

Visual-semantic conflicts in visualization: A case study using Patient-Reported Outcomes

Racquel Fygenon , Enrico Bertini , and Heather T Gold 

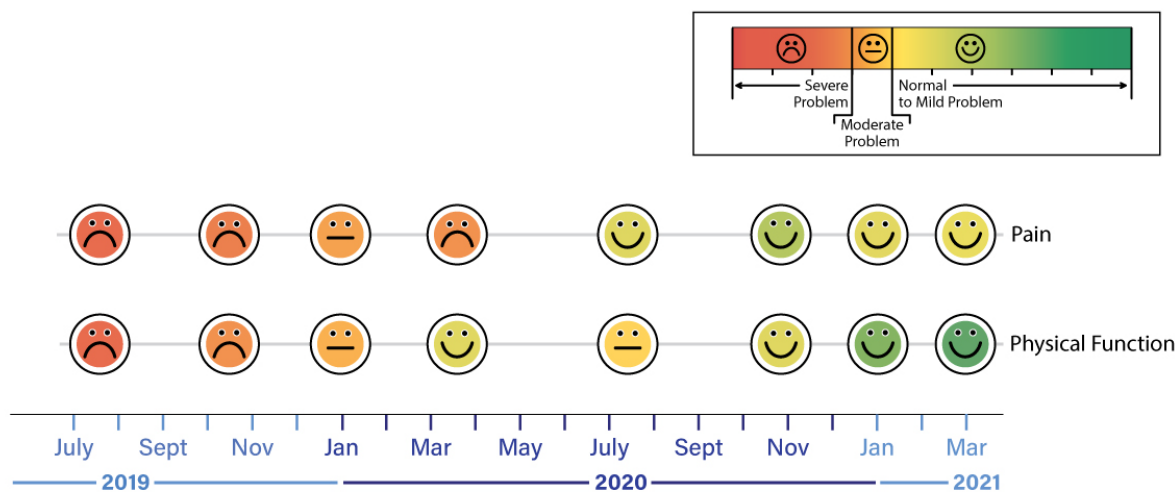


Fig. 1: The novel color-encoded smiley-timeline we test as an alternative to representing PRO scores via traditional line charts. The smiley-timeline encodes both symptoms (nonoptimal) and function (optimal) using color and pictographs to avoid visual-semantic conflicts incurred in the position encoding of line charts.

Abstract—Effective visualizations often employ directional paradigms to indicate improvement in alignment with viewers’ notions of optimality (e.g., y-axis numbers increase with higher positions). Occasionally, however, viewers’ visual and semantic paradigms of improvement conflict, leading to confusion. Such is the case with visualizations of the National Institutes of Health’s Patient-Reported Outcomes Measurement Information System (PROMIS), which consists of two types of Patient-Reported Outcome (PRO) scores, one ranging from 0 (“within normal limits”) to 100 (“severe”) for symptoms, and one ranging from 0 (“severe”) to 100 (“within normal limits”) for function. Current longitudinal visualizations of these scores rely on positional or numeric encoding to communicate changes in patient status. Semantically upwards directions are often associated with positive trends and more ideal situations (e.g., “I’m on the up-and-up”, “feeling 100”), while visually, upwards directions typically indicate more of something (e.g., stock price charts). When a line representing pain rises, it can be interpreted as pain getting worse or as pain being rectified. We design and test alternative visualizations to address the issue of directional incongruence and improve visualization efficacy. We do so via a case study of longitudinal PROMIS score interpretation, and present a novel “smiley”-timeline visualization that encodes scores via color and pictographs instead of position. We find that 1) color-encoded line charts that encode “up” as improvement over all variables and smiley-timelines result in the quickest response times, 2) visualizations with color encoding result in faster response times than their grayscale counterparts, and 3) smiley-timelines rank significantly higher than tested line charts in ease of use, intuitiveness, and likelihood of recommending them. Our findings support the rejection of strict adherence to the precision encoding hierarchy in circumstances in which visual and semantic directionality conflict. Finally, we present examples of other situations in which this conflict is present, providing the basis for future work where additional redesigns and evaluation are warranted. A free copy of this paper and all supplemental materials are available at <https://osf.io/kxzm3/>.

Index Terms—visualization, interpretability

- Racquel Fygenon and Enrico Bertini are with Northeastern University.
E-mail: fygenon.r@northeastern.edu | e.bertini@northeastern.edu
- Heather T Gold is with NYU Langone E-mail:
heather.gold@nyulangone.org

1 INTRODUCTION

Spatial-numeric associations are internalized by children who are influenced from the written direction of the language they speak and their experience in the physical world [50, 51]. These conventions strongly influence the interpretation of visualizations and thus inform many fundamental practices of data visualization (e.g., small-to-large x-axis units are plotted left-to-right) [28, 37, 50, 54]. While many visualizations succeed in using direction to make information readily accessible, it is possible for visual and semantic directional conventions to conflict [57], leading to confusion, frustration, and more seriously, miscommunication of crucial data [19, 43, 44].

This problem is acutely present in the communication of personal

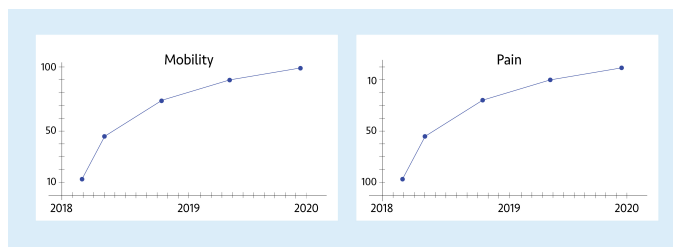


Fig. 2: With no other information, how has this patient been progressing? At first glance it may appear that the patient’s pain has increased, but the chart on the right depicts a patient whose pain has lessened over time.

health measures to patients. Health care in the United States has a long history of suffering from opaque and inaccessible personal health communication [23]. At the start of the 21st century, in a serious attempt to improve clinical management, health outcomes, and patient engagement, the National Institutes of Health (NIH) began a decade-long initiative to overhaul the systems in which clinical outcomes are documented and communicated to and between clinicians and patients [8, 20]. To do so, the NIH developed and tested the Patient-Reported Outcomes Measurement Information System (PROMIS) [20]. This system still suffers from miscommunication of collected data [19, 43, 44], and multiple surveys on Patient Reported Outcomes (PROs) find their use in care will remain limited if score data are difficult to interpret and integrate into decision making [16, 27].

A typical longitudinal visualization of PROMIS data is a line chart presenting time along the x-axis and scores along the y-axis [17–19, 35, 43–45, 48]. This design leads to well-documented confusion stemming from disagreement between semantic and visual cues [18, 19, 35, 44, 45, 48]. Figure 2 shows twin upwardly prone lines representing mobility and pain for a hypothetical patient, yet it is not clear how the patient is progressing. While upwardly trending lines are typically associated with improvement, suggesting both mobility and pain are “getting better,” English semantics dictate that “more” pain corresponds to a worsening condition. Numbering along the y-axis is also unhelpful; typical associations with number scales encounter the same semantic-directional conflict (e.g., “I don’t know whether 100 is severe or 0 is severe”) [19].

This paper seeks to address positional-semantic conflicts, using non-experts interpreting longitudinal trends in PROMIS score visualizations as a case study. First, via a series of qualitative pilot interviews, we provide a deep-dive into the reasoning behind the confusion that occurs when visualizations present positional and semantic incongruences. The results of this qualitative work are supported by numerous studies on the efficacy on PRO score visualizations. Next, we attempt to limit the possibility of this ineffectual reasoning through an alternate novel ‘smiley’-timeline visualization (Fig. 1), in which we replace all instances of conflicting positional encoding with color and pictographic encoding which align with semantic directional conventions. Finally, we test the interpretability of longitudinal trends, a key task in communicating patients’ overall health trajectory. We test our proposed smiley-timeline and four formats of line charts that are commonly used to communicate PROs to patients via electronic health record (EHR) systems. We also conduct a subjective evaluation of these five visualizations, gauging non-experts’ preferences for their use.

Patient-Reported Outcomes are a model example of positional-semantic conflict’s impact on accurate communication of critical information, though this same conflict can be found throughout other visualization types and applications. We finish this paper by zooming out, and presenting a broader view of other instances of common miscommunication, in which positional-semantic conflicts stand to be addressed in the future.

2 BACKGROUND & SIGNIFICANCE

2.1 Visual-Semantic Incongruence in Line Charts

Classically accepted paradigms in visualization define a hierarchy of visual encoding based on precision, placing the depicting of in-

formation via position as the strongest of all possible visual channels [14, 36, 42, 59]. Thus it is logical to hypothesize that line and bar chart visualizations may communicate informations, including PRO scores, more effectively than heat maps and pictograms. Yet, this hierarchy of visualization encodings is based on visual channels’ ability to convey precision [36]. Position encoding dominates in experiments that evaluate the precision of estimating a value, a difference between values, or an aggregate measure (e.g., average value) [29, 42, 59]. But highly accurate numeric interpretation is not always a visualization’s primary objective [15]. For PRO scores in particular, visualizations must be intuitive and easy to understand so that readers with varying graph literacy are able to understand key concepts in their depicted data. Thus, we assert that the accuracy of interpreted *amount* of change (e.g. “pain score changed by 2” vs “pain score changed by “10”) is secondary to the accuracy of communicated *direction* of change (e.g. “pain score went down” vs “pain score went up”). Additionally, position encoding does not reliably maximize reported intuitiveness and ease of use [15], which are integral to engaging patients and communicating PROs [52].

In visualization research, semantic-directional incongruence in line charts has only been lightly investigated. A 2022 analysis of line charts representing “negative-valence” (i.e., non-ideal) variables found that non-inverted graphs lead to higher performance than their inverted counterparts regardless of semantic convention [57].

2.2 Patient-Reported Outcomes

Since its development almost two decades ago, the NIH’s PRO system, PROMIS, has accrued hundreds of studies that suggest that documenting and referring to its patient-reported scores has significant positive effects on the clinical care and outcomes of patients that receive care throughout a wide range of medical fields, including oncology and arthritis [24, 25, 38]. With PROMIS becoming adopted in a range of clinical practices throughout the United States, Electronic Health Record (EHR) providers are integrating PROMIS score collection, analysis, and communication tools into their host of products [5, 10, 52, 58]. Beyond the US, other forms of PRO systems, like Australia’s PROMs and the European Organisation for Research and Treatment of Cancer’s Quality of Life (QoL) Questionnaires, are also advancing in popularity and in turn being recommended for integration into clinician- and patient-facing software [4, 9]. Even as EHR services increase access to PRO data via new dashboards and reporting features, lack of intuitive visualization techniques can lead to misinterpretation and restrict PRO data from being fully understandable [16].

PROMIS measures can quantify patient symptoms (e.g., fatigue) or function (e.g., mobility) and range from 0 to 100, using a normed T-score metric whereby “50 is the mean of the relevant reference population and 10 is the standard deviation” [6]. Cut-points for PROMIS measures vary in placement, number, and labeling (e.g., “within normal limits, mild, moderate, severe” and “excellent, good, fair, poor”) [7]. PROMIS numeric scoring is directly correlated to the amount of measure in question. A high score can indicate either a negative health outcome (e.g., worsening symptom) or a positive one (e.g., improving function) [7]. Understanding the evolution of patients’ scores over time is useful when managing health conditions [11].

2.3 Patient-Reported Outcome Visualizations

Past research maintains that line charts are a popular visualization among both patients and clinicians to communicate PRO scores over time [18, 19, 44, 45]. Line charts receive high scores of “usefulness” and “ease-of-understanding” and result in fairly accurate interpretation of function scores (e.g., mobility) [18, 19], despite leading to markedly lower accuracy when interpreting symptom scores (e.g., pain) [19]. Qualitative investigations indicate line charts’ most confusing attribute is directional inconsistency inherent in positional and numeric encoding [19, 44]. Several studies have investigated varying direction of health improvement in line charts [17, 45, 48], some of which suggest that interpretation accuracy increases when line charts encode higher positions as improvement in symptoms and function (e.g., Figure 2) (n=1113, n=1017) [45, 48].

To combat directional confusion and reduce demand on readers' working memory, the PRO Data Presentation Stakeholder Advisory Board developed guidelines that recommend visualizing electronic PRO scores using simple graphs with clear annotations, symbols to differentiate information, and color encoding to depict severity [1, 44], which is often facilitated via a traffic light red-yellow-green color scheme [1, 12, 49]. Such coloring leverages the familiarity bias of U.S. readers who are accustomed to red indicating danger and green indicating a lack thereof, but must be paired with redundant encodings via other mapping strategies (e.g., position, shape, text) to maintain accessibility for readers with color-vision deficiency [53]. We explore effects of color on trend interpretation accuracy in this paper. Prior research similarly explores visualizations that color-encode PRO severity [43, 45], although we are unaware of research that has examined similar graphs with and without color, as we present here.

A sparse amount of research has examined the effect of pictographs, sometimes referred to as "visual analogies", on PRO score comprehension. One study found pictographs led to worse PRO comprehension in comparison to line charts, but cautioned their results [31]. Another, found that pictographs based on the Wong-Baker FACES Pain Rating Scale overlaid onto bar charts resulted in better comprehension and was participants' preferred format for displaying longitudinal symptom data [46, 56]. A third study, determined that pictographs and color-encoded number lines resulted in more comprehension than line charts and simple text-based descriptions, but only examined data from two different points in time, failing to address how comprehension adjusted with multi-point longitudinal data [49]. Within general visualization research, pictographs have been shown to be more comprehensible than bar charts by readers with low literacy [34].

While research suggests that using line charts to depict symptom scores (e.g. pain) result in reader confusion [19], current guidelines, published as recently as 2021, still recommend the use of line charts to communicate such information [2, 12, 13]. Other recommended visualizations for longitudinal representation of PRO scores (including grouped bar charts, bubble/point plots, and tables) rely on the same numerical and positional encoding that triggers misunderstanding of line chart data [2, 12]. Heat maps are the sole recommended visualization that represent score changes via a non-positional and non-numerical encoding [2]. Still, heat maps are often tested and presented in guidelines with positional double encoding that couples a color scale to an x- or y-axis [2, 19]. For an overview of prior research on PRO score visualizations see Table 1.

To address the gap in research surrounding the interaction and effect of color-encoding and inverting y-axes of PRO score line charts, we developed four line chart stimuli. We also sought to explore the effect of removing y-axes encoding in multi-point (>2) longitudinal PRO score visualizations, by creating a color-encoded smiley-timeline visualization that is novel in this application. Figure 3 shows an overview of tested chart types. Below, we evaluate the efficacy and user preference of these 5 PRO score visualizations, specifically their effect on user comprehension of longitudinal trends.

3 MATERIALS & METHODS

We present a between-subjects study in which we develop a color-encoded smiley-timeline and test its effect on trend interpretation in comparison to four line charts typically used to depict PRO scores.

3.1 Visualization stimuli design

Informed by current literature, we hypothesize that removing positional encoding from PROMIS score visualizations will resolve readers' visual-semantic confusion around y-axis directionality. Our novel visualization has no y-axis, instead encoding PROMIS scores along a timeline via color and pictographs (Figure 3, bottom).

We use smiley-face pictographs due to their intuitive communication of positive/negative outcomes and their prevalence in clinical practice [55, 56]. To adhere as closely as possible to PROMIS's framework of continuous numeric scores (0-100) with discrete binned thresholds ("normal" to "severe"), we design these smiley-timelines to have continuous encoding using a color gradient and discrete encoding using three

mouth shapes. This design is also an approximation of the continuous positional and discrete color encoding of the line charts we test (Fig. 3, left column), allowing for a more valid comparison between the three conditions.

We design tested line charts to approximate typical PRO score visualizations and investigate if color and/or y-axis directionality changes general trend interpretation. The optimal direction of the y-axis for PRO visualizations of symptom scores is up for debate. Some research suggests directionality that adheres with common graphical paradigms (i.e. "down" corresponding to less pain) is more intuitive [54, 57]. Conflicting research on PRO score visualizations suggests the opposite is true (i.e., "down" corresponding to less optimal results like more pain) [45, 48]. Thus, we vary the direction of the y-axis on line charts that show a symptom—as done in [17, 45, 48]—and interspersed them with line charts that depict a function. To explore the effects of encoding symptom improvement in a non-positional manner, we compared smiley-timelines' performance to that of the four line charts: those with differing directions of improvement for symptoms and function and color (line-diff-c, top left), those with the same direction of improvement for symptoms and function and color (line-same-c, middle left), and their black-and-white equivalents (line-diff-bw, top right; line-same-bw, middle right).

We created all tested visualizations using Adobe Illustrator. Table 2 provides a summary of labeling conventions.

3.2 Pilot Study

We ran a series of initial pilot studies exploring participant performance for eight PROMIS score interpretation tasks, allowing us to fine-tune visualizations and survey wording for clarity.

In our final pilot study, we test chart performance for a single, widely applicable task: trend interpretation. This task, in which readers identify if a symptom or function is improving or worsening, is fundamental to many other ecologically valid tasks (e.g., finding maxima, estimating change over time) and is critical in assessing patient care over time. Limiting evaluated tasks allows us to concentrate resources and increase the statistical power of our analysis.

We recruited a sample of pilot participants from the platform Prolific¹ (n=142). Prolific connects potential human-studies participants with researchers, facilitating demographic screening, anonymization, and compensation. Each participant was randomly placed into groups of roughly equal size. Each group was assigned one of the five different graph types in Figure 3.

Participants were shown eight different trends, in random order, visualized via the graph type assigned to their group. Four graphs were titled "Pain" and four were titled "Physical Function". While "Pain" is qualified with additional verbiage (e.g., "pain interference") in PROMIS documentation, our goal is to test participant interpretation of a negative outcome. To reduce possible confusion, we opted for a simpler label.

Each participant saw stimuli consisting of two charts in which Pain was improving, two charts in which Pain was worsening, two charts in which Physical Function was improving, and two charts in which Physical Function was worsening. Participants were not instructed how to interpret the graphs and were asked to determine whether the graph showed the given variable getting better or worse.

3.2.1 Speed-accuracy tradeoff

Although it is typical to evaluate chart performance using response time and error rate as proxies for mental effort, the two metrics can confound each other. This "speed-accuracy tradeoff" occurs when participants rush through questions (high error rate and quick response times) or are exceedingly deliberate (low error rates and inflated response times) [30]. If either behavior is allowed, researchers are unable to determine which responses are skewed, and thus both metrics are undermined. Thus, researchers must decide to restrict one of the metrics to measure the other as a proxy for mental effort.

¹www.prolific.co

Paper	Visualization	Direction of encoding	Color*	Smiley pictograph	Longitudinal data
THIS PAPER	Line 	P & S		✓	✓
	Line 	P & S	✓	✓	✓
	Smiley-timeline 		✓	✓	✓
Austin et al, 2021	Line 	S			✓
	Bar 	S			✓
	Point 	S			✓
	Table 				✓
	Color bar 		✓		✓
Turichoe et al, 2020	Line 	S	✓		✓
	Color bar 		✓		✓
	Text 				✓
	Gauge icon 	S	✓		✓
Stonbraker et al, 2019	Line 	S			✓
	Smily line 	S		✓	✓
	Bar 	S			✓
	Smiley bar 	S		✓	✓
	Point 	S			✓
	Smiley point 	S		✓	✓
	Smiley icons 			✓	✓
	Sparkline table 	S			✓
Brundage et al, 2018	Line 	P & S			✓
Tolbert et al, 2018	Line 	P & S			✓
Snyder et al, 2017	Line 	P & S	✓		✓
Brundage et al, 2015	Line 	P & S			✓
	Point 	P	✓		
	Table 				✓
	Heat map 		✓		
Izard et al, 2014	Line 	F			✓
	Bar 	F			✓
	Table 				✓
	Smiley timeline 			✓	✓
Brundage et al, 2005	Line 	F			✓
	Text 	F			✓

Table 1: An overview of ten papers (including this paper) that investigate PRO score visualizations. Some of these papers also investigate visualizations that depict proportion of patients with PRO score changes. These visualization techniques are not included in the table. In the *Direction of encoding* column, P = Positional directionality (up is good for symptoms and functions), S = Semantic directionality (down is good for symptoms, up is good for functions), and F = only explored function score encoding, so semantic & positional directionality agree. *Color refers to only coloring associated with PRO score severity or amount. Coloring used to indicate different patients or symptoms is not included.

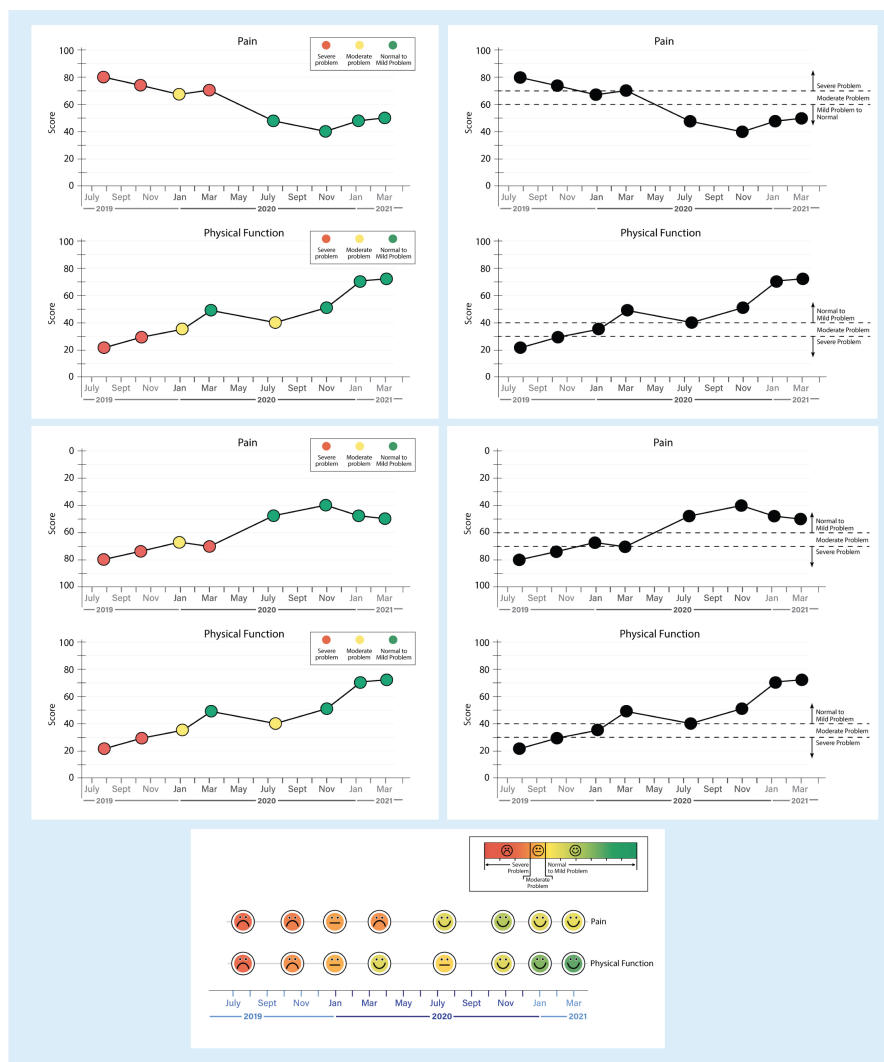


Fig. 3: Top left: Line chart with color [c] encoding and y-axes in the same [s] direction (line-s-c); top right: Line chart in black white [bw] and y-axes in the same direction (line-s-bw); middle left: Line chart with color encoding and y-axes in different [d] directions (line-d-c); middle right: Line chart in black white and y-axes in different directions (line-d-bw); bottom: Smiley-timeline with color encoding (smiley).

If two groups of participants are given sufficient time to read a graph, no significant difference in accuracy will be detected [30]. While accurately interpreting PROMIS scores is important in real-world scenarios, evaluating mental effort based on response accuracy may require restricting participants to see graphs for an unrealistically short amount of time.

Thus, to provide participants with a realistic amount of time ², in our final pilot study we restrict error rate and evaluate response time as a proxy for mental effort,

We do so by including a section where participants are presented with an untimed pre-survey question that shows both a PROMIS symptom and a PROMIS function score obviously improving. We ask participants to answer whether each graph shows its score getting better, or getting worse, and also allow participants to report if they are unsure³. If a participant cannot correctly interpret the obvious trends in an untimed setting, the survey ends. This pre-evaluation also allows us to prime participants on future charts without explicit instruction that might bias interpretation. All 142 participants who completed the full pilot survey made no more than two errors answering eight questions.

We also inform participants of a 10-second maximum per question, shown on countdown timers, to deter participants from answering

questions with an abundance of caution, which may cause artificially long response times [30].

The fastest pilot response times result from color-encoded line charts with y-axes encoding “up” as improvement for symptoms and function (line-same-c). Smiley-timeline visualizations result in the second fastest response times, while slowest times result from uncolored line charts with y-axes encoding improvement for symptoms and function in different directions (line-diff-bw)⁴.

3.3 Main study

Our main study compares response times, while also exploring user preference for the five different approaches to visualizing PROMIS scores. We test the same visualization stimuli as the final pilot study in a similar between-subjects experiment. Results from our pilot study inform an *a priori* sensitivity analysis and pre-registered hypotheses for our final study⁵. We hypothesize (H1) line charts with color encoding will have faster response times than their counterparts without color and (H2) smiley-timelines, which have color encoding, will have faster response times than line charts without color. We also hypothesize that charts that encode symptom and function improvement in the “up” direction would be interpreted more quickly, because they do not require

²Earlier pilot studies’ response time per question min 1.4s, median 3.7s, max 7.6s.

³Screening questions are in Figure SM1 in Supplementary Materials

⁴Pilot study data and analysis available at <https://osf.io/e2ghq>

⁵Hypotheses and final experimental design are at <https://osf.io/e2ghq>. Sensitivity analysis is in Table SM2 in Supplementary Materials.

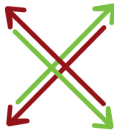

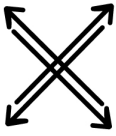
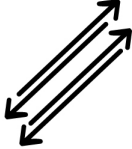

Label	line-diff-c	line-same-c	line-diff-bw	line-same-bw	smiley
Symbolic Encoding					
Directionality	Functions improve as line goes up. Symptoms improve as line goes down.	Functions and Symptoms improve as lines go up	Functions improve as line goes up. Symptoms improve as line goes down.	Functions and Symptoms improve as lines go up.	Functions and Symptoms improve as pictographs get greener and smile more.
Color	Traffic light color scheme	Traffic light color scheme	Black & white	Black & white	Traffic light color scheme

Table 2: Labeling convention for the four different line chart conditions is: line-[direction of improvement for Symptom and Function scores]-[color scheme].

participants to note changes in direction of improvement, resulting in **(H3 & H4)** line-same-c and line-same-bw would result in faster response times than line-diff-c and line-diff-bw, respectively. Lastly, we hypothesize **(H5)** smiley-timelines will not result in significantly slower participant response times than any line chart.

Due to low p-values from Shapiro-Wilk tests and our visual analysis of residual density and quantile-quantile plots of the pilot study data⁶, we perform a non-parametric sensitivity analysis via the software G*Power to determine an adequate sample size for our final study. The resulting effect size was $d < 0.25$ for 4 pair-wise conditions, $0.40 < d < 0.55$ for 4 pair-wise conditions, and $d > .75$ for 2 pair-wise conditions⁷. To explore only pairs of conditions that have "medium" to "large" effects [22], we focus solely on necessary sample sizes for studying six of the ten pairwise conditions, all of which had pilot-estimated effect sizes of $d > 0.40$. Of these six pairwise conditions, the largest necessary sample size for a desired power of 0.8 was 75 participants per condition, making our target sample size 375 participants. IRB approval was granted by our institution.

3.4 Participants

As in the pilot study, we recruit a sample of participants from Prolific⁸, balanced based on reported sex. We compensate participants \$10.23/hour with the expectation that survey completion would take less than 5 minutes. All participants are from the United States, fluent in English, and over age 45 years, so as to be at higher risk of experiencing pain symptoms and physical limitations.

3.5 Survey Design

Survey participants answer basic questions about personal demographics and history with clinical care before advancing to the untimed primer question⁹. If they complete the primer question successfully, we then inform participants of a 10-second time limit for all further questions and ask them to respond as quickly and accurately as possible. We place randomly place each participant into groups of roughly equal size via Qualtrics’s “Randomizer” and “Evenly Present Elements” features. We assign each group to one of the five different graph types in Figure 3. As in the pilot study, we show participants the same eight different trends, in randomized order using Qualtrics’s “Randomizer” feature to combat learning and order effects, visualized via the graph type assigned to their group. Participants are not instructed how to interpret the graphs and are asked to determine whether each graph shows the given variable improving or worsening.

⁶See Table SM1 in Supplementary Materials

⁷See Table SM2 in Supplementary Materials

⁸www.prolific.co

⁹see Section 3.2.1 for further details

All questions are accompanied by a countdown timer and phrased as “How has this patient’s pain [or physical function] been doing?” with the possible responses, “Getting better” and “Getting worse”. Response times are recorded from timer start to participants’ last click before submitting their answer.

Next, participants are asked to score the chart they saw based on how difficult it is to read, how quickly they feel they could read the chart, how intuitive the design is, and how likely they are to recommend using the chart. Scoring ranges from 1 (i.e., difficult, slow, unintuitive, unlikely to recommend) to 5 (i.e., easy, quick, intuitive, likely to recommend). Finally, participants answer questions about their previous use of PROs and online patient portals, and their preferences for using PROs in the future.

3.6 Analytic Methods

First, we discard any response times associated with an incorrect answer. Then, in accordance with our pre-registered analysis plan¹⁰, we conduct Shapiro-Wilk tests on each set of response times to assess skewness and kurtosis, rejecting all five null hypotheses that each sample came from a normally-distributed population¹¹. We apply pairwise Mann-Whitney tests¹² to determine significant differences between conditions’ response times, and correct for family-wise error rate using a two-stage Benjamini-Hochberg correction¹³. Next, we bootstrap confidence intervals¹⁴ for each condition’s response times using a bias-corrected and accelerated (BCa) bootstrap interval and visualized¹⁵ the resulting confidence intervals in Figure 4 [3, 26].

For the four subjective Likert-scaled questions on usability of visualizations, we conduct nonparametric analyses because participants may interpret Likert scales non-linearly [39]. After running pairwise, one-tailed Mann-Whitney tests on all visualizations, we apply the same methods as above to correct for family-wise error rate, and bootstrap and visualize confidence intervals¹⁶.

For both analyses, we consider $p < 0.05$ significant, and $p < 0.1$ to be of note.

Characteristic		Number (n=383)	%
Reported gender	Female	190	49.6%
	Male	193	50.4%
Age range	45-54	170	44.4%
	55-64	143	37.3%
	65-74	64	16.7%
	75+	6	1.6%
Ethnicity	White	342	84.9%
	Black	15	3.7%
	Asian	13	3.2%
	Mixed	10	2.5%
	Other	2	0.5%
	Data unavailable	1	0.3%
Reported having color-vision deficiency	No	371	96.9%
	Yes	8	2.1%
	I'm not sure	3	0.8%
	Prefer not to say	1	0.2%
History with pain: I have experienced...	Knee pain	195	50.9%
	Chronic pain anywhere	174	45.4%
	Hip pain	109	28.5%
	Hip/knee osteoarthritis	45	11.7%
	None of the above	96	25.1%
	Prefer not to say	1	0.3%
History with clinical care: I have experienced...	Surgery	207	54.0%
	Physical therapy	203	53.0%
	Physical/rehabilitation medicine	63	16.4%
	Occupational therapy	21	5.5%
	Rheumatology	16	4.2%
	Prefer not to say	1	0.3%
History with patient portals: I have...	Used a patient portal	262	68.4%
	Not used a patient portal	113	29.5%
	Not sure	8	2.1%
History with PROs:	PROs have been used in my care	42	11.0%
	PROs have not been used in my care	265	69.2%
	Not sure	76	19.8%

Table 3: Demographic information and participant history with clinical care and patient portals.

4 RESULTS

4.1 Participants

See Table 3 for demographic and clinical experience characteristics of main study participants.

We recruit 403 participants; 381 (94.5%) pass the required initial screening question¹⁷ and answer all 8 trend interpretation questions

¹⁰<https://osf.io/e2ghq>

¹¹via python package *scipy.stat*; see Table SM3 in Supplementary Materials

¹²via python package *pingouin*

¹³via *multipletests* function in python package *statsmodels*; see Table SM4 in Supplementary Materials for p-values

¹⁴via *bootstrap* function in the python package *scipy.stats*

¹⁵via javascript library *d3.js*

¹⁶see Figure 5

¹⁷See Figure SM1 in Supplemental Materials for screening questions; see Sec. 3.2 for justification

with 2 errors or fewer¹⁸. Our first round of recruitment resulted in one condition with a sample size of 74, so we reopened recruitment to ensure each chart type had at least 75 respondents, resulting in a total sample size of 383. Our pre-screening methodology excludes 3.7%, and an additional 1.2% are excluded for making more than two errors answering eight questions. Overall, we exclude 5.0% of participants. Of the remaining 383 participants, 91.4% report experiencing types of pain or clinical care which stand to benefit from the use of PROMIS [24,25].

4.2 Overall response time

For naming conventions, see Table 2.

Line-same-c leads to significantly faster response times than all other conditions (difference of means (DM) \approx 0.50 seconds), except for smiley-timelines. There is no significant difference in line-same-c and smiley-timeline response times (DM = 0.16 seconds). See Figure 4

¹⁸As based on performance in the final pilot study and defined by our pre-registered data exclusion criteria

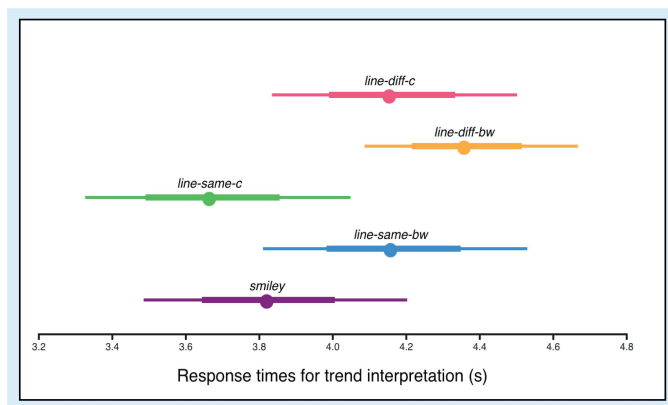


Fig. 4: A visual comparison of response times across conditions. Circles encode average response time, thicker bars represent bootstrapped 68% CI, and thinner bars represent bootstrapped 95% CI.

for confidence intervals¹⁹.

4.2.1 Color Encoding

The data provide evidence to support the hypothesis (**H1**) that color encoding decreases response time when interpreting trends. Line-same-c is significantly faster than line-same-bw ($p=0.020$, $DM=0.49$ seconds), and line-diff-c is faster than line-diff-bw ($p=0.098^*$, $DM=0.20$ seconds).

The data also provide evidence to support (**H2**) that color-encoded smiley-timelines result in quicker response times than black-and-white line charts. Smiley-timelines result in significantly faster times than line-diff-bw ($p=0.013$, $DM=0.54$ seconds), and faster times than line-same-bw ($p=0.07^*$, $DM=0.34$ seconds).

4.2.2 Line chart directionality

The data provide evidence to support (**H3 & H4**) that line charts that encode improvement for symptoms and function in the “up” direction result in quicker trend interpretation than line charts that do not. Response times from line-same-c are significantly faster than those from line-diff-c ($p=0.020$, $DM=0.49$ seconds). Response times from line-same-bw are slightly faster than those from line-diff-bw ($p=0.101^*$, $DM=0.20$ seconds).

4.2.3 Smiley-timelines

The data provide evidence to support (**H5**) that smiley-timelines do not result in significantly slower trend interpretation than line charts. Smiley-timelines are slightly slower than line-same-c ($p=0.145$, $DM=0.16$ seconds), but faster than line-same-bw ($p=0.07^*$, $DM=0.34$ seconds), line-diff-c ($p=0.07^*$, $DM=0.33$ seconds), and line-diff-bw ($p=0.013$, $DM=0.54$ seconds). See Figure 4 for visual comparison.

4.3 Preference rating

Smiley-timelines are rated on a 5-point scale as significantly easier to read (average $DMs = 0.33$ points), more intuitive (average $DMs = 0.60$ points), and more likely to be recommended (average $DMs = 0.55$ points) than all other visualizations. Smiley-timelines are also perceived as significantly quicker to read than line-diff-bw ($DMs = 0.37$). No other significant differences were found. See Figure 5

Lastly, we explore if and how participants may be interested in seeing their PROs in the future. Fifty-three percent are interested in clinicians using PROs to discuss their care, while 12% are not. We find no significant differences in interest in future use of PROs based on chart type shown.

¹⁹Adjusted p-values are available in Table SM4 in Supplementary Materials

*The adjusted p-value for this difference is up to double the $p<0.05$ threshold of significance specified in our pre-registered analysis plan: <https://osf.io/e2ghq>

*The adjusted p-value for this difference is up to double the $p<0.05$ threshold of significance specified in our pre-registered analysis plan: <https://osf.io/e2ghq>

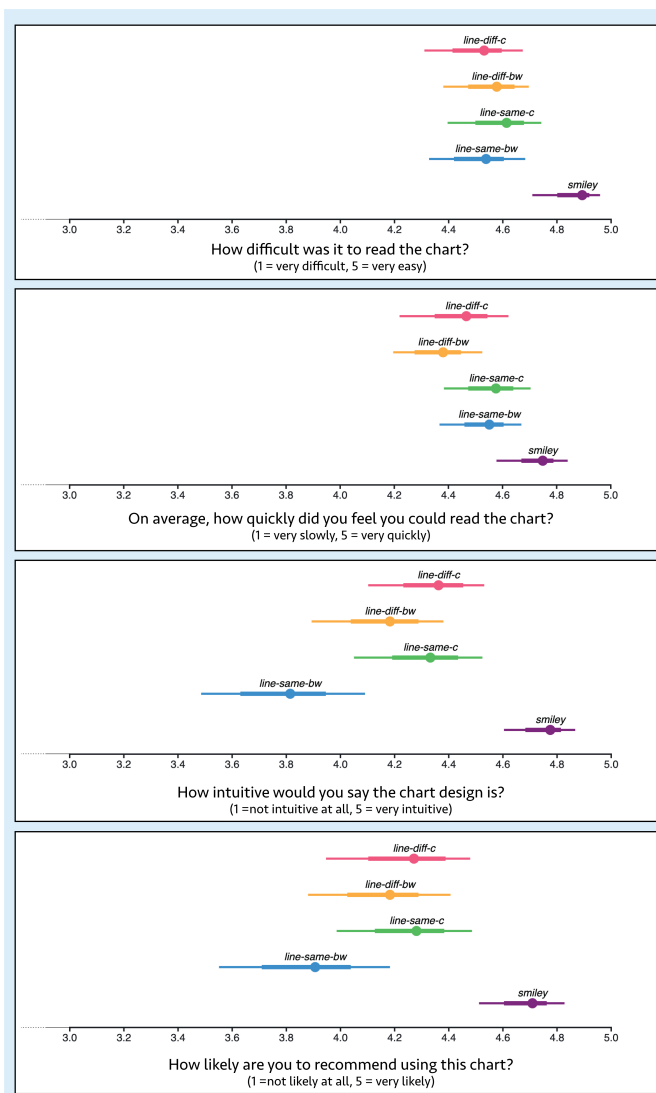


Fig. 5: A visual comparison of participant Likert responses to subjective questions across conditions. Circles encode average response time, thicker bars represent bootstrapped 68% CI, and thinner bars represent bootstrapped 95% CI.

4.4 Response Accuracy

While we do not examine response accuracy, nor have a pre-registered analysis plan to do so, percent of errors separated by chart type are available in Figure SM2 in Supplementary Materials.

5 DISCUSSION

We can address visual-semantic conflicts in visualizations by replacing or doubly encoding the channels that cause them. In the case study examined in this paper, we demonstrate that the positional-semantic conflict incurred by encoding PRO scores along the y-axis can be mitigated with color encoding and adhering to positional paradigms, or by entirely replacing y-axis encoding with colors and pictographs. We note that response times when interpreting the general trend of a patient’s scores indicate no significant difference between these two solutions, but qualitative responses from our participants exhibit that replacing the conflicting y-axis encoding completely leads to less perceived difficulty and more intuitiveness.

This study builds on previous examinations of the visual-semantic conflict inherent in visualizing PRO scores. Much of this current literature finds evidence for the y-axis conflict we address here, and some continues on to examine potential solutions. Notably, Stonbraker, et. al placed black-and-white smiley faces on line charts, among other

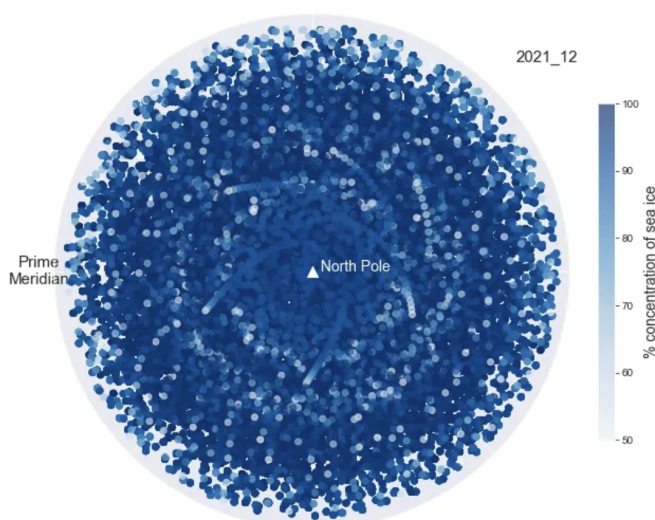


Fig. 6: Originally published on TowardsDataScience.com, the above scatter plot encodes the average percentage concentration of sea ice above the Arctic circle via a white-to-navy sequential color scheme [32]. Which areas of the scatterplot above show a higher concentration of ice? While dark colors, like navy, are often associated with more of something, white is typically associated with the color of ice.

standard graph types, and found PRO score graphs with smiley faces led to increased comprehension and were favored by readers with low graphic literacy [46]. Our study suggests that adding color to such stimuli may improve performance further.

5.1 Limitations

While pre-screening participants allowed us to address speed-accuracy tradeoff [30], it required the exclusion of important subsets of the population, namely readers with low graphic literacy. Additionally, our collection of participant sociodemographic data lacks indicators of graphic literacy, including education and income level. Future work would greatly benefit from recording these factors, addressing biases that might stem from skewed socioeconomic samples, and specifically examining solutions to visual-semantic conflicts for populations with low graphic literacy.

Although precision-based tasks, such as finding maximum score, are commonplace evaluating visualizations, we opted to evaluate general trend interpretation. This approach prioritizes interpretation of direction of change (i.e. pain improving or worsening) over estimation of amount of change (i.e., pain changing by a score of 20). Under this prioritization, smiley-timelines performed comparably to, if not better than, line charts.

The same relationship may not hold during the evaluation of precision-based tasks. The hierarchy of visual performance—as measured by precision—suggests that shape, color, and pictographic encodings are less precise than positional encodings [21, 36].

Further research is needed to determine how the visualizations presented in this study perform in other tasks. If a lack of precision is identified and conflicts with the objectives of a visualization, further precision can be communicated via other methods, like tool tips [1].

Additionally, we did not address familiarity bias within our experiments. The prevalence of line charts in daily life and education in the U.S. may aid participants in answering trend interpretation questions more quickly with line charts than they would if they were similarly exposed to smiley-timelines. At the same time, there is a strong familiarity associated with the smile pictography and red-yellow-green traffic light coloring used in the smiley-timeline [49, 53, 55, 56]. Thus, it is unclear whether familiarity biases likely affect response-time differences. Future research could further introduce participants to smiley-timelines before comparing visualization performance as a substitution for prolonged exposure.

5.2 Future work

The identification and potential remediation of visual-semantic conflicts presents plenty of opportunity for future work. Applied case studies, like the one we present here, provide nice contextual frameworks for evaluating the performance of redesigns.

Further investigation of the strength of visual-semantic conflicts when encoding data of different sizes and shapes is also warranted. The confusion behind visual-semantic conflicts in two line charts may change with a larger number of presented visualizations. Additionally the usability of redesigns to address conflicts may vary give the amount of data points visualized or the task at hand.

Despite the focus of this case study, visual-semantic conflicts are not restricted to y-axis encoding. Non-positional encodings, like color, shape, and borders are also susceptible to conflicting with semantics [33, 40, 41, 60–62]. Color, in particular, has been shown to result in semantic conflicts via the famous Stroop task [47], in which participants are asked to name the color of a word that spells out a different color (e.g., the word "BLUE" colored yellow), and visualization research maintains that it is best practice to choose semantically-appropriate colors when possible [40, 41]. At the same time, viewers tends to associate darker colors with more of something (i.e., a dark-is-more bias). Thus when visualizing a quantity of something that is associated strongly with light colors, a visual-semantic conflict may occur. For example, Figure 6 shows a color-encoded scatterplot of the percent concentration of ice in the arctic circle [32]. While, ice can be associated with a white/lighter color due to lived experience (e.g., an iceberg is white in a sea of blue), the dark-is-more bias conflicts with this semantic reasoning. In this case an alternate color scheme, or even a new encoding channel could increase comprehensibility.

6 CONCLUSION

Encoding severity via color in line charts can increase general trend interpretation of non-ideal variables. When visualizing PRO scores with the goal of improving interpretability, color-encoded smiley-timelines may be a useful alternative to line charts. Future research should explore novel approaches to reduce semantic confusion, within in and the beyond positional encoding of line charts.

SUPPLEMENTAL MATERIALS

All supplemental materials are available on OSF at <https://osf.io/kxzm3/>, released under a CC BY 4.0 license. In particular, they include (1) files containing the data and analyses for creating Tab. 3, Fig. 4 and Fig. 5, (2) files containing the data, analyses, and results for our pilot study, (3) examples of the stimuli used in the pilot and main studies, (4) screening questions and their answer keys, (5) a visual report of accuracy metrics from our main study.

ACKNOWLEDGMENTS

The authors wish to thank Nataliya Byrne for her research coordination and Raj Karia and Dr. Jim Slover for their early insights. We also thank Haifeng Zhang for helping with initial prototypes of PRO score visualizations. and Dr. Steve Haroz for his advisement on measuring proxies for mental effort.

This work was funded in part by the Agency for Healthcare Research Quality grant R21 HS027228 (MPI: HTG, EB)

REFERENCES

- [1] epros in clinical care - guidelines. <http://epros.becertain.org/reporting/guidelines/enhance-key-info>.
- [2] epros in clinical care - visualization library. <http://epros.becertain.org/tools-resources/applications/visualization-library>.
- [3] scipy.stats.bootstrap - scipy v1.11.1 manual. <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.html>.
- [4] Quality of life group website, Oct 2022. <https://qol.eortc.org/>.
- [5] Administration platforms, 3 2023. <https://www.healthmeasures.net/implement-healthmeasures/administration-platforms>.
- [6] Promis@, 2023. <https://www.healthmeasures.net/score-and-interpret/interpret-scores/promis>.

- [7] Score cut points, 2023. <https://www.healthmeasures.net/score-and-interpret/interpret-scores/promis/promis-score-cut-points>.
- [8] D. N. Ader. Developing the patient-reported outcomes measurement information system (PROMIS). *Medical Care*, 45(5):S1–S2, May 2007. doi: [10.1097/01.mlr.0000260537.45076.74](https://doi.org/10.1097/01.mlr.0000260537.45076.74)
- [9] A. Agarwal, T. Pain, J.-F. Levesque, A. Girgis, A. Hoffman, J. Karnon, M. T. King, K. K. Shah, R. L. Morton, and F. the HSRAANZ PROMs Special Interest Group. Patient-reported outcome measures (proms) to guide clinical care: recommendations and challenges. *Medical Journal of Australia*, 216(1):9–11, 2021. doi: [10.5694/mja2.51355](https://doi.org/10.5694/mja2.51355)
- [10] B. J. Ali J, Basch E. Users’ guide to integrating patient-reported outcomes in electronic health records. Technical report, Johns Hopkins University, 2017.
- [11] E. Austin, C. LeRouge, A. L. Hartzler, C. Segal, and D. C. Lavalley. Capturing the patient voice: Implementing patient-reported outcomes across the health system. *Quality of Life Research*, 29(2):347–355, 2019. doi: [10.1007/s11136-019-02320-8](https://doi.org/10.1007/s11136-019-02320-8)
- [12] E. J. Austin, C. LeRouge, J. R. Lee, C. Segal, S. Sangameswaran, J. Heim, W. B. Lober, A. L. Hartzler, and D. C. Lavalley. A learning health systems approach to integrating electronic patient-reported outcomes across the health care organization. *Learning Health Systems*, 5(4), Mar 2021. doi: [10.1002/lrh2.10263](https://doi.org/10.1002/lrh2.10263)
- [13] E. T. Bantug, T. Coles, K. C. Smith, C. F. Snyder, J. Rouette, and M. D. Brundage. Graphical displays of patient-reported outcomes (pro) for use in clinical practice: What makes a pro picture worth a thousand words? *Patient Education and Counseling*, 99(4):483–490, 2016. doi: [10.1016/j.pec.2015.10.027](https://doi.org/10.1016/j.pec.2015.10.027)
- [14] J. Bertin. *Semiology of Graphics*. University of Wisconsin Press, 1983.
- [15] E. Bertini, M. Correll, and S. Franconeri. Why shouldn’t all charts be scatter plots? beyond precision-driven visualizations, 2020. <https://arxiv.org/abs/2008.11310>. doi: [10.48550/ARXIV.2008.11310](https://doi.org/10.48550/ARXIV.2008.11310)
- [16] M. B. Boyce, J. P. Browne, and J. Greenhalgh. The experiences of professionals with using information from patient-reported outcome measures to improve the quality of healthcare: a systematic review of qualitative research. *BMJ Quality & Safety*, 23(6):508–518, Feb. 2014. doi: [10.1136/bmjqs-2013-002524](https://doi.org/10.1136/bmjqs-2013-002524)
- [17] M. Brundage, A. Blackford, E. Tolbert, K. Smith, E. Bantug, and C. Snyder. Presenting comparative study PRO results to clinicians and researchers: beyond the eye of the beholder. *Quality of Life Research*, 27(1):75–90, Nov. 2018. doi: [10.1007/s11136-017-1710-6](https://doi.org/10.1007/s11136-017-1710-6)
- [18] M. Brundage, D. Feldman-Stewart, A. Leis, A. Bezjak, L. Degner, K. Velji, L. Zetes-Zanatta, D. Tu, P. Ritvo, and J. Pater. Communicating quality of life information to cancer patients: A study of six presentation formats. *Journal of Clinical Oncology*, 23(28):6949–6956, Oct. 2005. doi: [10.1200/jco.2005.12.514](https://doi.org/10.1200/jco.2005.12.514)
- [19] M. D. Brundage, K. C. Smith, E. A. Little, E. T. Bantug, and C. F. Snyder. Communicating patient-reported outcome scores using graphic formats: results from a mixed-methods evaluation. *Quality of Life Research*, 24(10):2457–2472, May 2015. doi: [10.1007/s11136-015-0974-y](https://doi.org/10.1007/s11136-015-0974-y)
- [20] D. Cella, S. Yount, N. Rothrock, R. Gershon, K. Cook, B. Reeve, D. Ader, J. F. Fries, B. Bruce, and M. Rose. The patient-reported outcomes measurement information system (PROMIS). *Medical Care*, 45(5):S3–S11, May 2007. doi: [10.1097/01.mlr.0000258615.42478.55](https://doi.org/10.1097/01.mlr.0000258615.42478.55)
- [21] W. S. Cleveland and R. McGill. *Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods*, vol. 79. [American Statistical Association, Taylor Francis, Ltd.], 1984.
- [22] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 2 ed., 1988.
- [23] L. Cooper and D. Roter. Patient-provider communication: The effect of race and ethnicity on process and outcomes of healthcare. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*, pp. 552–593, 01 2003.
- [24] M. H. Crins, C. B. Terwee, R. Westhovens, D. Schaardenburg, N. Smits, J. Joly, P. Verschueren, K. Van der Elst, J. Dekker, M. Boers, and et al. First validation of the full promis pain interference and pain behavior item banks in patients with rheumatoid arthritis. *Arthritis Care and Research*, 72(11):1550–1559, 2020. doi: [10.1002/acr.24077](https://doi.org/10.1002/acr.24077)
- [25] R. A. Deyo, K. Ramsey, D. I. Buckley, L. Michaels, A. Kobus, E. Eckstrom, V. Forro, and C. Morris. Performance of a patient reported outcomes measurement information system (promis) short form in older adults with chronic musculoskeletal pain. *Pain Medicine*, 2015. doi: [10.1093/pm/pnv046](https://doi.org/10.1093/pm/pnv046)
- [26] B. Efron. *The bootstrap estimate of standard error*, p. 45–57. Chapman and Hall, 1 ed., 1993.
- [27] L. Grossman, S. Feiner, E. Mitchell, and R. M. Creber. Leveraging patient-reported outcomes using data visualization. *Applied Clinical Informatics*, 09(03):565–575, July 2018. doi: [10.1055/s-0038-1667041](https://doi.org/10.1055/s-0038-1667041)
- [28] M. Hartmann, V. Gashaj, A. Stahnke, and F. W. Mast. There is more than “more is up”: Hand and foot responses reverse the vertical association of number magnitudes. *J. Exp. Psychol. Hum. Percept. Perform.*, 40(4):1401–1414, Aug. 2014.
- [29] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’10, p. 203–212. Association for Computing Machinery, New York, NY, USA, 2010. doi: [10.1145/1753326.1753357](https://doi.org/10.1145/1753326.1753357)
- [30] R. P. Heitz. The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in Neuroscience*, 8, 2014. doi: [10.3389/fnins.2014.00150](https://doi.org/10.3389/fnins.2014.00150)
- [31] J. Izard, A. Hartzler, D. I. Avery, C. Shih, B. L. Dalkin, and J. L. Gore. User-centered design of quality of life reports for clinical care of patients with prostate cancer. *Surgery*, 155(5):789–796, 2014. doi: [10.1016/j.surg.2013.12.007](https://doi.org/10.1016/j.surg.2013.12.007)
- [32] B. K. Plotting sea ice concentration with 2 graphs using python., Jun 2022. <https://towardsdatascience.com/plotting-sea-ice-concentration-with-2-graphs-using-python-394bf4e8f361>.
- [33] P. C. L. F. B. A. P. e. a. Mariana Burca, Virginie Beauconsin. Is there semantic conflict in the stroop task? further evidence from a modified two-to-one stroop paradigm combined with singleletter coloring and cueing. *Experimental Psychology*, 68(5):274–283, 2021.
- [34] K. J. McCaffery, A. Dixon, A. Hayen, J. Jansen, S. Smith, and J. M. Simpson. The influence of graphic display format on the interpretations of quantitative risk information among adults with lower education and literacy: A randomized experimental study. *Medical Decision Making*, 32(4):532–544, 2012. PMID: 22074912. doi: [10.1177/0272989X11424926](https://doi.org/10.1177/0272989X11424926)
- [35] A. G. McNair, S. T. Brookes, C. R. Davis, M. Argyropoulos, and J. M. Blazeby. Communicating the results of randomized clinical trials: Do patients understand multidimensional patient-reported outcomes? *Journal of Clinical Oncology*, 28(5):738–743, Feb. 2010. doi: [10.1200/jco.2009.23.9111](https://doi.org/10.1200/jco.2009.23.9111)
- [36] T. Munzner. *Visualization Analysis and Design*. AK Peters Visualization Series. CRC Press, 2015.
- [37] A. V. Pandey, K. Rall, M. L. Satterthwaite, O. Nov, and E. Bertini. How deceptive are deceptive visualizations? an empirical analysis of common distortion techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI ’15, p. 1469–1478. Association for Computing Machinery, New York, NY, USA, 2015. doi: [10.1145/2702123.2702608](https://doi.org/10.1145/2702123.2702608)
- [38] F. Penedo, L. Oswald, J. Kronenfeld, S. Garcia, D. Cella, and B. Yanez. The increasing value of ehealth in the delivery of patient-centred cancer care. *The Lancet Oncology*, 21:e240–e251, 05 2020. doi: [10.1016/S1470-2045\(20\)30021-8](https://doi.org/10.1016/S1470-2045(20)30021-8)
- [39] H. C. Purchase. *Experimental Human-Computer Interaction: A Practical Guide with Visual Examples*. Cambridge University Press, 2012. doi: [10.1017/CBO9780511844522](https://doi.org/10.1017/CBO9780511844522)
- [40] K. B. Schloss, C. C. Gramazio, A. T. Silverman, M. L. Parker, and A. S. Wang. Mapping color to meaning in colormap data visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):810–819, 2019. doi: [10.1109/TVCG.2018.2865147](https://doi.org/10.1109/TVCG.2018.2865147)
- [41] V. Setlur and M. C. Stone. A linguistic approach to categorical color assignment for data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):698–707, 2016. doi: [10.1109/TVCG.2015.2467471](https://doi.org/10.1109/TVCG.2015.2467471)
- [42] P. M. Shah, R. E. Hegarty, and Mary. Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension. *Journal of Educational Psychology*, 91(4):690–702, 12 1999. doi: [10.1037/0022-0663.91.4.690](https://doi.org/10.1037/0022-0663.91.4.690)
- [43] K. C. Smith, M. D. Brundage, E. Tolbert, E. A. Little, E. T. Bantug, and C. F. Snyder. Engaging stakeholders to improve presentation of patient-reported outcomes data in clinical practice. *Supportive Care in Cancer*, 24(10):4149–4157, May 2016. doi: [10.1007/s00520-016-3240-0](https://doi.org/10.1007/s00520-016-3240-0)
- [44] C. Snyder, K. Smith, B. Holzner, Y. M. Rivera, E. Bantug, and M. Brundage. Making a picture worth a thousand numbers: recommendations for graphically displaying patient-reported outcomes data. *Quality of Life Research*, 28(2):345–356, Oct. 2018. doi: [10.1007/s11136-018-2020-3](https://doi.org/10.1007/s11136-018-2020-3)

- [45] C. F. Snyder, K. C. Smith, E. T. Bantug, E. E. Tolbert, A. L. Blackford, M. D. Brundage, and the PRO Data Presentation Stakeholder Advisory Board. What do these scores mean? presenting patient-reported outcomes data to patients and clinicians to improve interpretability. *Cancer*, 123(10):1848–1859, 2017. doi: [10.1002/cncr.30530](https://doi.org/10.1002/cncr.30530)
- [46] S. Stonbraker, T. Porras, and R. Schnall. Patient preferences for visualization of longitudinal patient-reported outcomes data. *Journal of the American Medical Informatics Association : JAMIA*, 27, 10 2019. doi: [10.1093/jamia/ocz189](https://doi.org/10.1093/jamia/ocz189)
- [47] J. Stroop. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18:643–662, 1935. doi: [10.1037/h0054651](https://doi.org/10.1037/h0054651)
- [48] E. Tolbert, M. Brundage, E. Bantug, A. L. Blackford, K. Smith, and C. S. and. Picture this: Presenting longitudinal patient-reported outcome research study results to patients. *Medical Decision Making*, 38(8):994–1005, Aug. 2018. doi: [10.1177/0272989x18791177](https://doi.org/10.1177/0272989x18791177)
- [49] M. Turchioe, L. Grossman, A. Myers, D. Baik, P. Goyal, and R. Masteron Creber. Visual analogies, not graphs, increase patients’ comprehension of changes in their health status. *Journal of the American Medical Informatics Association : JAMIA*, 27, 01 2020. doi: [10.1093/jamia/ocz217](https://doi.org/10.1093/jamia/ocz217)
- [50] B. Tversky. Visualizing thought. *Topics in Cognitive Science*, 3(3):499–535, 2011. doi: [10.1111/j.1756-8765.2010.01113.x](https://doi.org/10.1111/j.1756-8765.2010.01113.x)
- [51] B. Tversky, S. Kugelmass, and A. Winter. Cross-cultural and developmental trends in graphic productions. *Cognitive Psychology*, 23(4):515–557, 1991. doi: [10.1016/0010-0285\(91\)90005-9](https://doi.org/10.1016/0010-0285(91)90005-9)
- [52] N. W. Wagle. Implementing patient-reported outcome measures. *Catalyst Carryover*, 3(5), 2017. doi: [10.1056/CAT.17.0373](https://doi.org/10.1056/CAT.17.0373)
- [53] T. M. White, T. A. Slocum, and D. McDermott. Trends and issues in the use of quantitative color schemes in refereed journals. *Annals of the American Association of Geographers*, 107(4):829–848, 2017. doi: [10.1080/24694452.2017.1293503](https://doi.org/10.1080/24694452.2017.1293503)
- [54] B. Winter and T. Matlock. More is up and right: Random number generation along two axes. p. 3789–3974. Conference: Cognitive Science Society Meeting, 01 2013.
- [55] D. Wong and C. Baker. Smiling face as anchor for pain intensity scales. *Pain*, 89:295–297, 02 2001. doi: [10.1016/S0304-3959\(00\)00375-4](https://doi.org/10.1016/S0304-3959(00)00375-4)
- [56] D. L. Wong and C. M. Baker. Pain in children: Comparison of assessment scales. *Pediatric Nursing*, 14(1):9–17, 1998.
- [57] G. Woodin, B. Winter, and L. Padilla. Conceptual metaphor and graphical convention influence the interpretation of line graphs. *IEEE Transactions on Visualization and Computer Graphics*, 28(2):1209–1221, feb 2022. doi: [10.1109/TVCG.2021.3088343](https://doi.org/10.1109/TVCG.2021.3088343)
- [58] A. W. Wu, H. Kharrazi, L. E. Boulware, and C. F. Snyder. Measure once, cut twice—adding patient-reported outcome measures to the electronic health record for comparative effectiveness research. *Journal of Clinical Epidemiology*, 66(8, Supplement):S12–S20, 2013. Methods for Comparative Effectiveness Research/Patient-Centered Outcomes Research: From Efficacy to Effectiveness. doi: [10.1016/j.jclinepi.2013.04.005](https://doi.org/10.1016/j.jclinepi.2013.04.005)
- [59] J. Zacks and B. Tversky. Bars and lines: A study of graphic communication. *Memory cognition*, 27:1073–9, 12 1999. doi: [10.3758/BF03201236](https://doi.org/10.3758/BF03201236)
- [60] C. Ziemkiewicz and R. Kosara. The shaping of information by visual metaphors. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1269–1276, 2008. doi: [10.1109/TVCG.2008.171](https://doi.org/10.1109/TVCG.2008.171)
- [61] C. Ziemkiewicz and R. Kosara. Beyond bertin: Seeing the forest despite the trees. 30(5):7–11, 2010. doi: [10.1109/MCG.2010.83](https://doi.org/10.1109/MCG.2010.83)
- [62] C. Ziemkiewicz and R. Kosara. Implied dynamics in information visualization. p. 215–222, 2010.