A systematic review of measurement uncertainty visualizations in the context of standardized assessments

Heltne, Aleksander; Frans, Niek; Hummelen, Benjamin; Falkum, Erik; Germans Selvik, Sara; Paap, Muirne C.S.

Published in:
Scandinavian Journal of Psychology

DOI:
10.1111/sjop.12918

Scandinavian Journal of Psychology, 2023, 64, 595–608 DOI: 10.1111/sjop.12918

Review Article

A systematic review of measurement uncertainty visualizations in the context of standardized assessments

ALEKSANDER HELTNE,[1,2] NIEK FRANS,[3] BENJAMIN HUMMELEN,[1] ERIK FALKUM,[2] SARA GERMANS SELVIK[4,5] and MUIRNE C. S. PAAP[1,6]

[1]*Department of Research and Innovation, Clinic for Mental Health and Addiction, Oslo University Hospital, Oslo, Norway*
[2]*Institute of Clinical Medicine, University of Oslo, Oslo, Norway*
[3]*Department of Inclusive and Special Needs Education, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, the Netherlands*
[4]*Department of Psychiatry, Helse Nord-Trønderlag, Namsos Hospital, Namsos, Norway*
[5]*Department of Mental Health, Norwegian University of Science and Technology (NTNU), Trondheim, Norway*
[6]*Department of Child and Family Welfare, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, the Netherlands*

Heltne, A., Frans, N., Hummelen, B., Falkum, E., Germans Selvik, S. & Paap, M. C. S. (2023). A systematic review of measurement uncertainty visualizations in the context of standardized assessments. *Scandinavian Journal of Psychology*, *64*, 595–608.

This systematic review summarized findings of 29 studies evaluating visual presentation formats appropriate for communicating measurement uncertainty associated with standardized clinical assessment instruments. Studies were identified through systematic searches of multiple databases (Medline, Embase, PsycInfo, ERIC, Scopus, and Web of Science). Strikingly, we found no studies which were conducted using samples of clinicians and included clinical decision-making scenarios. Included studies did however find that providing participants with information about measurement uncertainty may increase awareness of uncertainty and promote more optimal decision making. Formats which visualize the shape of the underlying probability distribution were found to promote more accurate probability estimation and appropriate interpretations of the underlying probability distribution shape. However, participants in the included studies did not seem to benefit from the additional information provided by such plots during decision-making tasks. Further explorations into how presentations of measurement uncertainty impact clinical decision making are needed to examine whether findings of the included studies generalize to clinician populations. This review provides an important overview of pitfalls associated with formats commonly used to communicate measurement uncertainty in clinical assessment instruments, and a potential starting point for further explorations into promising alternatives. Finally, our review offers specific recommendations on how remaining research questions might be addressed.

Key words: Measurement uncertainty, visualization, standardized measurement, clinical decision making.

Niek Frans, Department of Inclusive and Special Needs Education, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, the Netherlands. E-mail: n.frans@rug.nl

INTRODUCTION

Clinicians use standardized assessment instruments, such as self-report measures, in their practice on a daily basis. Standardized clinical assessment instruments quantify patients' symptoms and experiences in standardized scores, which in turn can be used to inform a variety of decision-making tasks (Meyer, Finn, Eyde *et al.*, 2001). These different decision-making tasks can broadly be grouped into two overarching categories: (1) estimation (or mastery); and (2) classification[1] (Cronbach & Gleser, 1957; Eggen, 2010). For estimation tasks, the main goal is to determine where a patient falls on a clearly defined scale of measurement (Eggen, 2010; Linn, 1978). Examples include assessing cognitive functioning in a patient by administering an IQ-test, or assessing their symptom severity using a self-report inventory such as the Symptom Checklist−90-R (Vaurio, 2011). For classification tasks, the main goal is to assign the patient to the appropriate category (e.g., a diagnostic category or treatment program), based on one or more cut-off points (Cronbach & Gleser, 1957). Examples include assigning patients a psychiatric diagnosis if they qualify for at least a given number out of the total possible symptoms assessed by diagnostic interview, or referring patients to a given treatment if their self-reported level of symptom-load falls above a given cut-off value on a symptom severity test.

Whenever standardized measurements are used to inform decision making, it is important to acknowledge and consider that uncertainty and inaccuracy are inherent in any measurement. Standardized assessment instruments attempt to maximize accuracy by limiting the influence of external factors. Still, measurement uncertainty is unavoidably introduced, as factors other than the ones we want to measure impact the measured score (Meyer, 2007). A patient's score on a standardized assessment instrument could, for instance, be influenced by a pleasant encounter prior to filling out a self-report form, a slight misunderstanding during a clinical interview, or a bad night's sleep before a continuous attention test. It is well known that the impact of measurement uncertainty is especially high in decisions at the individual level, as opposed to assessment for group evaluations where the main focus is on an average group outcome (see for instance the EFPA guidelines for describing and evaluating psychological and educational tests; Evers, Hagemeister, Høstmælingen, Lindley, Muñiz & Sjöberg, 2013; Kruyen, Emons & Sijtsma, 2012).

Since the stakes for an individual patient are typically high in a clinical setting, it is of the utmost importance that high levels of reliability are obtained for test scores used for clinical purposes. However, several authors have warned that focusing on

optimizing reliability may come at the expense of predictive validity (see, for instance, Smits, van der Ark & Conijn, 2018). One way to mitigate this potential validity threat is to use multiple instruments (i.e., a test battery). This allows one to capture a broader range of features relevant to the construct being measured. This aligns well with clinical practice in psychiatry, where important decisions are ideally based on various sources of information. Combining this information is typically at the discretion of the clinician and can be a challenging endeavor. During such an integrative interpretation process, information concerning the reliability of each measure can be very useful in weighing information sources against each other. The level of uncertainty associated with standardized clinical assessment scores is therefore highly important for clinical decision making and represent an important informational aspect of individual instruments. Unfortunately, it is not yet common practice for clinicians to take into account the reliability of the test scores they are combining.

Depending on the decision context, measurement uncertainty affects how confidently one can interpret a patient's score, and thus how confidently one can make clinical decisions based on that score. In a classification-type decision, one would be interested in measurement uncertainty insofar as it would affect the classification outcome. For example, a patient's score might fall on either side of a classification cutoff point if that score is close to a cutoff point and/or its associated uncertainty is high. However, the field of psychiatry is increasingly transitioning towards dimensional models for diagnosis, where high importance is given to the estimated score itself (Zimmermann, Kerber, Rek, Hopwood & Krueger, 2019). In this context, the size of the measurement error, regardless of the score's position relative to some cutoff, becomes innately important.

Test publishers commonly express uncertainty associated with patients' scores by means of a reliability index (Charter & Feldt, 2001a, 2001b). Although widely used, it can be difficult to relate the meaning of such an index to the accuracy of an individual score. Other, related, statistics such as the standard error of measurement and confidence intervals, which are expressed on the same scale as the observed score, are often easier to interpret (Charter & Feldt, 2001b). Several studies have indicated, however, that these statistics can also lead to interpretational difficulties for clinicians (Belia, Fidler, Williams & Cumming, 2005; Goodwin & Goodwin, 1999; McManus, 2012). Hambleton and Zenisky (2013) argue that the communication of measurement error is one of the unique enduring challenges for score reports. If measurement uncertainty is not communicated in a clear and easily interpretable manner to stakeholders, some may ignore uncertainty information altogether (Hambleton & Zenisky, 2013). Providing clinicians with easily interpretable formats may therefore facilitate informed clinical decision making.

Visually displaying patient scores and the associated uncertainty may alleviate the interpretational difficulties associated with numerical expressions of uncertainty. Visual displays have been shown to facilitate understanding, engagement, and consideration of uncertainty in statistical data, in general (Simpkin & Armstrong, 2019). Furthermore, visual communication of measurement uncertainty has been found to draw attention to the level of uncertainty associated with a given measurement, and promote more optimal decision making considering uncertain information (Anic & Wallmeier, 2020).

In recent years, several studies exploring visual communication of measurement uncertainty have been published. Previous reviews of this literature have focused on the methodology used to evaluate these formats (Hullman, Qiao, Correll, Kale & Kay, 2019) or summarized results relevant to specific contexts, such as expressions of measurement uncertainty related to geographical/spatial data (Kinkeldey, MacEachren, Riveiro & Schiewe, 2017; Kinkeldey MacEachren & Schiewe, 2014). There are also reviews exploring the general effects of communicating uncertainty in clinical contexts, without a particular focus on visual communication, and documenting the importance of acknowledging uncertainty in clinical practice (Simpkin & Armstrong, 2019; Van der Bles, van der Linden, Freeman *et al.*, 2019). To our knowledge, no papers have yet been published summarizing studies that have explored visual communication of measurement uncertainty, with a focus on formats suited for use in a clinical context.

In the current study, we aimed to do so by exploring how different types of uncertainty visualizations affect outcome measures expected to be relevant to clinical decision making. These include the ability to identify key values (e.g., score estimates or averages) and associated uncertainty, as well as, estimating the likelihood of potential true scores given these values. Furthermore, we also explored misconceptions and misinterpretations associated with various visual formats, alongside the effect of presentation format on decision making. Lastly, we explored people's preferences for visual formats, as they may impact people's willingness to make use of displays to extract information.

We formulated the following research questions:

RQ1. To what degree does participants' ability to accurately identify and/or estimate key values vary across visual communication formats?

RQ2. What misconceptions are associated with various visual formats of uncertainty?

RQ3. To what degree does the type of visualization format used impact participants' ability to make decisions?

RQ4. To what degree do participants' preferences regarding visual formats differ?

We will summarize the findings of studies that evaluate visual representations of measurement uncertainty, for the purpose of relating these findings to the context of clinical test scores. In so doing, we may inform the development of score reports that incorporate measurement uncertainty in both existing and future standardized assessment instruments.

METHODS

Search strategy

A systematic search of the following databases was carried out in close collaboration with a senior librarian at the University of Oslo, Medical Library, in October 2020: Medline, Embase, PsycInfo, ERIC, Scopus, and Web of Science. In Scopus and Web of Science, the search was limited to subject areas deemed relevant to our purposes (see supplemental material for a detailed overview of included subject areas). In Medline, PsycInfo, and ERIC, no limitations were placed on the search. These databases

provide a wide coverage of publications within a variety of fields, that may have explored visual presentations of measurement uncertainty.

Before starting the structured search, an initial unstructured search for relevant publications was conducted by the first and second authors, in order to provide a basic overview of the available literature, to inform the development of the search strategy, and to refine the definition of relevant outcome criteria. The initial unstructured search yielded no studies exploring visual formats for communicating measurement uncertainty associated with standardized clinical measurements. The search did however identify 18 studies exploring visual communication of measurement uncertainty around a single point estimate (e.g., population means, sample means, and probability estimates). It was therefore decided to develop a systematic search strategy which included any single point estimate, and not just patient scores on standardized clinical assessments.

Titles and abstracts of these 18 articles identified in the initial unstructured search were screened by the first and second authors, in order to identify a list of keywords used to express: (1) visualizations; (2) uncertainty or measurement error; and (3) relevant outcome measures related to understanding, preference, accuracy, and decision making. This list was adapted and refined in collaboration with a senior librarian at the University of Oslo, Medical Library, who identified synonyms and related words using database thesauri. The 18 articles used to identify initial keywords were also used to test the search strategy, and to refine the inclusion and exclusion criteria.

The search terms and combinations used in Medline, Embase, PsycInfo, and ERIC were: "(((depict* OR display * OR glyph? OR graphical OR graphics OR represent * OR visual * OR imagery or graph? OR diagram? OR chart? or presentation*) ADJ5 (ambiguous data OR confidence interval? OR measurement error? OR uncertain * OR probability distribution*))) AND ((ability OR accura * OR comprehen * OR decision OR error rates OR reasoning OR inferen * OR judg * OR estimation OR prefer * OR perform * OR understand * OR utility OR appeal))." Searches were made in title, abstract, and keywords fields. A detailed outline of the search strategies has been provided in the supplemental material.

Inclusion and exclusion criteria

For this systematic review, we chose to limit our scope to studies published in peer reviewed journals from 1985 through October 2020. This time frame was chosen to narrowly match that of a previous review by Hullman *et al.* (2019), which included studies published between 1987 and 2018.

We included studies in which participants were shown at least one visualization of uncertainty around a point estimate. Any point estimate comparable in presentation to single test scores (e.g., population or sample means, and probability estimates) was accepted for inclusion in this review. Studies that explored uncertainty visualization of more complicated estimates, such as coefficients of statistical models (e.g., regression slopes, factor loadings) were not included in this review, since correct interpretations require a high degree of knowledge about the estimate being represented in the visualization. Additionally, we excluded studies in which participants were presented with visualizations of uncertainty in geolocation data (e.g., positions on two-dimensional maps).

Eligible studies needed to include one of the following outcome measures (described briefly here and in more detail in the data extraction section below): (1) accuracy (e.g., a quantitative measure of accuracy in mean or probability estimation); (2) understanding (e.g., qualitative or quantitative measures of participants' ability to reproduce and answer questions about understanding and/or internal representation of the information presented to them in a display); (3) decision process (e.g., think-aloud procedures) and decision quality (e.g., hypothetical decision-making scenarios); or (4) preference (e.g., ratings of visual appeal, readability or usability ratings).

Study selection

Following the completion of the search, retrieved articles were exported to an EndNote file, and duplicates were removed. An initial screening was then carried out by the first author, in order to remove gray literature, such as dissertations, conference papers, and other non-peer reviewed literature, as well as articles published before 1985.

After this initial screening, an independent screening of titles and abstracts was carried out in Rayyan (Ouzzani, Hammady, Fedorowicz & Elmagarmid, 2016), which is a web-based tool for organizing and screening records for systematic reviews or meta analyses. Title and abstract screening were carried out by the first author and a trained graduate student, according to the inclusion criteria specified for this study. To prevent exclusion of articles with a limited description of outcome measures in abstracts, inclusion criteria related to outcome measures were not considered at this stage. Any article labeled as "included" by at least one rater was taken forward for full-text screening. Articles labeled as "uncertain" by at least one rater were first discussed by both raters, and then passed to the full-text review stage, if a consensus to exclude the article could not be reached.

After title and abstract screening, independent full-text screening was carried out by the first and second authors. At this stage, inclusion criteria related to outcome measures were evaluated. Any disagreements between the two raters were discussed until consensus was reached for all articles. Reference lists for all included articles were screened manually by the first author for additional studies that fit the inclusion criteria, but were missed by the search algorithms.

Quality appraisal

In order to assess the potential risk of bias in the findings of the included studies, the design of each study was independently evaluated by the first and second authors with respect to 11 items related to sample size, sample representativeness, the reliability and validity of included measures, inclusion of confounding variables in statistical analyses, reporting of non-response, missing data and assumptions required for the statistical analyses chosen. For qualitative studies, items regarding the reliability of included measures, non-response, and missing data were not rated, leaving a total of eight items on which these studies were rated. These evaluation criteria were based on existing frameworks to evaluate risk of bias/methodological quality in experimental and/or qualitative studies, such as those published by the Joanna Briggs Institute (2020a, 2020b); the Critical Appraisal Skills Program (2018, 2020); and the Cochrane risk of bias assessment (Higgins, Altman, Gøtzsche *et al.*, 2011). The full set of evaluation criteria used in this review can be found in Tables S2.1–S2.4 of the supplementary materials. The specific selection of items was made to accommodate the wider range of study designs expected for studies relevant to this review. Any disagreements between raters were discussed until consensus was reached.

Data extraction

Included studies were grouped according to the outcome measures listed in the inclusion criteria section above. These were specifically defined as follows: (1) accuracy – studies in which researchers defined a quantitative measure of accuracy for participants' estimates of certain values based on the visualization format (e.g., estimates of a mean/best estimate, or estimates of the probability of a value higher/lower than a given value X); (2) understanding – studies in which the researchers evaluated participants' understanding and/or internal representation of the information presented to them in a visualization format, either through qualitative measures (e.g., think-aloud procedures or open-ended questions about the information presented), or through concrete tasks (e.g., interpreting the shape of the underlying probability distribution for a confidence interval); (3) decision making – studies in which participants were asked to make a decision, based on the information presented to them (e.g., choose between investment alternatives or treatment options in light of uncertain outcomes, or perform a preventive action in light of uncertain risk information); and (4) preference – studies in which participants ranked, rated, or gave qualitative feedback on their preference for various visual formats.

After grouping studies by outcome, the following data were transferred to Excel forms: (1) sample size; (2) sample composition; (3) visualization format(s) shown to participants; (4) context presented to participants; (5) covariates included in analysis; (6) tasks participants were asked to perform; and (7) reported findings.

RESULTS

Study selection

Figure 1 outlines the screening and selection process. Our database search identified 8,652 records. These were pooled with the 18 studies identified through the unstructured search used to generate initial search terms, adding up to a total of 8,670 studies. Duplicates were removed, leaving 4,598 unique records. After title, abstract, and full-text screening, a total of 29 studies were selected for inclusion in this review. Nineteen of these were identified through the systematic search of databases, and eight were identified through searching reference lists of included studies. Two additional studies were transferred from the list of 18 studies identified by the initial unstructured search, but which were not identified by the systematic database search. Table S1 of the supplementary materials gives a brief overview of: (1) sample

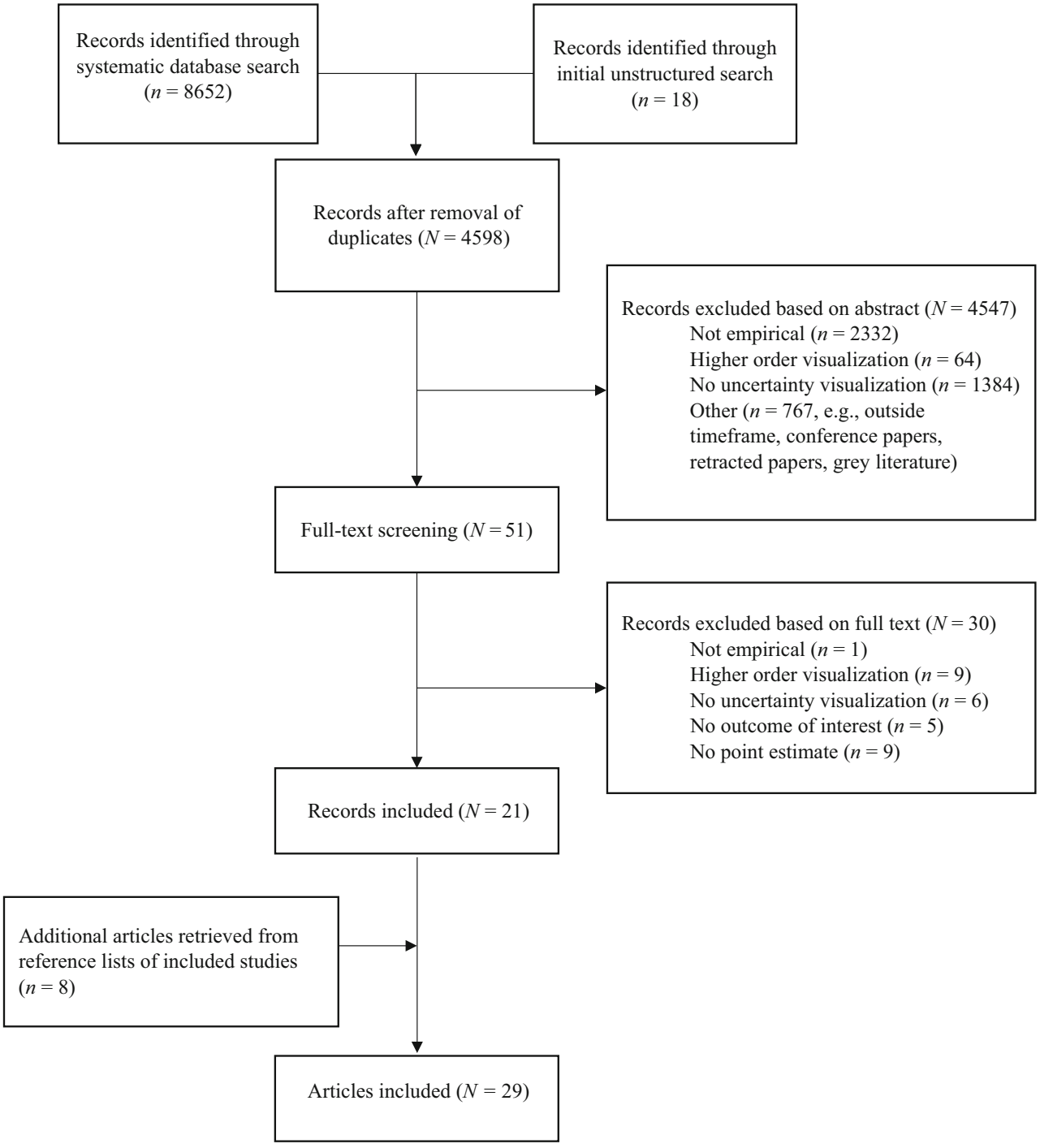

Fig. 1. Flowchart depicting the screening and inclusion process.

size; (2) sample composition; (3) visualization format(s) shown to participants; (4) context presented to participants; (5) covariates included in analysis; (6) tasks participants were asked to perform; and (7) reported findings of all included studies.

Quality appraisal

Quality appraisal showed that the studies satisfied between one and eight out of eleven quality appraisal items, with an average of 4.4 items. Several studies involved small samples from poorly defined populations (e.g., users, general online samples). The majority of the studies (66%) did provide a clear description of key sample characteristics, and included known confounders, such as education and numeracy, in their analyses (77%). However, only 34% of included studies provided a definition of the population they intended to sample/generalize their findings to. In addition, most of the studies used researcher-developed instruments with unknown reliability and validity. Even though 41% provided a clear theoretical rationale for these measures, the use of measures with unknown reliability and validity made it difficult to assess the quality of the results.

Two-thirds (67%) of the reviewed studies reported rates of missing data, but less than a quarter of these studies (23%) reflected on the potential impact of missing data on the outcomes. Non-response rates were only reported in 15% of the studies, possibly because sampling methods did not provide a clear insight into the non-response rate. Finally, only a small proportion of the reviewed articles (8%) reflected on the statistical assumptions of their main analyses (e.g., normality, linearity, and homogeneity of variance).

Three qualitative papers were included in this review. These satisfied on average 5.6 out of eight criteria. All these studies gave clear descriptions of their analyses and the study context, and at least partial descriptions of the key sample characteristics. The lowest scores were seen for criteria concerning the study sample, such as reasoning for sample size, and motivation for sampling strategy, which were only reported in one out of three studies.

The appraisal outlined here does not necessarily imply that the reviewed articles are of poor quality, but it indicates some potential sources of bias that will be addressed further in the discussion section. A full overview of the appraisal can be found in Tables S2.1–S2.4 of the supplemental material.

Study characteristics

The 29 included studies explored presentation of uncertainty information in a wide variety of contexts (e.g., clinical contexts [n = 4], meteorological forecasts [n = 9], and educational measurements [n = 3]). Only three studies, all in the field of educational measurement, presented participants with score reports from standardized assessments. Studies conducted in a clinical setting presented patients with uncertain risk estimates. In the meteorological forecast studies, formats representing projected temperatures, rainfall, or flood risk were used. Samples in included studies were often highly educated, with 19 out of 29 studies including samples where more than 50% of the participants had attended or finished undergraduate or graduate school. Furthermore, in six studies, samples were either partially or completely comprised of students or graduates from the fields of medicine and/or psychology. The included studies employed quantitative/mixed method designs (n = 26) and qualitative designs (n = 3). A total of 13 unique visual formats for communicating measurement uncertainty were examined in these studies. The formats were either simple displays that did not emphasize the shape of the underlying distribution (e.g., error bars) or displays that emphasized the shape underlying probability distribution (e.g., histograms, probability density functions, and violin plots). The four most commonly used formats are depicted in Fig. 2, whereas all remaining formats are depicted in Figs. S1–S9 of the supplemental materials.

Accuracy

Twelve studies included an accuracy outcome measure. The tasks most commonly used in these studies involved identifying the mean/best estimate (n = 7), or estimating the probability of a given value considering the presented display (n = 8).

Out of the seven studies with a mean estimation task, five found that mean estimates were more accurate for plots which explicitly marked the center of the probability distribution, such as error bars, boxplots, or other visual formats where the mean was superimposed on the display (Allen, Edwards, Snyder, Makinson & Hamby, 2014; Edwards, Snyder, Allen, Makinson & Hamby, 2012; Gschwandtner, Bögl, Federico & Miksch, 2016; Hullman, Resnick & Adar, 2015; Ibrekk & Morgan, 1987). Correll and Gleicher (2014) found no difference in mean estimation accuracy across visual formats. Nadav-Greenberg and Joslyn (2009) compared participants' mean estimates across several textual and numeric formats, as well as a combined numeric and visual format, where a boxplot accompanied a numeric format, and found no additional benefit of the boxplot compared to numeric formats alone.

Out of the eight studies with a probability estimation task, six studies compared accuracy across plots depicting the shape of the underlying probability distributions versus error bars that did not depict this shape (Allen *et al.*, 2014; Correll & Gleicher, 2014; Edwards *et al.*, 2012; Hullman *et al.*, 2015; Ibrekk & Morgan, 1987; Zwick, Zapata-Rivera & Hegarty, 2014). Five of these studies found that participants performed this task more accurately, when presented with the shape of the underlying distribution; whereas one study by Zwick *et al.* (2014) found no significant difference between standard error bars and variable width error bars. The remaining two studies compared either multiple formats that all depicted the underlying distribution (Lorenz, Dessai, Forster & Paavola, 2015), or compared a text format with combined textual and visual formats where a histogram accompanied the text (Gibson, Rowe, Stone & Bruin, 2013). Of these two studies, only Gibson *et al.* (2013) found significant differences in accuracy between included plots, with the combined textual and visual format outperforming the text-only format.

The specific displays presented to participants varied among the eight studies that included a probability estimation task. Table 1 shows which formats were included in the probability estimation studies that found significant differences across multiple visual formats. This table also shows the ranked performance of formats

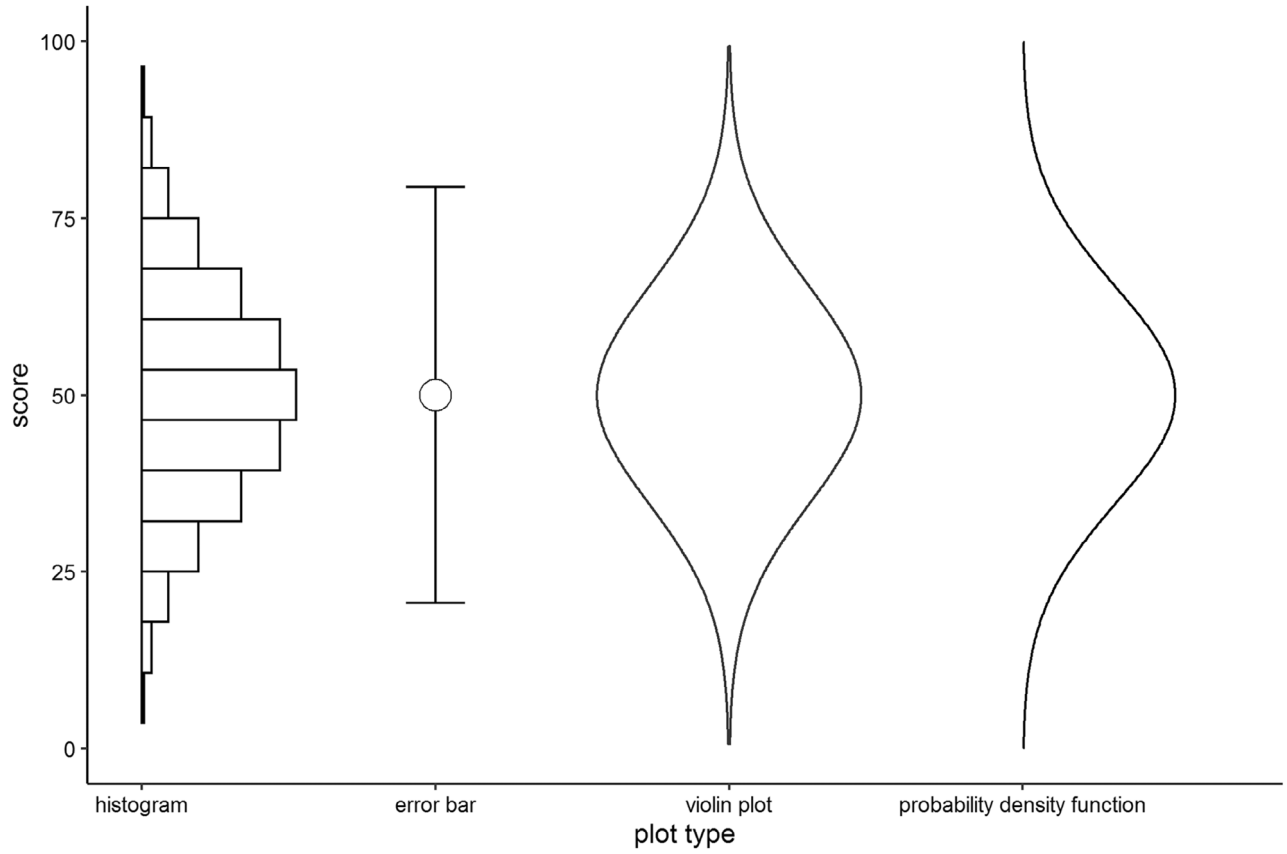

Fig. 2. Overview of the most commonly used formats.

Table 1. *Overview of comparisons of formats across studies which included a probability estimation task*

Study (Author, year)	Error Bars	Boxplot	PDF	CDF	CCDF	Gradient Plot	Violin Plot	HOP	Histogram
Allen *et al.* (2014)	4	–	1*	3	2*	–	–	–	–
Correll and Gleicher (2014)	2	1*	–	–	–	1*	1*	–	–
Edwards *et al.* (2012)	2*	3	4*	5	1*	–	–	–	–
Hullman *et al.* (2015)	2	–	–	–	–	–	2	1*	–
Ibrekk and Morgan (1987)	3	3	2	1	–	2	2	–	2
Gschwandter *et al.* (2016)	–	–	2	–	–	1*	3	–	–

Notes: The values in each column denotes, in descending order and based on participants accuracy on the probability estimation task, the ranked performance of each plot type in each study. The notation "–", denotes the plot was not included in the given study.
Plot type rankings marked with "*" indicate that the given plot type was explicitly reported to significantly outperform lower ranked plots in the given study.
Abbreviations: PDF = probability density function, CDF = cumulative probability density function, CCDF = compensated cumulative probability density function, HOP = hypothetical outcome plot.

across these studies. As can be seen from the table, only error bars were compared against each of the other plot types. Error bars were consistently outperformed by other plot types that depicted the shape of the underlying distribution. Across these other plot types (Boxplots, probability density function [PDF], cumulative probability density function [CDF], compensated cumulative probability density function [CCDF], gradient plots, violin plots, hypothetical outcome plot [HOP], and histograms), no clear pattern of performance emerged, that is, no single format consistently outperformed all others.

Two studies included in this review presented participants with unique tasks, not readily comparable with other included studies. Stock and Behrens (1991) explored participants' ability to estimate whisker length based on interquartile ranges for box-and-whisker plots and midgap plots (box-and-whisker plots where the box was left out, leaving a gap between the mean and the starting point of the whiskers. See Fig. S1 in the online supplement for an illustration of a box-and-whisker plot). They found that accuracy was higher for estimations of whisker length for boxplots than midgap plots, and concluded that whitespace may be difficult for participants to mentally manipulate, and is therefore considered potentially unsuited for communicating quantities. Hullman, Kay, Kim and Shrestha (2018) asked students to use the observed sample mean and associated confidence interval of a previous experiment to predict and draw the expected distribution of means, if the experiment was replicated several times. They found that participants made more accurate predictions, when they received a training task where the correct distribution was shown

with a discrete visualization (a quantile dot plot), as compared to a training task where the distribution was shown with a continuous visualization (probability density function). They also found that discrete depictions improved participants' graphical recall of the distribution.

The impact of confounding variables on accuracy was reported in six studies. Gibson *et al*. (2013) and Zwick *et al*. (2014) found that greater numerical and/or statistical experience was positively associated with performance on accuracy tasks. Ibrekk and Morgan (1987), on the other hand, found that participants' self-reported statistical experience had little or no effect on accuracy in these tasks. Edwards *et al*. (2012) found familiarity with the presentation format to predict performance on mean estimation tasks but not on probability estimation tasks, whereas Allen *et al*. (2014) found that familiarity impacted both mean and probability estimation, but only for PDFs. Finally, Hullman *et al*. (2015) found that participant accuracy in their study improved when using hypothetical outcome plots, but only when variance in the data was low.

Understanding

Fourteen studies explored understanding. Ten of these studies showed that participants often have trouble perceiving the shape of the underlying probability distribution, when presented with error bars or numerical uncertainty ranges (e.g., confidence intervals). Dieckmann, Gregory, Peters and Hartman (2017) and Dieckmann, Peters and Gregory (2015) found that perceptions of normality were significantly more frequent, when a normal distribution plot was presented alongside numerical ranges, as opposed to numerical ranges presented alone. Gschwandtner *et al*. (2016), similarly, found that participants were more likely to perceive that test scores were normally distributed, when presented with visualizations that showed the shape of the underlying distribution (histograms, PDFs, and violin plots), as compared with plots that did not show the shape of the underlying distribution (variations on error bars). Kalinowski, Lai, and Cumming (2018) showed that most of their participants failed to perceive normality for the underlying probability distribution, when shown standard error bars. A brief tutorial, where participants interacted with a novel visualization called the "cats' eye confidence interval" (similar to a violin plot), helped these participants understand the shape of the underlying probability distribution. Zapata-Rivera, Zwick, and Vezzu (2016) similarly reported that participants who attended a short tutorial, introducing them to variable-width error bars, were more aware of the shape of the underlying probability distribution than participants who had not participated in the tutorial.

Dieckmann *et al*. (2017), Han, Klein, Lehman, Massett, Lee and Freedman (2009), and Han, Klein, Lehman, Killam, Massett and Freedman (2011) found that regarding uncertainty ranges, participants may interpret preferred values as more likely than less preferred values. Han *et al*. (2009) found that, when participants were presented with both numerical and visual risk ranges, risk interpretations among participants with high dispositional optimism tended to fall toward the lower end of the risk range. Similarly, Dieckmann *et al*. (2017) found that political views could also impact interpretations of uncertainty ranges to be more in line with prior beliefs when interpreting uncertainty ranges.

One study reported a so-called "cliff effect," when participants were presented with error bars (Hoekstra, Johnson & Kiers, 2012). This entailed a misconception, where participants considered values outside a confidence range to be much less likely or even impossible, compared to values within the confidence interval. Belia *et al*. (2005) documented similar misconceptions associated with error bars in a sample of experienced researchers. In this study, participants were found to have misconceptions about how much overlap two uncertainty ranges could have, and still be significantly different from each other. Participants were asked to adjust the position of point estimates (indicating, e.g., a sample mean) and associated error bars (representing either 95% confidence intervals or standard error), to indicate significant difference. When the error bars represented 95% confidence intervals, participants adjusted the position of the means too much, indicating a misconception that confidence intervals of two population means could not overlap, if the population means were statistically different.

Correll and Gleicher (2014) documented a phenomenon called "within the bar bias," where participants presented with bar graphs perceived points that fell within the borders of the bar as more likely than those that fell outside of it, even when uncertainty regarding the size of the bar was presented with error bars superimposed on the plot.

Lastly, Padilla, Hansen, Ruginski, Kramer, Thompson and Creem-Regehr (2015) presented their participants with the task of indicating which of two temperature predictions was the most accurate, given the true temperature (i.e., the temperature that was actually measured). They found that, when two predictions were equidistant from the true temperature, participants tended to confuse precision with accuracy, believing that the estimate with the smallest confidence interval was the most accurate, which was not necessarily the case.

These studies all documented misconceptions associated with error bars which were not found for formats that emphasized the shape of the underlying probability distribution (e.g., probability density plots). It is also worth mentioning that two of the above-mentioned studies found that a generic drawing of a standard normal distribution function was sufficient to alleviate misconceptions about the underlying distribution (Dieckmann *et al*., 2017, 2015). On the other hand, Zwick *et al*. (2014) and Kalinowski *et al*. (2018) both found that the participants had erroneous ideas about the shape of the underlying distribution, even when they were shown visualizations that emphasized it. Zwick *et al*. (2014) showed participants both standard error bars and variable-width error bars. In think-aloud procedures, some of their participants erroneously stated that standard confidence bands would be better and more accurate than variable-width error bars, as they would not give the "false" impression that some values within the range were more likely than others. Similarly, a small proportion of the participants in the study of Kalinowski *et al*. (2018) failed to correct their initial misconception that all values within the confidence range were equally likely, even after getting the tutorial on the cat's eye confidence interval.

Three studies reported specific misinterpretations of other visualizations. Hopster-den Otter, Muilenburg, Wools, Veldkamp and Eggen (2019) found that the use of color coding or blur to indicate uncertainty was frequently misunderstood, as participants were unsure what color value and blur indicated. Ibrekk and Morgan (1987) noted that their participants had trouble extracting the mean from visualizations that did not mark the mean, although the underlying probability distribution shape was provided. Here, participants tended to select the highest point on the probability curve (i.e., the mode). For CDFs however, participants tended to select the maximum value as opposed to the mean, as this was the highest point on the probability curve. Savelli and Joslyn (2013) found that visualizations of range-based weather forecasts (error bars and gradient plots) led some participants to interpret the high and low points of the distribution as diurnal minimum and maximum temperatures. The authors referred to this misinterpretation as a "deterministic construal error," indicating a tendency that probabilistic information (e.g., an uncertainty range) is interpreted as indicating deterministic variance (e.g., daytime-high and nighttime-low temperatures). The number of misinterpretations diminished, when participants were presented with deterministic forecasts as compared to range-based forecasts. The number of misconceptions associated with range-based forecasts was only reduced when these were presented textually, without any visual aids.

Four studies reported statistically significant confounders which impacted understanding. Bruine de Bruin, Stone, Gibson, Fischbeck and Shoraka (2013) and Gibson *et al.* (2013) found that less numerate participants were less able to correctly interpret and understand uncertainty information. Schapira, Nattinger and McHorney (2001) similarly found that less educated participants were more skeptical of uncertainty information, leading them to consider the expression with uncertainty to be less trustworthy than a deterministic estimate. It should be noted, however, that Belia *et al.* (2005) documented clear misconceptions about uncertainty information among samples of published researchers within the fields of behavioral neuroscience, psychology, and medicine, indicating that even people who could be expected to be highly educated and to have statistical knowledge can still have trouble understanding uncertainty information.

Decision making

Out of the eleven studies that included decision making as an outcome measure, seven studies involved a behavioral choice task, where the authors had defined an optimal choice based on the demands of the task and the information given to participants (Allen *et al.*, 2014; Anic & Wallmeier, 2020; Durbach & Stewart, 2011; Edwards *et al.*, 2012; Nadav-Greenberg & Joslyn, 2009; Ramos, van Andel & Pappenberger, 2013; Savelli & Joslyn, 2013).

Four out of the seven studies with an optimal choice compared conditions in which participants were either provided or not provided uncertainty information. Nadav-Greenberg and Joslyn (2009) and Ramos *et al.* (2013) compared combined textual/numerical and visual displays which either did or did not express uncertainty. Both studies found that presenting uncertainty led more participants to choose the optimal decision. It is worth mentioning however, that Nadav-Greenberg and Joslyn (2009) found no additional benefit of a combined numerical and visual display (numerical expression of uncertainty range accompanied by a boxplot), compared to numerical displays alone. Anic and Wallmeier (2020) found that a histogram which emphasized the relative risk associated with various potential payoffs of investment alternatives produced more risk-aware investment decisions than displays depicting only expected outcomes. Authors attributed this to an increased awareness of uncertainty based on this more detailed display. Lastly, Savelli and Joslyn (2013) found that participants were more decisive and thus more prone to take appropriate preventative measures in light of range-based weather forecasts than deterministic forecasts.

Three out of the seven studies with an optimal choice compared multiple presentation formats. Allen *et al.* (2014) found that participants made the best decisions when presented with probability density function plots and CCDFs as opposed to other included plots (error bars, CDF, and scatter plots). Durbach and Stewart (2011) found that visualizations, such as a simplified PDF divided into quantiles and three-point approximations, led to better decisions than more complex visualizations, such as standard probability density functions, which appeared to overwhelm decision makers. Edwards *et al.* (2012) found that their participants made the best decisions, when presented with error bars, boxplots, and scatterplots, as opposed to cumulative distribution functions or various versions of probability density functions.

Among the four studies without an optimal choice, one found no significant effect of display type (Gibson *et al.*, 2013), whereas the other three did. Hopster-den Otter *et al.* (2019) found that teachers more often requested additional information in light of uncertain test scores, when scores were presented with error bars, as opposed to point estimates alone, or with color values or blur to indicate uncertainty. Bruine de Bruin *et al.* (2013) found that participants expressed greater support for proceeding with a construction project, when shown visual rather than textual presentations of risk associated with that project. Lastly, Bansback, Harrison, and Marra (2015) found that when patients were presented with uncertain risk estimates associated with treatment options, patients tended to choose the treatment option which had the lowest uncertain risk estimate.

Four studies reported on confounders which impacted decision making, regardless of presentation format. Hopster-den Otter *et al.* (2019) found that participants with more statistical experience were more inclined to ask for more information, when presented with uncertainty visualizations. Ramos *et al.* (2013), Allen *et al.* (2014), and Edwards *et al.* (2012) found that when participants were faced with monetary pressure (being low on funds when suboptimal choices lead to penalty cost), cognitive load (performing a concurrent memory task), or time pressure, they were more indecisive (Ramos *et al.*, 2013) or less likely to make optimal decisions (Allen *et al.*, 2014; Edwards *et al.*, 2012).

Preference

Twelve studies explored participants' preferences for presentations of uncertainty. Three of these studies explored participants' preferences for uncertainty to be communicated to them. Han

et al. (2009) found that participants preferred uncertainty about cancer risk to be communicated to them, and that they trusted risk ranges more than fixed risk estimates. Han *et al.* (2011) on the other hand, found no difference in perceived trustworthiness among three displays, of which one did not depict uncertainty information. Lastly, Schapira *et al.* (2001) found that participants with lower education reported they perceived formats which reported uncertainty as less trustworthy than deterministic displays. More educated participants on the other hand, expressed that they accepted that science involves uncertainty, and therefore preferred to be informed about the uncertainty in risk estimates. In light of this finding by Schapira *et al.* (2001), it is worth pointing out that Han *et al.* (2009) specified a minimum of high school-level education for participant inclusion, whereas Han *et al.* (2011) specified inclusion criteria to ensure that 30% of their sample would have a high school-level education or lower. In accordance with the findings by Schapira *et al.* (2001), it is possible that differences in sample education between the two Han *et al.* studies contributed to the contradictory findings.

Three studies compared textual/numerical communication formats with graphical formats. Bruine de Bruin *et al.* (2013) found that participants gave higher liking and trust ratings for graphical communication of uncertainty (histogram) than textual formats. Gibson *et al.* (2013), on the other hand, found no differences between participants' preference ratings for textual and graphical (histogram) formats for communicating uncertainty. The level of education of the participants and contexts for visualizations of the studies of Bruine de Bruin *et al.* (2013) and Gibson *et al.* (2013) were quite similar (less than 50% had college-level education, and participants were presented with risk measures associated with a construction project). Durbach and Stewart (2011) included a somewhat different sample (post-graduate students), and presented them with investment opportunities with varying projected earnings. The aforementioned authors reported that participants found numerical formats (three-point minimum, median and maximum) and fixed estimates (represented by bar graphs) easier to use than graphical formats (histogram, segmented probability distributions), when comparing investment alternatives.

Five studies compared participants' preferences across multiple visual formats. Four of these studies found that participants gave the highest preference ratings to the least detailed – and thus visually simplest – plots presented to them. Allen *et al.* (2014) and Edwards *et al.* (2012) found that their participants preferred error bars, boxplots and scatter plots, which they rated as easier to interpret and use than cumulative probability functions, complementary cumulative distribution function, and probability density functions. Similarly, Hopster-den Otter *et al.* (2019) found that their participants preferred error bars over plots that used color coding and blur to indicate uncertainty, which participants found to be confusing, and which were frequently misinterpreted. Lorenz *et al.* (2015) reported similar findings, with participants in this study consistently preferring the format they perceived as either most familiar or easiest to interpret. Gschwandtner *et al.* (2016) found that their participants gave much higher preference ratings to ambiguation plots (which used a lighter color to represent more uncertain regions on a bar graph representing the probability distribution) than they gave to violin plots, probability density functions, and gradient plots. Participants in this study, however, also preferred probability density functions over error bars, which goes against the indication that simpler plots are always the most preferred.

Three studies identified additional factors which may impact preference ratings. Bruine de Bruin *et al.* (2013) and Zwick *et al.* (2014) found that participants' numeracy impacted their preference ratings. In the study by Bruine de Bruin *et al.* (2013), low numerate individuals were found to prefer a textual explanation coupled with a histogram of risk over a text-only display. In Zwick *et al.* (2014), highly numerate individuals were found to prefer variable width confidence bands, while less numerate individuals preferred standard error bars. Lastly, Lorenz *et al.* (2015) found some indications that participants' preference ratings may depend on the purpose for which a certain plot was deemed useful, with participants sometimes rating certain plots (e.g., histograms) as most useful for planning, and other plots (bubble plots) as best suited for persuading others.

Four studies explored how preference ratings were associated with performance on certain tasks. Lorenz *et al.* (2015) found no association between participants' perceived understanding of and preferences for visual displays (which were highly correlated) and their actual comprehension of the displays. Zwick *et al.* (2014), and Durbach and Stewart (2011) found that participants tended to give the highest preference ratings to the plots with which they performed the best. Gschwandtner *et al.* (2016), on the other hand, found that participants least preferred gradient plots, compared to all other plots used in the study, notwithstanding that the gradient plot led participants to make the best probability estimates.

DISCUSSION

This systematic review explored the relationship between visual presentations of measurement uncertainty and participants' understanding of uncertainty information. Our goal was to translate findings of relevant studies into a clinical decision-making context. Our systematic search identified 29 studies which satisfied our inclusion criteria. Strikingly, we found no studies wherein clinicians were presented with visual presentations of measurement uncertainty associated with standardized test scores. The included studies did however contain important findings related to how presentation formats impact participants' ability to: (1) extract/estimate key information from presented formats; (2) understand key concepts and implications of uncertainty information; (3) make optimal decisions in light of uncertainty information; and (4) participants' preferences for various visual formats of communicating measurement uncertainty. In 19 out of the 29 included studies, participants were highly educated, with more than 50% of participants having either attended some college or completed undergraduate- or higher-level degrees. In six of these studies, samples were either entirely or partly comprised of students or graduates following undergraduate or graduate programs in psychology, medicine, or neuroscience. Some findings of this review may therefore generalize to other highly educated populations such as healthcare professionals, who were the target population of this review.

A consistent finding across included studies was that participants most accurately extracted and/or estimated key

information from a graphical display when the information they were asked to extract/estimate was emphasized in the display. For example, when asked to estimate probability, participants performed best when they were shown a format which emphasized the shape of the underlying probability distribution of uncertainty ranges (Allen *et al*., 2014; Correll & Gleicher, 2014; Edwards *et al*., 2012; Gschwandtner *et al*., 2016; Hullman *et al*., 2015; Ibrekk & Morgan, 1987; Zwick *et al*., 2014).

Several studies documented misconceptions regarding the underlying probability distribution when participants were shown error bars or numerical uncertainty ranges. In these studies, participants tended to perceive the underlying distribution as either uniform (i.e., believing all values within the uncertainty range are equally likely) or skewed (i.e., believing values on one side of the distribution are more likely than those on the other side) (Belia *et al*., 2005; Correll & Gleicher, 2014; Dieckmann *et al*., 2017, 2015; Gschwandtner *et al*., 2016; Han *et al*., 2009; Han *et al*., 2011; Kalinowski *et al*., 2018; Zwick *et al*., 2014). Presenting participants with a presentation format which emphasized the shape of the underlying distribution (e.g., PDFs, violin plots, or gradient plots) appeared to eliminate such misconceptions (Dieckmann *et al*., 2017, 2015; Han *et al*., 2009; Kalinowski *et al*., 2018; Zwick *et al*., 2014). Finding misconceptions associated with simplistic presentation formats is not surprising, given the vast body of research documenting how difficult it is for many to reason about statistical concepts (see for instance: Kahneman, 2003; Tversky & Kahneman, 1983). When communicating uncertainty about patient scores to clinicians, it is important to alert them to the wide range of biases which may influence their perception of this information. Findings of multiple included studies indicate that well-documented cognitive biases, such as confirmation bias, affect heuristic, and overconfidence may influence readers' perceptions of uncertain test scores. Furthermore, studies by Han *et al*. (2009, 2011) document how ambiguity aversion may lead readers to avoid, overlook or ignore uncertainty information altogether. These findings are consistent across several included studies and highlight the importance of clear communication of uncertainty in clinical score reports. Surprisingly, the included studies documented fewer biased interpretations and misconceptions related to complex formats such as probability density functions than they did for simplistic formats, such as error bars or numerical uncertainty ranges. More detailed plots may therefore be less vulnerable to these misconceptions.

Overall, presenting decision makers with uncertainty information was generally found to make them more aware of uncertainty and more likely to take uncertainty into account when making decisions (Anic & Wallmeier, 2020; Bruine de Bruin *et al*., 2013; Hopster-den Otter *et al*., 2019; Nadav-Greenberg & Joslyn, 2009; Ramos *et al*., 2013; Savelli & Joslyn, 2013). These findings contradict those of previous studies, indicating that decision makers are often unable to benefit from probability information (e.g., Tversky & Kahneman, 1974; Van der Bles *et al*., 2019). However, presenting decision makers with more detailed information, such as information about the shape of the underlying probability distribution, did not always lead to more optimal decision making. In fact, error bars, boxplots and three-point approximations were found to promote the most optimal decision making in two out of three studies comparing optimal decision making across visual formats (Durbach & Stewart, 2011; Edwards *et al*., 2012). Authors of these studies suggested this may be due to the fact that decision-making tasks are more complex and demanding than probability estimation tasks. Previous studies have suggested that, when faced with demanding tasks, people are more sensitive to informational overload, and are more prone to use automatic, heuristic information processing (Byström & Järvelin, 1995; Lurie, 2004). In line with these previous findings, Durbach and Stewart (2011) suggested that less detailed formats may be more helpful for decision makers, partly because they find them easier to process and interpret. Similarly, Edwards *et al*. (2012) suggested that their participants may have used heuristic approaches to error bars in order to obtain an approximate idea of the most likely outcomes. Using the best estimate and associated confidence intervals, participants may determine that the most likely values are the ones falling within the middle half of the confidence intervals on either side of the best estimate. This strategy does however require the participants to correctly perceive the underlying probability distribution as normally distributed and not skewed or uniform.

Studies examining participants' preferences regarding visual formats generally suggested simpler, less detailed plots such as error bars and box plots were preferred over more complex and detailed plots such as probability density functions and quantile dot-plots (Allen *et al*., 2014; Gschwandtner *et al*., 2016; Hopster-den Otter *et al*., 2019). This is in line with the statement of Hambleton and Zenisky (2013) that test users often indicate that detailed information on measurement error clutters score reports. Participants' preference ratings were also found to depend on their familiarity with a given display (Allen *et al*., 2014; Edwards *et al*., 2012; Lorenz *et al*., 2015). Interestingly, only one study found a relationship between preference ratings and performance on probability estimation (Zwick *et al*., 2014), and no studies found a relation between preference ratings and performance on decision-making tasks.

Clinical implications and recommendations for further research

This review suggests that communicating uncertainty may promote awareness of uncertainty and its implications. In a clinical context, such awareness may prompt clinicians to consider measurement uncertainty and its implications during clinical decision making, thus enabling them to make more optimal choices. Measurement uncertainty is not always communicated in standardized assessment instruments commonly used today. Given the overall positive effect of presenting uncertainty information on decision making (Anic & Wallmeier, 2020; Nadav-Greenberg & Joslyn, 2009; Ramos *et al*., 2013; Savelli & Joslyn, 2013), we strongly recommend that measurement uncertainty be communicated for standardized clinical assessment instruments.

Current clinical assessment instruments that report measurement uncertainty, often do so by means of either numerical confidence intervals or error bars. Results presented in this review suggest these formats may be poorly suited for informing probability estimation (Allen *et al*., 2014; Correll & Gleicher, 2014; Edwards *et al*., 2012; Gschwandtner *et al*., 2016;

Hullman *et al*., 2015; Ibrekk & Morgan, 1987; Zwick *et al*., 2014), and that these formats are associated with misinterpretations of the underlying probability distribution of uncertainty ranges (Belia *et al*., 2005; Correll & Gleicher, 2014; Dieckmann *et al*., 2017, 2015; Gschwandtner *et al*., 2016; Han *et al*., 2009; Han *et al*., 2011; Kalinowski *et al*., 2018; Zwick *et al*., 2014). These findings were consistent across several studies, some of which included samples comprised of final-year honor students in psychology, medicine, or neuroscience (Kalinowski *et al*., 2018) or experienced researchers in these fields (Belia *et al*., 2005). This highlights how difficult it may be to accurately interpret presentations of uncertainty and the importance of providing appropriate training to clinical healthcare professionals during their studies. In their book on developing statistical reasoning, Garfield and Ben-Zvi (2008) acknowledge the challenge of interpreting distributions among students, stating that decision making under uncertainty requires weighing the evidence to form a qualitative judgment, a skill that is much harder than interpreting the simple quantitative judgments made by statistical tests. Research on decision making sheds light on some of these challenges. Yaniv and Foster (1995), for example, noted that people have a tendency to avoid uncertainty at the cost of making a less accurate statement. This was reflected in the finding by Padilla *et al*. (2015) that participants tended to select the narrow distribution (i.e., the distribution with less variability/uncertainty) as more precise, even if it was less likely to contain the true value.

This review indicated that presentation formats that emphasize the shape of the underlying probability distribution may be better suited to promote accurate interpretations of measurement uncertainty associated with clinical assessment instruments. Whether such plots are also the best suited for promoting optimal clinical decision making remains unclear and should therefore be explored further in clinical contexts. Participants in decision-making studies included in this review were generally presented with unfamiliar decision-making scenarios such as construction projects on former military land, flood prevention, and so on (e.g., see Bruine de Bruin *et al*., 2013; Edwards *et al*., 2012). It is therefore unclear whether the finding that decision makers in these studies benefited the most from simple, less detailed plots would generalize to clinicians involved in clinical decision making. Clinicians are specifically trained in making clinical decisions. Their expertise and knowledge about the task at hand may impact how they experience the complexity of the decision-making tasks they face, potentially making them less likely to be overwhelmed by more detailed displays. Clinicians may therefore benefit from having more detailed information about the underlying probability distribution. Then again, it may be that the complexity of decision-making tasks in general makes it difficult even for clinicians to benefit from more detailed presentation formats.

As our systematic search retrieved no studies involving practicing clinicians who explicitly consider the uncertainty associated with test scores when making clinical decisions, we strongly recommend that this topic be further explored. More specifically, such studies could focus on comparing clinicians' ability to make optimal clinical decisions across various uncertainty display formats. In conducting such studies, it would be interesting to examine decision quality across different kinds of decision tasks, as the types of clinical decisions that are made may vary among different clinical settings and consequently require different uncertainty visualizations to promote optimal decision making. Estimation-type tasks, for instance, may be especially relevant to a neuropsychological setting (e.g., when estimating the neuropsychological functioning in a stroke patient). Such tasks may require visualizations which emphasize the relative likelihood of potential true scores given the observed score and associated standard error (such as probability density plots, for example). Classification tasks, which may be particularly relevant to an outpatient setting (e.g., in deciding whether to refer patients to inpatient care or assigning patients to diagnostic categories) may not require the same level of detail. For these tasks, a visualization which communicates the estimated patient score and some minimal information about the range of likely true scores (e.g., error bars) may be sufficient, as the main concern is to discern whether the patient's true score is likely to be above or below a given cut-off value.

Strengths and limitations

The inclusion of multiple databases for this systematic review (Medline, PsycInfo, ERIC, Scopus and Web of Science) ensured coverage of a wide range of potentially relevant research fields, from medicine and psychology to more technical fields such as computer science and meteorology. Furthermore, our comprehensive search strategy utilized words and phrases used to describe visualizations, measurement uncertainty, and relevant outcome measures in 18 studies identified in an initial unstructured search. We also screened the sources of all included studies for relevant records missed by our search strategy. This screen yielded eight additional studies. Checking the abstracts of these studies against the search terms in our search strategy, we found that these eight studies either: (1) did not include any of our search terms for uncertainty, instead using phrases such as "climate projections" or "predictive intervals" (n = 3); (2) did not include any of our search terms for visualizations, instead naming specific visualization formats, such as "error bars" or "cat's eye CIs" (n = 1); (3) violated the "adj5" combination rule between search terms expressing uncertainty and visualizations (n = 3); or (4) included none of our relevant search terms in their abstract (n = 1). Although it is possible that other relevant studies were not identified by our search terms, we believe our search strategy and approach ensured maximum inclusion of relevant studies.

This systematic review yielded no studies exploring visual presentation formats for measurement uncertainty in clinical contexts. The visualizations used in included studies are, however, appropriate for presenting measurement uncertainty in standardized clinical test scores. The focus on outcome measures considered relevant for clinical decision making allowed us to formulate broad recommendations relevant to clinical contexts. Furthermore, several included studies were conducted with samples matching our target population of healthcare professionals. Even so, there is a need for actively evaluating communication formats in the context of clinical assessment instruments and with the target audience of healthcare professionals. As Hambleton and Zenisky (2013) note, score

reporting has received less attention than other topics within psychometrics. Finding appropriate formats for communicating measurement uncertainty is an important step in developing useful reporting standards. The current article could serve as a starting point for research on this subject within the context of clinical measurement.

The quality appraisal framework developed for this systematic review provided the flexibility needed to accommodate the methodological heterogeneity of the included studies. This framework identified important limitations of the included studies – namely, a lack of specificity in study samples, often limited reflections on the impact of missing data/nonresponse, as well as the frequent use of measures with unknown psychometric properties developed specifically for single studies. These potential sources of bias had implications for this systematic review. Most included studies did not provide any reflections on the implications of missing data, or the statistical assumptions associated with their chosen analysis. Both these limitations may have introduced bias in the findings of the included studies. Furthermore, the frequent use of study-specific measures and lacking definition of sample populations made it difficult to explore potential explanations for discrepant findings of the studies. Such discrepancies were not uncommon across the outcome measures considered for this review (see for instance Table 1). Future explorations of the visualization of measurement uncertainty may benefit from defining a clear sample population and choosing previously established measures of outcomes, as this would enhance comparability across studies.

CONCLUSION

The studies included in this review reveal important limitations associated with error bars and numerical confidence intervals. Although these are the most commonly used formats to communicate measurement uncertainty in clinical contexts, these formats may promote erroneous interpretations of the underlying probability distribution and may be poorly suited to aiding probability estimation. Alternative formats which display the shape of the underlying probability distribution (e.g., PDF, gradient/violin plots, and HOPs) are generally found to counteract these misinterpretations, and facilitate accurate probability estimation. However, the lack of studies conducted within clinical decision-making contexts makes it difficult to establish whether these alternative formats would also be well suited to inform clinical decision making. As providing decision makers with uncertainty information was consistently found to promote optimal decision making, we strongly recommend further explorations as to how visual presentations impact clinicians' understanding of measurement uncertainty and their ability to make optimal clinical decisions. This review provides a potential starting point for such future studies by identifying appropriate presentation formats, as well as offering specific recommendations as to how remaining research questions might be addressed.

DECLARATION OF INTEREST STATEMENT

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a FRIPRO Young Research Talent grant, awarded to the last author (Grant no. NFR 286893) by the Research Council of Norway.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in published research articles retrieved through a systematic search of scientific databases. All reviewed records are cited and the search strategy is outlined in the article text and supplementary materials.

ENDNOTE

[1] Classification can be further subdivided into placement (i.e., classification based on univariate information) and selection (i.e., classification where rejection is a possible assignment outcome). However, a broad distinction between classification and estimation covers the most important difference in terms of the treatment of measurement error (Eggen, 2010). See Cronbach and Gleser (1957, p. 13) for a more detailed description of classification, placement, and selection.

REFERENCES

Byström, K. & Järvelin, K. (1995). Task complexity affects information seeking and use. *Information Processing & Management*, *31*, 191–213.

Charter, R.A. & Feldt, L.S. (2001a). Confidence intervals for true scores: Is there a correct approach? *Journal of Psychoeducational Assessment*, *19*, 350–364.

Charter, R.A. & Feldt, L.S. (2001b). Meaning of reliability in terms of correct and incorrect clinical decisions: The art of decision making is still alive. *Journal of Clinical and Experimental Neuropsychology*, *23*, 530–537.

Critical Appraisal Skills Programme. (2018). CASP qualitative studies Checklist. Retrieved May 10, 2021, from https://casp-uk.b-cdn.net/wp-content/uploads/2018/03/CASP-Qualitative-Checklist-2018_fillable_form.pdf.

Critical Appraisal Skills Programme. (2020). CASP randomised controlled trial standard Checklist. Retrieved May 10, 2021, from https://casp-uk.b-cdn.net/wp-content/uploads/2020/10/CASP_RCT_Checklist_PDF_Fillable_Form.pdf.

Cronbach, L.J. & Gleser, G.C. (1957). *Psychological tests and personnel decisions*. Champaign, IL: University of Illinois Press.

Eggen, T.J.H. (2010). Three-category adaptive classification testing. In W.J.van der Linden & C.A.W. Glas (Eds.), *Elements of adaptive testing* (pp. 373–387). New York: Springer.

Evers, A., Hagemeister, C. & Hostmaelingen, A. (2013). EFPA review model for the description and evaluation of psychological and educational tests (Tech. Rep. Version 4.2.6). European Federation of Psychology Associations.

Garfield, J. & Ben-Zvi, D. (2008). *Developing Students' statistical reasoning: Connecting research and teaching practice*. Dordrecht: Springer.

Goodwin, L.D. & Goodwin, W.L. (1999). Measurement myths and misconceptions. *School Psychology Quarterly*, *14*, 408–427.

Hambleton, R.K. & Zenisky, A.L. (2013). Reporting test scores in more meaningful ways: A research-based approach to score report design. In K.F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology, vol. 3: Testing and assessment in school psychology and education* (pp. 479–494). Washington, DC: American Psychological Association.

Higgins, J.P.T., Altman, D.G., Gøtzsche, P.C., Jüni, P., Moher, D., Oxman, A.D. *et al.* (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, *343*, d5928.

Hullman, J., Qiao, X., Correll, M., Kale, A. & Kay, M. (2019). In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE Transactions on Visualization and Computer Graphics*, *25*, 903–913.

Joanna Briggs Institute. (2020a). *Checklist for qualitative research. Joanna Briggs Institute*. Retrieved May 10, 2021, from https://jbi.global/critical-appraisal-tools

Joanna Briggs Institute. (2020b). Checklist for quasi-experimental studies (non-randomized experimental studies). Joanna Briggs Institute. Retrieved May 10, 2021, fromhttps://jbi.global/critical-appraisal-tools.

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American Psychologist*, *58*, 697–720.

Kinkeldey, C., MacEachren, A.M., Riveiro, M. & Schiewe, J. (2017). Evaluating the effect of visually represented geodata uncertainty on decision-making: Systematic review, lessons learned, and recommendations. *Cartography and Geographic Information Science*, *44*, 1–21.

Kinkeldey, C., MacEachren, A.M. & Schiewe, J. (2014). How to assess visual communication of uncertainty? A systematic review of geospatial uncertainty visualisation user studies. *The Cartographic Journal*, *51*, 372–386.

Kruyen, P.M., Emons, W.H.M. & Sijtsma, K. (2012). Test length and decision quality in personnel selection: When is short too short? *International Journal of Testing*, *12*, 321–344.

Linn, R.L. (1978). Demands, cautions, and suggestions for setting standards. *Journal of Educational Measurement*, *15*, 301–308.

Lurie, N. (2004). Decision making in information rich environments: The role of information structure. *Journal of Consumer Research*, *30*, 473–486.

McManus, I.C. (2012). The misinterpretation of the standard error of measurement in medical education: A primer on the problems, pitfalls and peculiarities of the three different standard errors of measurement. *Medical Teacher*, *34*, 569–576.

Meyer, G.J., Finn, S.E., Eyde, L.D., Kay, G.G., Moreland, K.L., Dies, R.R. *et al*. (2001). Psychological testing and psychological assessment. A review of evidence and issues. *The American Psychologist*, *56*, 128–165.

Meyer, V.R. (2007). Measurement uncertainty. *Journal of Chromatography A*, *1158*, 15–24.

Ouzzani, M., Hammady, H., Fedorowicz, Z. & Elmagarmid, A. (2016). Rayyan – a web and mobile app for systematic reviews. *Systematic Reviews*, *5*, 210.

Simpkin, A.L. & Armstrong, K.A. (2019). Communicating uncertainty: A narrative review and framework for future research. *Journal of General Internal Medicine*, *34*, 2586–2591.

Smits, N., van der Ark, L.A. & Conijn, J.M. (2018). Measurement versus prediction in the construction of patient-reported outcome questionnaires: Can we have our cake and eat it? *Quality of Life Research*, *27*, 1673–1682.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

Van der Bles, A.M., van der Linden, S., Freeman, A.L.J., Mitchell, J., Galvao, A.B., Zaval, L. *et al*. (2019). Communicating uncertainty about facts, numbers and science. *Royal Society Open Science*, *6*, 181870.

Vaurio, R. (2011). Symptom Checklist-90-revised. In J.S. Kreutzer, J. DeLuca & B. Caplan (Eds.), *Encyclopedia of clinical neuropsychology* (pp. 2447–2450). New York: Springer.

Yaniv, I. & Foster, D.P. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *Journal of Experimental Psychology: General*, *124*, 424–432.

Zimmermann, J., Kerber, A., Rek, K., Hopwood, C.J. & Krueger, R.F. (2019). A brief but comprehensive review of research on the alternative DSM-5 model for personality disorders. *Current Psychiatry Reports*, *21*, 92.

REVIEW RECORDS

Allen, P.M., Edwards, J.A., Snyder, F.J., Makinson, K.A. & Hamby, D.M. (2014). The effect of cognitive load on decision making with graphically displayed uncertainty information. *Risk Analysis*, *34*, 1495–1505.

Anic, V. & Wallmeier, M. (2020). Perceived attractiveness of structured financial products: The role of presentation format and reference instruments. *Journal of Behavioral Finance*, *21*, 78–102.

Bansback, N., Harrison, M. & Marra, C. (2015). Does introducing imprecision around probabilities for benefit and harm influence the way people value treatments? *Medical Decision Making*, *36*, 490–502.

Belia, S., Fidler, F., Williams, J. & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, *10*, 389–396.

Bruine de Bruin, W., Stone, E.R., Gibson, J.M., Fischbeck, P.S. & Shoraka, M.B. (2013). The effect of communication design and recipients' numeracy on responses to UXO risk. *Journal of Risk Research*, *16*, 981–1004.

Correll, M. & Gleicher, M. (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, *20*, 2142–2151.

Dieckmann, N.F., Gregory, R., Peters, E. & Hartman, R. (2017). Seeing what you want to see: How imprecise uncertainty ranges enhance motivated reasoning. *Risk Analysis*, *37*, 471–486.

Dieckmann, N.F., Peters, E. & Gregory, R. (2015). At home on the range? Lay interpretations of numerical uncertainty ranges. *Risk Analysis*, *35*, 1281–1295.

Durbach, I.N. & Stewart, T.J. (2011). An experimental study of the effect of uncertainty representation on decision making. *European Journal of Operational Research*, *214*, 380–392.

Edwards, J.A., Snyder, F.J., Allen, P.M., Makinson, K.A. & Hamby, D.M. (2012). Decision making for risk management: A comparison of graphical methods for presenting quantitative uncertainty. *Risk Analysis*, *32*, 2055–2070.

Gibson, J.M., Rowe, A., Stone, E. & Bruin, W.B.D. (2013). Communicating quantitative information about unexploded ordnance risks to the public. *Environmental Science & Technology*, *47*(9), 4004–4013.

Gschwandtner, T., Bögl, M., Federico, P. & Miksch, S. (2016). Visual encodings of temporal uncertainty: A comparative user study. *IEEE Transactions on Visualization and Computer Graphics*, *22*, 539–548.

Han, P.K.J., Klein, W.M.P., Lehman, T.C., Killam, B., Massett, H. & Freedman, A.N. (2011). Communication of uncertainty regarding individualized cancer risk estimates: Effects and influential factors. *Medical Decision Making*, *31*, 354–366.

Han, P.K.J., Klein, W.M.P., Lehman, T.C., Massett, H., Lee, S.C. & Freedman, A.N. (2009). Laypersons' responses to the communication of uncertainty regarding cancer risk estimates. *Medical Decision Making*, *29*, 391–403.

Hoekstra, R., Johnson, A. & Kiers, H.A.L. (2012). Confidence intervals make a difference: Effects of showing confidence intervals on inferential reasoning. *Educational and Psychological Measurement*, *72*, 1039–1052.

Hopster-den Otter, D., Muilenburg, S.N., Wools, S., Veldkamp, B.P. & Eggen, T.J.H.M. (2019). Comparing the influence of various measurement error presentations in test score reports on educational decision-making. *Assessment in Education: Principles, Policy & Practice*, *26*, 123–142.

Hullman, J., Kay, M., Kim, Y. & Shrestha, S. (2018). Imagining replications: Graphical prediction & discrete visualizations improve recall & estimation of effect uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, *24*, 446–456.

Hullman, J., Resnick, P. & Adar, E. (2015). Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PLoS One*, *10*, e0142444.

Ibrekk, H. & Morgan, M.G. (1987). Graphical communication of uncertain quantities to nontechnical people. *Risk Analysis*, *7*, 519–529.

Kalinowski, P., Lai, J. & Cumming, G. (2018). A cross-sectional analysis of students' intuitions when interpreting CIs [original research]. *Frontiers in Psychology*, *9*.

Lorenz, S., Dessai, S., Forster, P.M. & Paavola, J. (2015). Tailoring the visual communication of climate projections for local adaptation practitioners in Germany and the UK. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *373*.

Nadav-Greenberg, L. & Joslyn, S.L. (2009). Uncertainty forecasts improve decision making among nonexperts. *Journal of Cognitive Engineering and Decision Making*, *3*, 209–227.

Padilla, L.M., Hansen, G., Ruginski, I.T., Kramer, H.S., Thompson, W.B. & Creem-Regehr, S.H. (2015). The influence of different graphical displays on nonexpert decision making under uncertainty. *Journal of Experimental Psychology. Applied*, *21*, 37–46.

Ramos, M.H., van Andel, S.J. & Pappenberger, F. (2013). Do probabilistic forecasts lead to better decisions? *Hydrology and Earth System Sciences*, *17*, 2219–2232.

Savelli, S. & Joslyn, S. (2013). The advantages of predictive interval forecasts for non-expert users and the impact of visualizations. *Applied Cognitive Psychology*, *27*, 527–541.

Schapira, M.M., Nattinger, A.B. & McHorney, C.A. (2001). Frequency or probability? A qualitative study of risk communication formats used in health care. *Medical Decision Making*, *21*, 459–467.

Stock, W.A. & Behrens, J.T. (1991). Box, line, and Midgap plots: Effects of display characteristics on the accuracy and bias of estimates of whisker length. *Journal of Educational Statistics*, *16*, 1–20.

Zapata-Rivera, D., Zwick, R. & Vezzu, M. (2016). Exploring the effectiveness of a measurement error tutorial in helping teachers understand score report results. *Educational Assessment*, *21*, 215–229.

Zwick, R., Zapata-Rivera, D. & Hegarty, M. (2014). Comparing graphical and verbal representations of measurement error In test score reports. *Educational Assessment*, *19*, 116–138.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Supporting information S1 Documentation of literature search.

Table S1. Summary of included studies.

Table S2.1. Overview of quality appraisal.

Table S2.2. Overview of quality appraisal.

Table S2.3. Overview of quality appraisal.

Table S2.4. Overview of quality appraisal.

Figure S1. Boxplot.
Figure S2. Gradient Plot.
Figure S3. Quantile Dot Plot.
Figure S4. Cumulative Distribution Function.
Figure S5. Complementary Cumulative Distribution Function.
Figure S6. Color Coded Plot.
Figure S7. Blur Plot.
Figure S8. Bubble Plot.
Figure S9. Hypothetical Outcome Plot.

Received 26 April 2022, Revised 2 March 2023, accepted 16 March 2023