

Guidelines Are Not Rules: Characterizing Terminologies around Visualization Design Guidelines

Anna L. Chinni*

University of Wisconsin-Madison

Petra Isenberg§

Univ.-Paris-Saclay, CNRS, Inria, LISN

Karen B. Schloss††

University of Wisconsin-Madison

Md Dilshadur Rahman†

University of Utah

Kushin Mukherjee¶

Stanford University

Ghulam Jilani Quadri¶

University of Oklahoma

Daniel Weiskopf‡‡

University of Stuttgart

Bon Adriel Aseniero‡

Autodesk Research

Paul Rosen**

University of Utah

Our findings suggest that there is a divergence in how visualization authors understand guidance-related terms which we should address for the practical impact and progress of data visualization research.

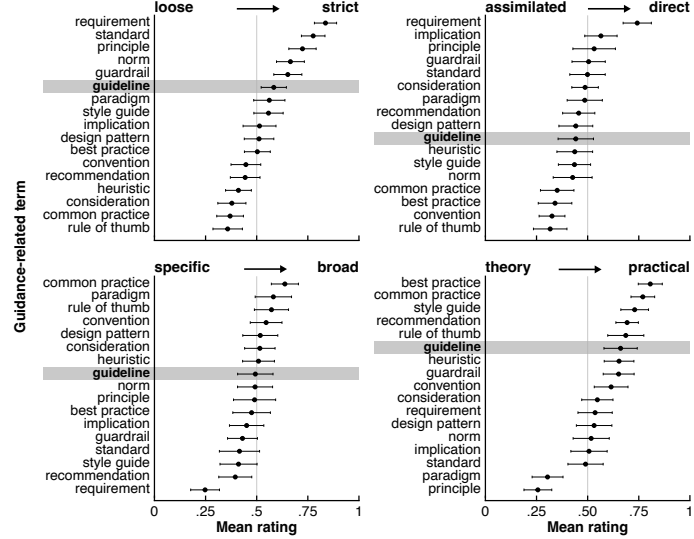


Figure 1: In visualization research, the term GUIDELINE is widely used to describe findings, yet the term is ambiguous and there is opportunity for more nuanced terminology. The ratings on the right show results from our crowdsourced study of guidance-related terms (see Sec. 3.2.3) (ordered by mean ratings averaged across participants with 95% confidence intervals) across 4 dimensions.

ABSTRACT

A common expectation in visualization research is that outcomes recommend how researchers and practitioners take action or make design decisions. We often express these as “guidelines.” Yet, the term “guideline” is both ambiguous and loosely defined, and what one researcher considers a guideline may be too broad, too loose, or too strict for another. We take a closer look at a broader set of terms that can express desirable results around visualization research, and untangle how these words are understood in the community in relation to other similar terms. We base our work on an exploratory study with experts, followed by a crowdsourcing study with a separate mapping phase ($n = 30$) and rating phase ($n = 42$) targeting input from the broader visualization community, and an analysis of

the use of terminology in 3,877 IEEE VIS papers published from 1990 to 2024. Based on our findings, we call for more nuanced, precise discussions of research outcomes and their communication to the broader community, including practitioners and students.

Index Terms: Guidelines, visualization, terminology.

1 INTRODUCTION

In the field of visualization, there exists a desire and need for generalizable visualization research findings that directly inform practice. Prescriptive advice on “what to do when” finds its way into introductory courses, textbooks, or tutorials, and is specifically common in practitioner-oriented visualization literature. It is, therefore, not surprising that an (erroneous) expectation is growing in the community that all forms of research should produce generalizable results that become quickly applicable and adoptable by the community. We have seen the expectation increasingly raised by reviewers who ask for guidelines for design—even when guidelines were inappropriate to propose from the current state of the work.

One reason for this mismatch in expectations may be a misunderstanding or disagreement about the meaning of guidance-related terminology in the community. Some may, for example, consider a GUIDELINE as a broad and loose RULE OF THUMB, while others see it as a strict rule to be followed [4, 34]. The importance of shared, consistently used terminology is often discussed in the study of science because it is crucial to how a community accumulates knowledge and community members learn to conduct science

*e-mail: alchinni@wisc.edu

†e-mail: dilshadur.rahman@utah.edu

‡e-mail: bon.aseniero@autodesk.com

§e-mail: petra.isenberg@inria.fr

¶e-mail: kushinm@cs.stanford.edu

¶e-mail: quadri@ou.edu

**e-mail: paul.rosen@utah.edu

††e-mail: kschloss@wisc.edu

‡‡e-mail: weiskopf@visus.uni-stuttgart.de

themselves [23]. Shared terminology allows work to be comparable; if one set of papers derives guidelines from authors’ opinions and another from rigorously validated empirical research, the resulting sets of guidelines become almost impossible to compare.

Ideally, authors would carefully choose or at least clearly define how they use the terms to describe the applicability of their research results, but we argue that this is rarely the case for the broader class of terms surrounding GUIDELINES. Yet, shared definitions would reduce conflicts between authors and reviewers and clarify expectations. Shared definitions would also further support cumulative knowledge building, literature searches, and meta-analyses of contributions. In discussions that began at a Dagstuhl Seminar [27], we have, therefore, set out to uncover the extent of the mismatch between community members’ understanding and use of guidance-related terms. We engaged in in-depth discussions among visualization experts on a set of identified words and their meanings. We also reviewed the literature on the use of the identified terms and conducted two community surveys to assess how other visualization experts understood them.

Through studies with crowdsourced participants and of visualization literature, we found that the terms vary in how they are understood, and that this variability carries over into practice. In our crowdsourced studies, guidance-related terms separated along several dimensions, including strictness, breadth, evidentiary basis, and theoretical vs. practical orientation, yet GUIDELINE, the term most often invoked in this debate, consistently fell near the center of these dimensions (Fig. 1), suggesting that the field’s most common term for guidance is also its least settled in meaning. A complementary analysis of three decades of IEEE VIS publications showed a parallel pattern: authors rely on a narrow handful of terms, chiefly GUIDELINE, RECOMMENDATION, and CONSIDERATION, to describe outcomes from their work. This narrow, ambiguous vocabulary has grown more common over the past decade rather than converging on more precise usage, underscoring that the mismatch between the nuance the community perceives and the words it actually uses is unlikely to resolve on its own.

Our findings suggest that, as we discuss ways to approach studying and producing effective visualizations, we should be mindful of the implications of the terms we use. We hope our results may serve as a starting point for more nuanced, precise discussions of research outcomes and their communication to the broader community, including practitioners and students, while also supporting the systematic and functional articulation of guidelines that inform system development, guide research, and improve teaching visualization. The landscape of guidance terminology that we propose, therefore, provides a practical scaffolding for communicating research results and their relative utility.

2 RELATED WORK

As a discipline at the intersection of computer science, art, design, mathematics, engineering, psychology, and many application domains, the visualization community has a very diverse relationship with “rules” or GUIDELINES. When students of visualization classes are trained in computer science, they are often happy to learn established guidelines like Cleveland and McGill’s ranking of elementary perceptual tasks [6] but frustrated when they hear that these rankings do not universally apply [7]. This frustration is perhaps unsurprising because one could argue that computer science training and practice are rule-based at their core; a focus on mathematics, algorithms, programming, or formal methods may instill a sense of certainty that rules are absolute, deterministic, and universally applicable. Art, and some areas of design, however, do not treat rules the same way. Here, rules are important for learning and mastering the craft but need to be broken to support creative expression [30]. Parsons and Shukla [32] argue that the visualization community is dominated by “Technical Rationality,” a shared view that practitioners’ application

of “general, codified knowledge” is important and researchers must deliver such knowledge. The authors further argue that the visualization field is predominantly positivist in nature, but that, to support design practice, both tacit and explicit knowledge must be better accounted for in the way visualization researchers communicate about and analyze their work.

Similar to Sedig and Parsons [34], we argue that visualization as a discipline grapples with these differing views but needs to agree on precise, shared terminology to prevent conflicts among readers, authors, reviewers, and, eventually, practitioners who seek to learn from visualization research. In contrast to their work, we focus here on the guidance landscape: a set of terms used to describe the practical or theoretical outcomes of visualization research.

2.1 Community Discourse on Guidelines

A visualization community discourse on guidelines has been fostered by the VisGuides workshop series, which started in 2016 under the name C4PGV [3]. The organizers identified a need to establish a centralized place to make guidelines and principles discoverable, describe their applicability, and record the evolution and improvement of principles and guidelines. By the fourth edition, the term PRINCIPLES was dropped from the workshop title [1] and the workshop description focused on guidelines with little mention of principles. One of the main outcomes of the workshop series is the accompanying VisGuides website <https://visguides.org/> [11]. Diehl et al. [12] later coded 248 posts that had been made on the site and found limited uptake and discussion, perhaps due to underlying difficulties with discussing details, searching for references, and abstracting and generalizing from specific empirical results to concrete advice. This issue recalls findings by Kim et al. [21], who documented a research–practice gap in visualization design guidelines through a mixed-methods study comparing 390 practitioner-facing guidelines with 235 empirical studies, and found frequent misalignment: empirical evidence often contradicted or only partially supported widely used guidelines, and researchers and practitioners prioritized different design attributes.

Another recent initiative, focused on guidelines, convened a Dagstuhl Seminar to unify visualization design recommendations [27]. The Dagstuhl Seminar was motivated by the question: “How do we formulate and integrate the knowledge we produce to best serve the visualization community and the world broadly?” and it included working groups on terminology (which is responsible for this paper), on practitioners, on competing value systems in visualization (this group held a panel at IEEE VIS 2025 [17]), and on guidelines+AI.

Given efforts such as these and the many researchers who participated in workshops, seminars, or panels on guidelines, it is safe to say that the visualization community has increasingly examined the landscape of guidance. We argue, however, that a critical lens on the terminology remains missing. In fact, few papers have provided definitions of guidance-related terms. A notable exception is the work by Chen et al. [4], who defined a guideline as “a process or a set of actions that may lead to a desired outcome or, alternatively, actions to be avoided to prevent an undesired outcome” and a principle as “a law or rule that must be followed and is usually expressed in a qualitative description.” These definitions broadly agree with ours, as offered in Table 1, but differ subtly in focus.

2.2 The Origin of Guidelines

Chen et al. [4] argued that guidelines for visualization stem from “accumulated experience and knowledge about some causal relations in a process.” If one views visualization research through a positivist lens, then this accumulated experience and knowledge must be grounded in objective empirical observation [32, 33] and may lead to guidelines that “can be systematized such that anyone can produce a high-quality, reliable visualization that shows the data without

extraneous distractors” [33]. If one takes a more interpretivist viewpoint, then the experience and knowledge Chen et al. discuss are acquired through critical reflection on situated practice: observing how practitioners actually design [33] or how people actually make sense of visualizations; and interpreting these observations in context. According to Meyer and Dykes [26], visualization research has questions and goals that are suited to each of these viewpoints. Yet, it is important to understand and question the evidentiary basis of visualization guidance and the processes by which empirical findings become broadly accepted [14, 16, 22, 24].

These arguments have long emerged in human-computer interaction. Greenberg and Buxton [16], for example, argued that empirical usability evaluations are often treated as a universal mechanism for validating research contributions, leading to design recommendations that are generalized beyond the scope of the supporting evidence. Similarly, Dourish [14] questioned the convention of translating empirical findings directly into design implications, arguing that empirical studies frequently contribute conceptual understanding rather than explicit design recommendations. According to Kosara [22], many canonical visualization principles originate from empirical findings, expert opinion, and common practice that are generalized into broadly accepted recommendations without clearly communicating their evidentiary basis or intended scope. More broadly, Lee et al. [24] emphasized that visualization knowledge is generated through diverse research paradigms, suggesting that design guidance may arise from multiple forms of research contributions rather than a single empirical methodology.

Collectively, these works demonstrate that visualization guidance varies in its origin, evidentiary support, and intended scope. However, they do not distinguish among or define terms that may be most suitable for describing types of guidance in the visualization literature. This problem is acutely felt by researchers who attempt to structure and define the guideline space [5, 11, 12], use it in recommendation systems [28], or derive theory from it.

3 STUDIES ON GUIDANCE TERMINOLOGY

We conducted two studies to deepen our understanding of the meaning of guidance terminology. The initial study took place during a Dagstuhl Seminar [27], in which we explored the space of guidance terminology and potential sources for such guidance (Sec. 3.1). The second study built on this initial approach but used more formal embedding and dimensionality reduction procedures, using data collected from the broader visualization community (Sec. 3.2).

3.1 Exploration of Guidance Terminology

This section describes an informal but structured exploration of guidance-related terminology. We aimed to build a shared foundation for understanding words used to describe different kinds of guidance and how those words vary in meaning among members of the visualization community. We also explored likely sources for providing different kinds of guidance.

3.1.1 Term Generation

To allow for nuanced and precise discussions of research outcomes, we need more than the term **GUIDELINE**. Six senior authors (BAA, PI, GJQ, PR, KBS, and DW) and one acknowledged contributor (CBX) discussed a purposefully large list of alternative terms to capture different aspects of the research results that might, for example, differ in their applicability, breadth, source, or strictness. We aimed to be forward-looking and include terms that are perhaps not yet used, but should be. To produce the definitions in Table 1 (also see the supplemental material for the preliminary definitions from the Dagstuhl discussion), we started with definitions from the Merriam-Webster English dictionary (online) and supplemented them with definitions from dictionary.com. We also asked ChatGPT for alternatives regarding specific nuances or variations of terms not

captured by the dictionary definition and picked those that captured our intent. We used AI when the dictionaries did not provide appropriate definitions; e.g., the two dictionaries defined **GUARDRAIL** only in its architectural sense. All authors reviewed and discussed the definitions and generated terms before inclusion in the final list.

3.1.2 Exploring the Space of Terms

As we generated our list of terms, we began exploring their similarity space on a large post-it pad. We placed terms that seemed similar in meaning close together, and terms that seemed different far apart. We iterated through this process, moving the terms around, to try to understand not only which terms were similar/different, but to start

Table 1: A list of definitions for the community to consider. The terms above the line were considered during the initial exploration (Sec. 3.1). These reflect our definitions, modified by our study results and rewritten with the help of AI (<https://chatgpt.com/share/6a469678-9714-83eb-9762-08f90c0671a3>). Icons indicate likely sources of each term, determined using survey data (see Sec. 3.1):  empirical evidence;  eminent leader;  art & design community;  emergent from unknown source (darker icons indicate more frequent responses:    , see Supplemental Fig. S.1 for data in matrix form). Terms below the line were added and used in Study 2 (Sec. 3.2) after the survey and, therefore, have no sources.

BEST PRACTICE	A practical, experience-based method derived from assimilated knowledge and widely accepted for reliably producing desirable results.	   
COMMON PRACTICE	A loose, broad, experience-based, and practical way of doing something, derived from assimilated knowledge and typically followed in practice.	   
CONSIDERATION	A loose factor or aspect to be kept in mind or given careful thought when thinking about or deciding something, without imposing any constraint or obligation.	   
CONVENTION	A rather loose, experience-based, and practical method of doing something, derived from assimilated knowledge and established in practice.	   
GUIDELINE	A moderately strict, practical recommendation derived from assimilated knowledge that helps steer actions or decisions without prescribing a rigid rule.	   
GUARDRAIL	A strict, relatively specific, practical constraint intended to prevent undesirable results.	   
NORM	A strict, widely accepted convention or practice grounded in assimilated knowledge and accumulated experience rather than formal theory.	   
PARADIGM	A theoretical, slightly broad framework or model that provides a modestly prescriptive way of approaching problems and developing solutions.	   
PRINCIPLE	A strict, theory-derived fundamental law or truth from which other concepts, rules, or conclusions are derived.	   
RECOMMENDATION	A practical, experience-based, and rather loosely binding yet specific suggestion derived from assimilated knowledge about actions, choices, or considerations.	   
RULE OF THUMB	A loose, broad, experience-based, and practical heuristic derived from assimilated knowledge and used for approximate decision-making.	   
STANDARD	A strict, moderately specific set of established criteria intended to ensure desirable results.	   
STYLE GUIDE	A practical, slightly strict, and rather specific set of rules derived from assimilated knowledge and established usage to ensure consistency.	   
DESIGN PATTERN	A structured, reusable solution to a recurring design problem, grounded rather in assimilated knowledge and practice than in theoretical derivation.	
HEURISTIC	A loose, experience-based, and practical approach derived from assimilated knowledge, used to guide discovery and problem-solving without strict formal rules.	
IMPLICATION	The rather specific practical consequences, effects, or recommendations that follow from a piece of guidance or a set of findings.	
REQUIREMENT	A strict and specific condition or obligation derived from empirical observation that must be satisfied.	

to understand the underlying dimensions of this space. Early on, we established that one important dimension was likely *strictness*, with terms like **GUARDRAIL** being near the *strict* endpoint of the dimension and terms like **RULE OF THUMB** being near the *loose* endpoint of the dimension. We explored various possibilities of what a second dimension might be, and settled on *coverage* ranging from *broad* to *specific*, with terms like **PRINCIPLE** being near the broad end and terms like **STYLE GUIDE** being near the specific endpoint.

3.1.3 Studying the Dimensionality of Terms and Sources

Having established our list of terms and identifying dimensions in which these terms might vary, we next conducted an informal survey to better understand these dimensions. We invited the other participants of our seminar, not involved in our working group; 13 responded. The survey had three parts. First, we presented participants with each of the 13 terms and asked them to rate each term on the following dimensions using a 5-point Likert scale: (1) *Strictness*: the degree to which the term implies loose/suggestive vs. strict/enforced guidance, and (2) *Coverage*: the degree to which the term spans the visualization design space, from very specific/narrow to very broad/general. Next, we gave the participants the list of terms again and asked them to categorize each term by its likely source, choosing from: empirical evidence (🎓), an eminent leader in the field (👑), the art and design community (🎨), or emergent from unknown sources (🤖). They were asked to select all sources that applied. Finally, we asked participants if there were any guideline-related terms that we left out that they thought we should consider.

The heatmaps in **Fig. 2** show the frequency ratings along the dimensions of strictness and coverage (see caption for details). Generally, judgments of the terms often covered the full spectrum, with **GUIDELINE** being a term that most participants judged in the center of both scales (i.e., the black center) but also varied widely across the full range of the dimensions (i.e., ratings spread across the entire heatmap). Other terms, like **STANDARD** and **NORM**, however, were interpreted more consistently as rather specific and strict.

To better understand how these terms relate to one another and are positioned in the Strictness/Coverage space, we superimposed the data from **Fig. 2** into a single plot in **Fig. 3**. We generated **Fig. 3** by (1) only including cells from **Fig. 2** with greater than 25% frequency for any term, (2) marking the highest frequency cells for each term with a dot (or dots for ties), (3) marking the remainder of cells which exceeded the 25% frequency threshold with lines connecting to their corresponding dots (we used dashed lines when connected cells were not immediately adjacent), (4) color-coding points and lines according to term strictness (cooler hues were looser, warmer hues were stricter) and broadness (lighter colors were more specific, darker colors were more broad). The resulting **Fig. 3** serves as an initial map of how guidance-related terminology varies from term to term, and also how much individuals vary in their interpretation of each term. For example, **GUIDELINE** landed near the middle of both dimensions with considerable variability in strictness, which may suggest high ambiguity in how the community interprets this term. **RULE OF THUMB** was considered relatively loose and broad, though some participants thought it was more specific. In **Fig. 3**, Strictness increased from **CONSIDERATION** to **RECOMMENDATION** to **STANDARD**, and all of these terms were moderate on coverage. And broadness increased from **GUARDRAIL** to **PRINCIPLE** to **PARADIGM**, but they were all relatively strict. These results suggest that when discussing guidance in the context of research, education, and design practice, there is opportunity to choose guidance terminology that aligns best with the intended strictness and coverage.

Table 1 shows the results regarding the likely source for each of the 13 terms. The source categorizations seemed to differ across terms, such as **PRINCIPLE** stemming most empirical evidence (🎓) and eminent leaders (👑) and **RULE OF THUMB** emerging from the

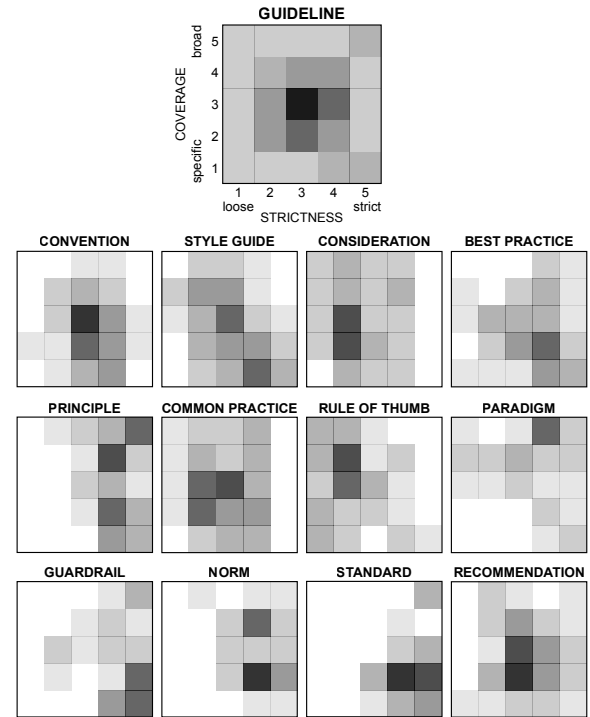


Figure 2: Frequency of term ratings along the dimensions of strictness (columns) and coverage (rows) from Study 1. Cell darkness indicates how many respondents chose that strictness–coverage combination (darker indicates greater agreement). A tightly clustered dark region indicates strong consensus (e.g., **STANDARD**); a diffuse, spread-out pattern indicates disagreement (e.g., **GUIDELINE**).

art and design community (🎨) or emerging from unknown sources (🤖). These results may be useful for aligning terminology choices with the source of information.

Finally, in response to our question about additional terms to consider, the participants offered **HEURISTIC**, **DESIGN PATTERN**, **REQUIREMENT**, and **JUSTIFICATION**. Through informal discussions, they also suggested **IMPLICATION**. For the crowdsourced studies, discussed in the next section, we decided to incorporate all of these suggestions except **JUSTIFICATION** because the term does not fit the guidelines space. A justification is not guidance about what should be done; it is an explanation of why something was done. We added the definitions for the new terms at the bottom of **Table 1** using the same methodology as above. However, as the dictionaries did not offer definitions for **DESIGN PATTERN**, we based our initial definition on Dearden and Finlay’s [10].

3.1.4 Limitations of This Initial Exploration

This study provided useful initial structure, but it was limited by the small, non-representative sample of workshop attendees, and the dimensions were researcher-defined rather than data-driven. Thus, we conducted two follow-up studies: a larger-scale crowdsourced experiment to map the space of guidance terminology and understand its dimensions (**Sec. 3.2**), and a corpus analysis to understand actual term usage in the literature (**Sec. 4**).

3.2 Crowdsourced Studies

To test the extent to which the space of guidance terminology from our exploratory study represents the perspectives of broader visualization community members, we conducted a study surveying a larger population of visualization researchers to (1) map the space of terminology (mapping phase) in which participants made triadic similarity judgments and (2) understand the nature of the space’s

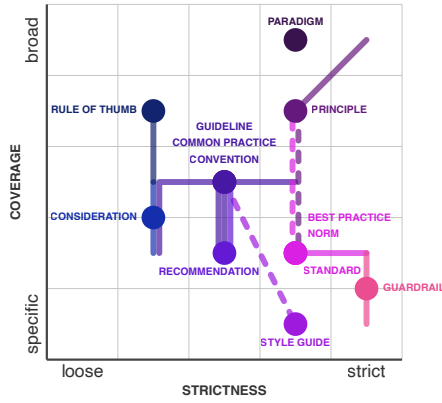


Figure 3: A visual map of how terms relate to one another, based on the data from the survey in Fig. 2, retaining the more frequent participant feedback (see text for details). Color coding: cooler colors = loose terms, warmer = strict, lighter = specific, darker = broad.

underlying dimensions by having participants rate terms on bipolar scales (rating phase). The experiment scripts, data, and analysis scripts for both phases of this study can be found in our GitHub repository (github.com/SchlossVRL/vis_guidelines).

3.2.1 Participants

To identify candidate participants, we first collected email addresses from IEEE VIS papers of the last five years using the *Grobid* tool. Next, we removed duplicates, incorrectly identified emails, and email lists. We considered excluding participants from Study 1, but opted to keep them in the pool as these studies consist of completely new tasks, and we did not want to limit input from members of the field. We then randomly divided the list into two halves: one that was invited to participate in the mapping phase, and the other that was subsequently invited to participate in the rating phase. We sent 387 invitations in the mapping phase and 395 in the rating phase (both numbers exclude bounced emails). A total of 30 participants completed the mapping phase, and 42 participants completed the rating phase. In both phases, we excluded data from participants who reported having fewer than 1 year of experience in the visualization community or did not complete all trials (mapping phase: $n = 1$ excluded, rating phase: $n = 0$ excluded). For participants in the mapping phase, their reported gender included: 10 women, 18 men, 1 mostly male, and 1 who chose not to specify, and their reported years in the VIS community included: 9 with > 1 , 7 with ≥ 10 , 9 with ≥ 10 , 4 with ≥ 30 , and 1 with $40+$ years (min = 5, max = $40+$). Reported gender identities in the rating phase included: 14 women, 26 men, 1 he/they, and one who chose not to specify, and reported years in the VIS community included: 15 with > 1 , 14 with ≥ 10 , 10 with ≥ 10 , and 3 with ≥ 30 (min = 3, max = 35). All participants gave informed consent, and the University of Utah IRB approved the experiment protocol.

3.2.2 Mapping Phase

This phase was analogous to the space exploration phase of Study 1 (Sec. 3.1.2), but using a more rigorous embedding procedure.

Procedure. The stimulus set included 17 guidance-related terms from our initial list (Sec. 3.1.1), expanded by terms suggested in survey feedback (Table 1). The final set included: DESIGN PATTERN, HEURISTIC, REQUIREMENT, IMPLICATION, PRINCIPLE, RULE OF THUMB, CONSIDERATION, COMMON PRACTICE, CONVENTION, RECOMMENDATION, BEST PRACTICE, STANDARD, NORM, GUARDRAIL, PARADIGM, STYLE GUIDE, and GUIDELINE.

We used a triadic similarity task to map the space of guidance terminology [15, 18, 20, 29]. While making their similarity judgments, participants were instructed to think back to times they had

read the results or discussions of a VIS paper and came across words describing the contributions or results of the work. On each trial, participants saw one target word and two choice words and selected the choice word that they felt was most similar to the target (see Supplemental Fig. S.2A for an example trial). They were informed there were no right or wrong answers for most trials and to go with their gut. The only trials with ‘correct’ answers were 10 catch trials in which one choice word was the same as the target. These catch trials were used to ensure participants were paying attention and understood their task, but were not included in analyses.

We collected 2,408 randomly sampled triplet trials for estimating our embeddings. Prior work established that at least $n \cdot d \cdot \log_2 n$ judgments are needed to estimate a stable d -dimensional embedding [20], where n is the number of items. We collected 10 times this number to account for noise in judgments to arrive at a total of 2,408 judgments. We distributed these trials across participants, each of whom completed 100 total trials—80 randomly sampled triplets used to fit the embedding, 10 validation triplets to estimate embedding fits, and 10 catch trials. All experiments were presented through the browsers on participants’ personal computers via jsPsych [8, 9].

Analysis and Results. We used the similarity judgments to estimate low-dimensional embeddings for each term such that terms frequently judged to be similar were embedded closer together in the space. We estimated embeddings using the crowd kernel learning (CKL) algorithm [37], which does not require similarity judgments for all possible pairwise sets of terms, unlike other classic approaches like multidimensional scaling (MDS) [38]. Instead, by collecting a sufficiently large set of ordinal judgments (e.g., a duck is closer to a canary than to a horse), modern implementations of CKL [35] can estimate a reliable embedding from a tractable number of human judgments. We fit embeddings in spaces of up to five dimensions and measured each solution’s performance with respect to dimensionality, thus identifying an appropriate dimensionality for the embedding (see Supplemental Sec. S.3 for details). We observed that all solutions performed similarly in predicting the held-out human judgments.

To further assess the dimensionality of the guidance terminology space, we conducted principal component analysis (PCA) on the five-dimension solution. PCA revealed that the first four components accounted for 95% of the variance in the embedding, with the fifth component contributing relatively little additional variance. These results suggested that the underlying structure of the guidance terminology space was likely 4D. We therefore retained a four-dimensional solution for subsequent analyses. To obtain an orthogonal set of components ordered by explained variance, we conducted PCA on the 4D embedding to interpret (Fig. 4a,b; additional results can be found in the Supplemental Material).

The first principal component (PC) seemed to align with the Strictness dimension from the exploratory study in Sec. 3.1. We also identified three potential dimensions that the additional components could represent: specific vs. broad (Coverage), theoretical vs. practical, and assimilated knowledge vs. direct evidence. To test how well these dimensions captured the variability in the guidance terminology space, we conducted the rating phase of the study.

3.2.3 Rating Phase

This phase was analogous to the dimension-rating task in Study 1 (Sec. 3.1.3), but with additional terms and scales.

Procedure. Participants rated the 17 guidance terms along the following dimensions: loose vs. strict (Strictness in Study 1), specific vs. broad (Coverage in Study 1), theoretical vs. practical, assimilated knowledge vs. direct evidence.

During the experiment, we instructed participants to think back to times when they had read the results or discussion of a VIS paper and encountered words describing the contributions or results of the work. Their task was to rate a list of such words, one at a time,

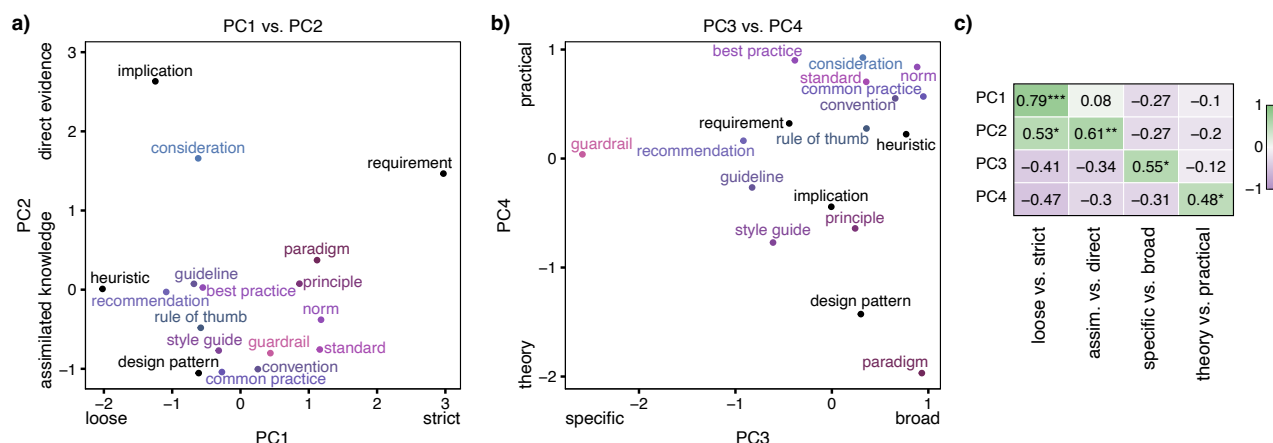


Figure 4: Guidance terms embedded in PC space for (a) PC1 with PC2 and (b) PC3 with PC4. The text color corresponds to Fig. 3 from Study 1 (new terms for Study 2 are in black). The axis labels of the PC plots are based on (c), which shows Pearson's r correlations between each rating scale and each PC: PC1 was strongly correlated with loose vs. strict, PC2 with direct evidence vs. assimilated knowledge, PC3 with broadness, and PC4 with theoretical vs. practical. * $p < .05$, ** $p < .01$, *** $p < .001$.

on a series of scales. They were then shown the full list of terms to be rated and the full list of scales on which each term would be rated. They were asked to anchor the endpoints of each scale in the context of the terms [31] by thinking about which terms they associate most strongly with each endpoint of each scale and to rate those words near the appropriate endpoints when encountered during the task. They were instructed to rate a word near the midpoint of a scale if they did not associate it strongly with either end of the scale. Participants would rate all of the words for a given scale, one at a time, before moving on to the next scale. Finally, before beginning, participants were told how to respond using the line-mark slider scale by sliding the cursor along the scale and clicking to record their response (see Supplemental Fig. S.2B for an example trial).

We presented trials in a block-randomized design—all terms for a given scale were presented in random order before moving on to the next scale, and the order of the scales was randomized. To reduce the risk of response-direction biases, we flipped the endpoints of the scales so participants were equally likely to rate terms on reversed versions of scales. For analysis, we flipped the ratings on reversed scales to be consistent with original scales and scaled all ratings from 0 to 1.

Analyses and Results. To test how well the scales described variability along the PCs from the mapping phase, we correlated the mean ratings for each term along each scale with the coordinates for each term along each PC. Fig. 4c shows the resulting correlation matrix. PC1 was strongly correlated with loose vs. strict, PC2 with direct evidence vs. assimilated knowledge, PC3 with specific vs. broad, and PC4 with theoretical vs. practical. Thus, similar to Study 1, strictness was the primary dimension of the term space (PC1), and coverage was also a relevant dimension (PC3), but this study also revealed two additional dimensions.

Fig. 1 (right) shows the mean ratings for each term along each of the four scales. For example, for loose vs. strict, looser terms included RULE OF THUMB and COMMON PRACTICE, whereas stricter terms included STANDARD and REQUIREMENT; for assimilated knowledge vs. direct evidence, assimilated knowledge terms included RULE OF THUMB and CONVENTION, whereas direct evidence terms included REQUIREMENT and IMPLICATION; for specific vs. broad, specific terms included REQUIREMENT and RECOMMENDATION, whereas broad terms included COMMON PRACTICE and PARADIGM; and for theory vs. practical, theory terms included PRINCIPLE and PARADIGM, whereas practical terms included COMMON PRACTICE and BEST PRACTICE.

These findings can be used to help researchers, educators, and

designers select terminology that best aligns with their intended message when communicating about visualization guidance.

4 TERM USAGE IN VISUALIZATION LITERATURE

We now move from perceived meaning to observed use, examining how terms appear in published VIS research. We focus on guidance attributed to a paper's own findings, analysis, or design process, and set aside prior work, technical meanings, and unrelated uses. We characterize this usage rather than judge whether an author selected the most appropriate term. The results can be explored at vis-guidelines.netlify.app.

4.1 Corpus and Method

We analyzed the full text of 3,877 IEEE VIS papers published from 1990 to 2024, collected from vispubdata.org [19]. We retrieved the exact singular and plural forms of the 18 terms in our terminology analysis and retained the surrounding passage, paper metadata, and document location for each match. This process yielded 32,629 occurrences across 10,838 term–paper pairs. We classified each occurrence as *guidance-adjacent*, *ambiguous*, or *likely non-guidance*. Guidance-adjacent cases described design or practice guidance attributed to a paper's work and presented as relevant beyond the immediate system or study. Ambiguous cases lacked enough context to determine the term's function, source, or intended scope. Likely non-guidance cases included technical or idiomatic meanings, bibliographic occurrences, guidance attributed to prior work, system-generated recommendations, and other unrelated uses.

We used Claude Opus 4.7 to assign an initial label to each occurrence according to these criteria. We then manually reviewed every occurrence initially labeled guidance-adjacent and corrected misclassifications. At the term–paper level, we assigned the strongest applicable label: guidance-adjacent if the pair contained at least one verified guidance-adjacent occurrence, ambiguous if it contained no guidance-adjacent occurrence but at least one ambiguous occurrence, and likely non-guidance otherwise. After classification and manual verification, we identified 625 guidance-adjacent occurrences across 319 of the 3,877 papers.

The supplemental website (vis-guidelines.netlify.app) provides the classification criteria, term-specific notes, representative and borderline examples, aggregation rule, and final occurrence-level classifications used in our analysis.

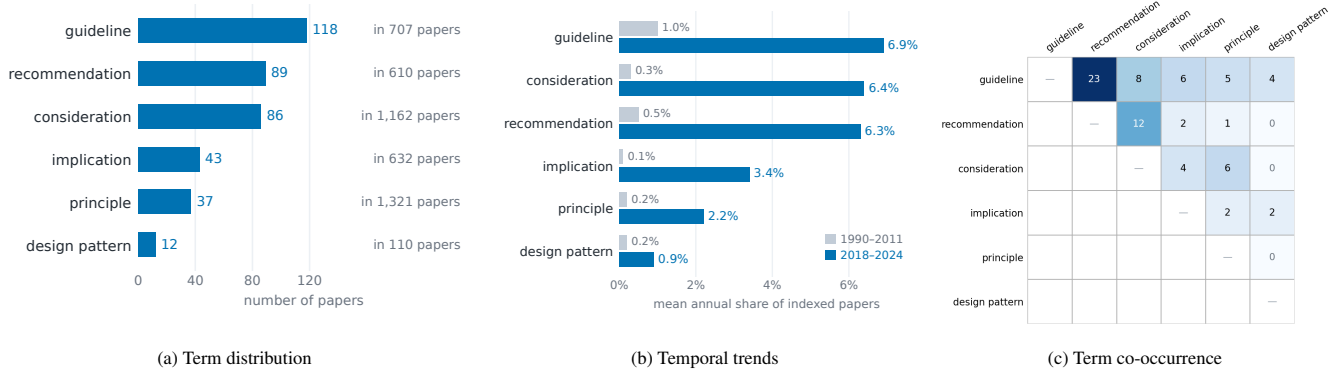


Figure 5: Guidance-adjacent term usage in the IEEE VIS corpus: (a) Guidance-adjacent use of the six most frequent terms, (b) their comparison for the 1990–2011 and 2018–2024 periods, and (c) pairwise co-occurrence of guidance-adjacent terms within papers.

4.2 Results

4.2.1 Distribution of Guidance-Adjacent Terms

The corpus of use was much narrower than our terminology landscape in Sec. 3. Authors used a small set of terms to describe guidance in their own work, even though those terms differ in how visualization researchers understand them. Overall word frequency also differed from guidance-adjacent use, and most papers relied on one guidance-related label rather than alternating among several.

Concentration in a small set of terms. We verified 625 guidance-adjacent occurrences in 319 papers, or 8.2% of the corpus. These occurrences formed 392 term–paper pairs, of which 385, or 98.2%, involved only six terms: GUIDELINE, RECOMMENDATION, CONSIDERATION, IMPLICATION, PRINCIPLE, and DESIGN PATTERN (Fig. 5a). GUIDELINE had the largest number of guidance-adjacent papers at 118, followed by RECOMMENDATION at 89, and CONSIDERATION at 86. BEST PRACTICE, HEURISTIC, and RULE OF THUMB together accounted for only seven additional pairs, while the remaining nine terms had no verified guidance-adjacent use. This concentration does not mean that the six terms describe the same kind of outcome. Sec. 3 shows that they differ in perceived strictness, breadth, evidentiary basis, and practical orientation. A CONSIDERATION derived from a design process, an IMPLICATION drawn from an empirical result, and a RECOMMENDATION directed toward practitioners can all appear as outcomes of a paper, but they communicate different expectations about how broadly, directly, or strongly the result should guide subsequent work. The corpus therefore shows that a small vocabulary is used to communicate several distinct forms of guidance-related knowledge.

Prevalence and guidance-adjacent use. A term’s overall prevalence did not indicate how often authors used it to present guidance from their own work. STANDARD appeared in 2,176 papers (56.1% of the corpus) and REQUIREMENT in 1,693 papers (43.7%), yet we retained no guidance-adjacent occurrence for either term. In the reviewed passages, REQUIREMENT usually referred to specifications for the system or study at hand rather than an outcome intended to inform work beyond that immediate context. By comparison, guidance-adjacent uses appeared in 16.7% of papers for GUIDELINE, in 14.6% for RECOMMENDATION, and in 10.9% for DESIGN PATTERN. The corresponding shares were lower for CONSIDERATION, IMPLICATION, and PRINCIPLE, even though some of these words appeared more often overall.

These differences separate the presence of a word from the role that it serves in a paper. Terms such as STANDARD and REQUIREMENT were common because they frequently described technical criteria, study conditions, or system specifications. Terms such as GUIDELINE and RECOMMENDATION were less common overall but more often used to present guidance-related outcomes. Whether a term signals a guidance contribution depends on the surrounding

claim, not simply on how often the word appears in the literature.

4.2.2 Temporal Patterns

Each of the six most frequent terms had a higher mean annual share of guidance-adjacent use in 2018–2024 than in 1990–2011 (Fig. 5b): from 1.0% to 6.9% for GUIDELINE, from 0.5% to 6.3% for RECOMMENDATION, and from 0.3% to 6.4% for CONSIDERATION. IMPLICATION, PRINCIPLE, and DESIGN PATTERN also had higher shares in the later period, although they remained less common.

These period averages do not establish a steady year-over-year increase, and the corpus does not explain why guidance-adjacent language became more common. The pattern is nevertheless broader than the increasing use of GUIDELINE alone. Several terms became more visible as labels for outcomes of a paper. The later literature continued to use multiple labels rather than converging on a single term for guidance-related contributions.

4.2.3 Within-Paper Patterns

Most papers used only one guidance-adjacent term. Of the 319 papers, 252 (79.0%) used one term, 61 used two, and six used three. The most frequent combination was GUIDELINE and RECOMMENDATION, found in 23 papers, followed by CONSIDERATION and RECOMMENDATION, found in 12 (Fig. 5c). The presence of two terms in a single paper does not imply that the authors treated them as synonyms, as the terms may describe different findings, inputs, or outputs. Still, co-use provides some evidence that authors alternated among several labels when discussing guidance-related outcomes. The corpus, therefore, does not point to indiscriminate switching as the main issue. Instead, a single label often serves as the primary cue to the status of a claim, even though Sec. 3 shows that readers may associate that label with different levels of strictness, breadth, evidence, and practical relevance.

4.3 Implications and Limitations

4.3.1 Term Use and Interpretation

Our corpus analysis shows that authors used only a narrow subset of terms when describing guidance from their own work. Six terms accounted for 98.2% of the guidance-adjacent term–paper pairs, and 79.0% of the papers used only one such term. For example, the varied uses of GUIDELINE illustrate the limits of relying on the label alone. Marschner and Lobb described quantitative metrics for evaluating reconstruction filters as providing “a useful guideline” [25], whereas Bederson and Boltman used the term for advice derived from an experiment, referring to “preliminary design guidelines” [2]. Djurcilov and Pang applied “a general guideline” to a conditional recommendation about when to use scattered-data methods [13]. Stasko and Zhang presented “design guidelines” developed from

system experience, experimental observations, and an analysis of related systems [36]. In these papers, GUIDELINE refers to evaluative criteria, tentative empirical advice, a conditional prescription, and guidance developed from several forms of evidence.

These examples do not establish that the term was used incorrectly. They show that GUIDELINE alone does not specify the form of the outcome, the evidence supporting it, its intended scope, or how strictly it should be followed. The same concern applies to CONSIDERATION, IMPLICATION, RECOMMENDATION, PRINCIPLE, and DESIGN PATTERN. Such terms can help characterize a contribution, but they do not replace an account of what the claim applies to, where it comes from, whom it is intended to inform, and how it should guide action. Our analysis therefore concerns the communication of guidance rather than the correctness of individual word choices. It also cannot explain why particular terms were favored; familiarity, disciplinary convention, reviewer expectations, and fit to the contribution remain possible explanations.

4.3.2 Sources of Guidance

We retained guidance presented as an outcome of a paper’s findings, analysis, or design process. We checked how the term was attributed, not whether it rested on a single form of evidence. Some cases drew on an experiment, while others combined empirical observations, design experience, and analysis conducted within the paper. We excluded passages that only attributed guidance to prior work, professional practice, community consensus, or another external source. The analysis, therefore, primarily captures the direct-evidence side of the distinction in Sec. 3, rather than the broader process by which evidence and experience accumulate into established guidance.

5 RECOMMENDATIONS

Our results highlighted a divergence in how visualization authors understand guidance-related terms. This divergence is problematic for research progress and the practical impact of visualization work. When researchers and authors have diverging understandings of what authors contributed and what they expected in terms of a contribution, frustration on either side is almost inevitable in the publication process. When students search for clear rules to follow but encounter only loose suggestions, they may feel lost. Furthermore, if guidance is understood primarily as a means of informing practice, researchers have argued that we first need to determine what type of guidance practitioners actually need [32].

It is clear that as we communicate about ways to approach studying and producing effective visualizations, we have to be mindful of the implications of the terms we use. Below, we outline four categories of recommendations which should be seen as “practical, experience-based, rather loosely binding yet specific suggestions” according to our study results:

Authors. Authors of visualization papers should think carefully about what they discovered and choose terms that really fit the contribution. For example, a paper describing a new design space might not be able to offer strict requirements or even loose considerations without direct evidence, but could easily describe common practice or make recommendations for new avenues to explore. Fig. 1 and Table 1 can help with choosing the right terminology—or authors could use them as a basis for adapting and describing a changed use of the term. In addition, our axes can serve as considerations for describing the scope of contributions: are they meant to be rather loose suggestions or strict requirements to follow, can they be broadly applicable or only to the specific context of the paper, or are they meant to advance theory or practice?

Reviewers. If reviewers are aware of the nuances of guidance concepts in visualization research and hold a shared understanding of terminology, that will help ensure their feedback aligns with the paper’s content and goals. Moreover, this understanding will help reviewers avoid asking for inappropriate revisions, such as asking for

stricter terms, such as GUIDELINES, even though only loose terms would make sense given the research presented. It is compelling to push authors to stricter, broader, and more actionable conclusions of their work. If the research outcomes of a paper under review provide evidence to support such conclusions, then, of course, they should be incorporated in the revision process. However, reviewers are advised not to push too far, which could result in published papers with potentially misleading or unreliable messages.

Research community. Reliable descriptions of research outcomes are key to progress in the research community as a whole, which continues to build on the body of published work. With a clearer, shared terminology, there will be more progress. It could be as simple as improving how we can assemble and merge knowledge from collections of papers (survey papers, textbooks, course notes, etc.). This effort can pave a way toward theory, following a path from single studies to meta-studies, to principles and a solid theory—at least for some facets of visualization. Some pathways toward visualization theory are discussed by Chen et al. [4]. The visualization research community is asked to adopt a shared, nuanced terminology to foster progress in this area. The findings of our paper can serve as a basis to inform terminology, but it will be a negotiation process until a shared understanding is established. This process might include an iterative approach of evolving reviewing and publication expectations, and could be fostered by future workshops, courses at conferences, or other community events.

Educators. Teachers or other educators of visualization research or practice are encouraged to clearly articulate what kind of messages they want to communicate, e.g., to students or outside practitioners. There is likely even less shared terminology and, even more importantly, shared background knowledge than in communication between researchers within the visualization research community. Therefore, even more effort should be placed on clarifying the meaning of the terminology used, e.g., by explaining or defining terms. With an appropriate understanding, it will be easier to manage the expectations of students or practitioners, thus fostering the actionability of visualization research findings.

6 CONCLUSION

The main goal of this paper is to raise awareness of the terminology used in the visualization guidance landscape and to shed some light on how terms are interpreted and used in the visualization research community. We based our discussion on a diverse and complementary set of sources: an exploratory study with experts at a Dagstuhl Seminar, a crowdsourcing study (with separate mapping and rating phases) that targeted input from the broader visualization community, and an investigation of the use of terminology in published VIS research. Our results are summarized in the ratings shown in Fig. 1 and our list of terms and their definitions in Table 1.

We decided to apply our findings to our own paper: After discussion within the group of authors, we converged to naming Sec. 5 RECOMMENDATIONS, based on our judgment regarding strictness (here being slightly toward the loose end) and being rather specific, practical, and mostly based on assimilated knowledge (observations from the literature and feedback from many experts in the field). However, we want to point out that there could have been other valid choices for the term used in our case. Also, we acknowledge that our list of terms is not comprehensive, e.g., additional useful terms could be RULES, CONSTRAINTS, or SUGGESTIONS.

Therefore, in general, we advocate for some flexibility, to be able to accommodate the many different facets and nuances in guidance results of visualization research, echoing the call for supporting diversity in visualization contributions by Lee et al. [24]. Through a negotiation process within the community, a better, shared understanding of terminology can be achieved, thus helping how we communicate guidance-related findings.

AI USE STATEMENT

Claude 4.8 Opus helped generate the illustrative artwork in Fig. 1 (the data plots were not AI-generated), assign labels (i.e., guidance-adjacent) to occurrences of terms in our corpus analysis in Sec. 4, and provided editorial and structural support in preparing the manuscript (this assistance did not generate the research content, findings, or conclusions). GPT 5.3 and Claude 4.8 Opus helped debug and write experiment code and debug figure style code. ChatGPT helped with definitions for guidance-related terminology in Sec. 3.1. All AI generated content was reviewed by the authors for accuracy.

ACKNOWLEDGMENTS

This work stemmed from Dagstuhl Seminar 25232, “Navigating the Maze of Guidelines to Unify Visualization Design Recommendations.” We thank Cindy Bearfield Xiong for participating in early stages of this work, and thank all seminar participants and crowdsourcing study participants. KM was supported by a seed grant from the Stanford Accelerator for Learning and National Science Foundation (NSF) award #2400471, KBS and ALC by NSF award #2419493, and DW by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)–Project-ID 251654672–TRR 161.

REFERENCES

- [1] B. Bach, A. Abdul Rahman, and A. Diehl. VisGuides: IEEE Workshop on Visualization Guidelines in Research, Design, and Education, 2022. https://visguides-workshop.github.io//2022_index.html, last visited: 2026-06-29. 2
- [2] B. B. Bederson and A. Boltman. Does animation help users build mental maps of spatial information? In *IEEE Symposium on Information Visualization*, pp. 28–35, 1999. doi: 10.1109/INFVIS.1999.801854 7
- [3] R. Borgo, N. Boukhelifa, K. Gaither, and M. Sedlmair. C4PGV: Creation, Curation, Critique and Conditioning of Principles and Guidelines in Visualization, 2018. <https://web.archive.org/web/20191208052319/http://c4pgv.swansea.ac.uk/index.html>, last visited: 2026-06-29. 2
- [4] M. Chen, G. Grinstein, C. R. Johnson, J. Kennedy, and M. Tory. Pathways for theoretical advances in visualization. *IEEE Computer Graphics and Applications*, 37(4):103–112, 2017. doi: 10.1109/MCG.2017.3271463 1, 2, 8
- [5] J. Choi, C. Oh, B. Suh, and N. W. Kim. Toward a unified framework for visualization design guidelines. In *ACM SIGCHI Conference on Human Factors in Computing Systems (Extended Abstracts)*, pp. 240:1–240:7, 5 2021. doi: 10.1145/3411763.3451702 3
- [6] W. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *American Statistical Association*, 79(387):531–554, 1984. doi: 10.2307/2288400 2
- [7] R. Davis, X. Pu, Y. Ding, B. D. Hall, K. Bonilla, M. Feng, M. Kay, and L. Harrison. The risks of ranking: Revisiting graphical perception to model individual differences in visualization performance. *IEEE Transactions on Visualization and Computer Graphics*, 30(3):1756–1771, 3 2024. doi: 10.1109/TVCG.2022.3226463 2
- [8] J. R. De Leeuw. jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47:1–12, 2015. doi: 10.1201/b19402-15 5
- [9] J. R. de Leeuw, R. A. Gilbert, and B. Luchterhandt. jspsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, 8(85):5351, 2023. doi: 10.21105/joss.05351 5
- [10] A. Dearden and J. Finlay. Pattern languages in HCI: A critical review. *Human-Computer Interaction*, 21(1):49–102, 2006. doi: 10.1207/s15327051hci2101.3 4
- [11] A. Diehl, A. Abdul-Rahman, M. El-Assady, B. Bach, D. Keim, and M. Chen. VisGuides: A forum for discussing visualization guidelines. In *Eurographics Conference on Visualization (EuroVis), Short Papers*, pp. 43–47, 2018. doi: 10.2312/eurovisshort.20181076 2, 3
- [12] A. Diehl, M. Kraus, A. Abdul-Rahman, M. El-Assady, B. Bach, R. S. Laramée, D. Keim, and M. Chen. Studying visualization guidelines according to grounded theory, 2020. arXiv:2010.09040 [cs.HC]. doi: 10.48550/arXiv.2010.09040 2, 3
- [13] S. Djurcilov and A. Pang. Visualizing gridded datasets with large number of missing values. In *IEEE Visualization*, pp. 405–408, 1999. doi: 10.1109/VISUAL.1999.809916 7
- [14] P. Dourish. Implications for design. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 541–550, 2006. doi: 10.1145/1124772.1124855 3
- [15] T. Giallanza, D. Campbell, J. D. Cohen, and T. T. Rogers. An integrated model of semantics and control. *Psychological Review*, 2024. doi: 10.1037/rev0000485 5
- [16] S. Greenberg and B. Buxton. Usability evaluation considered harmful (some of the time). In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 111–120, 2008. doi: 10.1145/1357054.1357074 3
- [17] L. Harrison, A. Kale, M. Meyer, C. Nobre, and A. Satyanarayan. Cognition to context: A conversation about technical approaches, social values, and tradeoffs in visualization. *IEEE VIS Panels*, 2025. 2
- [18] M. N. Hebart, O. Contier, L. Teichmann, A. H. Rockter, C. Y. Zheng, A. Kidder, A. Coriveau, M. Vaziri-Pashkam, and C. I. Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580, 2023. doi: 10.7554/eLife.82580 5
- [19] P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. D. Stolper, M. Sedlmair, J. Chen, T. Möller, and J. Stasko. vispubdata.org: A metadata collection about IEEE Visualization (VIS) publications. *IEEE Transactions on Visualization and Computer Graphics*, 23(9):2199–2206, 2016. doi: 10.1109/TVCG.2016.2615308 6
- [20] K. G. Jamieson and R. D. Nowak. Low-dimensional embedding using adaptively selected ordinal data. In *Allerton Conference on Communication, Control, and Computing*, pp. 1077–1084. IEEE, 2011. doi: 10.1109/allerton.2011.6120287 5
- [21] N. W. Kim, G. Myers, J. Choi, Y. Cho, C. Oh, and Y.-S. Kim. Understanding the research-practice gap in visualization design guidelines. *IEEE Transactions on Visualization and Computer Graphics*, 32(3):2467–2482, 2026. doi: 10.1109/tvcg.2025.3640072 2
- [22] R. Kosara. An empire built on sand: Reexamining what we think we know about visualization. In *Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization (BELIV)*, pp. 162–168. ACM, 10 2016. doi: 10.1145/2993901.2993909 3
- [23] T. S. Kuhn. *The Structure of Scientific Revolutions*, vol. 2 of *International Encyclopedia of Unified Science: Foundations of the Unity of Science*. University of Chicago Press, Chicago, 2nd ed., 1970. 2
- [24] B. Lee, K. Isaacs, D. A. Szafir, G. E. Marai, C. Turkay, M. Tory, S. Carpendale, A. Endert, and T. Rhyne. Broadening intellectual diversity in visualization research papers. *IEEE Computer Graphics and Applications*, 39(4):78–85, 2019. doi: 10.1109/MCG.2019.2914844 3, 8
- [25] S. R. Marschner and R. Lobb. An evaluation of reconstruction filters for volume rendering. In *IEEE Visualization*, pp. 100–107, 1994. doi: 10.1109/VISUAL.1994.346331 7
- [26] M. Meyer and J. Dykes. Criteria for rigor in visualization design study. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):87–97, 2020. doi: 10.1109/TVCG.2019.2934539 3
- [27] M. Meyer, G. J. Quadri, and P. Rosen. Navigating the maze of guidelines to unify visualization design recommendations (Dagstuhl Seminar 25232). *Dagstuhl Reports*, 15(6):32–50, 2026. doi: 10.4230/DagRep.15.6.32 2, 3
- [28] D. Moritz, C. Wang, G. L. Nelson, H. Lin, A. M. Smith, B. Howe, and J. Heer. Formalizing visualization design knowledge as constraints: Actionable and extensible models in Draco. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):438–448, 1 2019. doi: 10.1109/TVCG.2018.2865240 3
- [29] K. Mukherjee, L. Lessard, M. Gleicher, T. Rogers, and K. Schloss. Color-concept associations reveal a new conceptual space. *Proceedings of the National Academy of Sciences*, 123(38):e2526862123, 2026. doi: 10.1073/pnas.2526862123 5
- [30] C. Packard. *Art Rules: How Great Artists Think, Create and Work*. Quarto Publishing Group UK, Minneapolis, 1st ed., 2023. 2
- [31] S. E. Palmer, K. B. Schloss, and J. Sammartino. Visual aesthetics and

- human preference. *Annual Review of Psychology*, 64(1):77–107, 2013. doi: 10.1146/annurev-psych-120710-100504 6
- [32] P. C. Parsons and P. C. Shukla. Considering the role of guidelines in visualization design practice, 2022. OSF. doi: 10.31219/osf.io/mw376 2, 8
- [33] S. Saharan, I. Al-Hazwani, M. Meyer, and L. Garrison. A critical reflection on the values and assumptions in data visualization. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 1146:1–1146:8, 2026. doi: 10.1145/3772318.3791501 2, 3
- [34] K. Sedig and P. Parsons. *Design of Visualizations for Human-Information Interaction: A Pattern-Based Framework*. Springer International Publishing AG, Cham, 1st ed., 2016. doi: 10.1007/978-3-031-02602-7 1, 2
- [35] S. Sievert, R. Nowak, and T. T. Rogers. Efficiently learning relative similarity embeddings with crowdsourcing. *Journal of Open Source Software*, 8(84), 2023. doi: 10.21105/joss.04517 5
- [36] J. T. Stasko and E. Zhang. Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *IEEE Symposium on Information Visualization*, pp. 57–65, 2000. doi: 10.1109/INFVIS.2000.885091 8
- [37] O. Tamuz, C. Liu, S. Belongie, O. Shamir, and A. T. Kalai. Adaptively learning the crowd kernel. In *International Conference on Machine Learning*, pp. 673–680, 2011. doi: 10.48550/arXiv.1105.1033 5
- [38] W. S. Torgerson. Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4):401–419, 1952. doi: 10.1007/BF02288916 5

S SUPPLEMENTARY MATERIAL

S.1 Terms and Definitions from the Initial Exploratory Study

Table S.1 provides the terms and their preliminary definitions, according to the exploratory study at the Dagstuhl Seminar (see Sec. 3.1). The table also contains the surveyed sources for each of the terms. Fig. S.1 shows the frequency of responses for the likely sources of each term described in Sec. 3.1.3 and encoded in the icons in Table 1 and Table S.1.

Table S.1: A list of terms and definitions from our initial understanding of the landscape of guidance-related terminology. Icons indicate likely sources of each term, determined using survey data from Sec. 3.1, Exploration of Guidance Terminology: empirical evidence; eminent leader; art & design community; emergent from unknown source. Icon darkness reflects prevalence in survey responses from most to least prevalent (● ● ● ● ●).

Principle	A fundamental primary or general law or truth from which <i>others</i> are derived	
Rule of Thumb	A rough method of procedure based on experience and common sense	
Consideration	Something to be kept in mind or to be given careful thought	
Common Practice	A typical way of doing something	
Convention	An established and commonly accepted way of doing something	
Recommendation	A suggestion or piece of advice about what someone should do, choose, or consider	
Best Practice	A widely accepted method or technique for reliably producing desirable results	
Standard	An established set of criteria intended to ensure desirable results	
Norm	A commonly accepted convention or practice	
Guardrail	A constraint that helps prevent undesirable results	
Paradigm	A prevailing framework or model that shapes how problems are approached and solutions are developed	
Style Guide	A set of rules or guidelines that help maintain consistency in the way something is presented	
Guideline	Suggestion or recommendation to help steer actions or decisions without being a strict rule	

S.2 Example trials the Mapping and Rating Phases of the Crowdsourced Study

Fig. S.2 depicts example trials for both the triadic similarity task of the mapping phase and the rating task for the rating phase of the crowdsourced study. Descriptions of task instructions can be found in Sec. 3.2.2 (triadic similarity judgment task) and Sec. 3.2.3 (rating task) of the main text.

S.3 Embeddings from the Mapping Phase of the Crowdsourced Study

Here, we provide more information regarding the the mapping phase of the crowdsourced study (see Sec. 3.2 of the main paper).

We include the guidance terms plotted as seen in Fig. 4 (see Fig. S.6 and Fig. S.7), and provide Figs. S.8 to S.11, where we plot guidance terms in each remaining pairwise combination of the four principal components (PC1: loose vs. strict, PC2: assimilated knowledge vs. direct evidence, PC3: specific vs. broad, PC4: theory vs. practical) alongside the variance explained and model loss used to select the embedding dimensionality.

For the underlying analysis, we fit embeddings at five dimensionalities (one, two, three, four, and five where five was the highest

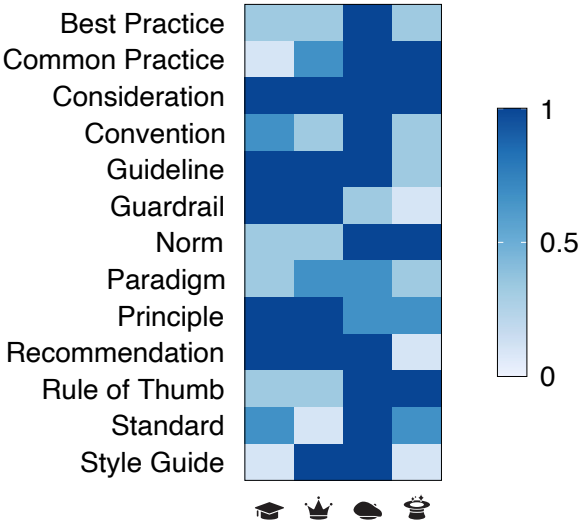


Figure S.1: Frequency of responses for the likely sources of each term from Sec. 3.1.3. The colors correspond to the icons encodings in Table 1 and Table S.1 (empirical evidence; eminent leader; art & design community; emergent from unknown source).

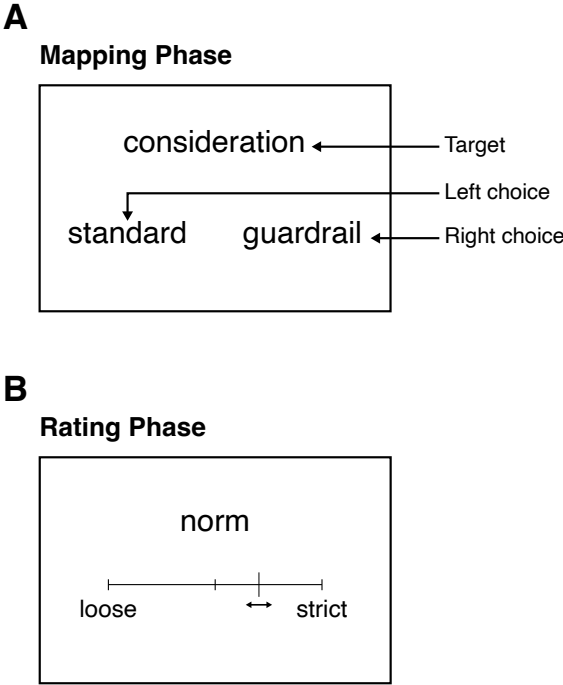


Figure S.2: Example (A) triadic similarity judgment trial and (B) rating trial for the crowdsourced study (Sec. 3.2.2 and Sec. 3.2.3).

dimensionality for which we could estimate a stable embedding) and computed how well each solution predicted held out similarity judgments. To measure each solution's performance, we first calculated the majority response for each triplet combination to serve as the human response. We then calculated the Euclidean distance for each embedding space between the target word and its two options for each trial. We defined the option with the smaller Euclidean distance from the target as the model's response for which option was "more similar." We then calculated the proportion of responses which aligned with the human majority response for each embedding solution. To establish a baseline for comparison with the variability of human responses, we calculated the proportion

agreement across participants with the majority response for a set of 10 held-out validation trials judged by all participants (see Fig. S.3).

We observed that all solutions performed similarly in predicting the held-out human judgments. When comparing the final loss of each model, we found that the 4-dimensional model had the best fit to the data (lowest final loss), and the addition of a 5th dimension did not improve performance (see Fig. S.4). To further assess the dimensionality of the guidance terminology space, we conducted principal component analysis (PCA) on the 5-dimension solution (see Fig. S.5). The PCA revealed the first four components accounted for 95% of the variance in the embedding, with the fifth component contributing relatively little additional variance. These results suggested that the underlying structure of the guidance terminology space was likely four-dimensional.

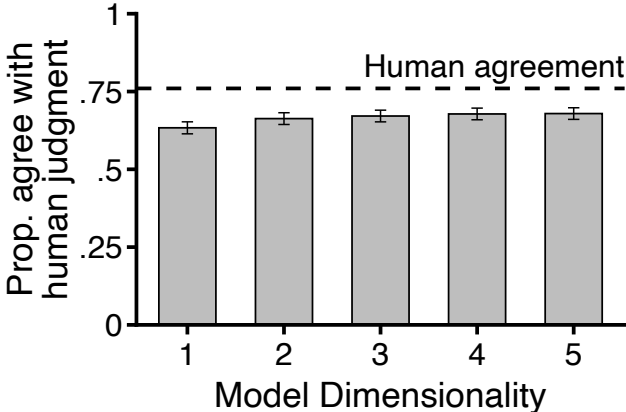


Figure S.3: Comparison of model agreement with the majority of human judgments across embedding dimensionalities. The dashed line indicates the proportion of human judgments that agreed with the majority response for a set of held-out validation trials consistent across participants.

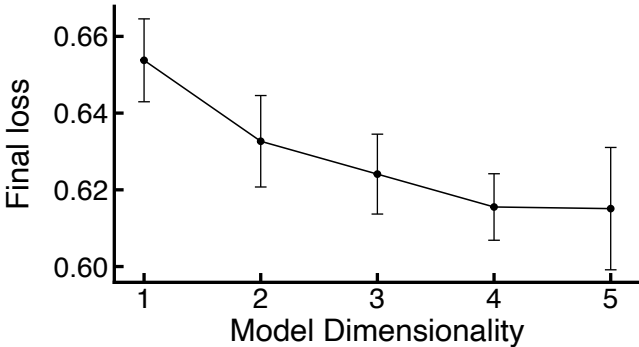


Figure S.4: Final model loss across embedding dimensionalities. Lower final loss indicates better model fit to the data. We estimated 10 embeddings for each dimensionality with stochastic sampling and computed 95% confidence intervals for the resulting average final loss scores.

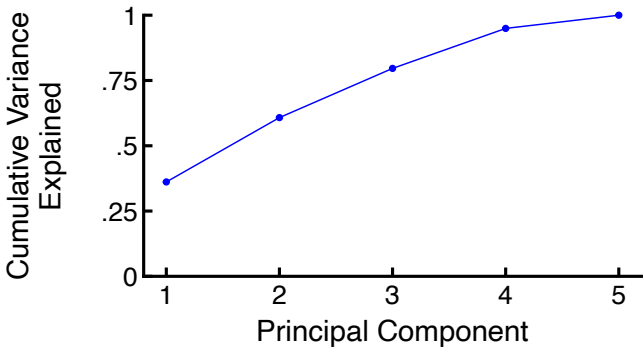


Figure S.5: Cumulative variance explained by each principal component when conducting PCA on a 5-dimension embedding solution.

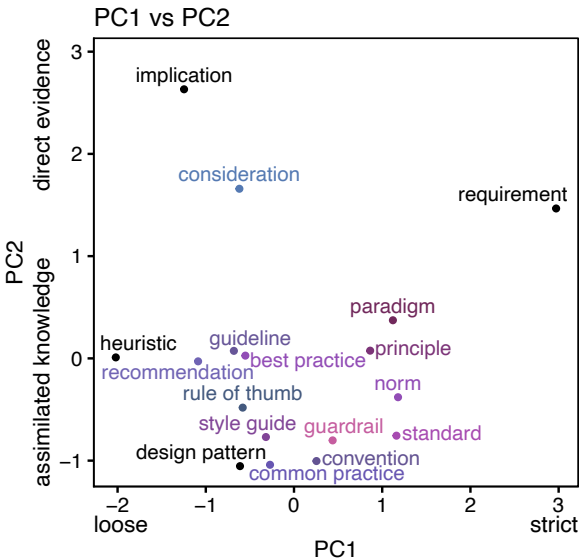


Figure S.6: PC1 (loose vs. strict) vs. PC2 (assimilated knowledge vs. direct evidence) as depicted in Fig. 4.

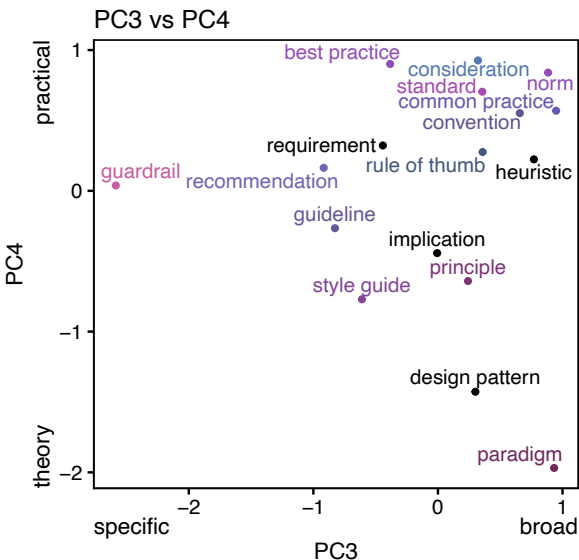


Figure S.7: PC3 (specific vs. broad) vs. PC4 (theory vs. practical) as depicted in Fig. 4.

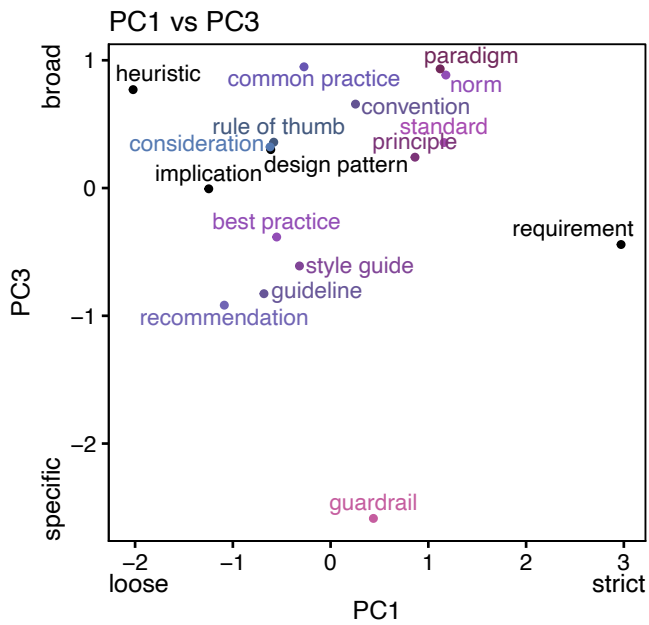


Figure S.8: PC1 (loose vs. strict) vs. PC3 (specific vs. broad).

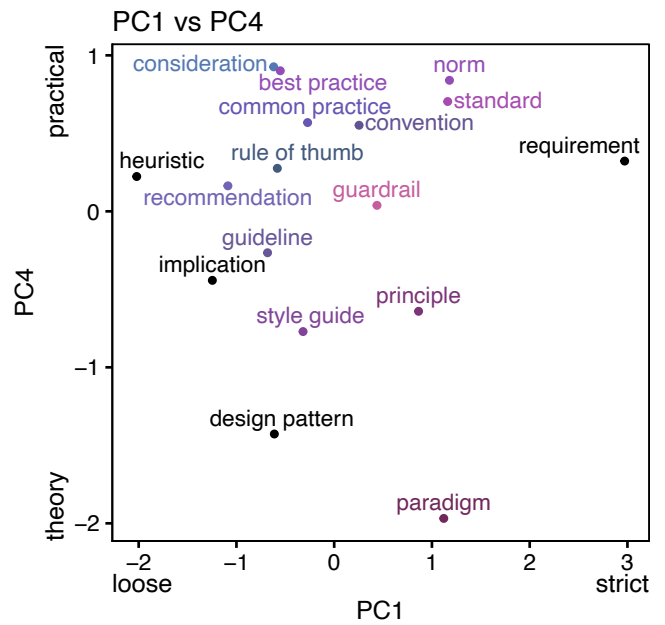


Figure S.9: PC1 (loose vs. strict) vs. PC4 (theory vs. practical).

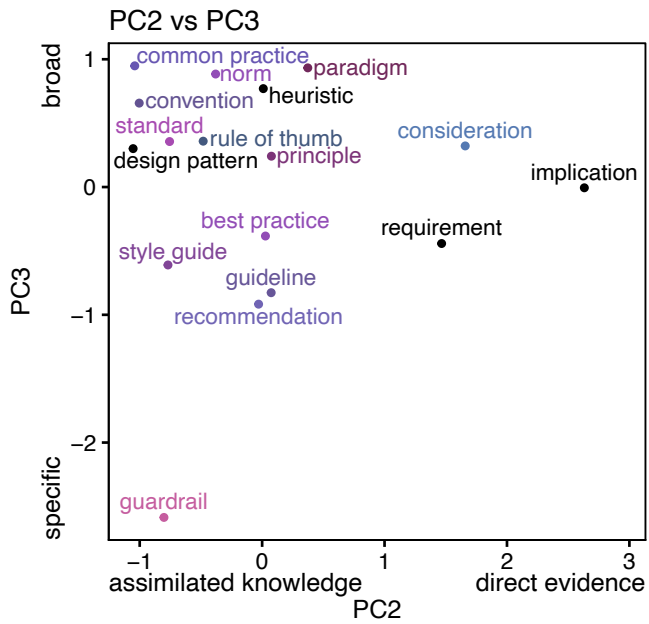


Figure S.10: PC2 (assimilated knowledge vs. direct evidence) vs. PC3 (specific vs. broad).

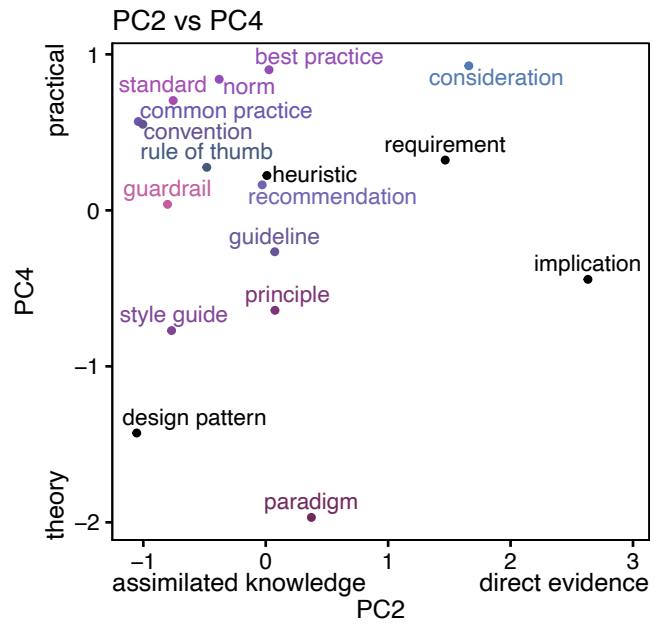


Figure S.11: PC2 (assimilated knowledge vs. direct evidence) vs. PC4 (theory vs. practical).