CITATION
Sauer, J., Baumgartner, J., Frei, N., & Sonderegger, A. (2020, September 30). Pictorial Scales in Research and Practice: A Review. European Psychologist. Advance online publication. http://dx.doi.org/10.1027/1016-9040/a000405

Pictorial Scales in Research and Practice: a review

Abstract

The present article is concerned with the theoretical foundations and practical aspects of developing pictorial scales. It aims to assess the potential of pictorial scales compared to verbal scales. The article provides a review of existing pictorial scales with a view to identifying suitable methodological approaches for developing such scales. The review showed that the development and especially validation of many previous pictorial scales did not follow a stringent methodological approach. A category system is proposed, which allows the classification of different types of pictorial scales. Finally, we present a first draft of a theoretical framework, which can provide guidance for the future development of pictorial scales. The present work carries the implication that a specific methodological approach is needed, which focuses more strongly on the particular needs of designing pictorial scales (e.g., testing the comprehensibility of pictures).

Keywords: pictorial scale; non-verbal scale, scale development, psychometrics

Rationale of article

Measuring subjective states of humans has been an important issue in psychological research. In order to obtain precise measures of the subjective state, measurement instruments with satisfactory psychometric properties are needed. The vast majority of these measurements were made by means of verbal questionnaires, ranging from very elaborate instruments to single-item scales. Pictorial scales represent an important supplement to verbal ones. Even if they cannot replace verbal scales in many application areas, their development should be considered and their usage should be encouraged in those application areas that provide favorable conditions to them (e.g., poor reading skills of respondents).

Due to the common use of verbal instruments, theoretical framework and detailed guidelines are available to support questionnaire development (e.g., Coolican, 2017; Hinkin, 1995). For the specific design of pictorial scales, no such guidelines are available. This may be due to the small number of such instruments compared to vast majority of verbal questionnaires.

There is some overlap between the requirements for verbal and pictorial scales, though there are very specific needs for the development of pictorial scales. These specific requirements are outlined in the present article. To this end, the following goals are pursued. (a) We aim to give a comprehensive overview of pictorial scales developed to measure constructs in different domains. (b) We aim to provide an overview of the advantages and drawbacks of using pictorial scales compared to verbal scales. (c) We aim to examine the methodological approach employed to develop pictorial scales. (d) We aim to offer researchers intending to develop pictorial scales a framework for the development and validation of such scales.

Issues in the design of pictorial scales

Generally, a pictorial scale may be defined as an instrument that makes use of image-based elements to convey the meaning of its items. The image-based elements may be used to represent the statement or the rating scale, or both of them. As our analysis of pictorial scales will show, there are many different types of such scales.

The most distinctive characteristic of a pictorial scales is that they are *mostly* free from text and therefore language independent (Betella & Verschure, 2016). The word 'mostly' is quite critical here since the use of some language-based elements in pictorial scales is not uncommon.

Many publications have addressed the advantages and disadvantages of pictorial scales, often by drawing a comparison to verbal scales (e.g., Bradley & Lang, 1994; Kunin, 1955). We will focus on issues that are mainly relevant to pictorial scales since general problems of questionnaire design have been discussed elsewhere (e.g., Choi & Pak, 2005). With regard to the design of pictorial scales, the following points have been referred to in the literature as being relevant: language independence, intuitive comprehension, respondent motivation, and pictorial aesthetics.

Language independence

The perhaps most salient feature of pictorial scales is that they are language independent (Betella & Verschure, 2016). However, it is presumably more accurate to state that they depend less on language rather than being full non-verbal scales (given that a number of pictorial scales use language-based elements in addition to pictorial elements). Therefore, for the scales with exclusively pictorial content, there is no need for item translation, making it possible to administer the scales across language borders (Desmet et al., 2016; Huisman et al., 2013; Paunonen, Ashton, & Jackson, 2001; Sonderegger et al., 2016).

Language independence may overcome measurement problems in application domains such as cross-cultural research (e.g., Boer, Hanke, & He, 2018). However, this is not to say that pictorial scales are free from measurement problems when collecting data in different cultural groups. While there is little research on this issue, we presume that such measurement problems will be smaller when cultural groups are compared, which are characterized by close cultural similarity (e.g. based on the dimensions of Hofstede, 1991). This primarily applies to multilingual countries, in which citizens have different mother tongues (e.g., Belgium, Switzerland). Conversely, the interpretation of some non-verbal signs may differ between cultures (e.g. ring gesture, in which the thumb and index finger are connected to form a circle, Müller, 2014). Of course, it is not advisable to use such cross-culturally ambiguous non-verbal elements in pictorial scales.

Furthermore, pictorial scales can be administered to respondents with insufficient competence in the target language (e.g., non-native speakers) as well as to children or adults with poor language skills or insufficient reading abilities (Buchanan & Niven, 2002; Cecil, McCrory, Holden, & Barker, 2016; Ghiassi et al., 2010; Paunonen et al., 2001; Sonderegger et al., 2016; Venham & Gaulin-Kremer, 1979).

Intuitive comprehension

When reviewing the literature, the issue of intuitive comprehension of pictorial scales features prominently but the picture emerging from empirical research is somewhat inconsistent. In the context of assessing affect and attitudes, it has been argued that pictorial scales are more intuitively comprehensible, since there is no ’necessity for translating feelings into words’ (Kunin, 1955, p. 66). Some researchers consider them as easier to understand since the measurement process is more direct using pictorial representation of emotions (Bradley & Lang, 1994). While words might fail to provide an exact description of

a subjective experience (Bradley & Lang, 1994; Haddad et al., 2012), pictures can represent such a subjective experience (Broekens & Brinkman, 2013; McGrath, Pianosi, & Buckley, 2005). There are cases in which pictorial representations can better express the different levels of an item the respondent can rate (e.g. levels of trunk deformity; Bago et al., 2010). This is because the pictorial scale permits a direct visual comparison of the object in question (e.g. trunk), whereas words only would increase complexity and ambiguity of the scale. Furthermore, pictorial scales are often shorter than verbal scales and their use is considered to be less mentally demanding (Wissmath, Weibel, & Mast, 2010). On the other hand, concerns have been expressed that pictorial scales may not be intuitively understandable because they are based on outdated graphical design principles (Betella & Verschure, 2016), or the graphic representations are perceived as oversimplified (Sonderegger et al. 2016). Due to these ambiguities, pictorial items often need specific (often written) instructions or verbal hints to avoid misinterpretation (Betella & Verschure, 2016; Broekens & Brinkman, 2013; Cecil et al., 2016).

The different views expressed in the literature may be related to the kind of construct to be measured. As our review of pictorial scales found in the research literature will show, there are certain constructs that are rather frequently measured by this type of scale. This is probably not a coincidence and reflects the presumption that some constructs are more amenable to measurement by pictorial scales than others (e.g., developing a scale measuring emotion or pain appears to be more promising than a scale aiming to measure the Big-5 personality trait).

Respondent motivation

An important aspect in the design of pictorial scales (and questionnaires more generally) is the motivation of the respondent to complete them. The question of respondent motivation

is also reflected in the creation of ‘questionnaire experience’ (QX) as a new term to emphasize the importance of such aspects in questionnaire design in general and, in particular, when designing pictorial scales (Baumgartner et al., in preparation). Pictorial scales are expected to increase motivation, focus attention and stimulate interest (Haddad et al., 2012; Valla et al., 1994). When completing them, they are less demanding, easier to understand and provide more pleasure (Desmet et al., 2011; Ghiassi et al., 2010). There is first evidence from a study showing higher scores of motivation when participants filled in a pictorial scale compared to a verbal one (Baumgartner et al., 2019b). The issue of respondent motivation may be of particular importance if the scale is administered several times to the same participants.

Aesthetics of pictorial stimuli

A very specific issue in the design of pictorial scales is related to potential influences of aesthetically very pleasing images or drawings. There have been concerns that the level of visual attractiveness may lead to the selection of the aesthetically most pleasing image instead of the one reflecting the most appropriate response (Haddad et al., 2012; Reynold-Keefer & Johnson, 2011). In particular, the degree of realism of the representation (e.g., how realistic the drawing of an item is) is expected to have an impact on the ratings given by respondents (Reynold-Keefer & Johnson, 2011). The results of the two studies reported suggest that there may be a respondent bias due to the attractiveness of certain pictures, which would result in measurement error.

Conclusion

The review of the literature revealed a considerable number of advantages for using pictorial scales while, at the same time, numerous publications also pointed out their

weaknesses. Although there is no unanimous agreement on the advantages of pictorial scales, there appears to be sufficient support in the research literature for the development and the use of such scales. The literature also revealed that despite considerable interest in the question surrounding the design of pictorial scales, there is very little empirical research that made comparisons between verbal and pictorial scales, which would provide more substantial evidence of their respective advantages and disadvantages. Regardless of these issues, a large number of image-based scales are used in practice and research. In the following section, a review of existing pictorial scales is presented with a view to gaining a better understanding of the methodological approaches used to develop them.

Pictorial scales: overview, classification and methodological approaches

Overview of existing pictorial scales

A literature review was carried out to identify and classify pictorial scales that had been published and used. The following databases were used: Google Scholar, PsycINFO, PSYNDEX, and PubMed. While an earlier review (Desmet et al., 2016) only examined instruments assessing affect, we included pictorial scales from all application domains. We used the following terms to search for pictorial scales across different domains: pictorial scale/questionnaire, non-verbal scale/questionnaire, image, visualization, cartoon-based scale/questionnaire, facial expressions, and face scale/questionnaire. Furthermore, we attempted to find scales by entering the following typical application domains of pictorial scales: affect, mood, pain, emotion, children, pediatrics, and physical illness. We also used the following psychometric terms: test validity, test reliability, construct validity, and test construction. We did not place any constraints on the literature search with regard to the type of the publication or its source (e.g. we included scholarly journals as well as conference

publications without peer review). We only searched for publications in English language and made extensive use of the backward search method (i.e. we searched the reference list of the relevant publications identified to locate further pictorial scales). The backward search method contributed substantially to the location of pictorial scales since the online search method proved to be difficult. This was because scales were linked to a wide range of keywords, often not making use of the most obvious ones (e.g. pictorial or non-verbal scale/questionnaire).

An exclusion criterion was that we limited our search to pictorial instruments that allowed respondents to make ratings, which excludes other picture-based instruments used in psychology such as intelligence tests measuring perceptual reasoning (e.g., WAIS-IV) or the Rorschach test (e.g., Urist, 1977).

In total, we identified 57 publications of instruments that provided a graphical illustration of the scale in the form of a figure. This list of instruments was compiled by using information from the online search, the backward search method, and suggestions made by other researchers. The pictorial scales are presented in table 1, following alphabetic order. Scales published without illustration were not listed in the table because it is difficult to assess the nature and the quality of a scale without knowing what it looks like.

The following results emerged from the in-depth analysis of published pictorial scales. (a) Most instruments are used in application areas such as assessment of emotional states and affect (n = 23), followed by medical diagnosis (n = 9). Research on emotion is a particularly suitable application area because it may be more difficult to describe the emotion verbally than to show it in the form of a drawing or manikin representing the facial expressions of the emotion. (b) It emerged that pictorial scales may be especially suitable for certain target groups. Children assessment is an important target domain because of the inability of younger children to read and the difficulties of older ones to understand complex language. Therefore,

many instruments were developed for children as the main target group. (c) Pictorial scales were generally shorter than typical verbal scales, with only a smaller number comprising more than 10 items. Interestingly, we found nearly half of the instruments to be one-item scales. (d) The literature review revealed a broad range of designs of pictorial scales. This includes exclusively pictorial scales of different types but also so-called hybrid scales that additionally contain different kind of non-pictorial content (e.g. verbal content in the form of keywords).

Classification of existing instruments

The pictorial scales identified in the literature were analyzed with regard to the kind of scale design used. This was done with a view to gaining an overview of the kind of pictorial scales used in research and practice and to creating a classification system.

The following approach was used for creating the categories of the classification system, involving all authors of the manuscript. All scales from table 1 were examined, focusing on common elements and salient differences. If the elements of two scales did not differ much from one another, it may suggest that both scales are similar in their underlying structure and should belong to the same category. If there were obvious differences with regard to the nature of a scale or its use compared to other scales, it should be assigned to a different category. Table 2 provides a listing of the different scale types and their characteristics. After a preliminary analysis of the 57 scales, the most frequent differences between scales were observed with regard to three main criteria. (a) *Conveyor of item meaning* refers to the way the meaning of the item is conveyed to the respondent. This could be in the form of exclusively pictorial content (e.g., scale #01 in table 2), combined verbal and pictorial content (e.g., scale #07 in table 2), and exclusively verbal content (e.g., scale #05 in table 2). (b) *Type of rating scale* refers to the way the different items can be rated and how the meaning of the

rating is conveyed to the respondent. This could be in the form of exclusively pictorial content (e.g., scale #01 in table 2), combined verbal and pictorial content (e.g., scale #10 in table 2), and exclusively verbal content (e.g., scale #08 in table 2). (c) The third criterion *level of integration* refers to the degree to which conveyor of item meaning and rating scale represent two separate entities (e.g., scale #02 in table 2) or an integrated single entity (e.g., scale #01 in table 2).

The use of the three criteria resulted in a 3 x 3 x 2 classification system being developed. To facilitate categorization, numerical elements such as scale numbers were considered as non-verbal elements. Based on the 3 x 3 x 2 classification, 18 categories were derived. However, two of the categories were not applicable because they would correspond to a full verbal scale, which was outside the remit of the present review. This resulted in a classification system with 16 suitable categories (see figure 1a).

The analysis revealed several interesting results. A broad distinction can be made between scales with exclusively pictorial content and hybrid scales containing both pictorial and verbal content. It showed that certain scale types were much more frequent than others. Slightly more than half of the scales (n = 30) were found to belong to the category *'pictorial conveyor of item meaning & pictorial rating scale & integrated entity'* (see figure 1a). They may be considered exclusively pictorial on both criteria (i.e. conveyor of item meaning and rating scale). Conversely, this shows that there were a considerable number of hybrid scales (n = 27), which may be indicative of the difficulties associated with the development of exclusively pictorial scales. There were 5 out of the 16 categories, to which none of the instruments could be assigned. The remaining 11 categories (representing hybrid scales) showed frequencies in the order of one to five. Finally, the results showed that exclusively verbal content on one criterion (i.e. either item meaning or scale rating) was rather rare, which may be indicative of the good intent of scale developers to minimize verbal content.

The categorization using supplementary classification criteria revealed the following picture (see figure 1b): (a) There is a large number of single-item scales (n = 25), being nearly as frequent in numbers as multi-item questionnaires (n = 32). This may suggest that pictorial scales may be particularly suitable when constructs are to be measured with very short scales rather than elaborate scales. (b) Pictorial scales for children are very frequent, with about half of the scales being developed for that target group. This may be due to children having generally poorer reading skills than adults, which does not matter if the scales are non-verbal. We acknowledge that the list of pictorial scales provided is probably not comprehensive. Given that we made extensive use of the backward search method and received additionally suggestions from other researchers about the existence of further pictorial scales, it suggests that an extensive keyword-based literature search is insufficient to identify all existing pictorial scales. Therefore, the non-comprehensiveness of our list may have somewhat biased the classification of instruments, which we carried out with regard to different properties.

Methodological approaches used for scale development

The pictorial scales identified were then examined with regard to the methodological approaches used for scale development. The scales that provided descriptions (even if very short) are presented in table 3.

The results of this analysis showed that the publications varied considerably with regard to the level of detail reported about the procedure of scale development. Similarly, the methods used for the development and validation of the scales, if reported, were very diverse. The pictorial items were often created in collaboration with graphic designers and experts from other fields. Furthermore, user feedback was obtained in various forms (e.g., thinking-aloud protocols, ranking of items, and suggestions for improvement). Of the scales presented

in table 3, for 18 of them the authors reported to have carried out a validation study in some form. For 11 scales, they reported some quantitative indicator of the psychometric quality of the scale (e.g. internal consistency, test-retest reliability) and for 5 scales, they reported one or more validity coefficients (e.g. convergent and divergent validity). Overall, it can be concluded that there is not a specific procedure of scale development that is widely used. Instead, a wide range of approaches is applied. Furthermore, the analysis showed that very few studies captured scale validity by reporting relevant validity coefficients and other psychometric indices.

Implications

Favorable conditions for developing and administering pictorial scales

Based on our analysis of existing instruments and our own deliberations, we believe that under the following circumstances, the development of a pictorial scale appears to be promising (see table 4a for a summary of the main points). (a) The language competence of respondents may be limited (be it their reading skills or their general understanding of the language). This includes target groups where respondents have different mother tongues in the same cultural space or at least in countries that are similar to one another in terms their cultural dimension (see, for example, the model of Hofstede, 1991). For example, multilingual countries such as Switzerland or Belgium would benefit from this. If the countries in which the same scale is applied are culturally very different from each other (in terms of Hofstede's model), this increases the risk that the same scale is interpreted in a different way. For example, the interpretation of some non-verbal signs differ between cultures (e.g., ring gesture). Of course, such ambiguous signs are to be avoided in pictorial scales.

Another important target group to which the argument about the limited language competence applies are children, though they may interpret a pictorial scale in a different way than adults. If a scale is used for both target groups, separate validation studies are required. (b) There is a need to measure a construct several times. The more often the same construct is measured, the more the pictorial scales will be able to reduce its disadvantage compared to a verbal scale, or increase its advantage over it, respectively. This is because in a repeated measurement context, a picture as a symbolic representation may allow a better and faster encoding of the information (similar to configural displays; Bennett & Walters, 2001). (c) There may be certain constructs that are more amenable to measurement by means of

pictorial scales. This refers to constructs that can be well represented by pictures. A prime example of such a construct is emotion because of the visual nature of emotional expressions (even if it acknowledged that certain emotion such as shame may not have unique facial expressions). (d) If there is a need to develop a short scale (in particular, a one-item scale), pictorial scales might be more advantageous than when an elaborate scale is required. This point was also supported by the rather high frequency of one-item scales emerging in the present review. Again, it has to be noted that verbal scales may also be highly suitable for single-item measures. Furthermore, it has to be acknowledged that there has been a long debate in more general terms about the suitability of single-item measures (i.e. verbal and non-verbal items) compared to multiple-item instruments (e.g., Gardner, Cummings, Dunham & Pierce, 1998; Venkatraman & Grant, 1998; Schmidt & Hunter, 1996).

All four points that we mentioned provide in our view favorable conditions that should be taken into account when considering the design and use of pictorial scales. To avoid any misunderstanding, we do not claim that under these circumstances pictorial scales are more suitable than verbal scales but rather that these circumstances increase the relative suitability of pictorial scales.

First draft of guidelines

Based on our experience in pictorial scale development (Baumgartner et al., 2019a; Baumgartner et al., 2019b; Baumgartner et al., in preparation), we developed a prototypical three-phase approach to the development of a pictorial scale. Figure 2 shows the three phases consisting of item generation, interpretation check and scale validation. This approach is based on best practices of scale development (Hinkin, 1995). Due to the visual nature of pictorial scales, these best practices had to be adapted, especially with regard to the first two phases. Several methods are proposed for each phase, of which some were inspired by

disciplines other than psychology (e.g., UX design, computer science). This set of methods is not exhaustive. The methods might be added or replaced depending on the application area of the instrument, the target group, or the skills of the researcher developing the pictorial instrument.

Item generation

The first phase is characterized by gathering ideas for an adequate representation of the items of a pictorial scale. In this phase, a group of experts needs to generate ideas for meaningful items and should be able to visualize the constructs of the scale while keeping the characteristics of the target population in mind (e.g. children, functional impairments). Therefore, a multidisciplinary team of experts is required, being able to offer the appropriate skills in the respective domains. For these goals, the following methods may be used: brainstorming, visualization (e.g., design studio) and rapid prototyping. Established verbal questionnaires may be used as a starting point for developing ideas and items. The goal of this phase is to develop the most promising ideas for a visual representation of a construct, and to create a first draft of items, which will be checked with test users in the next phase.

Interpretation check

The second phase involves assessing the visual drafts generated in the first phase by asking test users for feedback on how they interpret the items. Inspection methods such as interviews or the thinking-aloud technique (e.g., Nielsen, 1994) can be used to gain insights into the mental model of test users. Furthermore, expert discussions or expert surveys (e.g., Delphi method; Okoli & Pawlowski, 2004) can be used as additional methods to obtain valuable feedback ‘outside the box’. The second phase is crucial in laying the foundation for a scale with satisfactory psychometric properties. The steps in this phase should be iterative, that is, items have to be refined and re-evaluated until they are sufficiently well understood by respondents. Approaches such as the ISO 9186 comprehension test for graphical symbols

can be used as a yardstick to check if items are sufficiently comprehensible. The goal of this phase is to have a consolidated set of items with an adequate representation of the construct in question (high content validity), which is ready to be tested in a validation study for assessing psychometric criteria. Content validity may be assessed by the level of agreement between expert raters on the different dimensions of the construct. In addition to the methods proposed (which we consider particularly suitable for the development of pictorial scales because they had already put to the test in our research), the literature proposes further methods (e.g., Howard, 2018) that may be used for the development of pictorial items. Howard proposed quantitative methods such as item rating (i.e. participants rate the representativeness of each item against a definition of the construct to be measured), item sorting (i.e. participants assign each item to one of a set of constructs that they believe represents best what each item measures, with one of the constructs being the one to be measured), and combined methods of the two (i.e. participants rate each item for its representative for each construct, one of which is the construct to be measured). If this second phase is not successfully completed (e.g., the items developed do not reflect the underlying construct adequately), scale developers will need to move back to the previous phase.

Scale validation

The third phase involves an empirical assessment of the psychometric properties of the newly developed pictorial instrument by means of a validation study with the target group. These psychometric measures are calculated to evaluate the quality of the scale by examining 'its relationship with other variables of interest' (Hinkin, 1995, p. 971). This should include psychometric coefficients such as validity measures (convergent, divergent and criterion-related validity), reliability measures (e.g. internal consistency, re-test reliability) and sensitivity. Which of the psychometric coefficients should be used depend on the specific scale to be evaluated. For example, internal consistency cannot be determined for one-item

measures and re-test reliability should only be determined for trait measures but not for state measures.

The goal of the last phase is to determine whether the developed pictorial scale enjoys satisfactory psychometric properties or not. For this purpose, the psychometric properties of established verbal scales can be used as a yardstick. If the results of the validation study are not satisfactory, scale developers need to return to phase II, or even to phase I, if the underlying constructs are not adequately reflected in the pictorial scale.

The first draft of these guidelines to develop pictorial scales needs to be further elaborated, especially providing more detail on the different steps and methods. This may help us move towards a more standardized approach to the development of pictorial scales.

Future research and development of future pictorial scales

There is a need to address a number of issues with regard to future research but also future development of scales (see table 4b for a summary of the main points). (a) There is need for comparisons between pictorial scales and equivalent verbal scales to determine whether pictorial scales are able to match the psychometric properties achieved by verbal scales. This may include coefficients such as convergent validity, divergent validity, criterion-related validity, internal consistency, and sensitivity. These comparative studies should also include the measurement of supplementary evaluative criteria because they may not be of lesser importance. This refers to criteria such as scale completion time, participant dropout rate, and participant motivation. (b) If there are no equivalent verbal scales available for direct comparisons, benchmark testing could be used. This refers to a comparison of a pictorial scale with typical psychometric quality coefficients achieved by verbal scales in the same or a similar field. For example, if a good verbal scale measuring general life satisfaction achieves a convergent validity coefficient of r = .7, this score could be used as a benchmark

for a pictorial scale capturing the more specific construct of job satisfaction. Given the complexity of establishing benchmarks, an alternative method may represent the development of meaningful cut points (e.g., Hirschfeld & Thielsch, 2015). (c) Guidelines for developing pictorial scales need to be developed. These guidelines are needed to extend recommendations for the development of verbal scales, which have been widely published in textbooks and handbooks on psychometrics and questionnaire development. The pictorial nature of the scales requires additional methodological considerations that need to be included in the guidelines. For example, the use of visual design elements needs to be addressed. In addition to excellent verbal skills needed by test developers of verbal scales, artistic skills in drawing and design are required to be able to create meaningful and unambiguous pictorial items. (d) It is a difficult endeavor to make suggestions about which fields in psychology would benefit from an increased availability of pictorial scales. This task should be left to the experts of their respective fields. In the field of work psychology and ergonomics (which is most familiar to the authors), we feel that pictorial scales are clearly underrepresented. This was also confirmed by the literature review summarized in table 1. Therefore, we can see a number of areas, which would benefit from the development of pictorial scales, including the measurement of workload and perceived usability. (e) The qualitative approach of testing the comprehension of pictorial scales (e.g. by using thinking-aloud protocols) needs to be more standardized. For example, criteria should be defined when scale comprehension is at a sufficient level in order for the scale to be tested in a validation study. This could be a minimum percentage of a sample of a pre-defined size. The ISO 9186 comprehension test for graphical symbols may provide guidance for this purpose only to some extent. This is because this ISO norm has been designed for symbols that represent simpler concepts such as locations in buildings (e.g. toilet, check-in).

Conclusion and outlook

The present article provides a starting point for some reflection about the development and use of pictorial scales by integrating work from a wide range of application areas, which hitherto have not been considered together. These reflections need to culminate in a more structured methodological approach for the development of pictorial scales. Finally, we would like to stress that we do not believe that pictorial scales will to some extent replace verbal scales. However, we are convinced that the potential of pictorial scales is currently undervalued in research. Therefore, developing a structured methodological approach for the design of pictorial scales is expected to advance the field substantially.

References

Bago, J., Sanchez-Raya, J., Perez-Grueso, F. J. S., & Climent, J. M. (2010). The Trunk Appearance Perception Scale (TAPS): a new tool to evaluate subjective impression of trunk deformity in patients with idiopathic scoliosis. Scoliosis, 5(1), 6.

Barnett, L. M., Vazou, S., Abbott, G., Bowe, S. J., Robinson, L. E., Ridgers, N. D., & Salmon, J. (2016). Construct validity of the pictorial scale of Perceived Movement Skill Competence. *Psychology of Sport and Exercise*, *22*, 294–302. doi: 10.1016/j.psychsport.2015.09.002

Baumgartner, J., Frei, N., Kleinke, M., Sauer, J. & Sonderegger, A (2019a). Pictorial System Usability Scale (P-SUS): Developing an Instrument for Measuring Perceived Usability. CHI conference 2019.

Baumgartner, J., Sonderegger, A., & Sauer, J. (2019b). No need to read: Developing a pictorial single-item scale for measuring perceived usability. *International Journal of Human-Computer Studies, 122*, 78-89.

Baumgartner, J., Ruettgers, N., Hasler, A., Sonderegger, A., & Sauer, J. (in preparation). What you see is what you rate: developing pictorial instruments for measuring perceived usability.

Baxter, A. L., Watcha, M. F., Baxter, W. V., Leong, T., & Wyatt, M. M. (2011). Development and validation of a pictorial nausea rating scale for children. *Pediatrics, 127*(6), e1542-1549. doi: 10.1542/peds.2010-1410

Beasley, J. M., Davis, A., & Riley, W. T. (2009). Evaluation of a web-based, pictorial diet history questionnaire. Public health nutrition, 12(5), 651-659.

Bennett, K. B., & Walters, B. (2001). Configural Display Design Techniques Considered at Multiple Levels of Evaluation. *Human Factors*, *43*(3), 415–434. doi: 10.1518/001872001775898304

Betella, A., & Verschure, P. F. M. J. (2016). The Affective Slider: A Digital Self-Assessment Scale for the Measurement of Human Emotions. *PLOS ONE*, *11*(2), e0148037. doi: 10.1371/journal.pone.0148037

Bieri, D., Reeve, R. A., Champion, G. D., Addicoat, L., & Ziegler, J. B. (1990). The Faces Pain Scale for the self-assessment of the severity of pain experienced by children: development, initial validation, and preliminary investigation for ratio scale properties. *Pain*, *41*(2), 139–150.

Boer, D., Hanke, K., & He, J. (2018). On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology, 49*(5), 713-734.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59. doi: 10.1016/0005-7916(94)90063-9

Broekens, J., & Brinkman, W.-P. (2013). AffectButton: A method for reliable and valid affective self-report. *International Journal of Human-Computer Studies*, *71*, 641–667. doi: 10.1016/j.ijhcs.2013.02.003

Brumfitt, S. M., & Sheeran, P. (1999). The development and validation of the Visual Analogue Self-Esteem Scale (VASES). *The British Journal of Clinical Psychology*, *38*, 387–400.

Buchanan, H., & Niven, N. (2002). Validation of a Facial Image Scale to assess child dental anxiety. *International Journal of Paediatric Dentistry*, *12*(1), 47–52.

Büchi, S., & Sensky, T. (1999). PRISM: Pictorial Representation of Illness and Self Measure. A brief nonverbal measure of illness impact and therapeutic aid in psychosomatic medicine. *Psychosomatics*, *40*(4), 314–320. doi: 10.1016/S0033-3182(99)71225-9

Cecil, C. A. M., McCrory, E. J., Viding, E., Holden, G. W., & Barker, E. D. (2016). Initial Validation of a Brief Pictorial Measure of Caregiver Aggression: The Family Aggression Screening Tool. *Assessment*, *23*(3), 307–320. doi: 10.1177/1073191115587552

Choi, B. C. K., & Pak, A. W. P. (2005). A catalog of biases in questionnaires. Preventing Chronic Disease, 2(1), 1-13.

Coolican, H. (2017). Research Methods and Statistics in Psychology. Hodder & Stoughton: London.

Cowie, R., Douglas-Cowie, E., Savvidou, S., & McMahon, E. (2000). FEELTRACE: an instrument for recording perceived emotion in real time. 19-24. Paper presented at Speech and Emotion: Proceedings of the ISCA workshop, Newcastle, United Kingdom.

Desmet, P. (2003). Measuring Emotion: Development and Application of an Instrument to Measure Emotional Responses to Products. In Funology (S. 111–123). Springer, Dordrecht. Accessed on https://link.springer.com/chapter/10.1007/1-4020-2967-5_12

Desmet, P., Overbeeke, K., & Tax, S. (2001). Designing Products with Added Emotional Value: Development and Application of an Approach for Research through Design. *The Design Journal*, *4*, 32–47. doi: 10.2752/146069201789378496

Desmet, P., Vastenburg, M., & Romero, N. (2016). Mood measurement with Pick-A-Mood: review of current methods and design of a pictorial self-report scale. *Journal of Design Research*, *14*, 241–279. doi: 10.1504/JDR.2016.10000563

Döring, A. K., Blauensteiner, A., Aryus, K., Drögekamp, L., & Bilsky, W. (2010). Assessing values at an early age: The picture-based value survey for children (PBVS-C). Journal of personality assessment, 92(5), 439-448.

Dubi, K., & Schneider, S. (2009). The Picture Anxiety Test (PAT): A new pictorial assessment of anxiety symptoms in young children. Journal of Anxiety Disorders, 23(8), 1148-1157.

Frank, A. J. M., Moll, J. M. H., & Hort, J. F. (1982). A COMPARISON OF THREE WAYS OF MEASURING PAIN. *Rheumatology*, *21*(4), 211–217. doi: 10.1093/rheumatology/21.4.211

Fox, N.A., & Leavitt, L.A. (1995). The Violence Exposure Scale for Children (VEX). College Park, MD: University of Maryland.

Gardner, D. G., Cummings, L. L., Dunham, R. B., & Pierce, J. L. (1998). Single-item versus multiple-item measurement scales: An empirical comparison. *Educational and Psychological Measurement*, 58(6), 898-915.

Ghiassi, R., Murphy, K., Cummin, A. R., & Partridge, M. R. (2011). Developing a pictorial Epworth Sleepiness Scale. *Thorax*, *66*(2), 97–100. doi: 10.1136/thx.2010.136879

Girard, S., & Johnson, H. (2009). Developing affective educational software products: Sorémo, a new method for capturing emotional states. Journal of Engineering Design, 20(5), 493–510. doi: 10.1080/09544820903158827

Gueldner, S. H., Michel, Y., Bramlett, M. H., Liu, C.-F., Johnston, L. W., Endo, E., … Carlyle, M. S. (2005). The well-being picture scale: a revision of the index of field energy. *Nursing Science Quarterly*, *18*(1), 42–50. doi: 10.1177/0894318404272107

Haddad, S., King, S., Osmond, P., & Heidari, S. (2012, November). Questionnaire design to determine children's thermal sensation, preference and acceptability in the classroom. In Proceedings-28th International PLEA Conference on Sustainable Architecture+ Urban Design: Opportunities, Limits and Needs-Towards an Environmentally Responsible Architecture.

Harter, S., & Pike, R. (1984). The Pictorial Scale of Perceived Competence and Social Acceptance for Young Children. *Child Development*, *55*(6), 1969–1982. doi: 10.2307/1129772

Hinkin, T. R. (1995). A Review of Scale Development Practices in the Study of Organizations. *Journal of Management*, *21*(5), 967–988. doi: 10.1177/014920639502100509

Hirschfeld, G., & Thielsch, M. T. (2015). Establishing meaningful cut points for online user ratings. *Ergonomics*, *58*(2), 310-320.

Hofstede, G. (1991). Cultures and Organizations: Software of the Mind. London: McGraw.

Howard, M. C. (2018). Scale Pretesting. Practical Assessment, Research & Evaluation, 23(5).

Huisman, G., Van Hout, M., van Dijk, E., Geest, T., & Heylen, D. (2013). LEMtool – Measuring Emotions in Visual Interfaces. In C*onference on Human Factors in Computing Systems - Proceedings*. doi: 10.1145/2470654.2470706

Isbister, K., Höök, K., Sharp, M., & Laaksolahti, J. (2006). The Sensual Evaluation Instrument: Developing an Affective Evaluation Tool. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (S. 1163–1172). New York, NY, USA: ACM. doi: 10.1145/1124772.1124946

Jäger, R. (2004). Konstruktion einer Ratingskala mit Smilies als symbolische Marken. *Diagnostica*, 50(1), 31-38. https://doi.org/10.1026/0012-1924.50.1.31)

Johnstone, T., van Reekum, C. M., Hird, K., Kirsner, K., & Scherer, K. R. (2005). Affective speech elicited with a computer game. *Emotion (Washington, D.C.), 5*(4), 513–518. doi: 10.1037/1528-3542.5.4.513

Kunin, T. (1955). The construction of a new type of attitude measure. *Personnel Psychology, 8*, 65–77. doi: 10.1111/j.1744-6570.1955.tb01189.x

Lorish, C. D., & Maisiak, R. (1986). The Face Scale: a brief, nonverbal method for assessing patient mood. *Arthritis and Rheumatism*, *29*(7), 906–909.

Maćkiewicz, M., & Cieciuch, J. (2016). Pictorial Personality Traits Questionnaire for Children (PPTQ-C)-A New Measure of Children's Personality Traits. *Frontiers in Psychology*, *7*, 498. doi: 10.3389/fpsyg.2016.00498

Maldonado, C. C., Bentley, A. J., & Mitchell, D. (2004). A Pictorial Sleepiness Scale Based on Cartoon Faces. *Sleep, 27*(3), 541–548. doi: 10.1093/sleep/27.3.541

Manassis, K., Mendlowitz, S., Kreindler, D., Lumsden, C., Sharpe, J., Simon, M. D., … Adler-Nevo, G. (2009). Mood Assessment Via Animated Characters: A Novel Instrument to Evaluate Feelings in Young Children With Anxiety Disorders. *Journal of Clinical Child & Adolescent Psychology*, *38*(3), 380–389. doi: 10.1080/15374410902851655

McGrath, P. A., Seifert, C. E., Speechley, K. N., Booth, J. C., Stitt, L., & Gibson, M. C. (1996). A new analogue scale for assessing children's pain: an initial validation study. *Pain*, *64*(3), 435–443.

McGrath, P. J., Pianosi, P. T., Unruh, A. M., & Buckley, C. P. (2005). Dalhousie dyspnea scales: construct and content validity of pictorial scales for measuring dyspnea. *BMC Pediatrics*, *5*, 33. doi: 10.1186/1471-2431-5-33

Müller, C. (2014). Ring-gestures across cultures and times: Dimensions of variation. In: Body–language–communication: *An International Handbook on Multimodality in Human Interaction*, 2, 1511-1522. De Gruyter Mouton: New York. Editors: Cornelia Müller, Alan Cienki, Ellen Fricke, Silva H. Ladewig, David McNeill, Jana Bressem

Muris, P., Meesters, C., Mayer, B., Bogie, N., Luijten, M., Geebelen, E., … Smit, C. (2003). The Koala Fear Questionnaire: a standardized self-report scale for assessing fears and fearfulness in pre-school and primary school children. *Behaviour Research and Therapy*, *41*(5), 597–617. doi: 10.1016/S0005-7967(02)00098-0

Nielsen, J. (1994). Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human-Computer Studies*, *41*(3), 385-397.

Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: an example, design considerations and applications. *Information & Management*, *42*(1), 15-29.

Oros, L. & Richaud de Minzi, M. (2015). A Review Study of Psychometric Functioning of a Picture Scale to Assess Joy in Childhood. *Psychology, 6*, 223-233. doi: 10.4236/psych.2015.63022.

Paunonen, S. V., Ashton, M. C., & Jackson, D. N. (2001a). Nonverbal assessment of the Big Five personality factors. *European Journal of Personality*, *15*(1), 3–18. doi: 10.1002/per.385

Paunonen, S. V., Ashton, M. C., & Jackson, D. N. (2001b). Nonverbal assessment of the Big Five personality factors. *European Journal of Personality*, *15*(1), 3–18. doi: 10.1002/per.385

Paunonen, S. V., Jackson, D. N., & Keinonen, M. (1990). The Structured Nonverbal Assessment of Personality. *Journal of Personality*, *58*(3), 481–502. doi: 10.1111/j.1467-6494.1990.tb00239.x

Pollak, J. P., Adams, P., & Gay, G. (2011). PAM: A Photographic Affect Meter for Frequent, in Situ Measurement of Affect. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (S. 725–734). New York, NY, USA: ACM. doi: 10.1145/1978942.1979047

Read, J. (2008). Validating the Fun Toolkit: An instrument for measuring children's opinions of technology. *Cognition, Technology & Work, 10,* 119–128. doi: 10.1007/s10111-007-0069-9

Reynolds-Keefer, L., & Johnson, R. (2011). Is a picture is worth a thousand words? Creating effective questionnaires with pictures. Practical Assessment, *Research and Evaluation, 16*, 1–7.

Richters, J.E., Martinez, P., & Valla, J.P. (1990). Levonn: A cartoon-based structured interview for assessing young children's distress symptoms. Washington, DC: National Institute of Mental Health.

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect Grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, *57*(3), 493–502. doi: 10.1037/0022-3514.57.3.493

Sánchez, J. A., Hernández, N. P., Penagos, J. C., & Ostróvskaya, Y. (2006). Conveying Mood and Emotion in Instant Messaging by Using a Two-dimensional Model for Affective States. In *Proceedings of VII Brazilian Symposium on Human Factors in Computing Systems* (S. 66–72). New York, NY, USA: ACM. doi: 10.1145/1298023.1298033

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. Psychological Methods, 1(2), 199.

Schneider, E. (2004). Death with a Story: How Story Impacts Emotional, Motivational, and Physiological Responses to First-Person Shooter Video Games. *Human Communication Research - HUM COMMUN RES, 30*, 361–375. doi: 10.1093/hcr/30.3.361

Schubert, T. W., & Otten, S. (2002). Overlap of self, ingroup, and outgroup: Pictorial measures of self-categorization. Self and identity, 1(4), 353-376.

Sonderegger, A., Heyden, K., Chavaillaz, A., & Sauer, J. (2016). AniSAM & AniAvatar: Animated Visualizations of Affective States. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (S. 4828–4837). New York, NY, USA: ACM. doi: 10.1145/2858036.2858365

Stern, R. A., Arruda, J. E., Hooper, C. R., Wolfner, G. D., & Morey, C. E. (1997). Visual analogue mood scales to measure internal mood state in neurologically impaired patients: Description and initial validity evidence. *Aphasiology*, *11*(1), 59–71. doi: 10.1080/02687039708248455

Thielsch, M. T., Wetterkamp, M., Boertz, P., Gosheger, G., & Schulte, T. L. (2018). Reliability and validity of the Spinal Appearance Questionnaire (SAQ) and the Trunk Appearance Perception Scale (TAPS). *Journal of Orthopaedic Surgery and Research, 13*(1), 274.).

Truby, H., & Paxton, S. J. (2002). Development of the Children's Body Image Scale. *British Journal of Clinical Psychology*, *41*(2), 185–203. doi: 10.1348/014466502163967

Urist, J. (1977). The Rorschach test and the assessment of object relations. *Journal of Personality Assessment*, 41(1), 3-9.

Valla, J.-P., Bergeron, L., Bérubé, H., Gaudet, N., & St-Georges, M. (1994). A structured pictorial questionnaire to assess DSM-III-R-based diagnoses in children (6–11 years): Development, validity, and reliability. *Journal of Abnormal Child Psychology*, *22*(4), 403–423. doi: 10.1007/BF02168082

Valla, J. P., Bergeron, L., & Smolla, N. (2000). The Dominic-R: a pictorial interview for 6-to 11-year-old children. Journal of the American Academy of Child & Adolescent Psychiatry, 39(1), 85-93.

van Bakel, H. J., Maas, A. J. B., Vreeswijk, C. M., & Vingerhoets, A. J. (2013). Pictorial representation of attachment: measuring the parent-fetus relationship in expectant mothers and fathers. *BMC Pregnancy and Childbirth*, *13*(1), 138. doi: 10.1186/1471-2393-13-138

Venham, L. L., & Gaulin-Kremer, E. (1979). A self-report measure of situational anxiety for young children. *Pediatric Dentistry*, *1*(2), 91–96.

Venkatraman, N., & Grant, J. H. (1986). Construct measurement in organizational strategy research: A critique and proposal. *Academy of Management Review*, 11(1), 71-87.

Weibel, D., Schmutz, J., Pahud, O., & Wissmath, B. (2015). Measuring Spatial Presence: Introducing and Validating the Pictorial Presence SAM. *Presence: Teleoperators and Virtual Environments*, *24*(1), 44–61. doi: 10.1162/PRES_a_00214

Weibel, D., Wissmath, B., & Mast, F. W. (2010). Immersion in mediated environments: the role of personality traits. *Cyberpsychology, Behavior and Social Networking*, *13*(3), 251–256.

Wissmath, B., Weibel, D., & Mast, F. W. (2010). Measuring presence with verbal versus pictorial scales: a comparison between online-and ex post-ratings. *Virtual Reality*, *s*(1), 43-53.