Are Smileys Valid Answers? Survey Data Quality with Innovative Item Formats

Jacob Brauner

Capital Region Denmark

Author Note

This paper was published as a working paper in preparation for a Ph.D. research proposal. The author can be contacted at jjoe0255@regionh.dk.

Abstract

Introduction: Survey research is often designed based on multiple-choice questions although many other formats, also referred to as innovative item formats (IIF) exist, such as ranking, sorting, questions with pictures or smileys as response options. Research has suggested that IIF in a broad sense can strengthen data quality, but research is needed on a more specific level. Therefore, the purpose of this article is to present research for separate IIFs about the data quality for that type of item. Method: A literature study was conducted to identify articles that test the data quality of IIF. For each IIF research was discussed regarding aspects of data quality, such as reliability, validity and response rate. Results: A total of 166 research articles were identified with data from 218,532 participants revealing aspects of 22 IIFs with 13 subcategories. The type of evidence on data quality is quite varied and for some IIFs the evidence is supportive, for some it is not and for some it is inconclusive. With 6 IIFs the evidence was estimated in favour hereof, for 11 IIFs the evidence was inconclusive, 1 unfavoured and for 3 there was no evidence. With 6 IIFs potential confounders were identified. Discussion: The study suggests further research is needed where evidence is scarce. The present study could initiate more extensive systematic reviews within specific categories of IIF.

Keywords: innovative item formats, questionnaire, validity, review

Are Smileys Valid Answers? Data Quality with Innovative Item Formats

Mathews (1929) conducted probably one of the first studies on item format influence on data quality. Since then, numerous of such studies have been made. IIFs are being used extensively in questionnaires, especially within the use of tests. The higher use of IIFs over time may in part be due to pressure from politicians to use innovation to get better tests, since some national learning tests have been criticized for not having high enough validity and reliability (Crabtree 2016), but also due to higher popularity of their use among respondents (Wan & Henly 2012). Upon reflections over 100 years of survey research Krosnick said “These are exciting times for survey research.” (1999). One could add that the years are still exciting, since there is a pressure to design better tests and questionnaires. There is an extensive pool of knowledge about how tests based on MCIs perform, and a lot of evidence that tests perform well using MCIs. Yet, the question arises: would some tests and questionnaires perform even better using IIF or combinations of MCI and IIF? This is a key question since it could be argued that the item characteristics is the most basic element of a test (Thorndike 1967). This underline the importance of synthesizing research about IIF data quality.

Item formats are the types of questions or stimulus a researcher can choose from when a questionnaire is designed. Bennett et al (1990) suggested differentiating between seven types of item formats according to purpose: multiple-choice, selection/identification, reordering/rearrangement, substitution/correction, completion, construction, and presentation. The term Innovative Item Formats (IIF) was used by Zenisky & Sireci (2002) about 12 specific item types according to technical characteristics instead of purpose, to suggest everything that goes beyond the traditional multiple-choice, multiple answer, drop-down and standard open-ended items.

The name suggests innovation is an aspect that shapes how item types are evolved. Other researchers have later used the same term (Wan & Henly 2012) and some researchers have used the term Technology-Enhanced Items, (TEI, Scalise & Gifford 2006, Crabtree 2016). Zheng (2011) mentions several innovative item types: highlight, re-order or fill in blanks in text in different variations. Sireci & Zenisky (in Downing & Haladyna 2006) reviewed computerized item types and suggest a taxonomy of 21 item types within 10 main categories: drag-and-drop, hot-spot, reordering or rearrangement, completion, mathematical expressions, construction, formulating hypotheses, essay/short answer, passage-editing, and presentation. According to Sireci & Zenisky, compared to TEIs MCI items are often less expensive to develop than other formats and generally, they are found to be more reliable. Quite often tests have been developed using MCIs and therefore a much larger base of research on validity and reliability exists. Finally, survey software like Qualtrics and QuestionPro have helped identify relevant IIFs. The various typifications illustrate variation in IIF definitions and therefore explanations are given for each format in the present study.

In the present study the term IIF is used. Scalise & Gifford also suggested a taxonomy of item types, which has been useful in the present study, but the taxonomy was also expanded to include more item formats. In this article the term IIF will be used with an understanding like Zenisky & Sireci 2002, namely the understanding that IIF is everything except multiple-choice, multiple answer, drop-down and simple open-ended items found in most survey software. In this article response classifications are not defined by the trait they are to be used for, such as the bivariate Thurstone scale versus a discrete Likert or Juster scale¹. All these scales are, from a technological perspective, variations of multiple-choice items (MCI) that can be designed with

¹ For instance, behaviorally anchored ratings (Campbell et al 1973, Dickinson & Zellinger 1980) have not been explored, since this item type can be designed as a variation of semantic differentials.

multiple-choice format. Instead, the typification in this article is defined by the technical traits such as whether a respondent is faced with a slider or a discrete scale. The used definition also means that future innovation may result in more item formats included in the definition, potentially allowing for a more nuanced views on the social world.

Several experiments, meta studies and reviews have covered the quality of specific design features in questionnaires, such as questionnaire color, logo, time of delivery, length, wording and many other features of designs that are not related to item format (Haladyna et al 2010; Scherpenzeel 1997; Sørensen et al 2014). Also, synthesis studies have been conducted on specific IIF, namely randomization, draw function, time-limited answers and visual analogue scales (Knäuper 1999, Shulman 2000, Voyer 2011, Chyung et al 2018, Chyung et al 2018 II, Chiarotto et al 2019) but there is a lack of meta-knowledge about data quality across the different types of IIF as stressed previously by Wan and Henly (2012).

Research questions:

- Which innovative item formats can be identified?
- What does the evidence say about individual IIF data quality?
- What are the possible confounders on IIFs?

Existing knowledge about nonspecific Innovative Item Formats

A study conducted by Crabtree (2016) included $n=3813$ 7th to 12th grade students who had been tested with different versions of an End-of-Course algebra questionnaire. Each student received a test consisting of 30 multiple choice questions followed by five "technology enhanced" items, which is defined like IIFs. Students were assigned to one of 12 groups, each with their own combination of several TE items, not one type of item format in each group which means the study is focused on data quality with TE items on a general level. Questionnaires in each of the 12 groups were filled out by 288-355 students. This allowed for testing how TE items performed relative to the standard multiple-choice items alone. The following TEs were tested in different combinations in the 12 groups: open ended text input in a sentence, open ended number input, drawing a graph, drag and drop words or sentences into a sentence, point and click (text, images, numerals, expressions e.g.) and matching information. Overall, Crabtree found that TE items could provide more information, and measure the construct more efficiently than MCIs at higher levels of ability.

Jodoin (2003) studied data with an Item Response Theory (IRT) model, using previous data from two Microsoft Windows survey data collections ($2 \times 3,000+$ examinees²), both containing both innovative items and traditional ones. Jodoin refer to a drop-and-connect item type with a pool of possible entities that may be connected and a pool of possible connection links between the entities. The data was tested on IRT information and innovative items provided considerably more information than multiple choice, but also took more time.

Young & Willson (2012) compared two instruments. One was the 160 item E2 meant for nurse student final exam, including several item types such as multiple-choice, multiple-response, fill-in-the-blank, hot spot, chart/exhibit, and the Participant School Survey. Reliability

² Exact number of examinees is not expressed.

was determined based on comparison of Kuder-Richardson formula 20 for every examination returned, for three versions 0.91, 0.98 and 0.91 respectively. Validity was based on a sample of 37,184 students with associate degree, baccalaureate and diploma schools of nursing. The E2 was found to be between 96.36% and 98.30% accurate in predicting exam success.

These results (Crabtree 2016, Jodoin 2003, Young & Wilson 2012) exemplify the empirical rationale for research in the data quality of specific IIFs.

In the article the following 22 main question types with 13 identified subcategorizations are examined, where visual analogue scales could be divided further into subcategories:

- Battery: static vs greyout
- Battery, multientity scaling: drag-and-drop and clicking
- Constant sum: drag-and-drop and clicking
- Distance to nonsubstantive response
- Drag-and-drop sorting
- Draw function
- Finishing a sentence
- Image response, universal: discrete and slider
- Image response, specific
- Image, question / stimulus
- Image, dynamic / figural response: visible vs invisible categorization
- Image, video
- Interval item
- Randomization: nominal and ordinal
- Randomization, open ended response

- Randomization, separate non-random category
- Ranking
- Time limited answer: visible vs invisible time limit
- Uploading a file
- Visual analogue scale: several descriptors versus endpoints, stepless versus discrete, chromatic versus one colour, with and without graphics (such as smiley), with or without a marked neutral midpoint, have a default starting position or not, vertical or horizontal
- Visual analogue scale, self-anchoring
- Voice assistance / sound

Method

Relevant research articles were identified by searches on Google Scholar and EBSCO with limited results, and then by using chain sampling, i.e. by identifying additional studies through references in the first identified studies. The chain sampling limits the comparative element of the size of empirical research available for each IIF.

Relevance of included studies was limited to 1) any studies with comparative traits between one IIF and multiple choice or other item type, i.e. external validity and 2) aspects of psychometric scale validation, i.e. internal reliability and validity. For point one, simple contextual comparative setups were considered evidence of low order. Multitrait-multimethod (Campbell & Fiske 1959), factorial experimental setups (Fisher 1926) meta studies and reviews revealing the dynamics of effects when IIFs are used were considered of high order. For point two, contextual tests of reliability with Cronbach's alpha or test-retest and simple comparative studies without experimental setup were considered low order studies. Tests of construct validity,

which can take on many forms (latent traits, linearity, control of confounders, correlations between constructs e.g.) were considered being of high order. Direct measures of data quality such as response rate, dropout-rate, linearity was considered of higher order than indirect measures such as respondent satisfaction and usability. In other words, there are numerous empirical tests of different order that can illustrate the data quality of IIFs. Ideally, several studies test one IIF against several other formats and in several settings with different populations, themes and traits so that relevant influential factors on the data quality effect can be identified. Overall conclusions were based on methods, outcomes and sample sizes.

Analysis

In the following chapter each type of item and the evidens concerning data quality for that item type will be discussed. For each item type there is a final note on what the evidens points towards.

Battery / response matrix / grid. A battery, response matrix or grid is a table of questions that uses the same scale for all items. A dynamic batteri puts focus on one item at a time by greying out the items not in focus whereas a static batteri does not put focus on one item at a time. Usually batteries are with radio buttons but can also be with drag-and-drop.

Bell et al (2001) conducted a web experiment with n=1,615 participants distributed to either a battery-presentation or expanded presentation of items. Mean response time were 5.07 minutes for matrix and 5.22 minutes for expanded format. Health status and alpha reliability was not significantly influenced by format. Couper et al (2001) used data from a survey (n=665) with three design experiments, of which one was on several items per screen versus one per screen

and with 16 items total they found less completion time on grids (168 vs 194 seconds, $p < .01$). Iglesias et al (2001) tested two versions of SF-12 health scale, the classic in battery versus an adapted version with single-items, in a quasi-randomized trial ($n=422$) on item response rates and found no statistically significant difference but slightly better Cronbach's alpha for the single-item version (0.94 versus 0.91).

Tourangeau et al (2004) conducted six web survey experiments ($n=2,987 + 1,590$), one of these was an eight-item battery versus two four-item grids, confirming a hypothesis of short battery leading to higher consistency among answers (Cronbach's Alpha 0.621 versus 0.562). Galesic et al (2007) tested whether a dynamic visual element of changing answered items font to light grey can influence battery-based survey design regarding nonresponse, response pattern and dropout. The dynamic element gave lower nonresponse, higher differentiation of answers and lower dropout³.

In a split-panel experimental design Chesnut (2008) tested battery versus sequential response on a national stratified sample. The number of respondents was not reported but the survey was sent to a total of $n=30,000$. Chesnut tested several differences and concluded that sequential format gave higher response rate and had fewer missing data than grid, but no difference was found when data was restricted to large households. Overall, sequential layout either maintained or improved data quality.

Kaczmarek (2008) conducted a study on a total $n=1,561$ self-recruited respondents from several surveys using several graphical versions of a battery, one with white background, one striped, one with greyout after answer, one with crosses between boxes and one with greyout and cross. Highest completion rate was seen with greyout and lowest with plain white

³ This is based on a conference abstract. The paper could not be retrieved on request.

background. Response time was not significantly different between item types. Proportion of respondents who switched answers were highest with white background and lowest with both greyout and cross. Proportion who answered all items were highest with greyout alone and lowest with cross only.

Callegaro et al (2009) tested battery versus single-item presentation on two SF36v subscales (vitality and mental health), with four and five items respectively on a sample of n=360 respondents. The intervention groups got standard battery, shaded battery, one item per screen horizontally, one item per screen vertically and one item per screen with vertical shaded response. Callegaro et al did not find significant differences in alpha levels for each group, also not when all single-item and battery-experiments were compared as two groups. Correlations between reversed and straight items in battery and single-item experiment were insignificant except a vitality scale, first group, and nonresponse level was low in all groups. Level of difficulty and enjoyment scores were not different between groups.

Thorndike et al (2009) explored single item versus battery in data distribution and preference among n=710 participants seeking self-help treatment on the internet and filling out several scales: Beck Depression Inventory, Beck Anxiety Inventory, Quality of Life Index and Montgomery-Ålberg Depression Rating Scale. Based on factor models and a structural equation model, results suggest single item versus battery did not affect the construct. Preference was towards single items.

Derham (2011) tested completion levels, respondents' preferences for scale presentation type and influence of preferences on answers and future survey intentions. Sample size was not reported. The tested scales were word scales, number scales and emoticon scales. Questionnaires were based on battery-type designs. All tested item types were in the same survey, so that all

respondents answered all question types. The study revealed a higher rate of questions unanswered in emoticon surveys (24%) than in numbers format (3%) and words (4%). Respondents had highest preference towards word scales, found emoticons easiest to use, found word scales quickest to use, found word scales best expressed feelings, found word scales best gave true indication of answer, emoticons best made the respondent think about each question and overall survey enjoyment was highest with emoticons.

Kaczmirek (2011) conducted nine experiments ($n=2,243$) with a two-factorial design, three levels in each factor with highlight and grayout as dimensions. Each questionnaire had 10 items with 5-point fully labeled scale. Based on logistic and linear regression models Kaczmirek found no effect on lurking (browsing through questionnaire without filling it out substantially) by neither highlight or grayout. A combined pre- and postselection feedback emphasizing the item and not the answer category had the best data quality indicators in regards of missing items and changed answers.

Based on two experiments ($n=1,200 + 1,210$) based on battery-design, Couper et al (2012) tested breakoffs, missing data, and satisficing. One experiment tested dynamic elements to help the respondent get through the battery and the second manipulated or simplified the visual components of the battery. Couper et al found fewer missing data with dynamic feedback and that it helped reduce missing data and splitting the batteries into singular questions reduced missing data further and motivated underreporting, but visual elements had little effect.

To sum up, the studies 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), completion rate higher or insignificant (Kaczmarek 2008, Callegaro et al 2009) insignificant effect on construct (Thorndike et al 2009) and preference towards single item (Thorndike et al 2009).

Regarding confounders, switch rates were higher with white background and lowest with greyout and cross (Kaczmarek (2008), non-response and dropout was lower, differentiation with dynamic visual element of changing answered items font to light grey was higher (Galesic et al 2007), fewer missing data with dynamic feedback (Couper et al 2012), word scales in batteries preferred before emoticons and numbered scales, and overall higher survey enjoyment with emoticons (Derham 2011), no lurking-effect between highlight and grayout grids (Kaczmarek 2011).

Battery / response matrix / grid - multientity scaling. In some occasions the researcher may want to ask more than one question for each theme. As an example, in political science research, it can be used to rate aspects for political parties. Thus, the questionnaire battery contains more than one response row. Multientity batteries can be varied the same ways as single entity batteries with greyout and drag-and-drop.

Based on multiple discriminant analysis in a factorial setup Wong & Teas (2001) tested whether different types of multientity scaling would affect assessment differently based on data of perceptual maps of retail stores in an experimental study with five groups of 41 to 47 students (exact sample size not reported but distributed to N=222 students). They tested two types of multientity scaling, one with rating simultaneous aspects (entity-based) and one with rating one aspect at a time for several attributes (attribute-based) against standard single-entity batteries.

Based on MANOVA and F-test Wong & Teas found instability in both formats without significant differences in instability between the formats.

Borkenhagen et al (2005) explored body image disturbance of 32 anorexics and 30 in-vitro-fertilization patients using the Body Grid and the Body Identity Plot they developed themselves. The Body Grid was designed for psychotherapeutic work and deals with what the patient may consider "problematic zones", 14 categories such as belly, hip, thighs and bottom each rated on 13 aspects which makes a total of 182 ratings placed in a coordinate system, the Body Identity Plot. Borkenhagen et al found clear differentiation between real and imagined body image among anorectic patients opposed to in-vitro-fertilization patients as expected.

In a study of $n=1,082 + 838$ respondents Sikkel et al (2014), which is also mentioned under drag-and-drop, included 2D drag batteries among other item types consisting of dragging countries to cells of two-dimensional grids of quality and price, horizontally and vertically. They found low test-retest in both clicking and drag-and-drop version of 2D multientity in both horizontal and vertical parts compared to other clicking and drag-and-drop item formats, although slightly lower in drag-and-drop.

To sum up, a test based on multientity scaling discriminated clearly between subgroups (Borkenhagen et al (2005) and regarding confounders, entity-based versus attribute-based scaling both introduced instability and did not affect instability differently (Wong & Teas) and low test-retest scores in both clicking and drag-and-drop.

Constant sum. IIF with constant sum are characterized by respondents asked to distribute numeric shares of a predefined sum. The numbers typed in by the respondent need to

sum to a specific number to be considered a valid answer. Constant sum items are usually based on radio buttons but can also be drag-and-drop.

Louviere & Islam (2006) compared four measurement methods for product attribute importance among $n=32$ students and their preferences when buying food products to explore the concept of importance in questionnaire judgements. In this study they make a differentiation between measuring weight or importance: "direct" and "indirect". Direct being approaches with discrete choice or constant sum and indirect being approaches based on analyzing outcome measures such as choices. For constant sum design respondents were asked to allocate 100 points between features important when buying juice. For the discrete choice design 32 product options were set up. They conclude that different distributions are found for direct (best-worst Likert and constant sum) and indirect method and find higher context effects with indirect measures, but stronger correlations between the direct measures, suggesting Likert and constant sum measurement are somewhat comparable.

Conrad et al (2005) tested feedback on tally items with fixed sum about internet use in an experimental study with three groups ($n=951+924+864$): no feedback, delayed and concurrent feedback. They found feedback lead to more answers equal to the fixed sum than with no feedback. They also found that feedback was greater when feedback was displayed concurrent than when it followed submission of the whole series and that response time was lower with concurrent than delayed feedback.

Skedgel et al (2013) compared discrete choice items and constant-sum items in a survey based on societal preferences in a randomized fractional factorial trial ($n=154$). Respondents were divided into a group answering by discrete choice and one answering by constant sum. Respondents were asked to imagine themselves being a societal decision maker responsible the

allocation of a fixed budget between two health-care programs. In the discrete choice groups respondents were asked to choose between two answers in each of the questions and in the constant-sum survey they were asked to allocate a fixed budget to a preferred group using a slider that would change the distribution. After these choice tasks, respondents were asked to rate the importance of the attribute, the importance of an equal distribution of resources, using a 0-10 scale and finally to rate the difficulty of the questions. Alternatives in task 3 of each block were reversed and represented as task 8 to test preference consistency. Data was tested on completion rates, preference consistency, preferences and derived attribute importance. Skedgel et al found no significant difference in either of these aspects, except on completion rates which were higher with discrete choices. Skedgel et al were not able to identify other comparative studies between these item types.

To sum up, constant sum compared to Likert affect data distribution (Louviere & Islam 2006), significant differences were not identified on completion rates, preference consistency, preference and derived attribute importance (Skedgel et al 2013). Feedback, especially concurrent, gave more answers equal to fixed sum compared to no feedback (Conrad et al 2005).

Distance to nonsubstantive response. With distance to nonsubstantive response means that there is visibly higher distance to “don’t know” or “no answer” response option so that the respondent can clearly see that this option is not part of the continuum of for instance an ordinal Likert scale and therefore should not be considered part of the continuum.

Tourangeau et al (2004) conducted six experiments with web page user surveys, of which experiments 1 to 3 were relevant for distance to nonsubstantive response option (n=2,987 + 1,590). With experiments 1 and 2 Tourangeau et al were investigating effect of placement of

nonsubstantive response such as “no opinion” or “don’t know”, concluding midpoint was pushed in the direction of the nonsubstantive categories when these are not differentiated visually by line or space, presumably because the respondent is misled by the visual midpoint instead of the conceptual one. Similar result was found in experiment 3 ($n=2,722$), when distance was not equal, so that distance was gradually higher from left to right.

Drag-and-drop sorting. With drag-and-drop the respondent is asked to categorize answers in predefined categories, for instance by dragging the words with the mouse or by using a touchscreen.

In a study by Bennett & Sebrechts (1997), among $n=107$ undergraduate test takers, the ones who sorted well tended to have a higher test score and college grades. Test takers also preferred sorting to multiple-choice items and felt this type was a fairer measure of their ability.

Timbrook (2013) conducted a study with a comparison of two types of ranking format, one called “forced ranking” drawing lines and one with “unforced ranking” with drag-and-drop, where the respondent simply drags boxes into graphical representations of several numbered bins. Timbrook defined the unforced ranking as one where the respondent can place two answers in the same bin, although this might be considered an unconventional definition of a ranking item. On the other hand, the study is one of very few studies examining data quality of items with drag-and-drop traits. Drag-and-drop is often used as a sorting item, to allow the respondent to sort stimuli, and in this study the item was described as an unforced ranking or “stacked” ranking, meaning that several stimuli from the same question could have the same rank. 72 university students completed the Rokeach Value Survey Form D with traditional ranking items and ranked how important aspects of participant’s decision to attend University with the drag-

and-drop ranking. Timbrook analysed aspects of satisfaction and concluded that the used drag-and-drop variation was a clear improvement regarding usability, preference and number of repositionings.

Sikkel et al (2014) examined clicking versus dragging response options in web surveys with a Dutch non representative sample. Within the sample respondents were randomized between clicking and drag-and-dropping version. Sikkel et al used two versions of an instrument in an internet survey. The clicking version contained no questions with dragging response. The survey had n=1,082 respondents in first wave and n=838 of these responded in second wave with a 22.5 % dropout rate. In the survey Sikkel et al tested a survey including four types of sorting but also other item types (standard multiple choice for background, value slider, category slider and category slider with smileys). Card sorting consisted of pictures of holiday activities being dragged to “like very much”, “like a little” and so on, card sorting text consisted of text instead of pictures, ordering labels consisted of ordering countries according to attractiveness and finally 2-D drag grid consisted of dragging countries to cells of two dimensional grid of quality and price, horizontally and vertically. Sikkel tested test-retest at values between 0.463 and 0.718 ($p = 0.043 - 0.056$) for clicking items and 0.377 - 0.706 ($p = 0.045 - 0.06$) for drag-and-drop. Generally, clicking had better values and best results were seen with card sorting pictures and lowest with 2D drag grid vertical.

Based on a between-subjects factorial design with 15 experiments in three University Applicants web surveys and Panel Survey 2012 (n=846, 37.4% response rate), a University Applicants Survey 2012 n=5,977 (32.4% response rate) and a University Applicants Survey 2013 n=7,395 (40.4% response rate), Kunz (2015) tested different versions of rating scales compared to standard radio button items. The scales tested were five types of drag-and-drop,

namely grid type multiple item, drag-respons with the response dragged to the item being answered, drag-item with the stimulus dragged, single item vertical and single item horizontal. Kunz also tested drag-and-drop by item length (6, 10 and 16 items scales) to analyze likelihood of length influencing distribution and missing data. Respondents were randomly assigned to one of the six experimental groups. Respondents would either respond by dragging a response towards a response box or to drag answers away from the box. Kunz examined missing data, response time and respondent evaluation. Kunz tested several item traits and found more missing data and longer response time for both types of drag-and-drop but also that drag-and-drop promoted respondents' attentiveness and carefulness towards the task.

In a study conducted by Timbrook & Moroney (2016), a somewhat similar study as Timbrook 2013 with graphical bins, but this time with a higher interval, so that three stimuli would not just be placed in three bins, but in ten, allowing for the indication of distance between ranks. The experiment consisted of both a numbered and an unnumbered version of the bins. Participants reported high satisfaction with the ability to express answers with distance between ranks.

To sum up, test takers preferred sorting to multiple-choice (Bennett & Sebrechts 1997, Timbrook 2013), test-retest values were at same level as clicking but vertical part of 2-D grid drag-and-drop revealed lowest test results (Sikkel et al 2014), more missing data and longer response time was found for two types of drag-and-drop but respondent attentiveness and carefulness higher with drag-and-drop compared to radio-buttons and higher satisfaction was found with ability to express answers with distance between response options (Timbrook & Moroney 2016).

Draw function. With a draw function the respondent can give an answer by drawing on the screen via touchscreen or with the mouse.

Among $n=173$ women Björkstén et al (1999) used a scale with 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%). Björkstén concluded that the scale was valid.

One test that drawing is used for is screening for reduced cognition and dementia (Shulman 2000). Shulman tested psychometric properties based on literature study including 21 studies with a total of $n=2,071$. Shulman finds levels of validity too low for scientific use.

In a study by Elliot & Papadopoulos (2012) respondents were asked to provide top-of-mind image associations in addition to regular close ended questionnaire items. 1,147 respondents filled out the blank frames giving a response rate of 62,3%. The images were afterwards mapped to represent underlying structure using a qualitative grounded theory process. Respondents were asked what images would come to mind when they thought of U.S as a country in general and when thinking of U.S. as a travel destination.

Finishing a sentence. Finishing a sentence is an item type found in language and grammar tests. The respondent is asked to insert a word, sentence or symbol such as comma to make sense of the context. Finishing a sentence can be based on both a limited number of possible answers, and it can be design as an open-ended question. The item is sometimes based on multiple inserts in a longer narrative of multiple sentences. With limited number of answers this item type can be seen with dropdown and drag-and-drop. One could argue that this is technically just a variation of a dropdown item, but the response is placed within the stimuli.

Research about finishing a sentence item types have not been identified, but probably exists, since this IIF has been used extensively in learning tests.

Breland (1999) tested passage editing items used for assessment of writing skill among students ($n=52$). Passage editing item values correlated with self-assessments of ability, grade-point average and writing accomplishments and reliability was at 0.84.

Image response, universal (smiley, emoticon and similar). Universal image response can be used for expressing happiness or degree by for instance clicking or moving a graphical image with a smiley/emoticon, image of a glass of water (empty vs full). This response option is sometimes used when respondents are children or other populations with limited reading skills. Smileys-scales are usually represented as discrete response such as five faces but can also be slider scales where the facial expression of a smiley is changed when a slider is moved back and forth.

In a study by Buchanan & Niven (2002) two types of VAS dental anxiety scales were tested among 101 children ages 3-18. One had five round faces depicting unhappy, neutral and happy faces (The Facial Image Scale). The other consisted of eight cards, each with two figures of a child in different moods that could be interpreted as sad, angry, happy, scared, indifferent (The Venham Picture Test). The study showed a strong correlation ($r=0.7$, $p<0.001$) between the scales.

Buchanan & Niven (2003) propose a Facial Image Scale with five faces expressing moods to assess dental anxiety among children, based on a sample of 50 children. Effects of age or gender were insignificant and reasonable agreement between the child's rating of anxiety to

their parents rating of the same ($K=0.66$, $p<0.001$). Comparison of measure in weighting room and in the dental chair showed small rise in scores ($m=2.4$ vs $m=2.6$).

Castle & Engberg (2004) compared six questionnaires designed to test satisfaction among elder patients. The questionnaires were distributed randomly among elders ($n=2,450$, 78.5% response rate). One questionnaire had variations of response formats and the others had 5-item Likert, 5-item satisfaction format, 5-item valuation format, 4-item Chernoff faces and 10-item visual analogue format which allows for comparison of traditional formats, faces and VAS scales. Chernoff faces were liked the least (5%) and 10-item VAS was liked the most (39%). The variation coefficient was higher with VAS.

Buchanan (2005) has tested the reliability and validity of a four-item dental anxiety scale (Smiley Faces Program). In the questionnaire used all items are based on response using five-point smiley scales. Questionnaires were administered on computer with the scale starting out neutral allowing for the child to click on the screen to make the face happier or unhappier. The scale was tested on a convenience sample of $n=464$ children and demonstrated good internal consistency with Cronbach's alpha 0.8, test-retest (two weeks interval) had a correlation of 0.8 ($p<0.001$) and concurrent validity showed significant correlations between SFP and two other measures (both 0.6, sign. $p<0.01$). A sub-sample ($n=245$) was asked to also complete two existing paper-based dental anxiety measures, the Child Dental Anxiety Scale, MCDAS) and Children's Dental Fear Survey Schedule, DFSS, are both based on Likert scales. It is also noted by Buchanan that the children did not seem to need help or to have difficulty using the scale.

Desmet (2005) tested an instrument on $n=120$ participants to assess response to consumer products, based on 14 possible emotional responses by an animated PrEmo cartoon character with facial, bodily and vocal expressions. Desmet tested whether the single pictures were chosen

more often than could be expected from chance and concluded two needed further development. The PrEmo item was tested with test-retest against a verbal scale with high correlation (0.72 to 0.99, $p < 0.05$).

Derham (2011) has tested completion levels, respondents' preferences for scale presentation type and influence of preferences on answers and future survey intentions. The tested scales were word scales, number scales and emoticon scales. Questionnaires had several questions with same scale. All tested item types were in the same survey, so that all respondents answered all question types which could introduce bias in the comparison between item types. Questions were about banking, financial institution use, importance of, satisfaction with and likelihood of use of specified account types (90 items in 2009, 33 items in 2010). Respondents (unreported sample size) were also asked how they enjoyed the survey and how likely they were to complete similar surveys in the future. The study showed a higher rate of questions unanswered in emoticon surveys (24%) than in numbers format (3%) and words (4%). Respondents had highest preference towards word scales, found emotions easiest to use, found word scales quickest to use, found word scales best expressed feelings, found word scales best gave true indication of answer, emoticons best made the respondent think about each question and overall survey enjoyment was highest with emoticons. Derham notes that the survey program recorded whether the respondents used the slider or left it untouched, which is a relevant aspect because no answer can be confused with a neutral position if touch of slider is not recorded. In addition to this point one could argue that one could consider this difference being visible to the respondent, so that the respondent can see whether the item is left unanswered or not.

Emde & Fuchs (2012) carried out a survey among university journal readers and non-readers (n=611). Emde & Fuchs presented three face item types, one fixed, one dynamic facial expression items and one cognitively described. The dynamic version changed colour of the face from red (unhappy), to orange, yellow and green (happy) smiley and size of the smiley changed when mouse hovered over, and the cognitive version had text underneath. Emde & Fuchs did not find significant difference in distributions.

Broekens & Brinkman (2013) tested a method for measurement of human affect, the "AffectButton" in eight experiments (n=776 in total) to see if human affect could be measured more nuanced than happy versus sad and differ between pleasure, arousal and dominance as three dimensions of facial expression. The respondent provided affective feedback on the affective dimension's valence, arousal and dominance. The button presents a dynamically changing iconic facial expression changing based on the coordinates of the respondent pointer in the button. In other words, this scale is stepless whereas graphical smiley scales are usually discrete ordinal. Broekens analyse aspects of reliability, validity and usability across five previously published and 3 unpublished studies. They find reliable results of the AffectButton (0.81, 0.66 and 0.77 for pleasure, arousal and dominance, respectively). By correlation with other affect measure, validity was shown (0.93, 0.69 and 0.85). Regarding usability, preference was same between AffectButton and other ratings and AffectButton was quicker to use.

Bosch & Revilla (2020) tested the use of emojis in web surveys with 62 items among n=1,614 participants born between 1982 and 2003, also referred to as millenials. Bosch & Revilla also offered a response option with eight versions of a man showing both facial and body expression (smiling, angry, arms crossed e.g.). Respondents preferred emoji response compared to traditional scales, and overall, 88.1% (Mexico) and 75.5% (Spain) expressed they would like

to use emojis in surveys. 98.5% (Mexico) and 98.5% (Spain) of respondents used emojis in open-ended questions but several respondents pointed out the ambiguity of such responses. Compared to VAS scales, with the open-ended questions respondents could express with more variation such as face with sunglasses or heart-eyes.

To sup up, strong correlation was found between VAS and smiley scale (Buchanan & Niven 2002), unconfounded of age and gender, reasonable agreement between child and parent rating of a smiley scale (Buchanan & Niven 2003), Chernoff faces liked less than various text item formats and variation coefficient higher with VAS without faces (Castle & Engberg 2004), good internal consistency, test-retest and concurrent validity with a dental anxiety smiley scale (Buchanan 2005), high test-retest correlation for a multifaceted cartoon imagery item type (Desmet 2005) and a stepless smiley scale was found reliable and valid (Broekens & Brinkman 2013). In a study with possible order bias, Derham (2011) found higher rate of unanswered questions with emoticons, higher preference towards word scales and word items expressed feelings better, but emoticons easier to use and overall survey enjoyment highest with emoticons.

Between fixed, dynamic with change of colour and with supplemental text no significant difference in distribution was found (Emde & Fuchs 2012). Millenials preferred emojis to traditional scales but also expressed ambiguity when used in open-ended questions.

Image response, specific. As opposed to universal images with smileys, specific image response is used to give a better understanding of what is being answered by exemplifying with images alongside text. The evidence base is therefore more contextual.

Rofé (2004) introduces a method designed to map the sense of well being in neighbourhoods. The study consisted of two stages of surveying (n=743) and mapping people's

emotional response to a specific environment while the respondent and researcher walks around in the given environment. Rofé did not test aspects of data quality, and the study is therefore not included in the analysis, but Rofé suggested validation in a future study by correlating the map data with other related data.

Shamir & Kark 2004 present a method for measuring organizational identification with circles representing self and others, with seven versions, first one apart, then gradually closing and last with complete overlap. They tested different hypothesis on different samples and found test-retest correlation of 0.73 and 0.80 in two samples ($n=1,856$) and significant correlations with verbal scales.

Balram & Dragičević (2005) conducted a study with some similarities to Rofé 2004 with the purpose of measuring attitudes toward urban green spaces, using respondent plotting on geographical maps. 179 respondents (55.6% response rate) answered a questionnaire and factor analysis and reliability analysis was conducted on the data. Balram found a two-factor attitude structure towards green spaces.

Ijmker et al (2008) used items with additional pictures to provide visual guidance in a questionnaire about ergonomics. 84 office workers filled out an online questionnaire twice, with two weeks interval. For a subgroup of 38 on-site observations were made and goniometer measurements (angle measure) were performed. Ijmker concluded a percentage agreement between 71% and 100% for test–retest analysis, the comparison between questionnaire and on-site observation ranged from 31% to 100% for the comparison between questionnaire and on-site observation, and comparison of questionnaire and goniometer measures were between 26% and 71%. Ijmker concluded that the designed questionnaire was reliable but not useful to collect data for epidemiological field studies among office workers.

Shropshire et al (2009) used data from Survey2001 (n=1,679) with photographs alongside response options to ease responses about threatened animal species and concluded that support for preserving endangered species increased when pictures were added.

Xia et al (2011) tested reproducibility with intraclass correlation and validity of a food frequency questionnaire (FFQ) for female adolescents tested on 168 females ages 12 to 18 and tested against 24-hour dietary recalls. The panel filled out the 24-hour questionnaire three times in each of nine months and filled out FFQ twice during this period. Participants were asked to choose from photographs showing different-sized portions of 23 food items and in the 63 left items photographs of dishes, bowls and cups were used to represent portion size. Intake values were lower for recall than the questionnaire and intraclass correlation was moderate (0.4-0.8).

Waller et al (2012) have presented a research article about the validation of a foot and ankle screening survey, The Swindon Foot and Ankle Questionnaire, SFAQ. The 10-point items are based on discrete yes/no items and items with diagrams of feet and ankles and is for use in general rheumatology outpatients. SFAQ was tested with 301 patients. Correlation with clinical score and with a different screening scale, Manchester Foot Pain and Disability Index, MFPDI, showing correlations of 0.65 and 0.63, respectively.

In a study by Leutner et al (2016) the predictive validity of a creativity measure based on image response scales was covered with two versions of a 77 items questionnaire. It was designed to cover aspects of cognitive flexibility, curiosity and openness. Each item had a text-based question followed by between two and eight image response options. 964 respondents were divided in two groups, one was asked to assess varying levels of same trait and one was asked to assess different traits. Predictive validity was tested by asking all respondents to fill out both three existing tests, the Curiosity and Exploration Inventory-II, Cognitive Flexibility

Inventory and the Openness to Experience. Image responses were recoded into 321 dummy variables and added to a linear regression model as independent variables and creativity scores added as dependent variables. To compensate for over-fitting, Least Absolute Shrinkage and Selection Operator (LASSO) and Sequential Forward Selection (SFS). SFS was used additionally because LASSO is not able to account for some dummy variables coming from the same item. Leutner et al conclude that 14 significant items predicted outcome measured by existing Likert-scale creativity tests.

To sum up, Shamir & Kark (2004) found moderate test-retest values and correlation with verbal scales for an organizational identification based on images of circles, a two-factor structure was found with a scale based on geographical maps (Balram & Dragićević 2005), moderate test-retest values and correlation with observation for scale based on items with picture response (Ijmker et al (2008), data distribution was 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 found with image response (Leutner et al 2016).

Image, question / stimulus. This item type is based on text-based questions supported by images that help the respondent understand what the question is about. Items with image stimulus have traits somewhat like those of a vignette by framing a scenario.

Graybill & Heuvelman (1993) have tested the validity of hostility scores from Children's Form of the Picture-Frustration Study (P-F). It consists of 24 cartoon-like pictures depicting every day, interpersonally frustrating situations. Graybill & Heuvelman describe a cartoon shown making statements that either help to describe the frustration of the other or serve to

frustrate the child, and with a child depicted with an empty caption box. The child is then asked to fill out the empty question box based on how they think the depicted child would say, assuming the child respondent identifies with the frustrated child. In other words, the response type is basically an open-ended question with the stimulus supported by graphics. Graybill & Heyvelman found high correlations (0.82) between ratings from 237 children and ratings by graduate students.

Escalante (1995) used panel data with pictures in the McGill Pain Map questionnaire (MPM) to allow for the respondent location of body pain among 411 elderly community dwellers. Pictures were divided in 36 anatomical areas. MPM was tested with inter- and intra-rater agreement among three raters with kappa statistic. Escalante also tested intraclass correlation, internal validity with Spearman's rho and Present Pain Index and Pain Rating Index. External validity was tested with correlation with items from the Perceived Health, Amount of Bodily Pain, Pain Interference with Work scales. Escalante conclude that MPM is reliable and valid in the given sample.

Couper et al (2004) has explored the impact of questions presented as part of the item. These questions presented with a picture are divided into questions in which images play an essential role, in which images are supplemental and in which images are incidental, yet by some respondents experienced as part of the question. Data was based on an experiment among random samples within the Knowledge Networks panel of “approximately 100,000 from almost 50,000 households” with a 37% response rate in the USA. In the current study a sample of 2,385 (80%) was used. Questions were about travel, sporting events, music, dining out and shopping. For each topic four versions of the questions were designed: (1) one without picture, (2) a version featuring image of a salient but low frequency instance of the behavior in question, (3) a

version featuring image of a salient, high frequency instance and (4) a version that displayed both pictures. Experiment participants were randomly assigned to one of these four states. Data was tested with a comparative analysis using ANOVA and F-test. The highest number of reported events (such as shopping) by respondents were seen in the high frequency condition and with both high and low frequency condition. In other words, having pictures presented with the question seemed to boost the frequency. The number of pictures respondents saw did not relate to rating of satisfaction or length of survey.

To determine inter- and intrarater reliability on estimation of healing of femoral condyle OCD Wall et al (2017) conducted a cohort study among n=10 surgeons. The surgeons were asked to diagnose based on radiographic images. Some items had continuous response and some ordinal. The scale showed higher interrater and intrarater reliability with continuous response (0.94 and 0.84) compared to ordinal (0.61 and 0.68).

To sum up, high case correlations between picture ratings were identified (Graybill & Heyvelman 1993), a scale with pain mapping was found reliable and valid (Escalante 1995), higher respondent reporting found with images (Couper et al 2004) and a scale with image stimulus had higher interrater and intrarater reliability with continuous response than ordinal (Wall et al 2017).

Image, dynamic / interactive / figural response. With dynamic graphics or figural response, the respondent is supposed to click on certain parts of a picture, draw lines, click or drag interactive elements. The image area is usually mapped so that coding of response is possible according to where the respondent clicks or drags objects. Categorization of different areas of the dynamic image are not necessarily visible to the respondent, and whether the

categorization is visible is likely to affect data. The variation within this format is quite high and the purpose can for instance be to draw a line in a chart or click a point in an image of a human body with diagnostic purpose. This type of element is often found in tests. Huber et al 2009 describe various examples of using interactive images in health-related e-learning contexts where the respondent is asked to click specific parts of the picture as an answer.

Among 4th, 8th and 12th graders ($n=347+365+322$) Martinez (1991) randomly assigned participants and tested 25 figural response items against 25 multiple choice items with similar purpose. Results revealed higher difficulty for figural response (mean score 2.16 – 6.24 versus 3.82 – 8.85 for multiple choice, $p<0.05$), especially when designed to be difficult, slightly more discriminating for multiple choice (0.51 – 0.62 versus 0.45 – 0.46 for multiple choice), more reliable (0.35 – 0.71 versus 0.32 – 0.66 for multiple choice) and higher omit rates (4-6 – 6.8 % versus 2.1 – 4.4 % for multiple choice).

Based on 257 paid volunteers, Bennett et al (1999) studied the data quality of a scale with some items that had several correct possible answers, for instance, asking the respondent mathematical questions where correct answers would be within an interval. In other words, computational validation was needed to ensure correct test score. The scale had overall Cronbach alpha at .82.

Bennett et al (2000) tested graphical modeling among paid volunteers from higher education institutions ($n=243$), dividing participants in two groups, each got 18 graphical modeling items, but with differentiation in manipulation of features that could theoretically affect difficulty. One group got form A with under-determined word problems first and well-determined items after, and the other group got form B with under-determined word problems first and under-determined pure items after. Bennett et al did not find a significant difference in

performance. Reliability was high with 0.92 and 0.90 respectively overall. Performance on the test was not affected by the manipulations expected to cause difficulty.

Lim et al (2006) used a virtual reality (VR) computer graphic garden landscape as stimulus for landscape assessments and tested distribution in laboratory and internet samples (n=37). The landscape was represented with regular photos, static VR and walk-through VR images. Response was compared on profile analysis, Anova and t-test on respondent preference and same results were found between photos, static VR and walk-through VR.

Van Ooijen (2011) tested an image-based response system designed for radiology education where respondents click on X-ray images and concluded that the system was easy to use among n=57 students.

Wan & Henly (2012) tested reliability, efficiency and construct validity of “figural response” and “constructed response” (taxonomy from Bennett 1993) item types against multiple choice and found figural response values close to multiple-choice. Tests for 5th, 8th grade and highschool level all containing all three item types were used. The tests were computer-based with realistic representation of classroom experiments for and the figural response was graphics, audio, video and animation. The items required examinees to manipulate graphic elements of an item, click one or several hot spots on an illustration or complete a diagram by dragging and dropping elements. Sometimes response was open ended and sometimes selection. Across grade levels 5, 8 and highschool, time spent was slightly less with multiple choice (55, 52 and 44 seconds) compared to figural response (77, 73 and 59 seconds), (n=3,601). Information levels were similar across multiple choice, figural response and short constructed response. Factor models for grade 5, 8 and highschool (n=5,033 + 4,881 + 4,996) resulted in model fit for three factors with 5th grade, acceptable but rejectable single-factor model for 8th grade and acceptable

model fit for highschool. Factor loadings for multiple choice, figural response and constructed response were at similar levels across grade.

To sum up, higher difficulty, reliability and omit rates and lower discrimination with figural response than multiple-choice (Martinez 1991), difficulty not confounding performance on figural response and high reliability (Bennett et al 2000), interactive image versus static image and VR did not affect response (Lim et al 2006), easy usage of dynamic image (van Ooijen 2011) and with figural response had information levels and factor loadings similar to multiple choice (Wan & Henly 2012).

Image, video. Items with video have traits somewhat like items with pictures. Video can be used to exemplify or to present a case in the same way as a vignette.

In a study by Kersting (2008) 62 volunteer (56 completed criterion measure) math teachers completed a video-analysis assessment and meant as a proxy for teaching knowledge. Respondents watched 10 video clips of 1-3 minutes length and supplied answers in open-ended items which were then assessed by experts on four dimensions. The test was found reliable with interrater-reliability correlations of .79 - .85 (kappa 0.65 - 0.78), Cronbach's Alpha .90, good data fit, valid on discriminant validity (.22 - .60) and criterion validity showed two of four dimensions were statistically significant with a 32-item short-version of Mathematical Knowledge for Teaching instrument (Deborah Ball).

Interval item. With an interval item the respondents marks two points on a scale instead of one, thereby expressing an answer as an interval instead of a value.

McKelvie (1978) asked respondents to not only mark a value on a scale but to also mark a “confidence interval”, not in statistical sense, 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. No studies have been found that test data quality on items where respondents are asked to mark an interval on an ordinal or interval scale.

Randomization, nominal and ordinal. With nominal randomization all response options are randomized, which is meaningful when there is no natural given order of the response options. A question with simple nominal randomized sample space means that for each respondent the order of the possible answers varies randomly. A trait of many scales, such as Likert scales is that it has ordinal traits and, in this case, randomization should randomize the direction of the scale, not each response option. The mechanisms behind order bias has been theorized to be a matter of satisficing (Krosnick 1991), acquiescence or social desirability (disproportionate yay-saying), extremity (tendency to respond in scale extremes) or fatigue.

Based on data from 548 experiments with a total order effect was found in telephone surveys by the Gallup Organization Holbrook et al (2007), predominantly recency effect opposed to the primacy effect often seen in written surveys. Especially effects were found among populations with low education and with items difficult to comprehend, difficulties caused by wording or placement in the questionnaire. Interestingly, the effect was higher when questions explicitly or implicitly encouraged respondents to wait until all response options had been read out loud. Yan & Keusch (2015) also use experimental setup with 11-point end-labeled scale in telephone survey with a sample of $n=498$ and find that the direction significantly affects respondents' ratings. The study concludes anchoring-and-adjustment heuristic, not satisficing as

explanation for the bias. Since the evidence here is based on telephone surveys Holbrook et al and Yan & Keusch is not directly related to IIFs as defined here and therefore not included in the evidence base.

Among $n=130$ students, Mathews (1929), probably as one of the first, conducted a study of order bias by difference in reliability in an experiment based on five items, each with five ordinal response options, “like greatly” to “dislike greatly” - later referred to as Likert scales (1932) - and found an order effect.

Payne (1949) found higher impact of response order on binary and ordinal items in a study of two surveys with about 3,200 respondents (exact number not informed) each, when questions had more words and words that were more difficult to comprehend.

With a sample of $n=332$, Belson (1966) tested verbal rating with ascending versus descending order and found an order effect.

In Carp (1974) both effect of response option order in ordinal items and placement of item in questionnaire was tested among $n=899$ 65+ respondents. The effect of response order bias was not affected by where in the questionnaire the item was placed.

In an experiment based on a survey about child qualities ($n=1351$), Krosnick & Alwin (1987) identified bias in nominal items based on order and theorize on why differences are found, predicting where effect is strongest (low cognitive sophistication), which was confirmed.

Among a sample of $n=168$ beef producers, Israel et al (1990) tested ordinal response-order based on 10 items with four response options and a significant difference was not found.

Chan (1991) $n=102$ among students, using a test-retest design with same respondents, switching between ascending and descending order between the two surveys with five weeks interval, found difference in distribution.

Friedman et al (1993) found left-side selection bias $n=208$ on a 5-point Likert scale with either ascending or descending order.

In a systematic review based on experiments in 14 studies with a total of $n=1,047$ participants Knäuper (1999) found results suggesting age and education as confounders on response order effect. The 14 studies do not overlap with this study.

Barnette (2000) tested conditions in a factorial 2×3 setup ($n=915$) with highest reliability found when items were positively worded half-and-half ascending and descending order.

Weng & Cheng (2000) conducted a paper-based test-retest survey study among $n=858$ students with one-week interval finding no order effect.

Bishop & Smith 2001 found response order effects in a review based on $n=158$ auditory and $n=18$ visual existing Gallup experiments (number of respondents not reported) from 1930s up to early 1950s, finding order effect in 51% of auditory mode and 61% of visual mode experiments but they found higher effect in auditory mode when questions and response options were longer.

Attali & Bar-Hillel (2003) argue, that not only test takers tend to place their answer in the middle among response options, but that test makers also tend to place the right answer in the middle, strengthening the false positives even further. In a study of $n=127$ undergraduates, respondents were asked which of answer A, B, C, D was the right answer when the question was the capital of Norway. Despite no elaboration was given, forcing them to guess, there was strong bias towards answering in the middle (frequencies 7, 33, 28 and 1), similar tendencies on same sample with new item and with other experiments of $n=58$, $n=40$ (students guessing word explanation), $n=374$ (Berg & Rapaport 1954) and $n=35,560$ (PET examinees in a 25-item test).

They conclude, as an estimate across several experiments, that test takers have tendency to answer in the middle 75-80% of times, when they don't know the right answer.

Nicholls et al (2006) conducted a study of $n=292$ students finding higher scores in the group with descending-ordered scale.

In an item repetition setup among $n=156$ respondents Hofmans et al (2007) found small differences in data distribution between 5-point fully agree to fully disagree versus fully disagree to fully agree scales.

In a study of response direction of numbered and unnumbered items on Big Five Inventory, BFI-10 among $n=315$ students, Rammstedt & Krebs (2007) found an order bias when numbers were counterintuitive, that is, going from 8 at left and 1 at right, but did not identify a bias when categories were numbered going from 1 at left.

Based on $n=3,000+$ students (exact sample size not reported), Dillman (2009) tested satisfaction order effect on Likert scales and found no significant difference.

Based on a representative web survey among $n=397$ adults, Malhotra (2009) studied whether the order effect of five ordinal five-point scales is moderated by how much time spent on the questionnaire, based on regression analysis and found an effect. The highest primacy effect was found with respondents more than $1\frac{1}{2}$ standard deviation below mean completion time.

Based on two experiments with $n=2,182$ and $n=2,229$ respondents, Toepoel et al (2009) tested five-point rating in different formats: positive to negative vertical, negative to positive vertical, graphical scale horizontal, linear with numbers 1-5 and with numbers, two options. They found significant differences between verbal and graphical response. Additionally, based

on a regression model, age, education and gender were explored as confounders and found different distributions for all versions.

Hartley & Betts (2010) asked $n=465$ academic writers, reviewers and information scientists to rate abstracts on an 11-point semantic differential scale, with random assignment in a factorial setup testing a clear to unclear versus unclear to clear continuum and with ascending and descending numerical order. They found clear to unclear and descending numerical order had highest rating.

In a factorial 3×2 experimental study Salzberger & Koller (2013) tested three types of response instruction (asking to read carefully, reply spontaneously or no instruction) and two directions of response order on a sample of $n=1,644$. Salzberger & Koller identify a main effect by response order resulting in better model fit (Rasch and factor model) and interaction effect by instruction.

Based on a customer satisfaction survey ($n=27,406$), Stapleton (2013) found satisfaction was biased by response order.

Based on a random sample of $n=2,696$ with data from Time-sharing Experiments for the Social Sciences (TESS) panel, Garbarski et al (2014) conducted a 2×2 factorial study manipulating order of ordinal response options and order of self-rated health items and found an order bias.

Höhne & Lenzner (2015) explored response order effects on horizontal and vertical items using data from eye-tracking on $n=75$ test subjects. Results indicate that higher effect was found in vertical presentations than in horizontal and some respondents do not read all categories or pay equal attention to all categories but indicates relatively small effect sizes, higher in vertical

than in horizontal ratings. One could suspect a degree of Hawthorne-effect due to eye-tracking equipment.

Maeda (2015) tested vertical versus horizontal Likert response-order effect in with $n=1,693$ adult respondents and found left/top selection bias.

Liu & Keusch (2017) tested effects on scale direction using a latent class analysis model, based on data from the 2012 American National Election Studies before and after election with two modes of data collection (face-to-face and web), both modes in themselves representative ($n=11,424$). The study also included random assignment to one of two scale direction conditions used for 190 ordinal items with vertical response. Accounting for background variables as confounders, they found direction effect in both modes.

Yan et al 2018 use representative time series data drawn from American National Election Studies 2008 and 2012 ($n=2,322$) including face-to-face and web interview. The survey included an experiment varying direction of ordinal response scales. They explore in which circumstances bias is found and find stronger direction effect in non-attitudinal items, items placed early in survey, question location, based on scale length and on CAPI (Computer-Assisted Personal Interviewing) compared to self-administration.

To sum up, order effect was found in most studies (Mathews 1929, Payne 1949, Belson 1966, Carp 1974, Krosnick & Alwin 1987, Friedman et al 1993, Knäuper 1999, Bishop & Smith 2001, Nicholls et al 2006, Hofmans et al 2007, Toepoel et al 2009, Hartley & Betts 2010, Stapleton 2013, Garbarski et al 2014, Höhne & Lenzner 2015, Maeda 2015, Liu & Keusch 2017). Some did not find significant effect (Israel et al 1990, Weng & Cheng 2000, Dillman 2009). A study also found a difference in distribution (Chan 1991).

Confounding from question length (Payne 1949), age and education (Knäuper 1999), positive versus neutral wording (Barnette 2000), auditory versus visual mode (Bishop & Smith 2001), counterintuitive numbering (Rammstedt & Krebs 2007), in one study not by placement in questionnaire (Carp 1974) but in a later study with stronger design placement effect was found (Yan et al 2018), completion time (Malhotra 2009), graphical layout, age, education and gender (Toepoel et al 2009), clarity of wording (Hartley & Betts 2010), instruction (Salzberger & Koller 2013), vertical versus horizontal design (Maeda 2015, Höhne & Lenzner 2015), face-to-face versus web mode (Liu & Keusch 2017), attitudinal versus non-attitudinal items (Yan et al 2018).

Randomization, open ended response. Discrete variable items can sometimes have an added open-ended option such as “Other, specify:” that allows the respondent to give another response than the discrete choices allow. This item choice is traditionally put at the bottom or at the right side (left side in arabic). Therefore, a randomization of such an item should not include randomization of this choice option.

No studies exploring scales with open ended option were identified.

Randomization with separate non-randomized category. Questions of both ordinal and nominal traits sometimes have a “don’t know” or “no answer” choice option. Traditionally, these options, like randomization with open ended choice, have been put at the bottom or at the right side, or left side in arabic. And again, these choice options should not be included in the randomization.

In a 13-item survey among $n=436$ students with 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.

The previously mentioned study by Liu & Keusch (2017) reference “don’t know” being a response option in their study but did not report separately on this matter.

Ranking. With ranking items, 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. A challenge about ranking items is that the list of response options will necessarily have a starting point with a default rank that the respondent then can then rearrange. It could be argued that this could introduce a systematic bias by design if the default rank is not randomized. The ranking item represents a relative rank whereas separate multiple-choice questions, also called rating, is a respondent assessment of absolute position, with each assessment independent of the other assessments. In other words, with rating one respondent might choose to assess two topics equally important, whereas with ranking this is not possible.

Rokeach (1967) designed a value classification of two dimensions consisting of terminal (friendship, love, self-respect and more) versus instrumental (cheerfulness, ambition, love and more) values, named The Rokeach Value Survey (RVS). The RVS has been used in multiple later studies (Musik et al 2009, Timbrook 2013) and was designed to ask respondents to rank these 36 values. Rokeach did not find order effect (1967), found acceptable test-retest reliability (1973) and others have found acceptable construct validity (Penner 1971, Feather 1975 in Rokeach 1973).

Based on the Rokeach Value Survey participants ($n=236$) were asked to rank 36 values according to importance (Rankin & Grube 1980) with follow up after six weeks in a cross-over experiment. One version was the standard version and in the second version they were asked to rate separately from 1 to 99 on each item where 1 was “the most important guiding principle in my life” and 99 “not at all important to me. Nothing I believe or do is ever based on this value”. Rankin & Grube compared validity and reliability of the two item techniques and found correlations between the two versions between 0.3 and 0.8. Convergent and discriminant validity suggested both versions measure the same construct. Test-retest values were lower for the rating version than for individual rating, but they concluded that they did not differ because of expected difference due to interdependence in ranking items. Composite rank order was affected by measurement method. Individual rating was better predictor of attitude than ranking in 9 out of 15 comparisons. Based on several tests Rankin & Grube find that the results are inconclusive.

Alwin & Krosnick (1985) conducted a study showing that value rankings produced somewhat stronger relations than factors within the value ratings ($n=973$), although differences were not statistically significant. Also, they found, that, although results were somewhat similar, they reveal different results in regards of latent traits, i.e. a single factor ranking model and two or three factor rating model.

Based on the 1980 General Social Survey of survey response from 1,468 noninstitutionalized American adults, Krosnick & Alwin (1988) tested ranking technique using face-to-face interviews about parental values, , showing one card in which they were to assess whether ten child aspects, such as good manners, honesty e.g. were extremely important, very important e.g., and examined whether removing non-differentiating respondents from the sample affected correlations among the rating of child qualities and among ratings of child qualities and

educational attainment and find that a fairly large proportion assessed with little differentiation among ratings of child qualities, suggesting limitations of ranking items. Also, the experiment showed very low variation, where only 57.9% of respondents rated fewer than nine qualities equally, although it may also have been a tough topic to rank.

Research by Maio et al (1996) covers predictive validity for ranking items compared to individual ratings by comparing predictive validity of value rankings with value ratings on within-subject data. 211 psychology students ranked and rated importance of 42 values, indicated attitudes toward 30 controversial issues and assessed ethical acceptability of 74 behaviors. Participants were given three booklets, one for measurement of values, one for measurement of attitudes and one for measurement of judgement about ethical acceptability of various behaviors. The presentation order was counterbalanced between participants. Maio et al then tested correlations between values, such as honesty, and behaviors, such as example of dishonest behavior. Participants were asked to both rate and rank importance of values, then they rated and ranked importance of values on a 17-point scale. Rating of attitudes was based on a 9-point scale from “extremely unfavorable” to “extremely favorable”. They used Wilcoxon matched pair test to evaluate reliability of differences between correlation magnitudes. Ratings produced stronger correlations than rankings $Z=4.34$, $p<0.001$, thus exhibiting stronger predictive validity. Also, value ratings did not moderate relative validity and normal distribution scaling did not change the conclusion.

To sum up, a scale based on ranking items was found reliable and valid (Rokeach 1967) and same underlying construct was found between ranking and rating, but rating was better predictor of attitude (Rankin & Grube 1980), ranking exhibited stronger relations than rating

(Alwin & Krosnick 1985), ranking had more non-differentiating responses than rating (Krosnick & Alwin) and stronger predictive validity with rating than ranking (Maio et al 1996).

Time limited answers. Time limited answers makes it possible to define a timing interval for answers. This makes it possible to deny answers after a certain time limit or allowing the test points to be calculated based on how much time used by the respondent if the questionnaire has test-based characteristics. This question type is found in tests, such as IQ-tests, vocabulary tests, reaction tests and memory tests. Timing or countdown could by design be made visible or invisible to the respondent. It is relevant to point out that tests on timing have mostly been made on the questionnaire in full length and not per use of time limit on single items. IQ-tests with time limit has shown high validity (Elliott et al 2001; Bird et al 2004). Lu (2007) mentions speededness as the phenomenon of respondents spending too little time on tests, not thinking about the right answer and getting more wrong answers. Time limited answers are primarily used for psychological reaction tests such as Stroop test for mental rotation.

Weaver tested extended versus untimed testing and the normally achieving readers performed similarly in the timed and untimed conditions, respectively (Weaver 1993). All reading disabled respondents in the study had higher scores when given extra time (n=31 with learning disability and 40 without).

Stutts et al (1998) have analyzed the usefulness of five tests of cognitive function used to identify elderly drivers with increased risk of crash involvement and explore the connection between impaired cognitive function and increased crash risk. The study included the Trail Making Test parts A and B, the Short Blessed Orientation-Memory-Concentration test of cognitive impairment, a modified American Association of Retired Persons Reaction Time test,

and a timed Traffic Sign Recognition test (TSR). Data was collected from $n=3,238$ drivers at ages 65 and up, who applied for renewal of driver's license. None of the tests, including the timed test, were found effective as screening tools for identification of subsets of high-risk drivers.

In a thesis Huesman (2000) explores the validity of reading comprehension test scores and how it is affected by 20-minutes extended time limit given to those not completing within standard time. The study was based on samples of $n=526$ and found little difference. Two-factor model had best fit with standard time and one-factor best with with extended time.

Mullane & McKelvie (2000) tested english and french language with the Wonderlic Personnel Test among $n=133$ students with and without 12-minutes time limit and found highest increase in score with the french version when time constraints were removed.

Lesaux et al (2006) examined the effect of giving extra time to fill out a reading comprehension test among $n=64$ adults with reading disabilities. All participants took the test in both timed and untimed conditions. The study suggests extra time given during test is appropriate.

In a meta-analysis Voyer (2011) found 36 gender difference effect sizes based on 26 studies with a total sample of $n=3081$ on time limit constraints in tests. Voyer found linear but not homogenous effect weighted cohen's d of 0.7 when tests are administered with time constraints.

To sum up, timed IQ-tests have shown high validity (Elliott et al 2001; Bird et al 2004), insignificant difference of timed and untimed questionnaire (Caudery 1990), significantly higher test score with extra time given (Weaver 1993, Lesaux et al 2006), usefulness of five cognitive tests, one of them timed, were found useful for identifying elderly drivers with increased crash

risk (Stutts et al 1998), timing changed factor structure of a reading comprehension test (Huesman 2000), language (french versus english) confounding test score when timing was removed (Mullane & McKelvie 2000), gender confounding effect of timing (Voyer 2011).

Uploading files. Uploading a file can be used to attach a CV, a picture or other documentation used as part of an analysis. Without this option it would be necessary to ask respondents to upload documents elsewhere. No studies testing this IIF characteristic were identified but a meaningful design would be to text upload rate compared to upload rates external to the survey software used.

Visual analogue scales (sliders, graphical scales, semantic differential scales). 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 first had several descriptors along the line and later VAS versions only had endpoint descriptors. It is also sometimes referred to as semantic scales or semantic differential scales. These variations have been seen: several descriptors versus endpoints, stepless versus discrete, chromatic versus one colour, with and without graphics (such as smiley), with or without a marked neutral midpoint, have a default starting position or not and be vertical or horizontal. Each of these characteristics can be seen in various combinations although stepless slider with endpoint descriptors is the most common. Since the combinations are many it is difficult to cover evidence separately on each trait. Chyung et al (2018 II) and other researchers have pointed out the confusion of scale type names.

Champney & Marshall (1939) conducted a study rating parent behavior ($n=240$) using graphic rating scales with a varying number of intervals between 2 and 90 with two points, resulting in $r=0.56$ for 2 points and $r=0.76$ for 90 points. Reliability increased with the number of points on the scale, primarily rising to around 12 points and with stable r at higher number of points.

In a study by Matell & Jacoby (1972) 50 respondents were asked to respond by different scales ranging between two and 19 points in a total of 18 Likert type items. Matell & Jacoby tested proportion of scale used, proportion of uncertain response and mean testing time in minutes. The proportion of the scale used by respondents varied between 0.53 and 0.94, proportion of uncertain response was between 0.03 and 0.21 and mean testing time was between 7 and 11. Higher testing time was seen with higher number of points but otherwise no clear pattern seen. Although Likert type item is not what is covered here, the evidence suggests that near-stepless scales have higher response time.

Joyce et al (1975) compared a VAS scale and a 4-point Likert scale and concluded that VAS was more accurate, reliable and sensitive than Likert in the measurement of chronic pain ($n=30$).

Scott & Huskisson (1976) tested uniformity of distributions with six types of visual analogue and graphic rating scales, one vertical with only endpoints, one vertical with words, one vertical with numbers 1 – 20, and then horizontal versions of these three (different patient populations, $n=600$) in different variations in eight experiments. With all formats they found distributions not significantly different from uniform distribution, except with vertical numbered rating across two samples.

According to a study by Downie et al (1978) based on n=100 participants, good correlation between four types of rating scales for pain assessment were found. A 11-point numerical scale performed better than a 4-point descriptive scale and continuous visual analogue scale. Downie et al concluded that they measure the same underlying construct regardless of scale type.

Scott & Huskisson (1979) tested vertical visual analogue scales against horizontal scales among n=100 rheumatology outpatients. Each respondent got both scales and the study did not show significant difference ($\chi^2=2.92$, $p>0.1$) between the distributions and extremely high correlation between scores ($r=0.99$, $p<0.001$).

Littman et al (1985) tested relative performance between three single item analgesic rating scales, one VAS and two verbally expressed scales (verbal pain intensity and verbal pain relief) in n=1,497 patients and found strong correlations and minimal difference in sensitivity, yet overall verbal relief scale had highest sensitivity and is therefore recommended.

Aun et al (1986) tested usability of a linear analogue score (n=50) to be used among patients for pain assessment and found less error with vertical than horizontal presentation.

Guyatt et al (1987) compared VAS and 7-point Likert and concluded (n=28) based on pre-post test and comparison based on standardization to a 0-10 scale, that the difference was not statistically significant.

Nyrén et al (1987) compared a five-point Likert scale to a VAS scale in n=32 patients in a randomized controlled study and continuous four times daily measurement, concluding equal compliance and reliability.

Gift (1989) compared VAS and Likert on a social anxiety scale with completion by patients and doctors, concluding Spearman's ρ correlation 0.661 (n=50).

Jaeschke et al (1990) compared responsiveness and validity of a seven-point version of a Likert scale and a visual analogue scale to measure quality of life in chronic heart failure patients in a randomized controlled trial of administering digoxin versus placebo. Having Likert scale or VAS scale first was selected randomly among 20 patients. Jaeschke et al conclude that the scales had similar responsiveness and validity and that the VAS scale was less easily communicated and administered.

Tesler et al (1991) conducted a study to test validity, reliability, ease of use, preference and lack of background bias (age, gender, ethnic) on pain intensity scales among children aged 8 to 17 (n=1,223). Tesler et al tested a word-graphic rating scale, a visual analogue scale, a graded-graphic rating scale, a magnitude estimation scale and a color scale. The study suggests the word-graphic rating scale is used instead of the other tested scale types.

Among n=24 males, Neely et al (1992) tested correlation between a Likert and VAS version of a perception scale (Category Ratio Scale), concluding the standard CR-10 (nonlinear pain scale) was slightly more efficient, possibly due to greater discrimination ability at high scores.

Krosnick & Berent (1993) conducted eight experiments. Higher reliability was seen in a sample of n=494 when all the points on a scale are marked with words than if only some of them are (Krosnick & Berent, 1993). Partially labeled items took more time to answer (1 minute and 43 seconds versus two minutes and 13 seconds).

Among n=168 medical students, Pfennings et al (1995) tested functional health with a Likert and VAS version, finding higher responsiveness based on higher variance with the VAS version.

Brunier & Graydon (1996) find correlation between a Likert based fatigue sub-scale and a single-item VAS ($r=0.80$) among $n=43$ patients.

Hyland & Sodergren (1996) developed a global quality of life scale characterized by eight quantifiers on a VAS scale versus numbered VAS and a partially labeled scale and compared preference among several small subsamples of undergraduates, nursing students, patients and elderly residents (total $n=202$) for this in and a scale with 12 other scales. Four were categorial and four were VAS type scales. Correlations between self-rating and scenario ratings showed small significant differences. Intercorrelations were higher with the developed scale than within the group of categorial and VAS scales. The developed scale was most preferred among undergraduates and elderly respondents. Preferred starting point for the scale button was 20 or 100 rather than 0.

Bellamy et al (1999 I) tested eight self-rating pain scales among $n=214$ patients after a washout and a following oxaprozin treatment period. The scales were different versions of Likert and VAS, and the types contained continuous chromatic analogue, meaning a VAS scale that goes from one colour in one end to another colour in the other end. All scales detected improvements but differed in responsiveness. They concluded that standard VAS and Likert were more responsive than more complex measures, such as the chromatic analogue version.

In a different study by Bellamy et al (1999 II) with $n=124$ patients, the same conclusion was reached.

Paul-Dauphin et al (1999) compared six types of VAS, namely a classic standard VAS, one with a marked middle point, graphic rating, graduated, graduated-numbered and numerical rating, and horizontal scale was tested towards vertical. The scales were used to test ear, nose and throat infection at 2-month intervals in 870 subjects (77% response rate). Paul-Dauphin et al

compared variance and means and concluded that scale type influenced floor effect yet not the mean scores but had a significant effect on cross-sectional measurement of symptoms. Higher percentages of zero ratings was observed on graduated-numbered VAS and numerical rating scales, and lower percentages were seen on graphic rating scales and graduated VAS.

Grant et al (1999) tested three subjective pain rating scales, VAS, Borg CR-10 and five-point Likert against objective physiological fatigue among n=23 active males after steady-state exercise and concluded VAS performed best on reproducibility and sensitivity for breathlessness and Borg CR-10 was more sensitive for general fatigue.

Myles et al (1999) tested linearity of the pain VAS scale by asking n=52 post-operative patients of pain intensity, then treated with IV fentanyl and asked to consider when they had half as much pain. They found linearity, thus concluding it can be treated as a ratio scale.

Cook et al (2001) studied how reliability and how coarseness in slider scoring affects reliability in unnumbered graphic rating scales and Likert scales, using a select sample based on diversity of n=4,407. In this study the graphic rating scale was presented as a continuous line between two antonyms, and respondents were asked to draw a line through the continuum at the point most indicative of their view. They found alpha coefficients at same levels across IIFs.

Reips & Tourangeau (2002) conducted a web experiment with n=334 responses and manipulation of scale type with numbered labels or not along with other manipulations. The study did not reveal a significant effect of scale type manipulation.

Bolognese et al (2003) compared correlation, consistency and normality (Shapiro-Wilk test) of pain assessment by 100 mm VAS and Likert in a clinical trial of rofecoxib treatment versus placebo and found high correlation and normality but recommend Likert due to easier patient interpretation and administration.

Based on a n=98 patients in a 21-week randomized trial with naproxen sodium versus placebo, Averbuch & Katzper (2004) tested visual analog scales versus 5-point Likert scale and found correlation of >0.995 in both arms of the study.

Castle & Engberg (2004) randomly distributed six survey instruments with 2,450 responses (78.5 % response rate). They were designed with 5-item Likert, 5-item satisfaction, 5-item valuation, 4-item Chernoff faces and 10-item VAS. Chernoff faces were least liked and 10-item VAS was liked the most (39% preferred this design). The coefficient of variance was highest for the 10-item VAS, suggesting lower ceiling effect.

With n=85 chronic pain patients Cork et al (2004) tested VAS versus verbal rating and found a correlation of 0.906 ($p<0.0001$).

Flynn et al 2004 compared assessments with multi-item 7-point Likert scales and VAS scales in the 16 item Functional Dimensions of Coping scale by Ferguson and Cox, 1997 with four subscales on 102 under- and postgraduate psychology students answering both Likert and VAS versions of the scale. On all four subscales Likert scales yielded higher levels of functional coping and wider range of transactional coping patterns. The VAS version explained a slightly greater amount of the total variance (64% vs 61%). For each subscale correlations between Likert and VAS version was between 0.76 and 0.91. Kurtosis was higher on VAS scales. Based on these results Flynn et al recommend using Likert scales.

Van Laerhoven et al (2004) compared a Likert scale, an unnumbered VAS scale and a numbered VAS scale among 120 children to test preference and usability. Each child filled out three questionnaires with seven items in each questionnaire and the order of the questionnaire distributed randomly with the type of scale being the only difference. The children afterwards

rated the questionnaires and preferred the Likert scale and found this easier to fill out. Results between scales correlated strongly ($\rho=0.67-0.90$, $p<0.05$).

In a cross-sectional study based on $n=805$ participants, Hasson & Arnezt (2005) tested single-item VAS against five-item Likert construct on validity and reliability, namely intraclass correlation, marginal homogeneity test and found moderate to strong correlations (.42-.91) between VAS and Likert items and differences in seven out of eleven subscales (stress and healthrelated questions). In conclusion Hasson & Arnezt find, that single-item VAS is a useful substitute of multi-item Likert scales with uniform construct and that there is no conclusive evidence about the superiority of one scale before the other.

Myles & Urquhart (2005) did a test with half, baseline and double dosis of analgesic medication using VAS-scales on $n=22$ patients with severe acute pain. They found that VAS is distributed with linearity based on scores versus the varied medication.

Couper et al (2006) conducted experiments with several VAS features versus radio buttons, the use of midpoint versus no midpoint and numbered versus unnumbered scale points in a sample of $n=1,427$ respondents (10.8% response rate). The response distribution did not differ between scale types, but VAS generally had higher completion time and higher rate of missing data.

Friborg et al (2006) conducted a survey with completion by $n=334$ university students. The questionnaire had both Likert and Semantic differential items. Based on latent class analysis it was concluded that the semantic differential version had better data fit than the Likert scale. Friborg et al conclude that semantic differential formats could reduce acquiescence bias without lower psychometric quality. Internal consistency was lower for the semantic differential items than with Likert.

Hofmans & Theuns (2008) tested the effect of end anchors on linearity in VAS. In a 3 x 4 leadership styles versus paylevel factorial design and a 5 x 5 face attractiveness (picture) versus personality trait factorial design they tested two types of anchors (endpoints) based on n=62 in experiment 1 and n=62 in experiment 2 (same group) on undergraduates. The interventions were different imagined scenarios at the beginning of the questionnaire designed to allow for interventions of different dosage. For each of the two experiments participants got one of two conditions: a scale with fixed endpoints or a scale where respondents defined their own endpoints, also called a self-anchoring scale (Kilpatrick & Cantril 1960 in Hofmans & Theuns 2008). One could argue that this is a variation of VAS or that it is an IIF of its own kind. From a technological viewpoint the survey software must support the feature of allowing respondents to define anchors. This setup was used because earlier research has discussed whether endpoints are too extreme or not extreme enough which affects the distribution and at the same time different anchors can have different meanings to respondents. The results suggest parallelism for both VAS with equal weights and the ones with self-anchoring.

Reips & Funke (2008) tested different sizes in pixels of VAS scales on a computer screen, and had n= 439 respondents click at 20%, 50%, 1/5, 1/2 and so on, on the scale. Tendency to match actual value was stronger in short versions (150-400 pixels) of VAS, but difference in size is robust to measurement.

In an experimental study of n=436, Krebs & Hoffmeyer-Zlotnik (2010) asked respondents to reply using a scale from 0 to 8 and labeled endpoints and found an order effect and different distribution with higher number of observations in the middle with negative-positive format.

With $n=18$ physicians, Celenza & Rogers (2011) compared intraclass correlation and reliability measures (test-retest and Cronbach's alpha) for 100 mm VAS and Likert scales in a prospective crossover-study with cognitive self-assessment and found high intraclass correlation, highest test-retest and slightly lower alpha values for Likert.

Funke et al (2011) tested slider and Likert scales in horizontal and vertical version among 779 respondents in a 2x2 factorial experiment. Slider scales had higher drop-out rates and higher response time. Especially, problems with slider scales were seen among respondents with less than average education level.

Funke & Reips (2012) tested 5-point Likert scales against VAS, finding that $n=268$ respondents adjusted their ratings more often when using VAS, to maximize precision, with beneficial effect on data quality. They found no differences in means, dropout-rate, non-response or response time.

Abend et al (2014) used two experiments on undergraduates ($n=99+73$) to assess reliability, convergent validity and discriminant validity (experiment 1) and sensitivity (experiment 2) of single-item VAS compared to four Likert based scales and subscales of different anxiety-related aspects (trait anxiety, social anxiety, depression and state anger) with a total of 68 items. They found low to medium test-retest reliability, high correlation with STAI-State Likert based trait-anxiety scale (assumed gold standard), confirmed discriminant validity and sensitivity and Abend et al therefore recommend single-item VAS as complement to Likert-based anxiety scales in research.

Menold et al (2014) used eye-tracking technology to explore influence of scale length and labeling on order bias using two samples of $n=47$ each. The first experiment was based on varying whether four Likert items with seven response options had only endpoint labels or seven

labels and were chosen based on identifying a single factor in the ALLBUS 2008 survey and did not reveal any significant difference. With the second experiment results showed that 74% read both labels in the endpoint labels format opposed to format with seven labels where only 12.5% read all labels.

Rausch & Zehetleitner (2014) compared VAS and four-point discrete item (unspecified whether it is ordinal Likert or scale-type) with 20 participants on assessment of vision. Rausch & Zehetleitner used random dot kinematograms (a picture with random white dots moved on black background) and participants then assessed sense of movement with VAS and discrete item. They did not identify an effect of scale type.

Based on $n=187+144$ patients in an observational study and a randomized controlled trial with three arms consisting of three physiotherapist intervention types (specialized, non-specialized and group-based), Harland et al (2015) compared a VAS and Likert version of a back-pain scale. Convergent and discriminant validity was tested along with ceiling effect and missing values. They did not find significant differences in any of these measures, except fewer missing values in the Likert version.

Roster et al (2015) tested slider and categorical five-point radio-button response formats in a randomized controlled study among $n=39$ business managers (60 % response rate) to see if they give different measurements. Roster et al did not identify significant differences. Response rate was insignificant ($z = 0.288$, $p<0.78$). Response rate for slider and radio button, at average 5.93 and 4.95 respectively, was not significant ($t=0.593$, $p<0.56$). Two out of 12 items had differences in sensitivity score and no significant difference in score mean values ($p<0.09$). Finally, no significant difference was found regarding likelihood to respond, based on five randomly selected items ($p>0.15$).

Voutilainen et al (2015) tested patient satisfaction in quasi-experimental design with VAS vs Likert scales and n=150 respondents. VAS was less vulnerable to bias from confounding, avoided ceiling effect better and took shorter time to fill out.

Funke (2016) differs between discrete slider scales and VAS scales, where these are otherwise often seen as the same thing. Funke makes the point, that with sliders, respondents are asked to make a mark, whereas with VAS, respondents are asked to drag a slider button. Based on 1,807 responses he compared slider scale, VAS with graphics and radio button with three, five or seven response options. Funke found that slider scales affect response rate negatively, especially on mobile devices. Sliders led to 3 times higher dropout and confusion by respondents. Also, the difference of drag-and-drop and point-and-click affects data.

Hilbert et al (2016) tested the dutifulness NEO-PI-R subscale with a dichotomous, 5-point Likert and 100 mm VAS scale with repeated measures design on n=866 respondents. Results were analyzed on reliability with highest levels for VAS on Cronbach's Alpha and McDonald's omega, although insignificant for omega. Criterion validity by correlation with unexcused absent days showed slightly higher correlation with VAS, and convergent validity coefficients in a regression SEM pathmodel was used to test the relationship between scales showing high correlation between the scale types and highest correlations seen when same scale was measured with different response formats. Higher reliability was found with VAS and finally, ceiling effect was seen with dichotomous scale. Hilbert et al recommend choosing scale type based on practical considerations.

Maloshonok & Terentev (2016) completed four separate factorial randomized experiments, one of these related to innovative item formats, namely radio button versus slider or text box, concluding a reduction in number of respondents who choose "don't know" when radio

buttons are used. The overall sample was $n=3088$ but how many answering the radio button, slider and text box experiment was not specified.

Chyung et al (2018) conducted a literature study exploring research articles about four scale types for pain assessment, each with "no pain" and "worst pain possible" as endpoints but otherwise different traits: a Likert scale, a stepless semantic differential scale, semantic scale with 6 named steps and a scale with 11 numbered points. A total of 17 articles with an unspecified total sample size were identified but the study was inconclusive. It could be noted here that the visual analogue scale (VAS) within health science and the field of pain measurement is often considered as a scale specific for pain assessment, but that VAS is elsewhere not considered as a pain scale specifically.

Chyung et al (2018 II) synthesize studies that have studied data quality in continuous rating scales compared to other scale types, in other words a study like this one, except specifically on continuous scales. Chyung et al conclude that there is no one type being better than the other, between continuous and discrete scales and that researchers should choose based on usability. A total sample sized was not reported but the relevant studies have been included here. A table of pros and cons of discrete and continuous rating scales is presented. With discrete items are easier to use, generate higher completion rates, likely generate ordinal-type data, thereby limiting statistical usage. Continuous items allow more precise answers, improve reliability and allow for more statistical analysis. Caution should be used with paper-based continuous VAS-type items due to time consumption, difficulty used on mobile devices and with data loss as a result thereof. Chyung et al consider the conclusions of each included study but not the size of the studies and does not put focus on the methods used.

Bosch et al (2019) conducted two multitrait-multimethod experiments based on the representative Norwegian Citizen Panel survey (n=4,689), wave 7, divided in four subsets, and found similar data quality in slider and radio button if smartphone respondents are part of the sample. Without smartphone response the slider should be placed in the middle initially rather on left side.

In a systematic review of lower back pain rating scales, Chiarotto et al (2019) tested the standardized visual analogue scale against a numeric version where patients can select a number between the endpoints instead of marking a point on a line, and finally the BPI-PS 11-point pain severity scale. 23 studies in six languages were included with a total sample of n=4,083 (not overlapping other studies described here) of which 5 were on content validity and 18 focused on other aspects, and the result was inconclusive. According to Chiarotto the evidence was of low quality.

To sum up, higher reliability with higher number of points on scale (Champney & Marshall 1939), higher test time with higher number of points (Matell & Jacoby 1972), more accurate, reliable and sensitive (Joye et al 1975), VAS versus numerical and descriptive scale did not affect construct (Downie et al 1978), strong correlation and minimal difference in sensitivity for VAS against verbal scales (Littman et al 1985), 11-point numerical performing better than 4-point descriptive and continuous VAS (Downie et al 1978), VAS was less easily communicated and administered (Jaeschke et al 1990), word-graphic rating preferred over VAS (Tesler et al 1991), nonlinear scale more efficient than VAS (Neely et al 1992), higher responsiveness with VAS than Likert (Pfennings et al 1995), floor-effect in comparison of VAS versus other rating scales, VAS distribution linear (Myles et al 1999), lower ceiling effect in VAS than Likert and Chernoff faces (Castle & Engberg 2004), VAS explaining more variance than Likert but kurtosis

was higher with VAS (Flynn et al 2004), Likert preferred before numbered and unnumbered VAS (van Laerhoven 2004), VAS higher completion rate and more missing data than radio buttons (Couper et al 2006), better data fit with VAS than Likert (Friborg et al 2006), highest test-retest and lower alpha values for Likert compared to VAS (Celenza & Rogers 2011), VAS was found reliable, valid and sensitive (Abend et al 2014), VAS less vulnerable to confounding than Likert, avoided ceiling effect better and took shorter time (2015), VAS had lower response rate, especially on mobile device and higher dropout and sense of confusion (Funke 2016), higher reliability with VAS than Likert and dichotomous scale (Hilbert et al 2016), less “don’t know” with Likert than VAS (Maloshonok & Terentev 2016).

Insignificant difference was found between VAS and 7-point Likert on pre-post test (Guyatt et al 1987), insignificant difference in compliance between VAS and five-point Likert (Nyrén et al 1987), correlation between VAS and Likert (Gift 1989), VAS and seven-point Likert similar responsiveness and validity (Jaeschke et al 1990), correlation between Likert and VAS (Brunier & Graydon 1996, high intercorrelations between categorical and VAS (Hyland & Sodergren 1996), subscale performance varied by VAS versus Likert (Grant et al 1999), reliability at same level between VAS and Likert (Cook et al 2001), high correlation between VAS and Likert (Bolognese et al 2003), high correlation between VAS and Likert (Averbuch & Katzper 2004), high correlation between VAS and verbal rating (Cork et al 2004), strong correlation between VAS and Likert (van Laerhoven 2004), strong correlations between VAS and Likert but differences in subscales (Hasson & Arnezt 2005), VAS is linear (Myles & Urquhart 2005), no difference in distribution between VAS and radio buttons (Couper et al 2006), different distributions between numbered and standard VAS (Krebs & Hoffmeyer-Zlotnik 2010), high intraclass correlation, higher drop-out rate and response time with VAS compared to

Likert (Funke et al 2011), no difference in means, dropout-rate, non-response or response time between VAS and Likert (Funke & Reips 2012), vision assessment did not affect VAS and discrete response differently (Rausch & Zehetleitner), VAS versus Likert did not affect convergent and discriminant validity and ceiling effect and fewer missing values with Likert (Harland et al 2015), response time difference and likelihood to respond between VAS and Likert insignificant (2015), high correlation in pathmodel between VAS, Likert and dichotomous version of a scale (Hilbert et al 2016), discrete items easier to use, higher completion rates (Chyung et al 2018 II), study inconclusive on whether VAS or patients selecting number (Chiarotto et al 2019).

Distribution unconfounded by population (Scott & Huskisson 1976), vertical numbered VAS was significantly different from uniform distribution (Scott & Huskisson 1976), no significant difference and extremely high correlation between vertical and horizontal VAS (Scott & Huskisson 1979), less error with vertical than horizontal VAS (Aun et al 1986), higher reliability when scale is marked with words instead of only endpoint words (Krosnick & Berent 1993), 20 was preferred to 0 as starting point on a 0-100 VAS and age affected preference towards VAS (Hyland & Sodergren 1996), chromatic VAS (one colour in one end and another colour in the other end) was less responsive than classic VAS and Likert (Bellamy et al 1999 I + Bellamy et al 1999 II), higher percentage of zero-rating when VAS was numbered or graduated-numbered (Paul-Dauphin et al 1999), difference between numbered and unnumbered VAS was insignificant (Reips & Tourangeau 2002), parallelism found for VAS and self-anchoring VAS (Hofmans & Theuns 2008), validity higher with short VAS (Reips & Funke 2008), education level confounded drop-out and response time (Funke et al 2011), fewer respondents read labels when there are several labels (Menold et al 2014).

Visual analogue scale, self-anchoring. Self-anchoring means the respondent makes its own endpoint categories instead of prescribed endpoints made by the researcher. As mentioned under VAS-scales, Hofmans & Theuns (2008) tested self-anchoring scales (n=62) and results suggest parallelism for both VAS with equal weights and the ones with self-anchoring.

Voice assistance / sound. IIF with sound can be used for reading questions out loud as a supportive tool for respondents who are visually impaired or who have limited reading level, such as children, but could also be used simply to enhance the survey user experience or even as a stimulus like pictures that exemplifies or describes a case. Just like studies have shown that voice characteristics have relevance for data in CATI interviews, it is plausible that voice assistance in electronic surveys could influence data too (van der Vaart et al 2005).

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 cross-over trial of telephone versus written test among 64 subjects they found the test valid and reliable.

Yost et al (2009) tested item difficulty of 97 English- and 134 Spanish-speaking participants to assess item difficulty with a Talking Touchscreen health literacy measure where respondents would get multiple choice answers read out loud when touching the screen. They also asked respondents to estimate which of four icons best told that questions could be read out loud with preference toward the icon of a man talking. Data was validated with Rasch-model, reducing items from 98 English and 127 Spanish items through several rounds of subtraction and addition to a final 90 items.

Sinadinovic et al (2011) compared an interactive voice response system to online version of a questionnaire among 1861 participants and found higher response rate with online survey, a difference in distribution (reports of drug and alcohol use) and a difference in reliability slightly favouring the online version.

With 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 paper-and-pencil method among n=51 veterans 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 among 261 participants (children and adolescents) Medin et al (2016) used voice-assisted cartoon character helping participants completion. Medin et al found acceptable correlations on all six measures.

Results

Table 1 shows the number of included studies, total sample size, overall conclusion and references for each IIF.

In figure 1 the total sample size for each IIF and the number of studies identified is illustrated. Measured by total sample size and number of studies the evidence on randomization and VAS is quite extensive, whereas the evidence with other IIFs is much less. This probably reflects the much higher use of randomization and VAS, but it also illustrates that a stronger research focus on the other IIFs would be sensible.

Table 1

Overview of evidence

Item type	Number of studies	N	Overall conclusion (order of evidence + for/against)	References
Battery / response matrix / grid	12	14,563	Inconclusive Better data with grey-out Text versus emoticons in grids inconclusive	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 / response matrix / grid, multientity scaling	3	2,204	Inconclusive	Wong & Teas 2001, Borkenhagen et al 2005, Sikkel et al 2014
Constant sum	3	2,925	Inconclusive Possible gains for concurrent display (Conrad 2005)	Louviere & Islam 2006, Conrad et al 2005, Skedgel et al 2013
Distance to nonsubstantive response	1	4,577	In favour of IIF	Tourangeau et al 2004
Drag-and-drop sorting	6	16,317	Inconclusive	Bennett & Sebrechts 1997, Timbrook 2013, Sikkel et al 2014, Kunz 2015, Timbrook & Moroney 2016
Draw function	2 (22)	3,391	Inconclusive	Björkstén 1999, Shulman 2000, Elliot & Papadopoulos 2012
Finishing a sentence	1	52	In favour of IIF	Bleland 1999
Image response, universal (smiley, emoticon and similar)	12	6,186	In favour of IIF Chernoff faces unfavoured	Buchanan & Niven 2002, Buchanan & Niven 2003, Castle & Engberg 2004, Buchanan 2005, Desmet 2005, Derham 2011, Emde & Fuchs 2012, Broekens & Brinkman 2013, Bosch & Revilla 2020
Image response, specific	12	6,012	Inconclusive	Rofé 2004, Shamir & Kark 2004, Balram & Dragičević 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	Inconclusive	Graybill & Heuvelman 1993, Escalante 1995, Couper et al 2004, Wall et al 2017
Image, dynamic / figural response	10	20,139	In favour of IIF	Martinez 1991, Bennett et al 1999, Bennett et al 2000, Lim et al 2006, Van Ooijen 2011, Wan & Henly 2012
Image, video	1	62	In favour of IIF	Kersting 2008
Interval item	0	0	-	-
Randomization, nominal and ordinal	35	101,665	In favour of IIF Multiple confounders identified	Mathews 1929, Payne 1949, Belson 1966, Carp 1974, Krosnick & Alwin 1987, Israel et al 1990, Chan 1991, Friedman et al 1993, Knäuper 1999, Barnette 2000, Weng & Cheng 2000, Bishop & Smith 2001, Attali & Bar-Hillel 2003, Nicholls et al 2006, Hofmans et al 2007, Rammstedt & Krebs 2007, Dillman 2009, Malhotra 2009, Toepoel et al 2009, Hartley & Betts 2010, Salzberger & Koller 2013, Stapleton 2013, Garbarski et al 2014, Höhne & Lenzner 2015, Maeda 2015, Liu & Keusch 2017, Yan et al 2018
Randomization, open ended response	0	0	-	-
Randomization with separate non-randomized category	1	436	In favour of IIF	Krebs & Hoffmeyer- Zlotnik 2010
Ranking	4	2,888	Unfavoured of IIF	Rokeach 1967, Rankin & Grube 1980, Alwin &

				Krosnick 1985, Krosnick & Alwin 1988, Maio 1996
Time limited answer	6	4,040	Inconclusive Confounders: gender	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 (sliders, graphical scales, semantic differential scales)	54	36,898	Inconclusive Confounders: vertical vs horizontal, starting point, respondent age, chromatic versus no colour, length	Champney & Marshall 1939, Matell & Jacoby 1972, Joyce et al 1975, Scott & Huskisson 1976, Downie et al 1978, Scott & Huskisson 1979, Littman et al 1985, Aun et al 1986, Guyatt et al 1987, Nyrén 1987, Gift 1989, Jaeschke et al 1990, Tesler et al 1991, Neely et al 1992, Krosnick & Berent 1993, Pfennings et al 1995, Brunier & Graydon 1996, Hyland & Sodergren 1996, Bellamy et al 1999 I, Bellamy et al 1999 II, Paul-Dauphin et al 1999, Grant et al 1999, Myles et al 1999, Cook et al 2001, Reips & Tourangeau 2002, Bolognese et al 2003, Averbuch & Katzper 2004, Castle & Engberg 2004, Cork et al 2004, Flynn et al 2004, Van Laerhoven et al 2004, Hasson & Arnezt 2005, Myles & Urquhart 2005, Couper et al 2006, Friborg et al 2006, Hofmans & Theuns 2008, Reips & Funke 2008, Krebs & Hoffmeyer-Zlotnik 2010, Celenza & Rogers 2011, Funke et al 2011, Funke & Reips 2012, Abend et al 2014, Menold et al

				2014, Rausch & Zehetleitner 2014, Harland et al 2015, Roster et al 2015, Voutilainen et al 2015, Funke 2016, Hilbert et al 2016, Maloshonok & Terentev 2016, Chyung et al 2018, Chyung et al 2018 II, Bosch et al 2019, Chiarotto et al 2019
Visual analogue scale, self-anchoring	1	62	Inconclusive	Hofmans & Theuns 2008
Voice assistance / sound	5	2,468	Inconclusive	Turvey et al 2012, Lam et al 2009, Medin et al 2016, Yost et al 2009, Sinadinovic et al 2011
Total	166	218,532		

Figure 1
Distribution of evidence

Conclusion

A total of 166 research articles were identified with data from 218,532 participants revealing aspects of 22 IIFs with 12 further subcategories. The type of evidence on data quality is quite varied and for some IIFs the evidence is supportive, for some it is not and for some it is inconclusive. With 6 IIFs the evidence was estimated in favour hereof, for 11 IIFs the evidence was inconclusive, 1 unfavoured and for 3 there was no evidence. With 6 IIFs potential confounders were identified. The evidence base was extensive for randomization and VAS, fair with drag-and-drop and dynamic image, and limited for battery, constant sum, drag-and-drop, finishing a sentence, image questions, video, interval items, randomization with open ended response, randomization with separate non-randomized category, ranking, time-limited answer, file upload, visual analogue scale with self-anchoring and voice assistance.

Discussion

The study suggests further research is needed where evidence is scarce. Where strong evidence exists, such as for randomization and VAS, further evidence is less needed. Further research about the strengths and weaknesses of methods validating IIFs is also suggested. The inconsistent taxonomy found within several IIFs could suggest a more in-depth discussion about suggestions for a more systematic taxonomy for future use. Based on research like this future survey software will hopefully offer more IIF options to allow for stronger survey research.

When the specific evidence is seen in relation to more general evidence, like Crabtree 2016 described in the beginning of this paper, there is a discrepancy between evidence suggesting strengthened data quality with the use of several simultaneous IIFs and the more

varied evidence with the use of one specific IIF. This interesting difference suggests that heterogeneity may influence validity. This suggests that specific IIF should not only be seen isolated, but that it may influence data quality positively when a combination of IIFs are used because the respondent is more engaged. At the same time this paper has also illustrated that data quality varies between IIFs.

The evidence revealed a tendency to different IIF traditions within different research fields, such as VAS often used for pain rating and health science, dynamic images and time-limit often used in cognitive tests and multientity scaling used frequently for market research. These traditions may be meaningful, but it may also be healthy to challenge survey design traditions within each research field. A broader use of IIF may potentially even allow for a more nuanced perspective on the social world.

Although the iterative method of chain sampling has strengths it also has weaknesses against systematic review. The present study could therefore be a study initiating more extensive systematic reviews within specific categories of IIF.

References

- Abend, R., Dan, O., Maoz, K., Raz, S. & Bar-Haim, Y. (2014). Reliability, validity and sensitivity of a computerized visual analog scale measuring state anxiety. *Journal of Behavior Therapy and Experimental Psychiatry*, 45, pp. 447–453.
- Alwin, D.F. & Krosnick, J.A. (1985). The measurement of values in surveys: A comparison of ratings and rankings. *Public Opinion Quarterly*, 49, pp. 535-552.
- Attali, Y. & Bar-Hillel, M. (2003). Guess Where: The Position of Correct Answers in Multiple-Choice Test Items as a Psychometric Variable. *Journal of Educational Measurement*, 40(2), pp. 109-128.
- Averbuch, M. & Katzper, M. (2004). Assessment of visual analog versus categorical scale for measurement of osteoarthritis pain. *Journal of Clinical Pharmacology*, 44, pp. 368–372.
- Aun C., Lam Y.M. & Collect B. (1986). Evaluation of the use of visual analogue scale in Chinese patients. *Pain* 25, pp. 215-21.
- Balram, S. & Dragićević, S. (2005). Attitudes toward urban green spaces: integrating questionnaire survey and collaborative GIS techniques to improve attitude measurements. *Landscape and Urban Planning*, 71(2-4), pp. 147-162.
- 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), pp. 361-370.
- Bell, D.S., Mangione, C.M. & Kahn, C.E. (2001). Randomized testing of alternative survey formats using anonymous volunteers on the world wide web. *Journal of the American Medical Informatics Association*, 8, pp. 616–620.

- Bellamy, N., Campbell, J. & Syrotuik, J. (1999 I). *Comparative study of self-rating pain scales in rheumatoid arthritis patients*. Retrieved November 12, 2000 from <http://www.librapharm.co.uk/cmro/vol@@15/2001/main.htm>.
- Bellamy, N., Campbell, J. & Syrotuik, J. (1999 II). *Comparative study of self-rating pain scales in osteoarthritis patients*. Retrieved November 12, 2000, from http://www.librapharm.co.uk/cmro/vol_15/2000/main.htm.
- Belson, W.A. (1966). The effects of reversing the presentation order of verbal rating scales. *Journal of Advertising Research*, 6 (December), pp. 30-37.
- Bennett, R.E., Ward, W.C., Rock, D.A., & Lahart, C. (1990). *Toward a framework for constructed response items*. Princeton, NJ: Educational Testing Service.
- Bennett, R.E. & Sebrechts, M.M. (1997). A Computer-Based Task for Measuring the Representational Component of Quantitative Proficiency. *Journal of Educational Measurement*, 34(1), pp. 64-77.
- Bennett, R.E., Morley, M., Quardt, D., Rock, D. A., Singley, M. K., Katz, I. R. & Nhouyvanisvong, A. (1999). Psychometric and cognitive functioning of an under-determined computer-based response type for quantitative reasoning. *Journal of Educational Measurement*, 36, pp. 233–252
- Bennett, R.E., Morley, M., Quardt, D. & Rock, D.A. (2000). Graphical modeling: A new response type for measuring the qualitative component of mathematical reasoning. *Applied Measurement in Education*, 13, pp. 303-322.
- Berg, I.A. & Rapaport, G.M. (1954). Response bias in an unstructured questionnaire. *The Journal of Psychology*, 38, pp. 475-481.

- Bird, C.M., Papadopoulou, K., Ricciardelli, P., Rossor, M.N. & Cipolotti, L. (2004). Monitoring cognitive changes. *Psychometric properties of six cognitive tests*, 43(2), pp. 197-210.
- Bishop, G.F. & Smith, A. (2001). Response-Order Effects and the Early Gallup Split-Ballots. *Public Opinion Quarterly*, 65, pp. 479-505.
- Björkstén, M.G., Boquist, B., Talbäck, M. & Edling, C. (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), pp. 325-330.
- Bolognese J.A., Schnitzer T.J. & Ehrich E.W. (2003). Response relationship of VAS and Likert scales in osteoarthritis efficacy measurement. *Osteoarthritis Cartilage* 11(7), pp. 499–507.
- Borkenhagen, A., Klapp, B.F., Schoeneich, F. & Brähler, E. (2005). Differences in body image between anorexics and in-vitro-fertilization patients - a study with Body Grid. *Psychosoc Med*, 2: pp. 1-9.
- Bosch, O.J., Revilla, M., DeCastellarnau, A. & Weber, W. (2019). Measurement Reliability, Validity, and Quality of Slider Versus Radio Button Scales in an Online Probability-Based Panel in Norway. *Social Science Computer Review*, 37(1), pp. 119-132.
- Bosch, O.J. & Revilla, M. (2020). Using emojis in mobile web surveys for Millenials? A study in Spain and Mexico. *Quality and Quantity*.
- Breland, H.M. (1999). *Exploration of an automated editing task as a GRE writing measure (RR-99-9)*, Princeton, NJ: Educational Testing Service.

- Broekens, J. & Brinkman, W.P. (2013). AffectButton: A method for reliable and valid affective self-report, *Human-Computer Studies*, 71(6), pp 641-667.
- Brunier, G., & Graydon, J. (1996). A comparison of two methods of measuring fatigue in patients on chronic haemodialysis: Visual Analogue versus Likert Scale. *International Journal of Nursing Studies*, 33, pp. 338–348.
- Buchanan, H. & Niven, N. (2002). Validation of a facial image scale to assess child dental anxiety. *Int J Paed Dent*, 12: pp. 47-52.
- Buchanan, H. & Niven, N. (2003). Further evidence for the validity of the Facial Image Scale. *International Journal of Paediatric Dentistry*, 13(5), pp. 368-369.
- Buchanan, H. (2005). Development of a computerised dental anxiety scale for children: validation and reliability. *British Dental Journal*, 199, pp. 359-362.
- Callegaro, M., Shand-Lubbers, J., & Dennis, J.M. (2009). *Presentation of a single item versus a grid: Effects on the vitality and mental health subscales of the sf-36v2 health survey* [paper presentation]. Annual meeting of the American Association for Public Opinion Research, Hollywood, FL, May.
- Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. (1973). The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology*, 57, pp. 15-22.
- Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, 56, pp. 81–105.

- Carp, F.M. (1974). Position effects in single trial free recall. *Journal of Gerontology*, 29, pp. 581-587.
- Castle, N.G. & Engberg, J. (2004). Response Formats and Satisfaction Surveys for Elders, *The Gerontologist* 44 (3), pp. 358-367.
- Caudery, T. (1990). The Validity of timed essay tests in the assessment of writing skills. *ELT Journal*, 44(2), pp. 122-131.
- Celenza A. & Rogers I.R. (2011). Comparison of visual analogue and Likert scales in evaluation of an emergency department bedside teaching programme. *Emergency Medicine Australasia* 23(1), pp. 68–75.
- Champney, H. & Marshall, H. (1939). Optimal refinement of the rating scale. *Journal of Applied Psychology*. 23(3), pp. 323-331.
- Chan, J. (1991). Response-order effects in Likert-type scales. *Educational and Psychological Measurement*, 51(3), pp. 531-540.
- Chesnut, J. (2008). Effects of using a grid versus a sequential form of the ACS basic demographic data. (working paper)
- Chiarotto, A., Maxwell, L.J., Ostelo, R.W., Boers, M., Tugwell, P. & Terwee, C.B. (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), pp. 245-263.
- Chyung, S.Y., Kennedy, M., Campbell, I. (2018). Evidence-based survey design: The use of ascending or descending order of response options. *Performance Improvement Journal*, 57(9), pp. 9-16.

- Chyung, S.Y., Swanson, I., Roberts, K., & Hankinson, A. (2018 II). Evidence-based survey design: The use of continuous rating scales in surveys. *Performance Improvement Journal*, 57(5), pp. 38-48.
- Conrad, F.G., Couper, M.P., Tourangeau, R. & Galesic, M. (2005). *Interactive feedback can improve the quality of responses in Web surveys* [paper presentation]. Conference of the American Association for Public Opinion Research (AAPOR), Miami Beach, Florida.
- Cook, C., Heath, F., Thompson, R.L. & Thompson, B. (2001). Score Reliability in Webor Internet-Based Surveys: Unnumbered Graphic Rating Scales versus Likert-Type Scales. *Educational and Psychological Measurement*, 61(4), pp. 697-706.
- Couper, M.P., Traugott, M. & Lamias, M. (2001). Web survey design and administration. *Public Opinion Quarterly*, 65, pp. 230–253.
- Couper, M.P., Tourangeau, R. & Kenyon, K. (2004). Picture This! Exploring Visual Effects in Web Surveys. *The Public Opinion Quarterly*, 68(2), pp. 255-266.
- Couper, M.P., Tourangeau, R. & Conrad, F.G. (2006). Evaluating the Effectiveness of Visual Analog Scales - A Web Experiment. *Social Science Computer Review*, 24 (2), pp. 227-245.
- Couper, M.P., Tourangeau, R., Conrad, F.G., & Zhang, C. (2012). The design of grids in web surveys. *Social Science Computer Review*, 31(3), pp. 322-345.
- Cork, R.C., Isaac, I., Elsharydah, A., Saleemi, S., Zavisca, F., & Alexander, L. (2004). A comparison of the verbal rating scale and the visual analog scale for pain assessment. *The Internet Journal of Anesthesiology*, 8.
- Crabtree, A.R. (2016). *Psychometric properties of technology-enhanced item formats: an evaluation of construct validity and technical characteristics. Dissertation.* doi:

10.17077/etd.922fbj4d

- Davies, J. & Brember, I. (2006). The Reliability and Validity of the 'Smiley' Scale. *British Educational Research Journal*, 20(4), pp. 447-454.
- Derham, P.A.J. (2011). Using preferred, understood or effective scales? How scale presentations effect online survey data collection. *Australasian Journal of Market & Social Research*, 19(2), pp. 13-26.
- Desmet, P. (2005). Measuring emotions: development and application of an instrument to measure emotional responses to products. *Funology: from Usability to Enjoyment*, pp. 111–123.
- Dickinson, T.L. & Zellinger, P.M. (1980). A comparison of the behaviorally anchored rating mixed standard scale formats. *Journal of Applied Psychology*, 65, pp. 147-154.
- Dillman, D.A., Brown, T.L., Carlson, J.E., Carpenter, E.H., Lorenz, M.S. & R.L. Songster, R.L. (1995). Effects of Category Order on Answers in Mail and Telephone Surveys. *Rural Sociology*, 60, pp. 674–687.
- Dillman, D.A. & Christian, L.M. (2005). Survey mode as a source of instability in responses across surveys. *Field Methods*, 17(1), pp. 30-52.
- Dolan, R.P., Goodman, J., Strain-Seymour, E., Adams, J. & Sethuraman, S. (2011). Cognitive lab evaluation of innovative items in mathematics and English/language arts assessment of elementary, middle, and high school students: Research Report. Iowa City: Pearson. Retrieved from http://www.pearsonassessments.com/hai/images/tmrs/Cognitive_Lab_Evaluation_of_Innovative_Items.pdf.

- Downie W.W., Leatham P.A., Rhind, V.M. (1978). Studies with pain rating scale. *Ann Rheum Dis*, 37: pp 378-81. <https://www.ncbi.nlm.nih.gov/pubmed/686873>
- Downing, S.M. & Haladyna, T.M. (2006). *Handbook of Test Development*. New Jersey: Taylor & Francis.
- Elliott, J., Lee, S.W. & Tollefson, N. (2001). A reliability and validity study of the dynamic indicators of basic early literacy skills--modified. *School Psychology Review*, 30(1), p. 33.
- Elliot, S. & Papadopoulos, N. (2012). *Beyond Tourism Destination Image: Mapping country image from a psychological perspective* [paper presentation], 2012 ttra International Conference. Retrieved from:
<https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1722&context=ttra>
- Emde, M. & Fucks, M. (2012). Exploring Animated Faces Scales in Web Surveys: Drawbacks and Prospects. *Survey Practice*, 5(1), pp. 1-6.
- Escalante, A., Lichtenstein, M.J., White, K., Rios, N. & Hazuda, H.P. (1995). A method for scoring the pain map of the McGill pain questionnaire for use in epidemiologic studies. *Aging Clinical and Experimental Research* 7, pp. 358-366.
- Fisher, R. (1926). "The Arrangement of Field Experiments". Journal of the Ministry of Agriculture of Great Britain. London, England: Ministry of Agriculture and Fisheries. 33, pp. 503–513.
- Flynn, D., van Schaik, P. & van Wersch, A. (2004). A Comparison of Multi-Item Likert and Visual Analogue Scales for the Assessment of Transactionally Defined Coping Function, *European Journal of Psychological Assessment*, 20(1), pp. 49-59.
- Freyd, M. (1923). The graphic rating scale. *Journal of Educational Psychology*, 14, pp. 83-102.

Friborg, O., Martinussen, M. & Rosenvinge, J.H. (2006). Likert-based vs. semantic differential-based scorings of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience. *Personality and Individual Differences*, 40(5), pp. 873-884.

Flynn, D., Schaik, P. & Wersch, A. (2004). A Comparison of Multi-Item Likert and Visual Analogue Scales for the Assessment of Transactionally Defined Coping Function. *European Journal of Psychological Assessment*, 20(1), pp. 49-58.

Funke, F. & Reips, U.D. (2006). Visual Analogue Scales in Online Surveys: Non-Linear Data Categorization by Transformation with Reduced Extremes. 8th International Conference GOR06

Funke, F., Reips, U.D., Thomas, R.K. (2011). Sliders for the Smart: Type of Rating Scale on the Web Interacts With Educational Level. *Social Science Computer Review* 29(2), pp. 221-231.

Funke, F., & Reips, U.D. (2012). Why semantic differentials in web-based research should be made from visual analogue scales and not from 5-point scales. *Field Methods*, 24(3), pp. 310–327.

Funke, F. (2016). A Web Experiment Showing Negative Effects of Slider Scales Compared to Visual Analogue Scales and Radio Button Scales. *Social Science Computer Review*, 34(2), pp. 244-254.

Friborg, O., Martinussen, M. & Rosenvinge, J.H. (2006). Likert-based versus semantic differential-based scorings of positive psychological constructs: a psychometric

- comparison of two versions of a scale measuring resilience. *Personality and Individual Differences*. 40(5), pp. 873–884.
- Friedman, H.H., Herskovitz, P.J., & Pollack, S. (1993). *The biasing effects of scale-checking styles on response to a Likert scale*. In Proceedings of the American statistical association annual conference: Survey research methods, 2, pp. 792-795.
- Galesic, M., Tourangeau, R., Couper, M.P., & Conrad, F.G. (2007). *Using change to improve navigation in grid questions* [paper presentation]. General Online Research Conference (GOR'07), Leipzig.
- Garbarski, D., Schaeffer, N.C. & Jennifer, D. (2014). The Effects of Response Option Order and Question Order on Self-Rated Health. *Quality of Life Research* 24 (6), pp. 1443–53.
- Gift, A.G. (1989). Visual Analogue Scales: Measurement of a subjective phenomena. *Nursing Research*, 38, pp. 268–288.
- Grant, S., Aitchison, T., Henderson, E., Christie, J., Zare, S., McMurray, J. & Dargie, M.D. (1999). Acomparision of the reproducibility and the sensitivity to change of Visual Analogue Scales, Borg Scales, and Likert Scales in normal subjects during submaximal exercise. *Chest*, 116, pp. 1208–1217.
- Graybill, D. & Heuvelman, L.R. (1993). Validity of the Children's Picture-Frustration Study: A Social-Cognitive Perspective. *Journal of Personality Assessment*, 60(2) pp 379-389.
- Guyatt, G.H., Townsend, M., Berman, L.B. & Keller, J.L. (1987). A comparison of Likert and Visual Analogue Scales for measuring change in function. *Journal of Chronic Disability*, 40, pp. 1129–1133.

- Haladyna, T.M., Downing, S.M. & Rodriguez, M.C. (2010). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education* 15(3), pp. 309-333.
- Harland N.J., Dawkin, M.J. & Martin, D. (2015). Relative utility of a visual analogue scale vs a six-point Likert scale in the measurement of global subject outcome in patients with low back pain receiving physiotherapy. *Physiotherapy* 101(1), pp. 50–54.
- Hartley, J. & Betts, L. (2010). Four layouts and a finding: The effects of changes in the order of the verbal labels and numerical values on Likert-type scales. *International Journal of Social Research Methodology*, 13(1), pp. 17- 27.
- Hasson D. & Arnezt B.B. (2005). Validation and findings comparing VAS vs. Likert scales for psychosocial measurements. *International Electronic Journal of Health Education* 8(1), pp. 178–192.
- Hayes, M.H.S., & Patterson, D.G. (1921). Experimental development of graphic rating method. *Psychological Bulletin*, 18(2), pp. 98–99.
- Hilbert, S., Küchenhoff, H., Sarubin, N., Nakagawa, T.T. & Bühner, M. (2016). The influence of the response format in a personality questionnaire: An analysis of a dichotomous, a Likert-type, and a visual analogue scale. *TPM-Testing, Psychometrics, Methodology in Applied Psychology*, 23(1), pp. 3-24.
- Hofmans, J., Theuns, P., Baekelandt, S., Mairesse, O., Schillewaert, N. & Cools, W. (2007). Bias and changes in perceived intensity of verbal qualifiers effected by scale orientation. *Survey Research Methods*, 1(2), pp. 97- 108.
- Hofmans, J., & Theuns, P. (2008). On the linearity of predefined and self-anchoring visual analogue scales. *British Journal of Mathematical and Statistical Psychology*, 61, pp. 401–

413.

Höhne, J.K. & Lenzner, T. (2015). Investigating response order effects in web surveys using eye tracking. *Phihologija*, 48(4), pp. 361-377.

Holbrook, A.L., Krosnick, J.A., Moore, D. & Tourangeau, R. (2005). Response Order Effects In Dichotomous Categorical Questions Presented Orally: The Impact of Question and Respondent Attributes. *Public Opinion Quarterly* 71(3), pp. 325-348.

Holbrook, A.L., Krosnick, J.A., Moore, D., & Tourangeau, R. (2007). Response order effects in dichotomous categorical questions presented orally. *Public Opinion Quarterly*, 71(3), pp. 325–348.

Huber, S., Hess, F., Donkers, J., Heid, J., De Leng, Bas, Huwekdiek, S. & Haag, M. (2009) Enriching virtual patients with interactive images. *Bio-Algorithms and Met-Systems* 5(9) pp. 73-77.

Huesman, R.L. (2000). *The Validity of ITBS Reading Comprehension Test Scores for Learning Disabled and Non Learning-Disabled Students under Extended-Time Conditions* [paper presentation]. Annual Meeting of the American Educational Research Association.

Hyland, M.E. & Sodergren, S.C. (1996). Development of a new type of global quality of life scale, and comparison of performance and preference for 12 global scales. *Quality of Life Research* 5, pp. 469-480.

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, pp. 695-698.

Ijmker, S, Mikkers, J., Blatter, B.M., van der Beek, A.J., van Mechelen, W. & Bongers, P.M. (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): pp. 685-696. <https://doi.org/10.1016/j.apergo.2007.12.003>
- Israel, G.D. & Taylor, C.L. (1990). “Can Response Order Bias Evaluations?” *Evaluation and Program Planning*, 13(4), pp. 365–71.
- Jaeschke, R., Singer, J. & Guyatt, G.H. (1990). A comparison of seven-point and visual analogue scales Data from a randomized trial. *Controlled Clinical Trials*, 11(1), pp. 43-51.
- Jelínek, M., Květon, P. & Vobořil, D. (2015). Innovative testing of spatial ability: interactive responding and the use of complex stimuli material. *Cognitive Processing*, 16(1), pp. 45-55.
- Jodoin, M.G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, 40, pp. 1–15.
- Joyce, C.R.B., Zutshi, D.W., Hrubes, V. & Mason, R.M. (1975). Comparison of fixed interval and Visual Analogue Scales for rating chronic pain. *European Journal of Clinical Pharmacology*, 8, pp. 415–420.
- Kaczmirek, L. (2008). Human-Survey Interaction Usability and Nonresponse in Online Surveys. Dissertation, Universität Mannheim.
- Kaczmirek, L. (2011). Attention and usability in internet surveys: Effects of visual feedback in grid questions. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and Behavioral Research and the Internet*. pp. 191–214. New York, NY: Taylor and Francis.
- Kersting, N. (2008). Using Video Clips of Mathematics Classroom Instruction as Item Prompts to Measure Teachers' Knowledge of Teaching Mathematics, *Educational and Psychological Measurement*, 68(5), pp. 845-861.

- Knäuper, B. (1999). The Impact of Age and Education on Response Order Effects in Attitude Measurement. *The Public Opinion Quarterly*, 63(3), pp. 347-370.
- Krebs, D. & Hoffmeyer-Zlotnik, J. H. (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), pp. 118-127.
- Krosnick, J.A., Alwin, Duane F. (1987). An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. *Public Opinion Quarterly*, 51(2), pp. 201-219.
- Krosnick, J.A., & Alwin, D.F. (1988). A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly*, 52, pp. 526–538.
- Krosnick, J.A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, pp. 213–236.
- Krosnick, J.A. & Berent, M.K. (1993). Comparisons of party identification and policy preferences: The impact of survey question format. *American Journal of Political Science*, 37, 941-964.
- Krosnick, Jon A. 1999. —Survey Research. *Annual Review of Psychology*, 50, pp. 537-67.
- Kunz, T. (2015). Rating scales in web surveys: A test of new drag-and-drop rating procedures—WebSM. Dissertation. Retrieved oct 16th 2020 from http://tuprints.ulb.tu-darmstadt.de/5151/7/Kunz_2015_Rating_scales_in_web_surveys.pdf
- Van Laerhoven, H., van der Zaag-Loonen, H.J. & Derkx, B.H.F. (2004). A comparison of Likert scale and visual analogue scales as response options in children's questionnaires. *Acta Paediatr* 93, pp. 830-835.

- Lam, M.Y., Lee, H., Bright, R., Korzenik, J.R.S. & Bruce, E. (2009). Validation of Interactive Voice Response System Administration of the Short Inflammatory Bowel Disease Questionnaire. *Inflamm Bowel Dis*, 15(4), pp. 599-607.
- Lesaux, N.K., Pearson, M.R. & Siegel, L.S. (2006). The Effects of Timed and Untimed Testing Conditions on the Reading Comprehension Performance of Adults with Reading Disabilities. *Reading and Writing* 19, pp. 21-48.
- Leutner, F., Yearsley, A., Codreanu, S.C., Borenstein, Y., Ahmetoglu, G. (2016). From Likert scales to images: Validating a novel creativity measure with image based response scales. *Personality and Individual Differences*, 106.
- Lim, E.M., Honjo, T., Umeki, K. (2006). The validity of VRML images as a stimulus for landscape assessment. *Landscape and Urban Planning* 77(1-2), pp. 80-93.
- Littman, G.S., Walker, B.R. & Schneider, B.E. (1985). Reassessment of verbal and visual analog ratings in analgesic studies. *Clin Pharmacol Ther*, 38, pp. 16-23.
- Liu, M. & Keusch, F. (2017). Effects of Scale Direction on Response Style of Ordinal Rating Scales. *Journal of Official Statistics*, 33(1), pp. 137-154.
- Louviere, J.J. & Islam, T. (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, pp. 903-911.
- Lu, Y. (2007). Validity Issues in Test Speededness. *Educational Measurement*, 26(4).
- Maeda, H. (2015). Response option configuration of online administered Likert scales. *International Journal of Social Research Methodology*, 18(1), pp. 15-26.

- Maio, G.R., Roese, N.J., Seligman, C. & Katz, A. (1996). Rankings, Ratings, and the Measurement of Values: Evidence for the Superior Validity of Ratings. *Journal of Basic and Applied Social Psychology*, 18(2), pp. 171-181.
- Malhotra, N. (2009). Completion Time and Response Order Effects in Web Surveys. *Public Opinion Quarterly*, 72(5), pp. 914-934.
- Maloshonok, N. & Terenlev, E. (2016). The impact of visual design and response formats on data quality in a web survey of MOOC students. *Computers in Human Behavior*, 62, pp. 506-515.
- Martinez, M.E. (1991). A comparison of multiple-choice and constructed figural response items. *Journal of Educational Measurement*, 28, pp. 131-145.
- Mathews, C.O.P (1929). The Effect of the Order of Printed Response Words on an Interest Questionnaire. *Journal of Educational Psychology*, 30, pp. 128-134.
- Matell, M.S. & Jacoby, J. (1972). Is there an optimal number of Alternatives for Likert-scale Items? Effects of Testing Time and Scale Properties. *Journal of Applied Psychology*, 56, pp. 657-509.
- McKelvie, S.J. (1978). Graphic rating scales - how many categories. *British Journal of Psychology*, 69, pp. 185-202.
- McReynolds, P., & Ludwig, K. (1987). On the history of rating scales. *Personality and Individual Differences*, 8, pp. 281–283.
- Medin, A.C., Carlsen, M.H. & Andersen, L.F. (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 Nutr.* 19, pp. 3265-3275.

Menold, N., Kaczmirek, L., Lenzner, T. & Neusar, A. (2014). How Do Respondents Attend to Verbal Labels in Rating Scales? *Field Methods* 26(1), pp. 21-39.

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

Mullane, J. & McKelvie, S. J. (2000). Effects of Removing the Time Limit on First and Second Language Intelligence Test Performance. *Practical Assessment, Research, and Evaluation* 7(23), pp. 1-6.

Myles, P.S., Troedel, S., Boquest, M. & Reeves, M. (1999). The pain visual analog scale: Is it linear or nonlinear? *Anesthesia & Analgesia*, 89, pp. 1517–1520.

Myles, P.S. & Urquhart, N. (2005). The Linearity of the Visual Analogue Scale in Patients with Severe Acute Pain. *Anaesth Intensive Care*, 33, pp. 54-58.

Neely, G., Ljunggren, G., Sylven, C. & Borg, G. (1992). Comparison between the Visual Analogue Scale (VAS) and the Category Ratio Scale (CR) for the evaluation of leg exertion. *International Journal of Sports Medicine*, 13, pp. 133–136.

Nicholls, M.E., Orr, C.A., Okubo, M. & Loftus, A. (2006). Satisfaction guaranteed: The effect of spatial biases on responses to Likert scales. *Psychological Science*, 17(12), pp. 1027-1028.

Nyrén, O., Adami, O., Bates, S., Bergstrom, R., Gustavsson, S., Loof, L. & Sjoden, P.O. (1987). Self-rating of pain in non-ulcer dyspepsia. *Journal of Clinical Gastroenterology*, 9, pp. 408–414.

- van Ooijen, P.M.A., Broekema, A. & Oudkerk, M. (2011). Design and implementation of I2Vote – An interactive image-based voting system using windows mobile devices. *International Journal of Medical Informatics*. 80(8): pp. 562-569.
- Paul-Dauphin, A., Guillemin, F., Virion, J.M. & Briançon, S. (1999). Bias and precision in visual analogue scales: a randomized controlled trial. *Am J Epidemiol*, 15(10), pp. 1117-1127.
- Payne, S.L. (1949). Case study in question complexity. *Public Opinion Quarterly*, 13, pp. 653-658.
- Pfennings, L., Cohen, L., & van der Ploeg, H. (1995). Preconditions for sensitivity in measuring change: Visual Analogue Scales compared to rating scales in a Likert format. *Psychological Reports*, 77, pp. 475–480.
- Rankin, W.L. & Grube, J.W. (1980). A comparison of ranking and rating procedures for value system measurement. *European Journal of Social Psychology*, 10(3), pp. 233-246.
- Rammstedt, B. & Krebs, D. (2007). Does Response Scale Format Affect the Answering of Personality Scales? *European Journal of Psychological Assessment*, 23(1), pp. 32-38.
- Rausch, M., & Zehetleitner, M. (2014). A comparison between a visual analogue scale and a four point scale as measures of conscious experience of motion. *Consciousness and Cognition*, 28, pp. 126–140.
- Reips, & Tourangeau, R. (2002) *Visual Context Effects in Web Surveys*. Online Social Sciences. In Batinic, B., Reips, U.-D. & Bosnjak, M., Seattle: Hogrefe & Huber.

- Reips, U.D., & Funke, F. (2008). Interval-level measurement with visual analogue scales in Internet-based research: VAS generator. *Behavior Research Methods*, 40(3), pp. 699–704.
- Rokeach, M. (1967). Value Survey, Sunnyvale, California, Halgren Tests (873 Persimmon Avenue, 94087).
- Rofé, Y. (2004). Mapping the sense of well being in a neighborhood: survey technique, and analysis of agreement and variation. Downloaded april 26th 2020 from https://www.researchgate.net/publication/316213103_Mapping_the_sense_of_well_being_in_a_neighborhood_survey_technique_and_analysis_of_agreement_and_variation
- Roster, C.A., Lucianetti, L. & Albaum, G. (2015). Exploring Slider vs. Categorical Response Formats in Web-Based Surveys. *Journal of Research Practice*, 11(1), pp. 1-19.
- Salzberger, T. & Koller, M. (2013). Towards a New Paradigm of Measurement in Marketing. *Journal of Business Research* 66, pp. 1307-1317.
- Scalise, K. & Gifford, B. (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, A.C. & Saris, W.E. (1997). The Validity and Reliability of Survey Questions - A Meta-Analysis of MTMM Studies. *Sociological Methods & Research*.
- Schubert, T.W. & Otten, S. (2002). Overlap of Self, Ingroup, and Outgroup: Pictorial Measures of Self-Categorization. *Self and Identity*, 1, pp. 353-376.
- Scott, J. & Huskisson, S.C. (1976). Graphic representation of pain. *Pain*, 2, pp. 175-184.
- Scott, J. & Huskisson, E.C. (1979). Vertical or horizontal visual analogue scales. *Ann Rheum Dis*, 38, p. 560.

- Shamir, B. & Kark, R. (2004). A single-item graphic scale for the measurement of organizational identification. *Journal of Occupational and Organizational Psychology*, vol. 77(1), pp. 115-123.
- Shulman, K.I. (2000). Clock-drawing: is it the ideal cognitive screening test? *International Journal of Geriatric Psychiatry*, 15(4), pp. 548-561.
- Sikkel, D., Steenbergen, R. & Gras, S. (2014). Clicking vs. dragging: Different uses of the mouse and their implications for online surveys. *Public Opinion Quarterly*, 78, pp. 177–190.
- Sinadinovic K, Wennberg, P., Berman, A. H. (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 Alcohol Depend* 114, pp. 55-60.
- Sireci, S.G. & Zenisky, A.L. (2006). Innovative Item Formats. In Steven M. Downing & Thomas M. Haladyna (Eds.) *Handbook of Test Development*. Lawrence Erlbaums Associates, Publishers: London. <https://fatihegitim.files.wordpress.com/2014/03/hndb-t-devt.pdf>
- Skedgel, C.D., Wailoo, A.J., Akehurst, R.L. (2013). Choosing vs. allocating: discrete choice experiments and constant-sum paired comparisons for the elicitation of societal preferences. *Health Expectations* pp 1-14, doi: 10.1111/hex.12098
- Stapleton, C.E. (2013). The Smartphone Way to Collect Survey Data. *Survey Practice* 6(2).
- Stutts, J.C., Stewart, J.R. & Martell, C. (1998). Cognitive test performance and crash risk in an older driver population. *Accid. Anal. And Prev.*, 30(3), pp. 337-346.
- Svalastoga, K. (1959) *Prestige, class and mobility*. New York: Arno.
- Sørensen, J.L., Thellesen, L., Strandbygaard, J., Svendsen, K.D., Christensen, K.B., Johansen, M., Langhoff-Roos, P., Ekelund, K., Ottesen, B., & Vleuten, C. Van der (2014).

- Development of knowledge tests for multi-disciplinary emergency training: a review and an example. *Acta Anaesthesiologica Scandinavica*, 59(1), pp. 123-33.
- Tesler, M.D., Savedra, M.C., Holzemer, W.L., Wilkie, D.J., Ward, J.A. & Paul, S.M. (1991). The word-graphic rating scale as a measure of children's and adolescents' pain intensity. *Research in Nursing & Health*. 14(5), 361-371.
- Timbrook, J.P. (2013). A Comparison of a Traditional Ranking format to a Drag-and-Drop Format with Stacking. Dissertation.
http://rave.ohiolink.edu/etdc/view?acc_num=dayton1367241685
- Timbrook, J. & Moroney, W.F. (2016). Ranking: Perceptions of Tied Ranks and Equal Intervals on a Modified Visual Analog Scale. Psychology Faculty Publications.
https://ecommons.udayton.edu/psy_fac_pub/26
- Thorndike, R.L. (1967). Reliability. In D. N. Jackson & S. Messick (Eds.), *Problems in Human Assessment*, pp. 217-240. New York: McGraw-Hill.
- Thorndike, F.P., Carlbring, P., Smyth, F.L., Magee, J.C., Gonder-Frederick, L., Ost, L.G. & Ritterband, L.M. (2009). Web-based measurement: Effect of completing single or multiple items per webpage. *Computers in Human Behavior*, 25, pp. 393-401.
- Toepoel, V., Das, M. & van Soest, A. (2009). Design of Web Questionnaires: The Effect of Layout in Rating Scales. *Journal of Official Statistics* 25(4), pp. 509–28.
- Tourangeau, R., Couper, M.P. & Conrad, F. (2004). Spacing, Positioning, and Order. Interpretative Heuristics for Visual Features of Survey Questions. *Public Opinion Quarterly* 68, pp. 368-393.

- Turvey, C., Sheeran, T., Dindo, L., Wakefield, B. & Klein, D. (2012). Validity of the Patient Health Questionnaire, PHQ-9, administered through interactive-voice-response technology. *Journal of Telemedicine and Telecare*, 18, pp. 348-351.
- Van der Vaart, W., Ongena, Y., Hoogendoorn, A. & Dijkstra, W. (2005). Do Interviewers' Voice Characteristics Influence Cooperation Rates in Telephone Surveys? *International Journal of Public Opinion Research*, 18(4), pp. 499-499.
- Voutilainen, A., Pitkääho, T., Kvist, T., & Vehviläinen-Julkunen, K. (2015). How to ask about patient satisfaction? The visual analogue scale is less vulnerable to confounding factors and ceiling effect than a symmetric Likert scale. *Journal of Advanced Nursing*, 72(4), 946–957.
- Voyer, D. (2011). Time limits and gender differences on paper-and-pencil tests of mental rotation: a meta-analysis. *Psychonomic Bulletin & Review*, 18, pp. 267-277.
- Wall, E.J., Milewski, M.D., Carey, J.L., Shea K.G., Ganley, T.J., Polousky, J.D., & Zbojniec, A. (2017). The reliability of assessing radiographic healing of osteochondritis dissecans of the knee. *The American Journal of Sports Medicine*, 45(6), pp. 1370–1375.
- Waller, R., Manuel, P. & Williamson, L. (2012). The Swindon Foot and Ankle Questionnaire: Is a Picture Worth a Thousand Words? *International Scholarly Research Network*, pp. 1-8.
- Wan, L. & Henly, G.A. (2012). Measurement Properties of Two Innovative Item Formats in a Computer-Based Test, *Applied Measurement in Education*, 25(1), pp. 58-78.

- Weaver, S.M. (1993). The Validity of the use of extended and untimed testing for postsecondary students with learning disabilities. Thesis. National Library of Canada.
- Weng, L. & Cheng, C. (2000). Effects of response order on Likert-type scales. *Educational & Psychological Measurement*, 60(6), pp. 908-924.
- Wewers, M.E. & Lowe, N.K. (1990). A critical review of visual analogue scales in the measurement of clinical phenomena. *Res Nurs Health*, 13(4), pp. 227-236.
- Wong, J.K. & Teas, R.K. (2001). A test of the stability of retail store image mapping based on multientity scaling data, *Journal of Retailing and Consumer Services*, 8(2): pp. 61-70.
- Xia, W., Sun, C., Zhang, L., Zhang, X., Wang, J., Wang, H. & Wu, L. (2011). Reproducibility and Relative Validity of a Food Frequency Questionnaire Developed for Female Adolescents in Suihua, North China. *PloS One* 6(5).
- Yan, T. & Keusch, F. (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
- Yan, T., Keusch, F. & He, Lirui (2018). The Impact of Question and Scale Characteristics on Scale Direction Effects. *Survey Practice*, 11(2).
- Yost K. J., Webster, K., Baker, D. W., Choi, S. W., Bode, R. K., Hahn, E. A. (2009) Bilingual health literacy assessment using the Talking Touchscreen/la Pantella Parlanchina: development and pilot testing. *Patient Educ Couns*, 75, pp. 295-301.
- Young, A., Willson, R. N., Pamela, R. N. (2012) Predicting NCLEX-RN Success: The Seventh Validity Study HESI Exit Exam. *Computers, Informatics, Nursing* 30(1) pp 55-60. doi: 10.1097/NCN.0b013e3182343edf

Zenisky, A. L., & Sireci, S. G. (2002). Technological Innovations in Large-Scale Assessment.

Applied Measurement in Education, 15(4), pp. 337-362. doi:

10.1207/S15324818AME1504_02

Zheng, Y. (2011) Research Note: Establishing Construct and Concurrent Validity of Pearson

Test of English Academic

<https://pdfs.semanticscholar.org/04e0/27c798140fb269bd43338bd972b2a0b3cabe.pdf>