys, and Other Survey Item Formats

Jacob Brauner
Copenhagen Center for Arthritis Research, Center for Rheumatology and Spine Diseases, Rigshospitalet, Glostrup, Denmark
jjoe0413@regionh.dk

Abstract:

Technological development has allowed researchers to apply numerous item formats in web-based surveys. A growing body of research suggests that the use of formats for web and paper other than multiple choice, such as ranking, sorting, questions with pictures e.g., may offer relevant alternatives that can strengthen data quality. These formats are referred to as Innovative Item Formats (IIF). Existing literature in the field is not able to present a systematic overview and functional typology of IIFs and their impact on data quality. Therefore, a review of the research is needed for each IIF.

This review is designed with the purpose of covering which typers of IIF that exist and what type of evidence there is about data quality on these IIFs. Based on a scoping review, this article presents the existing research literature on specific IIFs. A total of 62 research articles with data from 89,365 participants were identified. A more extensive typification of IIFs than previously used, one that includes a total of 23 IIFs and 13 subcategories, is suggested. Researchers designing questionnaires can use this knowledge to obtain higher-quality data.

Keywords:

Innovative item formats, survey, questionnaire, validity, reliability, scoping, review

ISSN: 2245-3083

Introduction

Numerous studies of item format data quality in questionnaires for tests and surveys have been conducted. Yet synthesized knowledge, i.e. reviews and meta-analysis, across the different types of Innovative Item Formats (IIFs) is sparse, as stressed previously by Wan and Henly (2012). IIFs are being used extensively in questionnaires, especially in psychometric tests, and increasingly over time as technology evolves. Researchers strive for valid and reliable tests and the increasing use of IIFs over time may also, in part, be due to pressure from politicians to use innovation to develop better tests. Thus, some tests have been criticized for not having sufficiently high validity and reliability, and evidence suggests that the use of IIFs may strengthen data quality (Crabtree 2016). Additionally, the rise may be due to higher popularity among respondents (Wan & Henly 2012). As the demand for the use of IIFs in surveys rises, so does the supply of formats in survey software and the need to synthesize research about each single IIF to ensure optimal questionnaire designs.

Several experiments, meta-studies, and reviews have outlined the quality of other survey design features than item formats, such as questionnaire color, logo, time of delivery, length, wording, and many other features (Haladyna et al. 2010; Scherpenzeel 1997; Sørensen et al. 2014). Research was also previously synthesized on single IIFs, namely, randomizations, draw functions, time-limited answers, and visual analog scales (Knäuper 1999; Shulman, 2000; Voyer 2011; Chyung et al. 2018; Chyung et al. 2018 II; Chiarotto et al. 2019). There is a lack of synthesized knowledge about data quality across the different types of IIF.

Various researchers have tested questionnaires based on multiple IIFs against traditional questionnaires and found that IIF provided more information, allowed for higher efficiency, or that it gave results similar to classical formats, but also that IIFs took more time to fill out (Crabtree 2016; Jodoin 2003; Young & Wilson 2012). For instance, Crabtree tests questionnaires containing multiple IIFs versus a traditional format without IIFs. Conclusions like these substantiate a review on each specific IIF.

Scoping reviews offer identification and mapping of existing evidence before a systematic review is completed. Based on a scoping review of existing literature, this paper will explore which IIFs can be identified from a comprehensive list of IIFs and which types of evidence about data quality that exists. The scoping review is designed to allow systematic review within separate IIFs based on identified evidence and terminology.

1. Method

The scoping review, which has been used extensively in recent years, is an approach intended to identify sufficient evidence as a precursor to a full synthesis such as a

systematic review and/or meta-analysis, or to determine that further primary research will be needed before further synthesis (Tricco et al. 2016). This has value 1) when no prior reviews are available and 2) when the methods and topics are broad and heterogenous. The scoping review is exploratory in its nature and is based on iterative methods as opposed to systematic reviews, which involves enhancing search criteria repeatedly to broaden the search (Armstrong et al. 2011).

By design, scoping reviews have little restrains on characteristics in the identified studies. Instead, the strength of the scoping review is a mapping process to identify what exists so that future systematic reviews can be designed with meaningful systematic restrictions regarding which topics and methods to include. Since little knowledge could be identified beforehand, starting out with a systematic review would be of little value, since search terms would be impossible to define. According to the standards of a systematic review, the systematic review has a better foundation as a successor to scoping reviews than being the first review (Munn et al. 2018). The scoping review is designed to map the evidence and not to draw conclusions on outcomes such as the effect of using IIFs. In this sense the scoping review differs from a narrative literature review.

The review process in this study began with the identification of key concepts. In this case, semantic differentials, ranking, sorting, and smiley scales were the starting concepts. Since these item formats are often found in survey software, we expected to identify further formats by exploring features in such software. Specifically, Qualtrics and Question Pro were chosen as software that could potentially reveal additional item formats. Relevant research articles were then initially identified in Google Scholar, EBSCO, JSTOR, and Rex (the Danish National Library research database). For several item formats, the searches in these databases revealed very few relevant results. It was therefore decided to include snowball sampling, also known as chain sampling, based on references in the already identified articles. This allowed for the identification of a much higher number of articles and a further expansion of the typification. The inclusion of item formats was initially determined by definitions adopted in previous research articles. However, we realized in the process that this made for an unclear demarcation between item formats for inclusion and exclusion.

A flow diagram of the research process is presented in Figure 1. The review includes inferential and psychometric studies on each of the IIFs. The research was identified using database searches, snowball sampling, and survey software. Since the scoping review is done as a predecessor for systematic review, the inclusion criteria are broad. Articles in peer-reviewed journals, books, dissertations, conference papers, and online resources were included. In the end of the process, a full review of two journals in years 2019 – 2021 were included. A full review of select journals have been suggested as a way of strengthening the review since the iterative method may produce inexhaustive results (University Libraries Health Science Library, 2023). If further

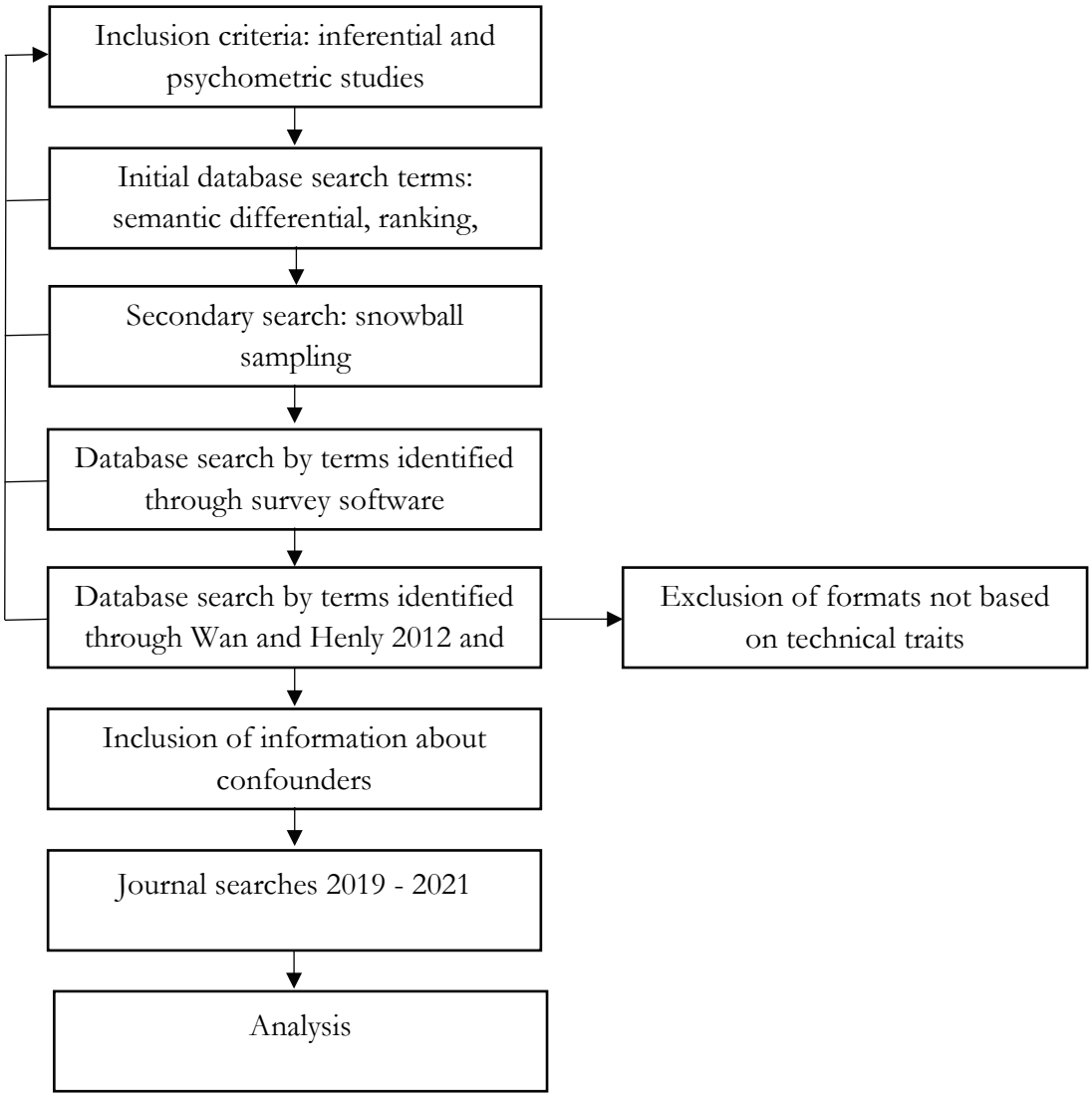
results are found through a full review of a journal, this suggests more evidence may be found elsewhere also. The journals were chosen based on frequency of existing results and impact factor.

The studies included fell into three categories: 1) studies with inferential traits regarding IIF impact on an outcome, 2) studies with scale construct traits, i.e. IIF ability to measure a given trait, and 3) studies with mixtures of inferential traits and scale construct traits. Standards have been suggested for both inferential and scale construct traits, and our inclusion criteria were based on a combination of these standards. Standards for inferential traits are based on an overall hierarchy of methods, whereas standards for scale constructs are based on a hierarchy and specific guidelines. The complexities of each category of standards suggest the need for a scoping review to allow consideration of each research method.

Several general standards for causal inference methods, regardless of field, have been suggested (Guyatt et al. 1995; Murad et al. 2016). These are generally designed for the purpose of estimating causal effects and the limitations of the methods through the theoretical introduction of systematic bias. Systematic reviews and meta-studies of randomized controlled trials are usually placed at the top of the hierarchy, followed by single randomized controlled trials (RCT), then cohort studies on the next level, and finally case-control studies, cross-sectional surveys, and case reports on the last level (Guyatt et al. 1995). The present review will include systematic reviews, meta-studies, randomized controlled studies cohort studies, and cross-sectional surveys.

Various suggested standards for construct traits also exist. The inclusion criteria in the present study for construct traits are guided by American Psychological Association (APA) Standards (American Educational Research Association 2014). Although the APA Standards have a slight overlap with inferential standards, their purpose is different. They are intended to assess whether a latent construct can be measured with a given method and research evolve around validity and reliability aspects within Item Response Theory and 'classical' theory. There are many test aspects, including confirmatory factor analysis inspecting the existence of latent constructs through tests of factor analysis appropriateness, dimensionality, factor loading, communality, variance explained, and model fit. Reliability is often tested by Cronbach's α often referred to as a measure of internal consistency based on item covariances divided by total variance i.e. a measure of fluctuations. Other aspects of reliability can be tested by test-retest, interrater reliability, and intrarater reliability. Psychometric tests cover various aspects of validity and reliability and include a large selection of statistical methods such as structural equation models, Rasch-models, principal component analysis, multitrait multimethod, and regression models, which were not identified in this scoping review. Interpretation of psychometric tests can be a complex matter but for instance, Eigenvalue factor loading above 1 and Cronbach's α above .7 and below .9 is usually considered acceptable.

Figure 1 Flow diagram



IIF Typification

Several sources were considered in the identification of item format typification that each include some IIFs. Instead of typification based on function Bennett et al. (1990) suggested differentiating between purpose: multiple choice, selection/identification, reordering/rearrangement, substitution/correction, completion, construction, and presentation. Zenisky and Sireci (2002) used the term innovative item formats (IIF) to identify 12 specific item types by their purpose in a test. The term includes everything that goes beyond the traditional multiple choice, multiple answers, dropdown, and standard open-ended items.

Other researchers have subsequently used the same term (Wan & Henly 2012), which is why it is also used here. Still, others have used the term technology-enhanced items (TEI) (Scalise & Gifford 2006; Crabtree 2016), with an emphasis on technological development. Zheng (2011) mentioned several innovative item types, including highlighting, reordering, or filling in blanks in text in different variations. Sireci and Zenisky (in Downing & Haladyna 2006) reviewed computerized item types and suggest a taxonomy of 21 item types in 10 main categories: drag-and-drop, hotspot, reordering or rearrangement, completion, mathematical expressions, construction, formulating hypotheses, essay/short answer, passage editing, and presentation. Finally, the survey programs Qualtrics and QuestionPro and research articles have helped identify relevant IIFs. The partial overlap between formats from these sources confirm the need of a complete typification.

In the present article, the term IIF is used partly in accordance with the understanding of Zenisky and Sireci (2002) that IIF covers every format except multiple choice, multiple answers, dropdown, and simple open-ended items found in most survey software. However, the typification used here is based on technical traits instead of the purpose in a test, since the purpose is expected to represent a gray area that makes it difficult to differentiate between studies for inclusion and exclusion. The typification is expanded to include more types of IIF and subcategories based on technical traits found in survey software and through snowball sampling and database searches. Thus, in this article, 22 main question types with 13 identified subcategorizations are examined (see table 1). Categories in this typification can be combined, for example, by finishing a sentence with drag-and-drop sorting or randomizing order of response options in a battery.

Table 1 Types of item formats

Item format	Description
Battery	With gray out, the text color is changed from black to gray once the item is answered. This makes it clear to the participant that this response option is no longer in focus.
Battery, multi-entity scaling	Batteries consist of multiple questions represented with the same response options. Questionnaire batteries can be designed with more than one parameter for each item, such as measuring both respondent satisfaction and importance for each topic. With multiple boxes vertically these are sometimes referred to as grids or a response matrix. If two aspects are measured for each item, the respondent will see too columns for each item.
Constant sum	Respondents are asked to distribute numeric shares of a predefined sum. The numbers typed in by the respondent

Item format	Description
	need to add up to a specific number to be considered valid answers.
Distance to nonsubstantive response	Distance to a non-substantive response means that there is a graphically visible higher distance to a "don't know" or "no answer" response. This allows the respondent to see clearly that this option is not part of a substantive response option interval.
Drag-and-drop sorting	The respondent is asked to categorize answers in predefined categories, by dragging the words with a mouse or by using a touchscreen, for instance.
Draw function	With a draw function, the respondent can give an answer by drawing on the screen via a touchscreen, use a mouse or stylus to supply handwriting. Draw function can be used for a signature or a drawing.
Finishing a sentence	The respondent is asked to insert a word, a sentence, or symbol such as a comma to make sense of the context, based on discrete choice or open-ended response.
Image response, universal	Universal image response can be used for expressing happiness or degree by clicking or moving a graphical image with a smiley/emoticon or image of a glass of water (empty vs full). Smiley scales are usually represented as discrete response options, such as a succession of five faces, but can also be continuous slider scales on which the facial expression of a smiley is changed when a slider is moved back and forth.
Image response, specific	As opposed to universal images with smileys, specific image responses are used to give a better understanding of what is being answered by exemplifying it with images alongside text. Therefore, the evidence is more contextual.
Image, question/stimulus	Text-based questions supported by images that help the respondent understand what the question is about.
Image, dynamic/figural response	The respondent clicks on certain parts of a picture, draw lines on it, or click or drag interactive elements. The image area is usually mapped so that coding of the response is possible according to where the respondent clicks or drags objects.
Image, video	As image stimulus but with video, often to exemplify or present a case in the same way as a vignette.

Item format	Description
Interval item	The respondents mark two points on a scale instead of one, thereby expressing an answer as an interval instead of a single fixed value.
Randomization: nominal and ordinal	With nominal randomization, all response options are randomized in random order. With ordinal randomization, the direction of the item, rather than that of each response option, is randomized, such as a Likert-scale changing between agree-disagree and disagree-agree
Randomization, open-ended response	Discrete variable items can have an added open-ended option such as "Other, please specify:" that allows the respondent to give another response than the discrete choices allow. This format is often used when the listed response options are assumed to be non-exhaustive. This item choice is traditionally put at the end. Therefore, the randomization mechanism is not applied to this response option.
Randomization, separate non-random category	A randomized item that includes a response option that is not randomized such as "don't know" or "no answer".
Ranking	The respondent is asked to arrange a given set of choices in a specific order, which can be done with drag-and-drop, dropdown, or typing.
Time-limited answer: visible vs. invisible time limit	The respondent is given limited time to answer or is shown a time count.
Uploading a file	Uploading a file can be used to attach documentation used as part of an analysis as an alternative to supplying documents separately.
Visual analogue scale: several descriptors vs. endpoints, stepless vs. discrete, chromatic vs. one color, with and without graphics (such as smiley), with or without a marked neutral midpoint, having a default starting position or not, vertical or horizontal	In this article, graphic scales (Hayes & Patterson 1921; Freyd 1923—later referred to as visual analogue scales or VAS scales (Wewers & Lowe 1990)—were defined as any scale with a slider. Both terms describe analogue, stepless scales represented by a line. Graphical scales originally had several descriptors along the line, while later VAS versions only had endpoint descriptors. These are also sometimes referred to as semantic scales or semantic differential scales. These variations were identified: several descriptors versus endpoints, stepless versus discrete, chromatic versus one color, with and without graphics (such as smileys), with or without a marked neutral midpoint, with or without a default starting position,

Item format	Description
	and vertical or horizontal. Chyung et al. (2018 II), among others, have pointed out the confusion of scale type names.
Visual analogue scale, self-anchoring	Self-anchoring means the respondent makes their own endpoint categories instead of prescribed endpoints being made by the researcher.
Voice assistance/sound	Sound or voice-assistance supports the text form of the question for visually impaired or child respondents. They can also be used simply to enhance the survey user experience or as a stimulus.

The scoping review of the literature of IIFs will be conducted in the following sections including all IIFs in the typology, except nominal/ordinal randomization and visual analogue scale (sliders, graphical scales, semantic differential scales). The extensive research on these IIFs suggests systematic review instead.

2. Results

In this section results are reported alphabetically by the name of the IIF. Table 2 shows the number of included studies, total sample size, the overall conclusion, and references for each IIF. In Figure 2, the total sample size for each IIF and the number of studies identified are illustrated. High variation in the number of studies and sample size is seen within each type of IIF. These results lay a foundation for systematic review feasibility within each type of IIF.

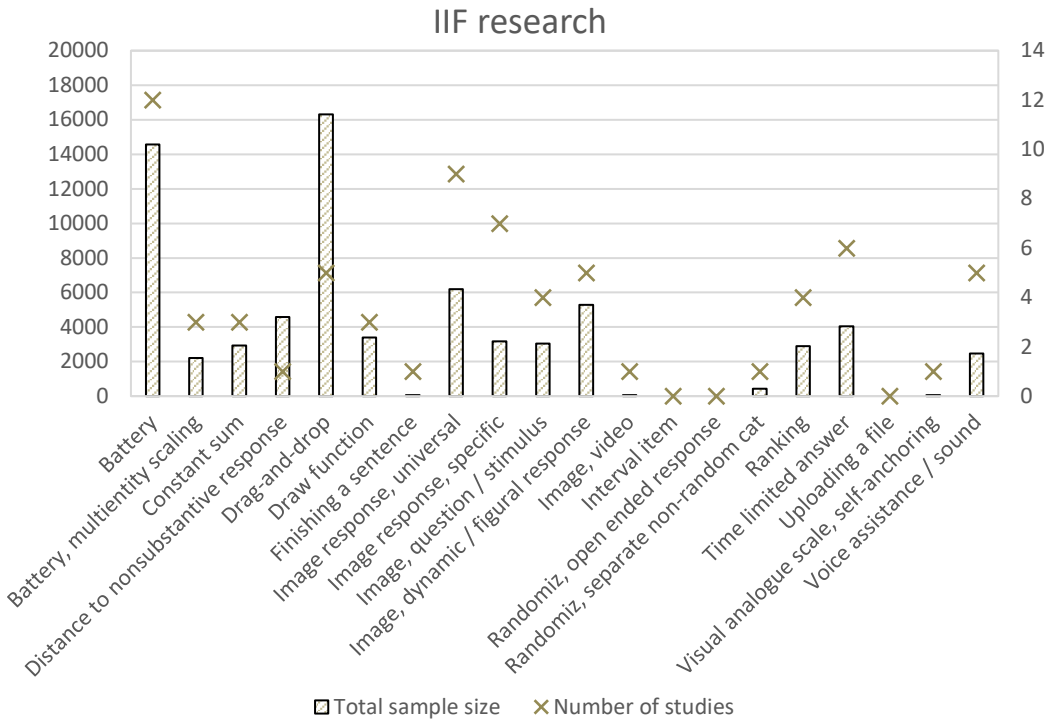
Table 2 Overview of evidence

Item type	Number of studies	N	References
Battery/response matrix/grid	12	14,563	Bell et al. 2001; Couper et al. 2001; Iglesias 2001; Tourangeau et al. 2004; Galesic et al. 2007; Chesnut 2008; Kaczmirek 2008; Callegaro 2009; Thorndike et al. 2009; Derham 2011; Kazmirek 2011; Couper et al. 2012
Battery, multientity scaling	3	2,204	Wong & Teas 2001; Borkenhagen et al. 2005; Sikkel et al. 2014
Constant sum	3	2,925	Louviere & Islam 2006; Conrad et al. 2005; Skedgel et al. 2013

Item type	Number of studies	N	References
Distance to nonsubstantive response	1	4,577	Tourangeau et al. 2004
Drag-and-drop sorting	6	16,317	Bennett & Sebrechts 1997; Timbrook 2013; Sikkel et al. 2014; Kunz 2015; Timbrook & Moroney 2016
Draw function	2 (22)	3,391	Björksten 1999; Shulman 2000; Elliot & Papadopoulis 2012
Finishing a sentence	1	52	Bleland 1999
Image response, universal	12	6,186	Schwarz et al. 1998, Buchanan & Niven 2002; Buchanan & Niven 2003; Castle & Engberg 2004; Buchanan 2005; Desmet 2005; Derham 2011; Reynolds-Keefer & Johnson 2011, Emde & Fuchs 2012; Broekens & Brinkman 2013; Toepoel et al 2019, Bosch & Revilla 2020, Gummer et al 2020
Image response, specific	12	6,012	Rofé 2004; Shamir & Kark 2004; Balram & Dragićević 2005; Ijmker 2008; Shropshire et al. 2009; Xia et al. 2011; Waller et al. 2012; Leutner et al. 2016
Image, question/ stimulus	4	3,043	Graybill & Heuvelman 1993; Escalante 1995; Couper et al. 2004; Wall et al. 2017
Image, dynamic/ figural response	10	20,139	Martinez 1991; Bennett et al. 1999; Bennett et al. 2000; Lim et al. 2006; Van Ooijen 2011; Wan & Henly 2012
Image, video	1	62	Kersting 2008
Interval item	0	0	-
Randomization, open-ended response	0	0	-
Randomization with separate non-	1	436	Krebs & Hoffmeyer-Zlotnik 2010

Item type	Number of studies	N	References
randomized category			
Ranking	4	2,888	Rankin & Grube 1980; Alwin & Krosnick 1985; Krosnick & Alwin 1988; Maio 1996
Time-limited answer	6	4,040	Weaver 1993; Stutts et al. 1998; Huesman et al. 2000; Mullane & McKelvie 2000; Lesaux et al. 2006; Voyer 2011
Uploading a file	0	0	-
Visual analogue scale, self-anchoring	1	62	Hofmans & Theuns 2008
Voice assistance/ sound	5	2,468	Turvey et al. 2012; Lam et al. 2009; Medin et al. 2016; Yost et al. 2009; Sinadinovic et al. 2011
Total	**166**	**89,365**	

Figure 2



FULL REVIEW OF JOURNALS 2019–2021

Public Opinion Quarterly and Social Science Computer Review were chosen for full review for the years 2019-2021 since several of the identified articles were published here and because these journals have high impact factors. Based on this review one additional study was identified in Social Science Computer Review (Hu 2019). In a sample of 5,599 Hu found horizontal ranking scales to have higher response time than vertical ranking but no significant difference on primacy effect, missing items and reliability.

RESULTS FROM ITERATIVE REVIEW

Below, each IIF and the related evidence concerning data quality will be reviewed.

Battery

The studies reviewed found small or insignificant time- differences (Bell et al. 2001; Couper et al. 2001; Kaczmirek 2008), insignificant or ambiguous results on reliability (Bell et al. 2001, Iglesias et al. 2001, Tourangeau et al. 2004, Callegaro et al. 2009), ambiguous results on response rate (Iglesias et al. 2001; Chesnut 2008), fewer missing data with sequential items (Chesnut 2008), higher or insignificant completion rates (Kaczmirek 2008; Callegaro et al. 2009) insignificant effect on the construct

(Thorndike et al. 2009), and respondent preference toward single item (Thorndike et al. 2009).

Switch rates were highest with a white background and lowest with gray out and cross (Kaczmirek 2008), non-response and dropout rates were lower, differentiation with the dynamic visual element of changing the color of answered items to light gray was higher (Galesic et al. 2007), dynamic feedback had fewer missing data (Couper et al. 2012), word scales in batteries were preferred before emoticons and numbered scales, and overall higher survey enjoyment was found with emoticons (Derham 2011). No lurking effect was found between highlight and grayout grids (Kaczmirek 2011).

Battery – multi-entity scaling

A test based on multi-entity scaling discriminated clearly between subgroups (Borkenhagen et al. 2005). As far as confounders were concerned, entity-based and attribute-based scaling both introduced instability and did not affect instability differently (Wong & Teas 2001), while low test-retest scores were found in both clicking and drag-and-drop.

Constant sum

Louviere and Islam (2006) found that constant sum affected data distribution compared to Likert. Significant differences in completion rates and preference consistency were not identified (Skedgel et al. 2013). IIFs with feedback, especially the concurrent type, produced more answers equal to fixed sums than IIFs with no feedback (Conrad et al. 2005).

Distance to nonsubstantive response

Tourangeau et al. (2004) conducted six experiments with webpage user surveys, of which experiments 1 to 3 were relevant for the distance to non-substantive response option. With experiments 1 and 2, the researchers investigated the effect of placement of a non-substantive response such as "no opinion" or "don't know," concluding that the midpoint was pushed in the direction of the non-substantive categories when these were not differentiated visually by line or space, presumably because the respondent was misled by the visual midpoint instead of the conceptual one. A similar result was found in experiment 3 when the distance was not equal, so the distance was gradually higher from left to right.

Drag-and-drop sorting

Test takers preferred sorting to multiple-choice (Bennett & Sebrechts 1997; Timbrook 2013). Test-retest values were at the same level as with clicking, but the vertical part of a 2-D grid drag-and-drop revealed the lowest test results (Sikkel et al. 2014). More missing data and longer response times were found for two types of drag-and-drops, but respondent attentiveness and carefulness were higher with drag-and-drop than

with radio buttons, and higher satisfaction was found with the ability to express answers with the distance between response options (Timbrook & Moroney 2016).

Draw function

Björksten et al. (1999) used a scale with visual analogue scale (VAS) items and respondent drawings to validate a musculoskeletal pain and conditions scale designed for clinical assessment. The scale had high scores on sensitivity (95%) and specificity (88%). Shulman (2000) tested psychometric properties based on a literature study that included 21 articles and found levels of validity too low for scientific use. In a study by Elliot and Papadopoulos (2012), respondents were asked to provide top-of-mind image associations in addition to regular close-ended questionnaire items. Respondents filled out the blank frames, giving a response rate of 62.3%.

Finishing a sentence

Research about finishing a sentence-item types has not been identified.

Image response, universal

Based on experimental data Schwarz, Grayson, and Knäuper (1998) found an effect of graphical scales using ladder, pyramid and onion-shaped box-distributions confirming to implied distributions. A strong correlation was found between the classical VAS and smiley scales (Buchanan & Niven 2002). There were no confounding effects of age and gender, with a reasonable agreement between child and parent ratings of a smiley scale (Buchanan & Niven 2003). Chernoff-face items were liked less than various text item formats, and the variation coefficient was higher with VAS without faces (Castle & Engberg 2004). There was good internal consistency, test-retest, and concurrent validity with a dental anxiety smiley scale (Buchanan 2005). A high test-retest correlation was found for a multifaceted cartoon imagery item type (Desmet 2005), Reynolds-Keefer & Johnson (2011) found similar patterns between four smiley scales among 15 child respondents, and correlations with different formats, other constructs, and outcome prediction were found with a step-less smiley IIF (Broekens & Brinkman 2013). In a descriptive study, Derham (2011) found a higher rate of unanswered questions with emoticons and a higher preference for word scales. Word items expressed feelings better, but emoticons were found to be easier to use, and overall survey enjoyment was highest with emoticons.

No significant difference in distribution was found between fixed format, dynamic smileys with a change of color, and those with supplemental text (Emde & Fuchs 2012). Respondents preferred emojis to traditional scales, but also expressed ambiguity when they were used in open-ended questions. In a sample of 7,096 participants Toepoel, Vermeeren, & Metin (2019) found lower average scores on graphical scales with hearts and stars compared to other formats, whereas smileys received scores in line with radio buttons. Based on experimental data (n = 2,247+3,993+3,993).

Gummer et al 2020 explored smiley face IIF. In one experiment they found no significant difference between using smiley faces and text except an increased response time and less response time when smiley face scales were fully or end-point text labeled. They found small pseudo R^2 differences between verbal end-labeled, smiley fully labeled and smiley end-labeled models on midpoint response, extreme response, number of clicks and response time. In another experiment part of the same study, they also did not find significant difference between verbally labeled smiley scale and other scales, fewer midpoint responses, and higher response time.

Image response, specific

Shamir & Kark (2004) found moderate test-retest values and correlation with verbal scales for organizational identification based on images of circles. In other studies, a two-factor structure was found with a scale based on geographical maps (Balram & Dragićević 2005), moderate test-retest values and correlation with observation were noted for a scale based on items with picture response (Ijmker et al. 2008), data distribution was found to be dependent on image response (Shropshire et al. 2009; Xia et al. 2011), a partly image-based scale had moderate correlation with clinical score (Waller et al. 2012) and partial prediction of outcome was found with image response (Leutner et al. 2016).

Image, question/stimulus

High correlations between picture ratings were identified (Graybill & Heyvelman 1993). A scale with pain mapping was found reliable and valid (Escalante 1995). Higher respondent reporting was found with images (Couper et al. 2004), and a scale with an image stimulus had higher interrater and intrarater reliability with continuous than with ordinal responses, although multidimensionality was not tested (Wall et al. 2017).

Image, dynamic / figural response

Higher difficulty, reliability, and omit rates, along with lower discrimination rates, were found with figural response than with multiple choice (Martinez 1991). The difficulty did not confound performance on figural response, and high reliability (Bennett et al. 2000), interactive images versus static images, and VR did not affect response (Lim et al. 2006). Easy usage of dynamic images was found (van Ooijen 2011), and figural response had information levels and factor loadings similar to multiple choice (Wan & Henly 2012).

Image, video

In a study by Kersting (2008), volunteer math teachers completed a video-analysis assessment which was interpreted as a proxy for teaching knowledge. Respondents watched 10 video clips of 1–3 minutes in length and supplied answers in open-ended items, which were then assessed by experts on four dimensions. The test had interrater-reliability correlations of .79 –.85 ($\varkappa$ 0.65 - 0.78) and Cronbach's α .90. It also had good

data fit and was found valid on discriminant validity (.22–.60). Criterion validity showed two of four dimensions were statistically significant with a 32-item short-version of Mathematical Knowledge for Teaching instrument.

Interval item

McKelvie (1978) asked respondents to not only mark a value on a scale but also to mark a "confidence interval." This was not a statistical interval, but a subjectively estimated margin of confidence on a Likert-type scale. McKelvie concluded that a scale of five points was the most preferred by respondents.

Randomization, open-ended response

No studies exploring scales with open-ended options were identified in our research.

Randomization with separate non-randomized category

In a 13-item survey of 436 students that included an 8-point ordinal importance scale with a "can't tell" option, Krebs & Hoffmeyer-Zlotnik (2010) found order bias, although not systematically in one direction. Liu & Keusch (2017) mentioned in reference to nominal and ordinal randomization that "don't know" was a response option in their study, but did not report related results.

Ranking

The same underlying construct was found between ranking and rating, but the rating was a better predictor of attitude (Rankin & Grube 1980). It also exhibited stronger relations and had more non-differentiating responses than ratings (Krosnick & Alwin 1985). Stronger predictive validity was found with rating than with ranking (Maio et al. 1996).

Time-limited answer

Timed IQ tests have shown high validity (Elliott et al. 2001; Bird et al. 2004), the insignificant difference between timed and untimed questionnaires (Caudery 1990), and significantly higher test scores with extra time given (Weaver 1993; Lesaux et al. 2006). Five cognitive tests, one of them timed, were found useful for identifying elderly drivers at increased crash risk (Stutts et al. 1998). Timing changed the factor structure of a reading comprehension test (Huesman 2000). Language (French versus English) confounded test scores when the timing was removed (Mullane & McKelvie 2000), and gender was identified as a confounder of timing (Voyer 2011).

Uploading a file

Uploading a file can be used to attach documentation used as part of an analysis as an alternative to supplying documents separately. No studies testing this IIF characteristic were identified in our research.

Visual analogue scale, self-anchoring

Hofmans & Theuns (2008) tested self-anchoring scales (n=62), and the results suggested parallelism for both VAS with equal weights and the ones with self-anchoring.

Voice assistance/sound

Lam et al. (2009) designed and tested the validity of a self-administered questionnaire for bowel disease assessment administered via touchtone telephone and prerecorded questions. With a randomized controlled crossover trial of telephone versus written test responses among 64 subjects, they found the test valid and reliable. Yost et al. (2009) tested item difficulty with a Talking Touchscreen health literacy measure in which respondents would get multiple choice answers read out loud when touching the screen. They also asked respondents to estimate which of the four icons best indicated that questions could be read out loud, and preference toward the icon of a man talking was shown. Sinadinovic et al. (2011) compared an interactive voice response system to an online version of a questionnaire and found a higher response rate with the online survey, a difference in distribution (reports of drug and alcohol use), and a difference in reliability slightly favoring the online version.

In a randomized controlled trial, Turvey et al. (2012) tested a 9-item Patient Health Questionnaire to assess depressive symptoms using interactive voice response compared to the paper-and-pencil method and found differences in mean scores and reliability and higher sensibility with paper-and-pencil. In a cross-sectional external validation study of a food intake recall questionnaire tested against plasma concentrations of β-carotene, α-carotene, β-cryptoxanthin, lycopene, lutein, and zeaxanthin, Medin et al. (2016) used a voice-assisted cartoon character to help participants complete the survey. Medin et al. found acceptable correlations on all six measures.

3. Discussion

The purpose of the scoping review was to map IIF typification and identify types of evidence. For some IIFs, namely finishing a sentence, randomization with open-ended response, ordinal randomization, single-item time-limitation, interval items, uploading files, and images as question/stimulus, little research was identified. More primary research of these IIFs is needed instead of systematic reviews. These IIFs are being offered in survey software and they are therefore used extensively in questionnaires. This makes it necessary to understand the impact of these IIFs on data quality. As comparative evidence between other IIFs proves, acceptable data quality cannot be assumed. Based on the identified evidence systematic review is more feasible with the other IIFs (battery, constant sum, distance to nonsubstantive response, drag-and-drop

sorting, draw function, image response, nominal randomization, ranking, uploading, VAS, and voice assistance).

More research is needed about how the use of IIFs collectively impacts data quality and why. Research that tests questionnaires containing traditional item formats versus multiple IIFs by impact on construct validity traits, non-response, and distributions of data impact is suggested.

Including survey software as a source of IIF typology had an impact regarding graphical scales and ranking, and full journal searches of Public Opinion Quarterly and Social Science Computer Review 2019 - 2021 had an impact on identified research regarding ranking.

TYPIFICATION AND SEMANTICS

The snowball/chain sampling method, using the references from each paper to identify other relevant research, enhanced the search results. This also meant that less evidence may have been identified for IIFs when there was little evidence identified during the initial iterations. For instance, finishing a sentence is a type of IIF often used in tests and the frequent use of this IIF suggests that research on data quality could exist. The lack of identified research may be because relevant search terms were not identified. Therefore, further exploration of typologies may be relevant.

Some identified IIFs seem to have a single name used by most researchers, such as "randomization", which allows for easy identification of relevant research. Other identified IIFs have several names. For instance, item format labels such as "graphic scale", "slider", "visual analogue scale", and "semantic differential scale" are sometimes referred to as being the same IIFs, other times not. Sometimes labels change over time because going from paper to electronic format changes what the formats are called. In some of the identified research the traits of the IIFs are not described, such as not specifying whether a graphic scale has endpoints, is discrete or continuous, has more labels than the endpoints, has smileys instead of text e.g. For all of the identified IIFs it will be necessary to deal with such unclarities in future systematic reviews. Typification used here can provide improved search terms for reviews, especially when several labels are used for a type of IIF.

The term "innovative" in IIF suggests these items are more innovative than "classical" formats, although this may be questionable. The original graphic scale (Hayes & Patterson 1921) later referred to as a visual analogue scale, was designed before the ordinal Likert scale, and both have since been used extensively in research. It is difficult to say why one is seen as more innovative than the other. Still, as some research suggests (Crabtree 2016), IIF may have a positive impact on data quality when these formats are used extensively instead of classical formats. If there is higher data quality with IIFs this could be because respondents are more motivated because of the

variation in questionnaire design, although other factors may also be relevant. On top of potentially higher data quality, the extended use of IIFs may broaden researcher perspectives on measured phenomena. If IIFs allow for higher data quality, it may be considered relevant to act innovatively in the questionnaire design process.

FRAMEWORKS OF FUTURE SYSTEMATIC REVIEWS

Some IIFs are used more often in questionnaires meant for tests with right and wrong answers than in questionnaires meant for polls and vice versa. The evidence covered in this paper does not explore whether a specific type of IIF has a different impact in tests than in polls. Usage is often a matter of tradition more than a matter of purpose. For instance, Likert scales are seen more in polls whereas VAS items are seen more in clinical tests although it might as well be the other way around. VAS items have a theoretical advantage in equidistance between response options making them more suitable for analysis as interval data, whereas Likert scales are considered ordinal. Finishing a sentence is used more in tests than in polls, although it doesn't mean that it can't be used in polls. Most formats can be used in both types of surveys which does not suggest differentiation between IIFs for tests and IIFs for polls. Since polls often don't contain several types of IIFs, more research on data quality impact when more IIFs are used, is suggested.

IDENTIFIED METHODS

The mapped methods used to explore data quality in existing research can be considered when decisions for inclusion criteria are made. The identified research includes descriptive and inferential methods such as correlations, factor models, regression models, randomized controlled studies, and more. Studies based on other construct validity methods than factor analysis, such as structural equation models, Rasch-models, and multitrait multimethod were not identified. The studies based on testing IIF psychometric traits are often focused around basic aspects of validity and reliability. Therefore, research about how IIFs affect constructs is sparse.

The focus on mapping means less room for the pros and cons of each study design. The identified research is heterogenous and suggests a focus on methods with reasonable comparability in future systematic reviews. This could include tests of significant differences in response rate or satisfaction in randomized controlled studies. Although studies of construct validity with factor analysis and factorial studies of the impact on data quality may also reveal aspects of data quality with high significance, these studies may be too few and be too incomparable to include in systematic reviews.

OUTCOMES

The scoping review provides suggestions for the choice of outcomes in systematic reviews. Finding different distributions or different constructs such as factor structures

in different types of IIFs doesn't necessarily mean that one IIF is better than the other, except if a distribution or construct is expected. Even if two factor analysis studies are comparable in terms of topic and sample, comparison is still difficult since psychometric tests often consist of several test aspects. Correlations between types of IIFs don't say if one is better than the other. High correlations are interesting only if one IIF already has known high-validity evidence. Therefore, it is difficult to conduct a systematic review based on psychometric traits.

Outcomes such as respondent's preference, happiness, enjoyment, usage, or sense of difficulty can be difficult to compare. Although high respondent satisfaction has suggested high validity in responses due to previous research, this outcome measure is less comparable between studies due to different phrasings and satisfaction could be influenced by more than the IIFs.

Response time is not recommended as an outcome since interpretation is ambiguous. Long response time could suggest high strain on respondents leading to less focus and lower data quality. On the other hand, short response time could also suggest respondents skip through items without reading the questions properly.

The item-level and questionnaire-level response rates are outcomes that are comparable across item formats. Interpretation of response-rate is nonambiguous in the sense that researchers would generally want to maintain high response-rate for representativeness and internal validity. Therefore, response rate would be a relevant outcome in future systematic reviews.

4. Conclusion

The study identified a total of 62 research articles with data from 89,365 participants, revealing aspects of 22 IIFs with 13 further subcategories.

The results of this analysis suggest that more research is necessary, especially on types of IIF for which evidence is scarce, which is finishing a sentence, randomization with open-ended response, ordinal randomization, single-item time-limitation, interval items, uploading files, and images as question/stimulus. Systematic review is more feasible with the following IIFs: battery, constant sum, distance to nonsubstantive response, drag-and-drop sorting, draw function, image response, nominal randomization, ranking, uploading, VAS, and voice assistance. For other IIFs primary research was identified and for some IIFs research has been synthesized (Knäuper 1999; Shulman, 2000; Voyer 2011; Chyung et al. 2018; Chyung et al. 2018 II; Chiarotto et al. 2019).

Since additional results regarding ranking was found based on full searches from journals in the years 2019 – 2021, this suggests more relevant evidence may exist.

No systematic reviews were identified which suggests systematic reviews could be considered for the IIFs where research was identified.

References

Alwin, Duane F. and Jon A. Krosnick. 1985. "The measurement of values in surveys: A comparison of ratings and rankings". *Public Opinion Quarterly* 49: 535-552. doi: 10.1086/268949

American Educational Research Association. 2014. *Standards for Educational and Psychological Testing*. Washington DC, Washington: American Educational Research Association.

Armstrong, R., Hall, B. J., Doyle, J., & Waters, E. (2011). "'Scoping the scope' of a Cochrane review". *Journal of Public Health*, 33(1):147-150. https://doi.org/10.1093/pubmed/fdr015

Balram, Shivanand and Suzana Dragićević. 2005. "Attitudes toward urban green spaces: integrating questionnaire survey and collaborative GIS techniques to improve attitude measurements". *Landscape and Urban Planning* 71(2-4):147-162. doi: 10.1016/j.landurbplan.2004.02.007

Bell, Douglas S., Carol M. Mangione, and Charles E. Kahn. 2001. "Randomized testing of alternative survey formats using anonymous volunteers on the world wide web". *Journal of the American Medical Informatics Association* 8:616-620. doi: 10.1136/jamia.2001.0080616

Bennett, Randy E., William C. Ward, Donald A. Rock, and Colleen LaHart. 1990. "Toward a framework for constructed response items." *ETS Research Report Series* 1990:1-66. doi: 10.1002/j.2333-8504.1990.tb01348.x

Bennett, Randy E. and Marc M. Sebrechts. 1997. "A Computer-Based Task for Measuring the Representational Component of Quantitative Proficiency." *Journal of Educational Measurement* 34(1):64-77. doi: 10.1111/j.1745-3984.1997.tb00507.x

Bennett, Randy E., Mary Morley, Dennis Quardt, Donald A. Rock, Mark K. Singley, Irvin R. Katz, and Adisack Nhouyvanisvong. 1999. "Psychometric and cognitive functioning of an under-determined computer-based response type for quantitative reasoning." *Journal of Educational Measurement* 36:233-252. doi: 10.1111/j.1745-3984.1999.tb00556.x

Bennett, Randy E., Mary Morley, Dennis Quardt, and Donald A. Rock. 2000. "Graphical modeling: A new response type for measuring the qualitative component of mathematical reasoning." *Applied Measurement in Education* 13:303-322.

Berg, Irwin A. and Gerald M. Rapaport. 1954. "Response bias in an unstructured questionnaire". *The Journal of Psychology* 38:475-481. doi: 10.1080/00223980.1954.9712954

Bird, Chris M., Kyriaki Papadopoulou, POL Ricciardelli, Martin N. Rossor, and Lisa Cipolotti. 2004. "Monitoring cognitive changes. Psychometric properties of six cognitive tests." British Journal of Clinical Psychology 43:197-210. doi: 10.1348/014466504323088051

Björkstén, M. G., B. Boquist, M. Talbäck, and C. Edling. 1999. "The validity of reported musculoskeletal problems. A study of questionnaire answers in relation to diagnosed disorders and perception of pain." *Applied Ergonomics* 30(4):325-330. doi: 10.1016/s0003-6870(98)00033-7

Borkenhagen, Ada, Burghard F. Klapp, Frank Schoeneich, and Elmar Brähler. 2005. "Differences in body image between anorexics and in-vitro-fertilization patients - a study with Body Grid." *Psychosoc Med* 2:1-9.

Bosch, Oriol J and Melanie Revilla. 2020. "Using emojis in mobile web surveys for Millenials? A study in Spain and Mexico." *Quality and Quantity 55:39-61.*

Broekens, Joost and Willem-Paul Brinkman.(2013. "AffectButton: A method for reliable and valid affective self-report." *Human-Computer Studies 71*(6):641-667.

Buchanan, H. & N. Niven. 2002. "Validation of a facial image scale to assess child dental anxiety." *International Journal of Paediatric Dentistry* 12:47-52.

Buchanan, Heather and N. Niven. 2003. "Further evidence for the validity of the Facial Image Scale." *International Journal of Paediatric Dentistry* 13:368-369. doi: 10.1046/j.1365-263X.2003.00488.x

Buchanan, H. 2005. "Development of a computerised dental anxiety scale for children: validation and reliability." *British Dental Journal* 199:359-362. doi: 10.1038/sj.bdj.4812694

Callegaro, Mario, Jeffrey Shand-Lubbers, and J. Michael Dennis. 2009. *Presentation of a Single Item versus a Grid: Effects on the Vitality and Mental Health Scales of the SF36v2 Health Survey.* Retrieved June 1, 2022 (https://www.researchgate.net/publication/253454379_Presentation_of_a_Single_Item_versus_a_Grid_Effects_on_the_Vitality_and_Mental_Health_Scales_of_the_SF36v2_Health_Survey)

Castle, Nicholas G. and John Engberg. 2004. "Response Formats and Satisfaction Surveys for Elders." *The Gerontologist 44*(3):358-367. DOI: 10.1093/geront/44.3.358

Caudery, Tim. 1990. "The Validity of timed essay tests in the assessment of writing skills." *ELT Journal* 44:122-131.

Chesnut, John. 2008. "Effects of using a grid versus a sequential form of the ACS basic demographic data." Retrieved June 1, 2022 (https://www.census.gov/content/dam/Census/library/working-papers/2008/acs/2008_Chesnut_01.pdf)

Chiarotto, Alessandro, Lara J. Maxwell, Raymond W. Ostelo, Maarten Boers, Peter Tugwell, and Caroline B. Terwee. 2019. "Measurement Properties of Visual Analogue Scale, Numeric Rating Scale, and Pain Severity Subscale of the Brief Pain Inventory in Patients With Low Back Pain: A Systematic Review." *The Journal of Pain* 20(3):245-263. doi: 10.1016/j.jpain.2018.07.009

Chyung, Seung Y., Megan Kennedy, and Ingrid Campbell. 2018. "Evidence-based survey design: The use of ascending or descending order of response options". *Performance Improvement Journal* 57(9):9-16. doi: 10.1002/pfi.21800

Chyung, Seung Y., Ieva Swanson, Katherine Roberts, and Andrea Hankinson. 2018 II. "Evidence-based survey design: The use of continuous rating scales in surveys." *Performance Improvement Journal* 57(5):38-48. doi: 10.1002/pfi.21763

Conrad, Frederick G., Mick P. Couper, Roger Tourangeau, and Mirta Galesic. 2005. "Interactive feedback can improve the quality of responses in Web surveys". Retrieved June 1, 2022 (https://www.researchgate.net/publication/228689449_Interactive_feedback_can_improve_quality_of_responses_in_web_surveys).

Couper, Mick P., Michael W. Traugott, and Mark J. Lamias. 2001. "Web survey design and administration." *Public Opinion Quarterly*, 65, pp. 230-253. DOI: 10.1086/322199

Couper, Mick P., Frederick G. Conrad and Roger Tourangeau. 2002. "Visual Context Effects in Web Surveys." *Online Social Sciences*. Batinic, B., U. D. Reips, and M. Bosnjak, Seattle: Hogrefe & Huber.

Couper, Mick P, Roger Tourangeau, and Kristin Kenyon. 2004. "Picture This! Exploring Visual Effects in Web Surveys." *The Public Opinion Quarterly 68*(2):255-266. doi: 10.1093/poq/nfh013

Couper, Mick P., Roger Tourangeau, Frederick G. Conrad, and Chan Zhang. 2012. "The design of grids in web surveys." *Social Science Computer Review* 31(3):322-345. doi: 10.1177/0894439312469865

Crabtree, Ashleigh R. 2016. "Psychometric properties of technology-enhanced item formats: an evaluation of construct validity and technical characteristics." Retrieved June 1, 2022 (https://iro.uiowa.edu/esploro/outputs/doctoral/

Psychometric-properties-of-technology-enhanced-item-formats/9983777220902771) doi: 10.17077/etd.922fbj4d

Davies, Julie and Ivy Brember. 2006. "The Reliability and Validity of the 'Smiley' Scale". *British Educational Research Journal* 20(4):447-454. DOI: 10.1080/0141192940200406

Derham, Philip A. J. 2011. "Using preferred, understood or effective scales? How scale presentations affect online survey data collection." *Australasian Journal of Market & Social Research* 19(2):13-26.

Desmet, Pieter M. A. 2005. "Measuring emotions: development and application of an instrument to measure emotional responses to products." Pp. 111-124 in *Funology: from Usability to Enjoyment*, edited by M.A. Blythe and A. Monk. Switzerland: Springer, Cham.

Dolan, Robert P., Joshua Goodman, Ellen Strain-Seymour, Jeremy Adams, and Sheela Sethuraman. 2011. *"Cognitive lab evaluation of innovative items in mathematics and English/language arts assessment of elementary, middle, and high school students: Research Report"*. Retrieved June 1, 2022 (http://www.pearsonassessments.com/hai/images/tmrs/Cognitive_Lab_Evaluation_of_Innovative_Items.pdf).

Downing, Steven M., and Thomas M. Haladyna. 2006. *Handbook of Test Development*. Mahwah, NJ: L. Erlbaum.

Elliott, Jacquelyn, Steven W. Lee, and Nona Tollefson. 2001. "A reliability and validity study of the dynamic indicators of basic early literacy skills—modified". *School Psychology Review* 30:33. doi: 10.1080/02796015.2001.12086099

Elliot, Statia and Nicolas Papadopoulos. 2012. "*Beyond Tourism Destination Image: Mapping country image from a psychological perspective"* Retrieved November 12, 2021 (https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1722&context=ttra)

Emde, Matthias and Marek Fuchs. 2012. "Exploring Animated Faces Scales in Web Surveys: Drawbacks and Prospects." *Survey Practice 5*(1):1-6. DOI: 10.29115/SP-2012-0006

Escalante, A., M. J. Lichtenstein, K. White, N. Rios, and H. P. Hazuda. 1995. "A method for scoring the pain map of the McGill pain questionnaire for use in epidemiologic studies." *Aging Clinical and Experimental Research* 7:358-366. doi: 10.1007/BF03324346

Fisher, Ronald (1926). "The Arrangement of Field Experiments." *Journal of the Ministry of Agriculture of Great Britain* 33:503-513. doi: 10.23637/rothamsted.8v61q

Freyd, M. 1923. "The graphic rating scale." *Journal of Educational Psychology* 14:83-102. doi: 10.1037/h0074329

Funke, Frederik and Ulf-Dietrich Reips. 2006. "Visual Analogue Scales in Online Surveys: Non-Linear Data Categorization by Transformation with Reduced Extremes." Retrieved June 1, 2022 (http://www.frederikfunke.de/papers/Funke%20&%20Reips%20-%20VAS%20-%20GOR06.pdf)

Galesic, M., R. Tourangeau, M. P. Couper, and F. G. Conrad. 2007. "*Using change to improve navigation in grid questions*." Leipzig: General Online Research Conference (GOR'07).

Graybill, Daniel and Lorene R. Heuvelman. 1993. "Validity of the Children's Picture-Frustration Study: A Social-Cognitive Perspective." *Journal of Personality Assessment* 60(2):379-389. doi: 10.1207/s15327752jpa6002_13

Gummer, T., Vogel, V., Kunz, T., & Roßmann, J. (2020). "Let's put a smile on that scale: Findings from three web survey experiments". *International Journal of Market Research*, 62(1):18–26.

Guyatt G. H., D. L. Sackett, J. C. Sinclair, R. Hayward, D. J. Cook, R. J. Cook. 1995. "Users' guides to the medical literature. IX. A method for grading health care recommendations. Evidence-Based Medicine Working Group". *JAMA* 274:1800–1804. doi:10.1001/jama.1995.03530220066035. PMID 7500513

Haladyna, Thomas M., Steven M. Downing, and Michael C. Rodriguez. 2010. "A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment." *Applied Measurement in Education* 15:309-333. doi: 10.1207/S15324818AME1503_5

Hayes, M. H. S. and Patterson, D. G. 1921. "Experimental development of graphic rating method". *Psychological Bulletin* 18:98-99.

Hofmans, Joeri and Peter Theuns. 2008. "On the linearity of predefined and self-anchoring visual analogue scales." *British Journal of Mathematical and Statistical Psychology* 61:401-413. doi: 10.1348/000711007X206817

Holbrook, Allyson L., Jon A. Krosnick, David Moore, and Roger Tourangeau. 2005. "Response Order Effects In Dichotomous Categorical Questions Presented Orally: The Impact of Question and Respondent Attributes." *Public Opinion Quarterly* 71(3):325-348. doi: 10.1093/poq/nfm024

Holbrook, Allyson, Jon Krosnick, and Roger Tourangeau. 2007. "Response order effects in dichotomous categorical questions presented orally." *Public Opinion Quarterly* 71(3):325-348. doi: 10.1093/poq/nfm024

Huesman, Ronald L. 2000. "The Validity of ITBS Reading Comprehension Test Scores for Learning Disabled and Non Learning-Disabled Students under Extended-Time Conditions". Annual Meeting of the American Educational Research Association.

Hu, Jingwei. (2019). "Horizontal or Vertical? The Effects of Visual Orientation of Categorical Response Options on Survey Responses in Web Surveys". *Social Science Computer Review*, 779-792. doi: 10.1177/0894439319834296

Iglesias, C. P., Birks, Y. F. & Torgerson, D. J. 2001. "Improving the measurement of quality of life in older people: the York SF-12." *QJM*, 94:695-698. doi: 10.1093/qjmed/94.12.695

Ijmker, S., J. Mikkers, B. M. Blatter, A. J. van der Beek, W. van Mechelen, and P. M. Bongers. 2008. "Test–retest reliability and concurrent validity of a web-based questionnaire measuring workstation and individual correlates of work postures during computer work." *Applied Ergonomics* 39(6):685-696. doi: 10.1016/j.apergo.2007.12.003

Jelínek, Martin, Petr Květon, and Dalibor Vobořil. 2015. "Innovative testing of spatial ability: interactive responding and the use of complex stimuli material." *Cognitive Processing* 16(1):45-55.

Kaczmirek, Lars. 2008. "Human-Survey Interaction Usability and Nonresponse in Online Surveys". Retrieved June 1, 2022 (https://d-nb.info/992375924/34)

Kaczmirek, L. (2011). "Attention and usability in internet surveys: Effects of visual feedback in grid questions". Pp. 191-214 in *Social and Behavioral Research and the Internet,* edited by M. Das, P. Ester and L. Kaczmirek. Routledge.

Kersting, Nicole. 2008. "Using Video Clips of Mathematics Classroom Instruction as Item Prompts to Measure Teachers' Knowledge of Teaching Mathematics." *Educational and Psychological Measurement* 68(5):845-861. doi: 10.1177/0013164407313369

Knäuper, Bärbel. 1999. "The Impact of Age and Education on Response Order Effects in Attitude Measurement." *The Public Opinion Quarterly* 63(3):347-370.

Krebs, Dagmar and Juergen H. P. Hoffmeyer-Zlotnik. 2010. "Positive First or Negative First? Effects of the Order of Answering Categories on Response Behavior." *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 6(3):118-127. doi: 10.1027/1614-2241/a000013

Krosnick, Jon A. and Duane F. Alwin. 1988. "A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values." *Public Opinion Quarterly* 52:526-538. doi: 10.1086/269128

Krosnick, Jon A. 1991. "Response strategies for coping with the cognitive demands of attitude measures in surveys." *Applied Cognitive Psychology* 5:213-236. doi: 10.1002/acp.2350050305

Kunz, T. 2015. "*Rating scales in web surveys: A test of new drag-and-drop rating procedures.* Dissertation". Retrieved October 16, 2021 (http://tuprints.ulb.tu-darmstadt.de/5151/7/Kunz_2015_Rating_scales_in_web_surveys.pdf).

Lam, Manuel Y., Hang Lee, Renee Bright, Joshua R. Korzenik, and Bruce, E. Sands. 2009. "Validation of Interactive Voice Response System Administration of the Short Inflammatory Bowel Disease Questionnaire". *Inflammatory Bowel Diseases* 15(4):599-607. doi: 10.1002/ibd.20803

Lesaux, Nonie K., M. Rufina Pearson, and Linda S. Siegel. 2006. "The Effects of Timed and Untimed Testing Conditions on the Reading Comprehension Performance of Adults with Reading Disabilities." *Reading and Writing* 19:21-48.

Leutner, Franziska, Adam Yearsley, Sonia-Cristina Codreanu, Yossi Borenstein, Gorkan Ahmetoglu. 2016. "From Likert scales to images: Validating a novel creativity measure with image based response scales." *Personality and Individual Differences* 106:37-40. doi: 10.1016/j.paid.2016.10.007

Lim, En-Mi, Tsuyoshi, Honjo, Kiyoshi Umeki. 2006. "The validity of VRML images as a stimulus for landscape assessment." *Landscape and Urban Planning* 77(1-2):80-93.

Liu, Mingman and Florian Keusch. 2017. "Effects of Scale Direction on Response Style of Ordinal Rating Scales." *Journal of Official Statistics* 33(1):137-154. doi: 10.1515/jos-2017-0008

Louviere, Jordan J. and Towhidul Islam. 2006. "A comparison of importance weights and willingness-to-pay measures derived from choice-based conjoint, constant sum scales and best–worst scaling." *Journal of Business Research* 61:903-911. doi: 10.1016/j.jbusres.2006.11.010

Lu, Ying and Stephen G. Sireci. 2007. "Validity Issues in Test Speededness". *Educational Measurement* 26:29-37. doi: 10.1111/j.1745-3992.2007.00106.x

Maio, Gregory R., Neal J. Roese, Clive Seligman, and Albert Katz. 1996. "Rankings, Ratings, and the Measurement of Values: Evidence for the Superior Validity of Ratings." *Journal of Basic and Applied Social Psychology* 18(2):171-181. doi: 10.1207/s15324834basp1802_4

Martinez, Michael E. 1991. "A comparison of multiple-choice and constructed figural response items." *Journal of Educational Measurement* 28:131-145. doi: 10.1111/j.1745-3984.1991.tb00349.x

McKelvie, Stuart J. 1978. "Graphic rating scales - how many categories." *British Journal of Psychology* 69:185-202. doi: 10.1111/j.2044-8295.1978.tb01647.x

McReynolds, Paul and Klaus Ludwig. 1987. "On the history of rating scales". *Personality and Individual Differences* 8:281-283. doi: 10.1016/0191-8869(87)90188-7

Medin, Anine C., Monica H. Carlsen, and Lene F. Andersen. 2016. "Associations between reported intakes of carotenoid-rich foods and concentrations of carotenoids in plasma: A validation study of a web-based food recall for children and adolescents." *Public Health Nutrition.* 19:3265-3275. doi: 10.1017/S1368980016001622

Mills, C.Wright (1961). *The sociological imagination*. New York, NY: Grove Press.

Mullane, Jennifer and Stuart J. McKelvie. 2000. "Effects of Removing the Time Limit on First and Second Language Intelligence Test Performance." *Practical Assessment, Research, and Evaluation* 7(23):1-6. doi: 10.7275/ph8y-yz89

Munn Zachary, Micah D. J. Peters, Cindy Stern, Catalin Tufanaru, Alexa McArthur, and Edoardo Aromataris. 2018. "Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach." *BMC Medical Research Methodology* 18:1-7. doi: 10.1186/s12874-018-0611-x

Murad, M. H., Asi, N., Alsawas, M., & Alahdab, F. (2016). New evidence pyramid. Evid Based Med. 21(4):125-127. doi: 10.1136/ebmed-2016-110401

van Ooijen, P.M.A., A. Broekema, and M. Oudkerk. 2011. "Design and implementation of I2Vote – An interactive image-based voting system using windows mobile devices." *International Journal of Medical Informatics* 80(8):562-569. doi: 10.1016/j.ijmedinf.2011.05.002

Rankin, William L. and Joel W. Grube. 1980. "A comparison of ranking and rating procedures for value system measurement." *European Journal of Social Psychology* 10(3):233-246. doi: 10.1002/ejsp.2420100303

Reynolds-Keefer, L., and Johnson, R. (2011). "Is a Picture Worth a Thousand Words? Creating Effective Questionnaires with Pictures". *Practical Assessment, Research & Evaluation,* 16(8):1-7.

Rofé, Yodan. 2004. "Mapping the sense of well being in a neighborhood: survey technique, and analysis of agreement and variation". Retrieved April 26, 2021

(https://www.researchgate.net/publication/316213103_Mapping_the_sense_of_well_being_in_a_neighborhood_survey_technique_and_analysis_of_agreement_and_variation)

Scalise, Kathleen and Bernard Gifford. 2006. "Computer-Based Assessment in E-Learning: A Framework for Constructing 'Intermediate Constraint' Questions and Tasks for Technology Platforms." *The Journal of Technology, Learning and Assessment* 4(6).

Scherpenzeel, Annette and Willem Saris. 1997. "The Validity and Reliability of Survey Questions - A Meta-Analysis of MTMM Studies". *Sociological Methods & Research* 25:341-383. doi: 10.1177/0049124197025003004

Schubert, Thomas and Sabine Otten. 2002. "Overlap of Self, Ingroup, and Outgroup: Pictorial Measures of Self-Categorization." *Self and Identity* 1:353-376. doi: 1529-8868/2002

Schwarz, Norbert, Carla E. Grayson, and Barbel Knäuper. (1998). "Formal Features of Rating Scales and the Interpretation of Question Meaning". *International Journal of Public Opinion Research,* 10:177–183.

Shamir, Boas and Ronit Kark. 2004. "A single-item graphic scale for the measurement of organizational identification." *Journal of Occupational and Organizational Psychology* 77(1):115-123. doi: 10.1348/096317904322915946

Shulman, K. I. 2000. "Clock-drawing: is it the ideal cognitive screening test?" *International Journal of Geriatric Psychiatry* 15:548-561. doi: 10.1002/1099-1166(200006)15:6<548::aid-gps242>3.0.co;2-u

Sikkel, Dirk, Reinder Steenbergen, and Stoerd, Gras. 2014. "Clicking vs. dragging: Different uses of the mouse and their implications for online surveys." *Public Opinion Quarterly* 78:177-190. doi: 10.1093/poq/nft077

Sinadinovic Kristina, Peter Wennberg, and Anne H. Berman. 2011. "Population screening of risky alcohol and drug use via Internet and Interactive Voice Response (IVR): a feasibility and psychometric study in a random sample." *Drug and Alcohol Dependence* 114:55-60. doi: 10.1016/j.drugalcdep.2010.09.004

Sireci, Stephen G. & Zenisky, April L. (2006). "Innovative Item Formats." Pp. 329-348 in *Handbook of Test Development,* edited by Steven M. Downing and Thomas M. Haladyna. London, UK: Lawrence Erlbaums Associates. Retrieved October 1 2021 (https://fatihegitim.files.wordpress.com/2014/03/hndb-t-devt.pdf)

Skedgel, Chris D., Allan J. Wailoo, Ron L. Akehurst. 2013. "Choosing vs. allocating: discrete choice experiments and constant-sum paired comparisons for the

elicitation of societal preferences." *Health Expectations* 18:1-14, doi: 10.1111/hex.12098

Stutts, Jane C., J. Richard Stewart, and Carol Martell. 1998. "Cognitive test performance and crash risk in an older driver population." *Accident Analysis & Prevention* 30(3):337-346. doi: 10.1016/s0001-4575(97)00108-5

Svalastoga, Kaare. 1959. *Prestige, class and mobility*. New York, NY: Arno.

Sørensen, J.L., L. Thellesen, J. Strandbygaard, K. D. Svendsen, K. B. Christensen, M. Johansen, P. Langhoff-Roos, K. Ekelund, B. Ottesen, and C. van der Vleuten. 2014. "Development of knowledge tests for multi-disciplinary emergency training: a review and an example." *Acta Anaesthesiologica Scandinavica* 59:123-33. doi: 10.1111/aas.12428

Thorndike, Frances P., Per Carlbring, Frederick L. Smyth, Joshua C. Magee, Linda Gonder-Frederick, Lars-Göran Ost, and Lee M. Ritterband. 2009. "Web-based measurement: Effect of completing single or multiple items per webpage". *Computers in Human Behavior* 25:393-401.

Timbrook, Jerry P. 2013. "A Comparison of a Traditional Ranking Format to a Drag-and-Drop Format with Stacking." Retrieved June 1, 2022 (http:// rave. ohiolink.edu/etdc/view?acc_num=dayton1367241685)

Timbrook, Jerry and William F. Moroney. 2016. "Ranking: Perceptions of Tied Ranks and Equal Intervals on a Modified Visual Analog Scale". Retrieved June 1, 2022 (https://ecommons.udayton.edu/psy_fac_pub/26)

Toepoel, V., Vermeeren, B., & Metin, B. (2019). "Smileys, stars, hearts, buttons, tiles or grids: Influence of response format on substantive response, questionnaire experience and response time". *Bulletin of Sociological Methodology*, 142:57–74.

Tourangeau, R., M. P. Couper, and F. Conrad. 2004. "Spacing, Positioning, and Order. Interpretative Heuristics for Visual Features of Survey Questions." *Public Opinion*

Quarterly 68:368-393. doi: 10.1093/poq/nfh035

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K., Colquhoun, H., Kastner, M., Levac, D., Ng, C., Sharpe, J. P., Wilson, K., Kenny, M., Warren, R., Wilson, C., Stelfox, H. T., & Straus, S. E. (2016). A scoping review on the conduct and reporting of scoping reviews. *BMC Medical Research Methodology,* 16(15). https://doi.org/10.1186/s12874-016-0116-4

Turvey, Carolyn, Tom Sheeran, Lilian Dindo, Bonnie Wakefield, and Dawn Klein. 2012. "Validity of the Patient Health Questionnaire, PHQ-9, administered

through interactive-voice-response technology". *Journal of Telemedicine and Telecare* 18:348-351. doi: 10.1258/jtt.2012.120220

University Libraries Health Science Library, "Scoping Reviews: Step 3: Conduct Literature Searches", accessed Nov 28th, 2023: https://guides.lib.unc.edu/scoping-reviews/search#s-lg-box-29819393

Voyer, Daniel. 2011. "Time limits and gender differences on paper-and-pencil tests of mental rotation: a meta-analysis." *Psychonomic Bulletin & Review* 18:267-277. doi: 10.3758/s13423-010-0042-0

Wall, Eric J., Matthew D. Milewski, James L. Carey, Kevin G. Shea, Theodore J. Ganley, John D. Polousky, Nathan L. Grimm, Emily A. Eismann, Jake C. Jacobs, Lucas Murnaghan, Carl W. Nissen, Gregory D. Myer. 2017. "The reliability of assessing radiographic healing of osteochondritis dissecans of the knee." *The American Journal of Sports Medicine* 45(6):1370-1375. doi: 10.1177/0363546517698933

Waller, Rosemary, Peter Manuel, and Lyn Williamson. 2012. "The Swindon Foot and Ankle Questionnaire: Is a Picture Worth a Thousand Words?" *International Scholarly Research Network* 1:1-8. doi: 10.5402/2012/105479

Wan, Lei and George A. Henly. 2012. "Measurement Properties of Two Innovative Item Formats in a Computer-Based Test" *Applied Measurement in Education* 25(1):58-78.

Weaver, Susan M. 1993. "The Validity of the use of extended and untimed testing for postsecondary students with learning disabilities." *Dissertation Abstracts International* 55.

Wewers, M. E. and N. K. Lowe. 1990. "A critical review of visual analogue scales in the measurement of clinical phenomena." *Research in Nursing & Health* 13:227-236. doi: 10.1002/nur.4770130405

Wong, John K. and R. Kenneth Teas. 2001. "A test of the stability of retail store image mapping based on multientity scaling data." *Journal of Retailing and Consumer Services* 8(2):61-70.

Xia, Wei, Caihong Sun, Li Zhang, Xin Zhang, Jiajia Wang, Hui Wang, and Lijie Wu. 2011. "Reproducibility and Relative Validity of a Food Frequency Questionnaire Developed for Female Adolescents in Suihua, North China". *PLOS ONE* 6(5). doi: 10.1371/journal.pone.0019656

Yan, Ting and Florian Keusch. 2015. "The Effects of the Direction of Rating Scales on Survey Responses in a Telephone Survey". *Public Opin Quarterly* 79(1):145-165. doi: 10.1093/poq/nfu062

Yost, Kathleen J., Kimberly Webster, David W. Baker, Seung W. Choi, Rita K. Bode, Elizabeth A. Hahn. 2009. "Bilingual health literacy assessment using the Talking Touchscreen/la Pantella Parlanchina: development and pilot testing." *Patient Education and Counseling* 75:295-301. doi: 10.1016/j.pec.2009.02.020

Zenisky, April L. and Stephen G. Sireci. 2002. "Technological Innovations in Large-Scale Assessment". *Applied Measurement in Education* 15(4):337-362. doi: 10.1207/S15324818AME1504_02

Zheng, Ying. 2011. "Research Note: Establishing Construct and Concurrent Validity of Pearson Test of English Academic". Retrieved June 1, 2022 (https:// pdfs.semanticscholar.org/04e0/27c798140fb269bd43338bd972b2a0b3cabe.pdf).