Validation of SplitVectors Encoding for Quantitative Visualization of Large-Magnitude-Range Vector Fields

Henan Zhao, *Student Member, IEEE*, Garnett W. Bryant, *Member, IEEE*, Wesley Griffin, *Student Member, IEEE*, Judith E. Terrill, *Member, IEEE*, and Jian Chen, *Member, IEEE*

Abstract—We designed and evaluated SplitVectors, a new vector field display approach to help scientists perform new discrimination tasks on large-magnitude-range scientific data shown in three-dimensional (3D) visualization environments. SplitVectors uses scientific notation to display vector magnitude, thus improving legibility. We present an empirical study comparing the SplitVectors approach with three other approaches - direct linear representation, logarithmic, and text display commonly used in scientific visualizations. Twenty participants performed three domain analysis tasks: reading numerical values (a discrimination task), finding the ratio between values (a discrimination task), and finding the larger of two vectors (a pattern detection task). Participants used both mono and stereo conditions. Our results suggest the following: (1) SplitVectors improve accuracy by about 10 times compared to linear mapping and by four times to logarithmic in discrimination tasks; (2) SplitVectors have no significant differences from the textual display approach, but reduce cluttering in the scene; (3) SplitVectors and textual display are less sensitive to data scale than linear and logarithmic approaches; (4) using logarithmic can be problematic as participants' confidence was as high as directly reading from the textual display, but their accuracy was poor; and (5) Stereoscopy improved performance, especially in more challenging discrimination tasks.

Index Terms—Vector field, scientific visualization in virtual environments, quantum physics, visual encoding, large-range data

1 Introduction

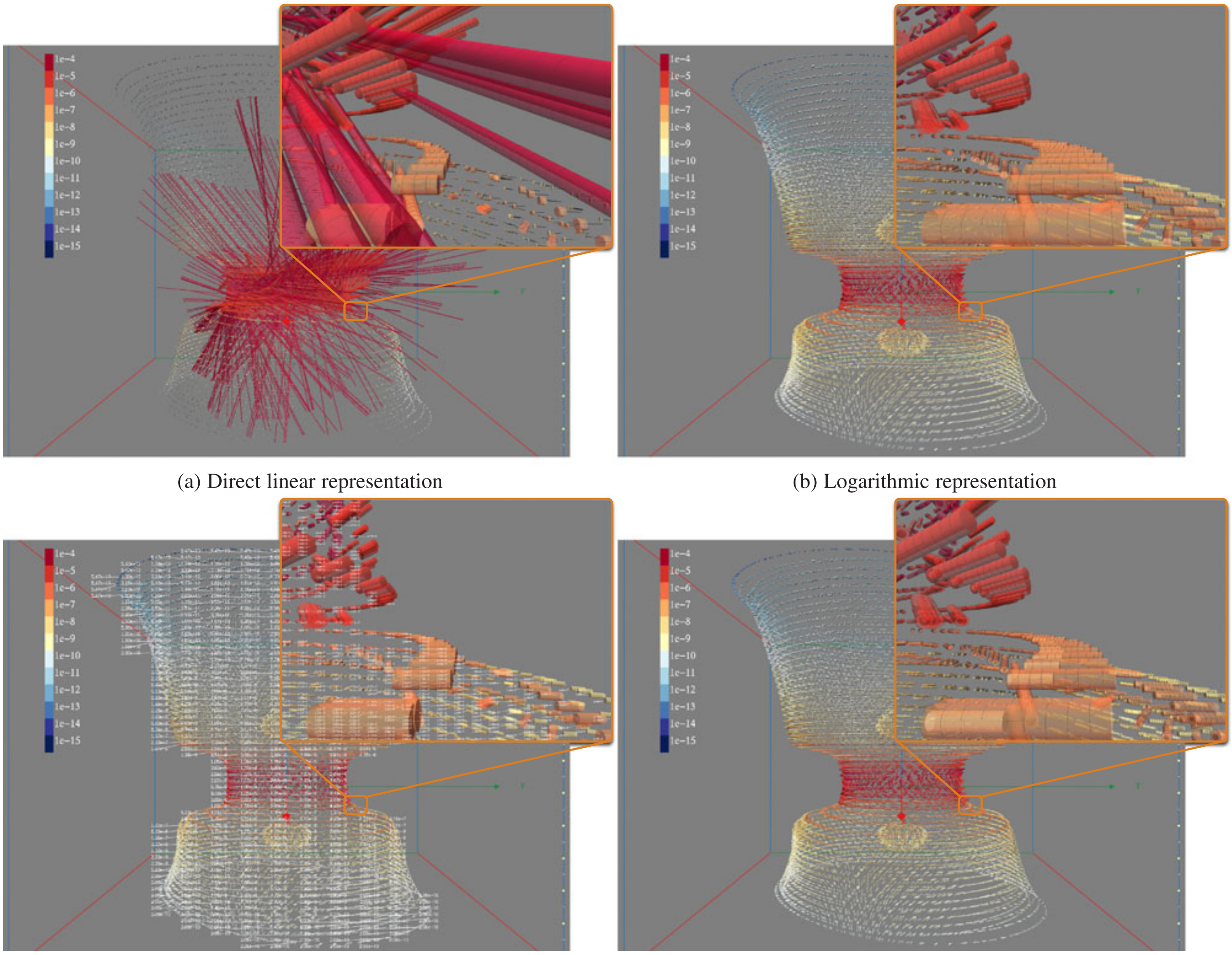
Quantum physics measurements and simulations produce data with large ranges: the order of magnitude differences between the largest and smallest numbers can be 10^9 or more [1]. Linear mappings of the numerical values to a vector in a visualization environment lead to small nonzero-data that are invisible and to large data too large to be seen (Fig. 1a). The information-rich virtual environment (IRVE) approach can successfully display numerical values as text attached to each vector location to produce more accurate reading [2]; however, for dense scientific data, displaying more than a few hundred labels can be prohibitive (Fig. 1c). The logarithmic approach is common in engineering and science. Logs, however, are not linear and require some mental calculation of the real values, which may produce a greater mental workload (Fig. 1b). Fig. 1 illustrates problems with these conventional approaches that motivate our new visualization (Fig. 1d), which remains robust for any magnitude range.

- H. Zhao, W. Griffin, and J. Chen are with the Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, Baltimore, MD 21250.
  E-mail: {henan1, jichen}@umbc.edu, wesley.griffin@nist.gov.
- G.W. Bryant, W. Griffin, and J.E. Terrill are with the National Institute of Standards and Technology, Gaithersburg, MD 20899.
  E-mail: {garnett.bryant, wesley.griffin, judith.terrill}@nist.gov.

Manuscript received 8 June 2015; revised 1 Mar. 2016; accepted 5 Mar. 2016. Date of publication 9 Mar. 2016; date of current version 3 May 2017.
Recommended for acceptance by H. Theisel.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2016.2539949

Data encoding needs to address analysis tasks. Scientists' comprehension of complex visualization requires more than just *qualitative* tasks of pattern detection: they also must perform *quantitative* discrimination tasks, e.g., reading particular numerical values and seeing how much one value differs from another [3] originally introduced by [4]. Traditional tasks in pattern detection are qualitative, requiring only two or three choices (larger, smaller or equal), and are, therefore, relatively easy to visualize. In analyzing quantum physics simulation results, physicists go through the following analysis stages: 1) develop an understanding of the large-scale global overview; 2) build a qualitative understanding of pattern distributions; 3) study the differences between or within datasets to understand extremes, ratios, and value distributions in certain regions; and 4) extract scientific insights about their data. Therefore, a visualization must support both *detection* (1 and 2) and *discrimination* (3) in order to support discovery tasks (4). This process is also similar to the process of overview-and-details commonly found in other applications [5].

Very little design and empirical work has been done on complex visualization discrimination and quantitative measure in large-magnitude-range data. In addition, the visualization researchers or practitioners currently have little guidance on which encodings are more appropriate for which tasks and on whether encoding methods can adapt to hardware display features. Though there is extensive research on textual display in IRVEs [6], [7], [8] and glyph designs for two-dimensional (2D) data comprehension [9], [10], [11], [12], the perception of data in three-dimensional (3D) space is different. For example, perspectives in 3D



(a) Direct linear representation

(b) Logarithmic representation

(c) Text labeling representation

(d) Our SplitVectors representation

Fig. 1. Four encoding approaches for the same large-range vector fields. In (a), smaller values are invisible and patterns are occluded; (b) shows patterns but introduces mental workload to learn the precise values of the visualized data; (c) uses textual display which causes occlusion; while our approach in (d) shows patterns without introducing occlusion.

change size perceptions [13] that may or may not be corrected [14]; also, occlusion influences encoding approaches and the choices of display characteristics [15]. Given this complexity, it is unclear whether or not visual encodings produced for 2D will be usable in 3D.

The goals of this research are to (1) explore new encoding approaches (Fig. 1) to reduce clutter and to support real-world discrimination tasks and (2) investigate whether a new vector field visualization can lead to more accurate feature analysis in complex physics simulation results. Since this is the first study to compare styled encoding effects, we focus on real-world uses in desktop mono and stereo conditions, so that our results will be directly applicable to quantum physics simulation results analysis tasks. We test both stereo and mono conditions because our collaborators believe that stereo will assist spatial data understanding and have remarked on how much stereoscopic effects can clarify structure.

More specifically, we consider: are there differences in accuracy and efficiency when novel encodings are used compared to the classical linear, logarithmic, and textual approaches? Can the encoding be combined with both stereo and mono display conditions, i.e., stereo and mono, to further improve users' task efficiency and effectiveness? Does the choice of encoding influence discrimination for increasing scientists' confidence in their judgments? Answering these questions will have implications for using encoding and cues in the visualizations of dense simulation results.

Our research contributes to the following:

- A new design that lets scientists perform change detection and discrimination tasks for more precise numerical data measurement,
- Statistical and anecdotal evidence, including time, accuracy, and confidences on the performance of encoding styles and stereoscopy that will be useful in designing future visualization for accurate measurement of complex spatial structures,
- Explanations of observed performance patterns and implications for design from task-dependent performance measurement, and
- A first look into the broader issues of display and encoding for showing explicit numerical values and task performance.

2 Related Work

This section presents research that influences our work in quantitative visualization in metrology or measurement science, visual content or encoding, and stereo-occlusion tradeoffs.

2.1 Quantitative Visualization in Metrology

Quantitative visualization is visualization that shows numbers applicable to complex data visualization in scientific domains, as first proposed by Trafton et al. [16]. Quantitative visualization is a central theme in scientific measurement or metrology. Because taking measurements is not always possible in physical experiments, quantification in simulated environments assists scientific validation by enabling many different conditions to be tested and examined [17].

Some tools incorporate interactive two-dimensional (2D) charts and use brushing and linking to link the charts to 3D counterparts that select features and therefore assist data selection [18]. Our purpose here has a similar theme of effective tool-building, but exploits the fundamental effectiveness of encoding approaches instead of the interaction method, following systematic design and validation approaches [19], [20].

2.2 Visual Content Representation: Perceiving Numerical Values

Discriminating numerical values is fundamentally different from change detection. Tasks in Forsberg et al. included pattern detection to compare large or small and marking particle movement by tracking streamlines [15]. Chen et al. studied quantitative comparison in tensor fields and found that the error rate in quantification remained high [8]. Despite research on vector field analysis, design and validation for discrimination tasks is rare.

The ability to show precise numerical simulation results has been one of the design goals of information-rich virtual environments. Polys et al. showed molecular simulation results that display text labels for association-occlusion tradeoffs [6]. Our work has a similar goal: enabling precise information presentation in 3D environments. Yet, existing approaches will not work in dense visualizations, simply because textual labels in our vector field of study are several orders of magnitude more dense than any of the scenes in previous studies. In the studies by Poly's et al. and Chen et al., about 20 labels are shown on the screen while ours introduces around a thousand.

2.3 Glyph Design

A recent review article and associated IEEE VIS 2014 tutorial provide fruitful resources for glyph-based visualization [11], [21]. Among the various glyph designs, one overarching goal is to display patterns or structures in vector fields common in flow and tensor field analytics. Only recently have researchers begun to consider quantitative representation beyond patterns. Pilar and Ware have provided the most thorough analysis of wind barb glyphs designed to convey precise wind speed and patterns [12]. Accuracy at intervals of five knots is possible. The glyphs coupled streamline width, curvature, and oriented arrows to compose a quantitative glyph with continuous streamlines (versus the classic straight tip lines that show patterns and quantitative information).

Ropinski et al. present a comprehensive survey of 3D glyph design in medical data visualization according to their pre-attentive attributes [22]. This taxonomy is useful to differentiate *basic shapes* (by changing geometric properties, e.g., size or orientation) and *composite glyphs* (by specialized mapping often for encoding multivariate attributes). Our design is necessarily related to composite glyphs, as our technique maps a vector magnitude into a two-channel composition such that large-range data can be properly displayed.

Our design idea, to use scientific notation to represent 3D vector fields to generate new glyphs, is mostly inspired by the 2D flow-glyph approaches [23]. A similar design idea of using scientific notation to represent large range data was recently studied by Borgo et al. [10]. However, there has been no direct transfer of empirical study results from 2D to 3D to validate similar visual effects. For example, length is considered one of the most accurate visual attributes for representing quantities in 2D. Length is, however, unstable in 3D, making estimation difficult, as evidenced by some of the distance measures in several solid empirical studies [24], [25], [26]. Built on top of existing knowledge, our goal in this work is to measure encoding effectiveness in 3D.

2.4 Stereo-Occlusion Tradeoffs

Scientific visualization in Virtual Environments (VEs) has studied display characteristics to answer questions such as how immersion, tracking, and 3D interaction help interpret data. Our work is different in that we explore the visual encoding to learn exactly what must be shown to enable more effective visual discovery in stereoscopic settings. Though it is believed that stereo can assist flow field visualization [15], [27], little is known about whether or not stereo affects encoding choices. Forsberg et al. compare glyphs and tube renderings in both stereo and mono conditions and found glyphs are better [15], but that study only focuses on direct vector field visualization. Novel encoding designs like ours have not been tested in stereoscopic conditions.

Ware and Franck compared rendering in desktop VEs and mono for graph visualization and found stereo and tracking improved performance threefold [27]. Ware's more recent work suggests that humans perceive 2.05D (a bit more than 2D, but not 2.5D or 3D yet) [28]. Using this theory, Alper et al. empirically validated 2.05D design and suggested 2.05D could be sufficient for understanding graph structure [29]. Built on top of this work, our study expands the understanding of new tasks in stereoscopic conditions.

Clutter management has been a fruitful approach to improving usability [30]. To understand legibility, Chen et al. studied pattern recognition in dense diffusion tensor MRI datasets and suggested that participants would not be able to see patterns when density was high, implicating the stereo-occlusion tradeoff [31]. That study expands consideration of the dense conditions. Our research further enlarges the task space to address new discrimination tasks in which users are asked not only which vectors at spatial locations were faster but also how much faster. This expansion may help broaden the task space in many scientific visualizations in stereo and desktop use.

3 Data and Task Abstractions in Quantum Physics

This section presents our first contribution: characterizing the *data* and *tasks* in the quantum physics simulations that drive the design and validation process for our new encoding design. We begin with a domain analysis and state the analytical tasks abstracted from the quantum physics simulations in order to improve reusability of our techniques, following user-centered design practices [32] and visualization design methodologies [20]. Our first step is to establish a task taxonomy for activities of scientific interest in metrology in physics; these are driven by uses in quantum physics simulations and are useful for us in systematically designing and validating our encoding methods and guiding our empirical studies.

3.1 Data Characteristics

Quantum physicists conduct simulations in order to understand behaviors of quantum dots, which are 10 nanometer (nm) scale crystals carrying a vector quantity called electron spin [33]. Spin is a property of electrons in atoms and is a vector with magnitude and direction. Scientists are interested in quantifing electron states and in how the electrons interact with a magnetic field or other electrons with spin.

In the dataset used in our study, local motion near each atom is computed by solving the quantum mechanical Schrödinger equation and is described using the local orbitals of each atom in the computational volume. This computational method provides atomic-scale information about the quantum states. A typical computational volume with the quantum dot and the material surrounding it may contain a million atoms. With this basis, the Schrödinger equation is converted from its differential form into a matrix eigenvalue equation. Iterative diagonalization techniques are used to find the eigenstates and energy eigenvalues of interest. Once the eigenstates are found, the spin content of each atom can be determined.

The computational volume used in this study is $X : [-25a, 25a], Y : [-25a, 25a]$, and $Z : [-10a, 17a]$. The distance between two adjacent z layers is $0.25a$ and in each layer the distances between two adjacent points along the x axis or y axis are $1a$ (here a is the lattice constance of the material.) Our visualization goal is to show these spins at the sampling mesh.

A major challenge in interpreting the spin attributes lies in the large ranges of spin magnitudes. The range in magnitude for the vector field used in this study is $[10^{-15}, 10^{-4}]$. For a sense of how different the values can be: if a $1\,\mathrm{cm}$ visual edge length represents the smallest magnitude (10^{-15}) then the largest vector magnitude (10^{-4}) must be the distance from the Earth to the moon.

Physicists are often interested in data in the range of the level of 10^4 smaller than the maximum to the maximum. And physicists still want to see those smaller vectors in order to validate their simulation results. The large-ranges of magnitude make perceiving the spatial spin distribution difficult. Spin orientation can also be difficult to perceive because the large-magnitude spin vectors can occlude the small-magnitude spin vectors. As a result, this paper focuses on showing magnitude to produce legible vector fields.

3.2 Analytical Workflow and Task Characteristics

Using task analysis [34] to ensure ecological or external validity of studies [35], [36], we obtain the following measurable low-level tasks:

1) *Change detection*: where do the spins change orientation? What are the patterns of changes, in one direction or in multiple directions?
2) *Comparison of numerical values*: is the magnitude greater at A or B? Is the orientation change from its neighbors larger at A or B? Where are the regions with the largest spin magnitudes?
3) *Locating extreams in a certain region*: where are the largest and smallest regions in the field? What is the maximum and minimum magnitude in the field?
4) *Boundary specification*: where does the boundary between regions of orientation change?
5) *Change discrimination*: how much change has occurred? How much larger is the value at point A than the entire field or another point in space?
6) *Length discrimination*: what is the magnitude at point A? What is the ratio between the vector magnitudes at points A and B?

Our tasks (1)-(3) are classical pattern detection and comparison tasks, while our tasks (5)-(6) are scientific-use-specific pattern discrimination tasks involving the estimation and comparison of quantitative values. Our task (4) is related to both discrimination and detection.

4 SplitVectors Design

Our second contribution is a way of showing large-range data that allows scientists to perform quantitative discrimination.

4.1 Design Aims

We arrived at the following set of encoding aims while designing encoding for large-range data:

- *Aim 1. Represent the entire data range.* The data are typically distributed over a large range with a large variance. Encoding and display choices should show all the data. This will allow viewing all of the data in a set of numbers regardless of the order of magnitude.
- *Aim 2. Pixel efficacy.* Data complexity may require a properly designed encoding in which pixels used for drawing must be employed economically and without redundancy.
- *Aim 3. Discrimination task effectiveness.* Physicists perform both quantitative discrimination and qualitative detection.
- *Aim 4. Ability to use only a few visual channels.* Many simulations produce heterogeneous data that may need more than one encoding (e.g., color, size, texture) to show various attributes. We thus attempted to use as few visual dimensions as possible in order to accommodate other variables.

Traditional approaches to presentation do not meet all the above requirements. For example, textual display in an IRVE may meet Aims 2 or 3 depending on the visualization but not Aim 1. A conventional linear approach supports Aims 3 and 4 but could be poor for Aims 1 and 2. Logarithmic display

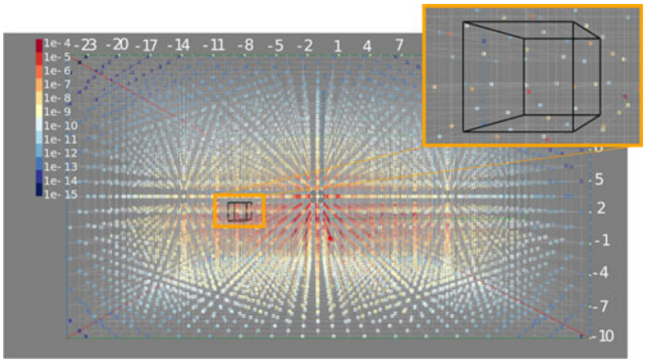


Fig. 2. The data distribution. Due to high data density, this figure marks one point out of every seven points. In other words, the actual data is seven times more dense along each of the three axes. We also color sampling points to show the magnitude distribution.

supports Aims 1 and 2, but may be poor on Aim 3 due to the loss of linear mapping.

4.2 SplitVectors Design Choices

Basic Concept (Figs. 3 and 4). SplitVectors is an encoding method for showing large-range data that satisfies all the design aims listed above. SplitVectors displays the spin magnitude using the two scientific notation terms: the digit term and the exponent term. For example, a magnitude of 1,200,000,000 can be written in scientific notation as 1.2×10^9, where 1.2 is the digit term and nine is the exponent. Instead of directly drawing a length 1,200,000,000, we represent a vector with the two values 1.2 and 9. Fig. 3 shows these two numerical values encoded as cylinder heights: the large outer cylinder represents the exponent and the inner thinner cylinder represents the digit term. Since we separate the visual representation of the digit and the exponent terms. We call our design SplitVectors. Both negative and positive exponents can be drawn.

Using scientific notation, the digit term is bounded to a real number in the range [1, 10) and the exponent is always an integer. Fig. 4 shows an example representation of our approach supporting the display of large-range data, where the data contains two cutting planes of data in the range $[10^{-14}, 10^{-4}]$. Both small-range and large-range data are visible when displayed with our SplitVectors approach. With the traditional linear approach, the large vector is displayed in a reasonable length but the smaller one is invisible.

Encoding choices (Fig. 5). There are many possibilities for displaying the digit and exponent terms, given the common visual dimensions of shape, hue, texture, brightness, and orientation. The design of line encoding initially followed the finding of Forsberg et al. [15] that line-arrow glyphs led to more accurate answers than tubes. Our first encoding is called line-arrow-sphere, where the lengths of the center line represent the digit and the exponent terms. To differentiate these two terms, an *arrow* is put at the end of the *digit* and a *sphere* at the end of the *exponent* (Fig. 5a). We quickly found two drawbacks of this line-arrow-sphere method. First, the lines had lower data-ink ratio [37] and tended to

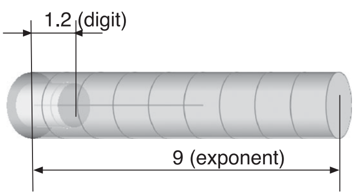
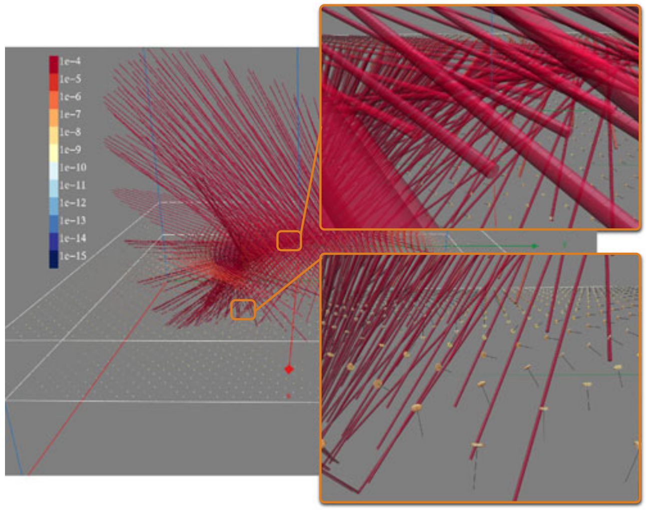
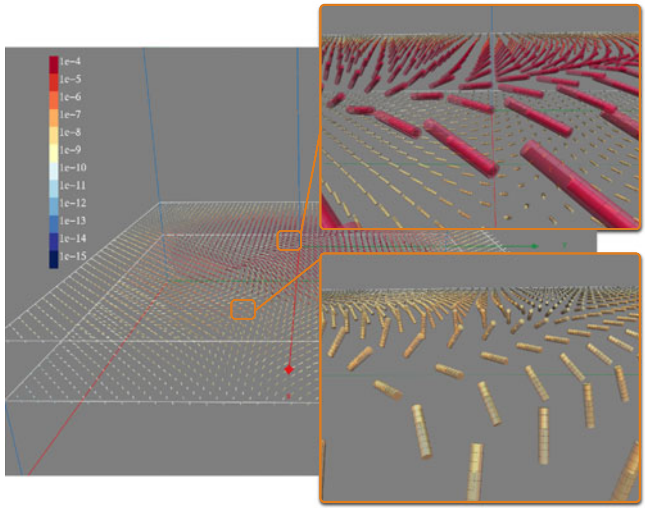


Fig. 3. SplitVectors design: A vector magnitude is represented in two terms by its scientific notation.



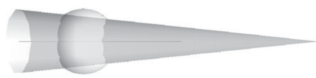
(a) Linear representation



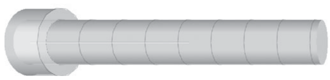
(b) Our SplitVectors representation

Fig. 4. Two cutting planes in 2z: (a) Linear representation (b) Our SplitVectors representation. Colors are mapped to different exponents of magnitude.

(a) Line-arrow-sphere



(b) arrow-sphere



(c) tube-tube

Fig. 5. Encoding choices.

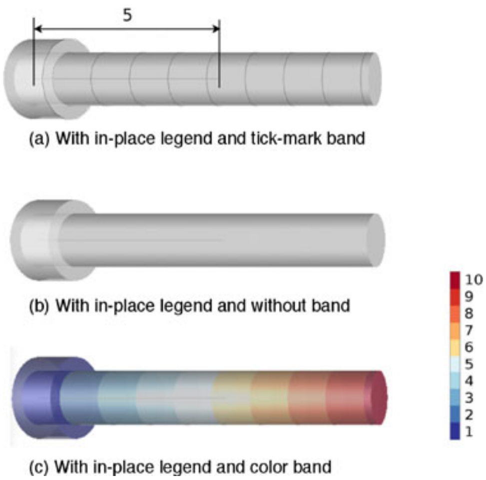


Fig. 6. Legend and tick mark band.

diminish, making the arrows and spheres stand out more and violating Aim 2. Second, our vector field is denser than that in Forsberg et al. [15] and lines could also be occluded by arrows and spheres.

Our second approach is to remove the lines and encode the entire arrow as a long, thin arrow (Fig. 5b). For this approach, however, human eyes are more sensitive to slopes and viewers might misread the edge of the cones as the vector size. Instead, following the results of Ware and Franck [27], we used tubes to increase the amount of useful ink (Aim 2) to represent magnitude for our third encoding: the tube-tube design (Fig. 5c). Here, the two terms in the SplitVectors are depicted using two embedded tubes; the height of the outside tube represents the exponent and the inside the digit term. A pilot study showed that tubes were better than the other two approaches for most tasks, and thus our experiment used this third solution of tube-tube drawing.

In-place legend (Fig. 6). Because we are interested in showing length and legend location affects perceived length estimates in 3D, we need to place a legend in the field that maps to magnitude. This also allows us to meet Aim 3 above so that users can perform discrimination tasks beyond pattern detection. It is also interesting to note that a length legend is rarely provided in 3D visualization tools (e.g., Paraview, a common visualization tool, lacks length legends [38]).

Our original design choice was to align the legend in 3D space with the center plane of the vector field. A pilot study suggested that this was difficult to use, because 3D perspective distorts length, thus making length comparison extremely challenging. We then designed an in-place legend, i.e., placing a legend line in the center line of each vector cylinder glyph, making its color subtly darker than the cylinder glyph. As Fig. 6 shows, the glyphs are transparent and the legends are visible. A legend always represents the 5-unit length. In this way, participants can tell the length more precisely and can also ignore the legend from afar if needed Participants in our pilot studies found this method intuitive and felt the legend did not interfere with the real data.

Color choices (Fig. 6). We also use color to double-encode length (Fig. 6c). The colors are chosen from the colorbrewer [39] of the common red-blue colors to differentiate the upper and lower half of the entire data range. Our collaborator liked color representation in the design process due to its clear support of visual grouping. An advantage of our encoding is that colors create *banding* effects so that each individual unit is clearer.

We did not, however, observe any significant effect of this color band in a pilot study looking at the benefits of the banding design, though participants and our collaborator liked this approach. In the empirical study, we therefore use a single gray color and draw some subtle lines, similar to the 2D tick marks around each tube or cylinder to indicate a unit length (Fig. 6a).

5 EVALUATION

We now discuss our experiment on determining quantitatively how effective this new type of visualization is and whether or not it is useful for clutter reduction when coupled with stereo.

5.1 Hypotheses

Our hypotheses were:

- H1. For large-range data in discrimination tasks, SplitVectors can lead to more accurate readings than linear and log approaches, especially for large-range values in discrimination tasks.
- H2. For small dynamic range data, SplitVectors and linear methods are likely to show similar accuracy, while log will lead to the least accurate answers due to its nonlinearity.
- H3. In both stereo and mono conditions, SplitVectors will lead to good task performance in task completion time and accuracy.
- H4. For discrimination tasks, users' confidence level will be greatest in SplitVectors and text.
- H5. For discrimination tasks, users' confidence level will be greater in stereo than mono.
- H6. For discrimination tasks, participants will prefer stereo to mono, especially for the highly cluttered linear approach.
- H7. For detection tasks, all visualizations will perform equally well regardless of encoding approaches.
- H8. For detection tasks, all visualizations will perform equally well regardless of stereoscopic conditions.

5.2 Independent and Dependent Variables

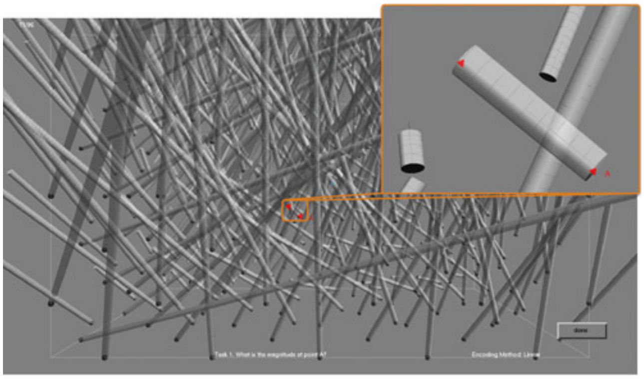
Our experiment used a $4 \times 2 \times 3$ within-subject design with three independent variables: visual encoding method (linear, log, text, and SplitVectors), display conditions (mono and stereo), and tasks (three types). Dependent variables include error and accuracy, task completion time, confidence, data-range sensitivity, and subjective ratings.

5.2.1 Tasks in Empirical Study

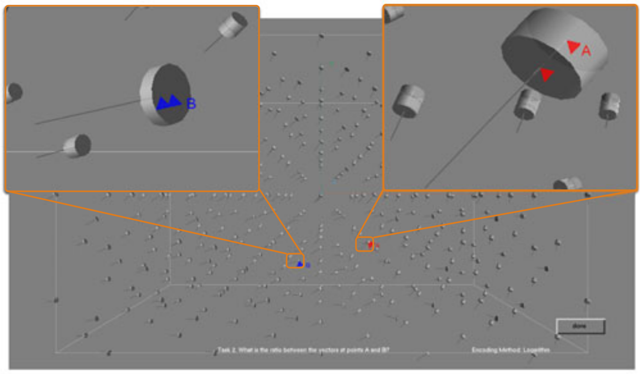
Three tasks are selected, the first two tasks are *discrimination* and the last is *detection* of magnitude only.

Task 1 (MAG, Fig. 7a): Magnitude reading at A. An example task is *What is the magnitude at point A?*

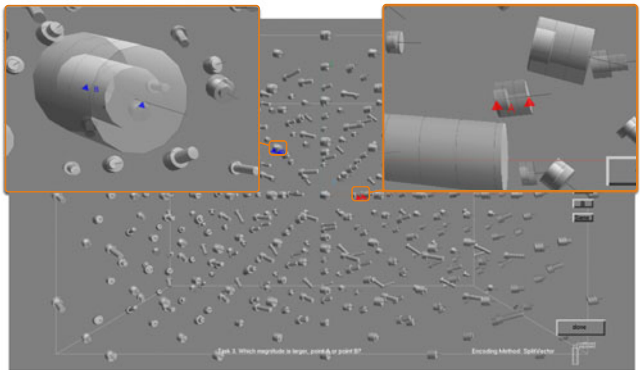
Task 2 (RATIO, Fig. 7b): Magnitude differences at A and B. An example task is *What is the ratio between the vectors at points A and B?*



(a) Task 1 (MAG): What is the magnitude at point A? (answer: 9.1)



(b) Task 2 (RATIO): What is the ratio between the vectors at points A and B? (answer: 3.6)



(c) Task 3 (COMP): Which magnitude is larger, point A or point B? (answer: B)

Fig. 7. The three tasks used in this study. The callouts show the task-relevant glyphs.

Task 3 (COMP, Fig. 7c): Which magnitude is greater? An example task is *Which magnitude is larger, point A or point B?*

The task taxonomy was presented in Section 3.2. We excluded those tasks that (1) might be irrelevant, e.g., "define boundaries" would need orientation information, or (2) might be not interesting or might take too long to finish, e.g., "detect extremes" where participants would need to search for the largest value. We also removed some tasks after a series of pilot studies due to study length (we wanted to limit the study to two hours to avoid fatigue.)

5.2.2 Measurement Metrics

In addition to task completion time and confidence, we did two further types of analysis to help understand the pros and cons of these four encoding approaches and their uses, relevant to measurement science at the National Institute of Standards and Technology (NIST).

Accuracy measure studies how sensitive a method is to error uncertainty based on the relative error (RE) or fractional uncertainty [40]. The RE is calculated as $RE = |CorrectAnswer - Participant\ Answer|/CorrectAnswer$. RE takes into account the value of the quantity being compared and thus provides a more accurate view of the errors. For example, judging numerical values off by 10 would be different when the baseline value is 20 than when it is 1,000: the first should lead to an error of 50 percent but the second to an error of only 2 percent.

The benefit of measuring REs is that it gives a more accurate interpretation of our design methods than correctness, since we can perturb REs when scientists need some uncertainty in answers. Participants' answers are considered "correct" if the relative error falls in a certain range of uncertainty. For example, when the correct answer is 10, an answer of 9 or 11 is considered correct when the allowable relative error is 10 percent.

Data-range sensitivity analysis computes the sensitivity of a visualization method to the data range. For example, we expect that reading 9.2×10^{10} will take no longer than reading 9.2×10^{1} using our SplitVectors approach; yet a traditional approach may need much longer time when reading the larger value. Analyzing the time differences in the dynamic range can help us understand how accuracy is affected by the scale of data.

This measure helps address our hypothesis H1, as we did not anticipate that SplitVectors would increase time for smaller ranges and would significantly improve task performance for large dynamic range data, since we are interested in methods that work in large range conditions. To validate our approach, we want to demonstrate that it would behave well with a more normal range as well. In other words, we hope that our approach will scale well with the data range.

5.3 Experimental Design

5.3.1 Trial Order

Table 1 shows all of the conditions the participants see during the experiment and the order of execution. The twenty participants in the study (See Section 5.3.2) were assigned to four blocks, with each block including five participants, as shown in the four rows in the *Participant* column. Each participant saw all encoding approaches and the four encodings were ordered in a Latin square design; the order is shown in the second column in balanced order of the four visual encoding approaches: SplitVectors, Linear, Log, and Textual Display.

Using each of the four encodings, each participant performed eight sub-tasks with 8 different datasets, as shown in the third column in Table 1. Each participant saw half mono and half stereo conditions of the eight listed conditions. The physicist suggested using the range of power four magnitude differences, where most regions of interest are represented. Here, numerical values of 0 to 3 mean that the dynamic range is the same (0) or differed by 10^1, 10^2 or 10^3. These eight conditions are randomly ordered. Therefore, each participant performs 32 subtasks under each task condition or 96 tasks for all three task conditions. We have carefully selected 96 datasets (see Section 5.3.3) for these

TABLE 1
Experiment Design: Each Participant Uses Four Encoding Techniques in Both Stereo and Mono Conditions of Data in Four Different Dynamic Ranges, with a Dynamic Range No More than 10^4

Participant	Encoding	Data (**stereo-data range** (randomly ordered)
P1, 5, 9, 13, 17	SplitVectors	on-0, off-0; on-1, off-1; on-2, off-2; on-3, off-3
	Linear	on-0, off-0; on-1, off-1; on-2, off-2; on-3, off-3
	Log	on-0, off-0; on-1, off-1; on-2, off-2; on-3, off-3
	Text	on-0, off-0; on-1, off-1; on-2, off-2; on-3, off-3
P2, 6, 10, 14, 18	Linear	$<same\ as\ above>$
	SplitVectors	
	Text	
	Log	
P3, 7, 11, 15, 19	Log	$<same\ as\ above>$
	Text	
	SplitVectors	
	Linear	
P4, 8, 12, 16, 20	Text	$<same\ as\ above>$
	Log	
	Linear	
	SplitVectors	

tasks so that no participant sees the same data twice to avoid learning effects.

5.3.2 Participants

Twenty participants (16 male and 4 female) of mean age 25.2 (standard deviation = 5.0) participated in the study. Their majors were: three in physics, five in mechanical engineering, one in math, three in chemistry, one in materials engineering, three in electrical engineering, and four in computer science. All participants have normal or corrected-to-normal vision. The four females were placed in each of the four blocks in Table 1.

5.3.3 Dataset Selection

We carefully selected the data to avoid introducing a confounding factor of dataset difficulty. We generated 1,000 samples of quantum-dot simulation results and then selected a subset of 96 from datasets for use in the study.

A random walk with box size of $5 \times 3 \times 3$ (black box in Fig. 2) was used to create the 1,000 random samples. There were 445 to 455 vector sampling locations in each selected data region.

The 96 datasets satisfy the following conditions. We require answers to be at locations where context information is available, i.e., they cannot be at the boundary of the selected region. We also disallow repetition to avoid learning effects: no data can be re-used for the same participant. Since data must include a broad measurement: for each task, we selected the magnitude from each exponential term exactly once in order to have a balanced design. We represented the exponent of the minimum data by 0, so that the range of the exponential terms was from 0 to 3. This lets us observe the effects related to hypothesis H1 where we want to measure the sensitivity of our technique to the data range.

For task 1 (*What is the magnitude at point A?*), point A is near the center of the data, meaning that its x, y, and z coordinates are in the range $[-1/3, 1/3]$ of the center of the bounding box in the selected region. There were four instances per visualization method for each of the two display conditions, either stereo or mono ($4 \times 4 \times 2 = 32$ trials). Of the four instances, one dynamic range for each magnitude from 10^0 to 10^4 is used.

For task 2 (*What is the ratio between the vectors at points A and B?*), points are near the center of the dataset volume, and there are four instances per visualization method ($4 \times 4 \times 2 = 32$ trials). Difference in speed between query points is in range $[1.1 \times Min_{magnitude}, 0.9 \times Max_{magnitude}]$.

For task three (*Which magnitude is larger, point A or point B?*), points near the center of the dataset volume are selected, so that enough contextual information is available. There are four instances per visualization method ($4 \times 4 \times 2 = 32$ trials). Differences in speed between query points are in range $Max_{magnitude} \times [0.2, 0.5]$, where $Max_{magnitude}$ is the maximum speed for that specific data volume.

5.3.4 Procedure and Other Factors

Participants first completed an Institutional Review Board (IRB) consent form and pre-questionnaire on their demographic information and background in quantum physics simulations and computer use. Participants were then trained on the four approaches and the stereoscopic uses. The formal study began, and they went through the 96 trials and interaction events were logged to a data file. Participants completed a post-questionnaire followed by a debriefing. At the beginning of each experiment and at breaks, participants were asked to confirm that they could see the stereo methods properly and were not seeing double images. The supplementary material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TVCG.2016.2539949, contains the training documents.

For Tasks 1 and 2, where the answers were numerical values and ratios, participants were asked to click a "done" button when they decided on an answer and then type in the answer and select the confidence level on the next screen. They were told they could not go back to see the data after clicking done to avoid confounding the study result by different typing speeds.

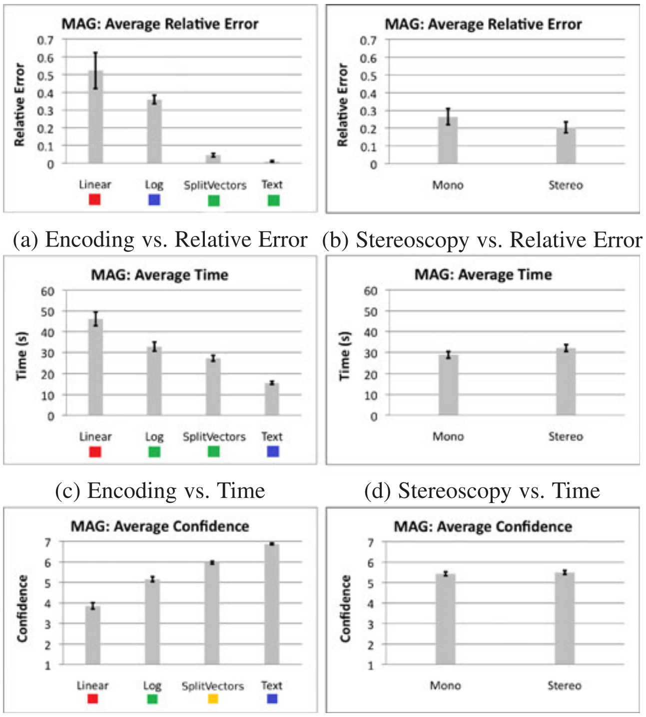
The room had natural light and the display was a BenQ GTG XL 2720Z 27" stereoscopic, gamma-corrected display (resolution $1{,}920 \times 1{,}080$). The stereo rendering was implemented using OpenGL quad-buffered stereo viewed with NVIDIA nVision2.

Participants sat in a standard desk chair. Since they were free to move their head closer or further away; we estimate that the viewing distance varied between 0.3 m and 0.6 m. Participants were required to wear 3D glasses during the study to ensure equal brightness in the mono and stereo conditions. Participants could rotate the data using the trackball interaction; zoom in-out was also supported.

The textual displays are drawn in screen space and used a label placement algorithm [41] to ensure that no labels overlap. An explicit line links each label with the corresponding 3D vector.

6 Results

In this section, we first show summary statistics and then give a statistical analysis by visual encoding and stereo conditions.

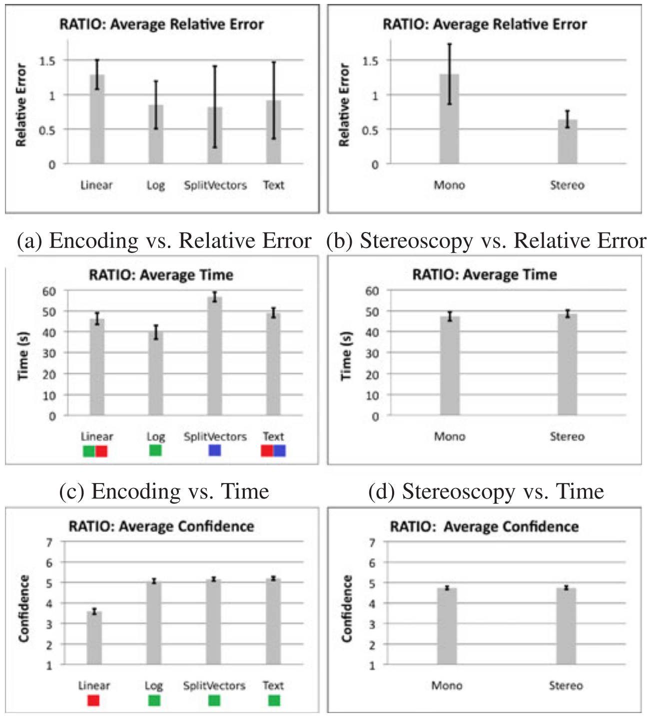


(a) Encoding vs. Relative Error (b) Stereoscopy vs. Relative Error

(c) Encoding vs. Time (d) Stereoscopy vs. Time

(e) Encoding vs. Confidence (f) Stereoscopy vs. Confidence

Fig. 8. Task 1 (MAG) results. Error bars represent standard error. Same colors represent the same Tukey group.



(a) Encoding vs. Relative Error (b) Stereoscopy vs. Relative Error

(c) Encoding vs. Time (d) Stereoscopy vs. Time

(e) Encoding vs. Confidence (f) Stereoscopy vs. Confidence

Fig. 9. Task 2 (RATIO) results. Error bars represent standard error. Same colors represent the same Tukey group.

6.1 Overview

We collected 1,920 data points (96 from each of the 20 participants). There are 640 data points from each of the three tasks. All hypotheses were supported except H5 and H7.

Our results clearly demonstrate the benefits of our new SplitVectors design in both mono and stereo conditions. SplitVectors reduced error by 10 times and used only 40 percent of overall task completion time for the magnitude reading tasks compared to the traditional linear approach (Fig. 8). The accuracy of using SplitVectors was about the same as text. Log was the worst: it led to more errors yet the participants' confidences were among the highest. The linear approach worked well when participants judged a single-digit value but the relative errors increased significantly when the data range increased, meaning the log technique was not scalable to data range in discrimination tasks (Fig. 9).

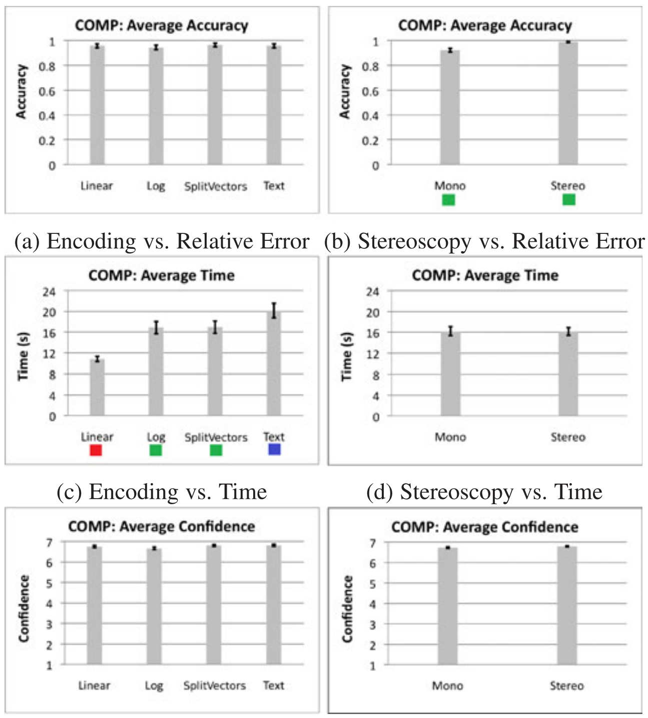
6.2 Analysis Approaches

We now discuss the details of the analysis including thresholds and significance. Table 2 shows F and p values computed with SAS's general linear model (GLM) procedure. Tukey pairwise comparisons among dependent variables (e.g., stereo, visualization, and ranking) are computed in post-hoc analysis. All error bars represent standard error.

6.3 Visual Encoding Versus Time, Relative Error, and Confidence

Relative errors (Figs. 8a, 9a and 10a). We observed a significant main effect of encoding on relative errors for the discrimination tasks of MAG only. In the MAG tasks, SplitVectors improved accuracy about tenfold over linear and sevenfold compared to log (Fig. 8a). SplitVectors and text were in the same accuracy group, as tested by Tukey post-hoc analysis.

In the RATIO tasks, linear led to the largest errors (almost double) of the four approaches while log, SplitVectors, and text had similar errors (Fig. 9a). The lack of



(a) Encoding vs. Relative Error (b) Stereoscopy vs. Relative Error

(c) Encoding vs. Time (d) Stereoscopy vs. Time

(e) Encoding vs. Confidence (f) Stereoscopy vs. Confidence

Fig. 10. Task 3 (COMP) results. Error bars represent standard error. Same colors represent the same Tukey group.

TABLE 2
Statistics for the Various Analyses by Tasks

Task	Variable	Significance
MAG	**Rel. Error (encoding)**	$F(3,639)$ **= 22.5,** $p < 0.0001$
	Mean time (encoding)	$F(3,639)$ **= 35,** $p < 0.0001$
	Confidence (encoding)	$F(3,639)$ **= 146.8,** $p < 0.0001$
	Rel. Error (stereo)	$F(1,639) = 1.2$, $p = 0.27$
	Mean time (stereo)	$F(1,639) = 2.02$, $p = 0.16$
	Confidence (stereo)	$F(3,639) = 0.25$, $p = 0.62$
RATIO	Rel. Error (encoding)	$F(3,639) = 0.23$, $p = 0.87$
	Mean time (encoding)	$F(3,639)$ **= 6.98,** $p < 0.0001$
	Confidence (encoding)	$F(3,639)$ **= 56.9,** $p < 0.0001$
	Rel. Error (stereo)	$F(1,639) = 2.1$, $p = 0.15$
	Mean time (stereo)	$F(1,639) = 0.6$, $p = 0.24$
	Confidence (stereo)	$F(3,639) = 0.006$, $p = 0.94$
COMP	Rel. Error (encoding)	$F(3,639) = 0.23$, $p = 0.88$
	Mean time (encoding)	$F(3,639)$ **= 12.2,** $p < 0.0001$
	Confidence (encoding)	$F(3,639) = 1.87$, $p = 0.13$
	Accuracy (stereo)	$F(1,639)$ **= 16.3,** $p < 0.0001$
	Mean time (stereo)	$F(1,639) = 0.002$, $p = 0.96$
	Confidence (stereo)	$F(3,639) = 1.36$, $p = 0.24$

significance may be due to the large variances in the data. As expected, all methods worked equally well for COMP, the detection task, leading to lower error rates using all approaches (Fig. 10a).

Completion time (Figs. 8c, 9c and 10c). For task completion time, the encoding approach was a significant mean effect in all three tasks (Table 2). In the MAG tasks, each of *linear* and *log* was in separate Tukey group compared to *SplitVectors* and *text* in terms of task completion time (Fig. 8c). In the RATIO tasks, *SplitVectors* and *text* led to longer reading time (Fig. 9c). For detecting "which is larger" in COMPs, *linear* required the least time due to its intuitiveness, *log* and *SplitVectors* were about the same, and *text* took the longest perhaps because it took participants some time to read numbers (Fig. 10c).

Confidence (Figs. 8e, 9e and 10e). Confidence levels were collected after participants finished each task on a scale of one (least confident) to seven (most confident). Confidence levels revealed that in MAG, participants had about the same confidence when using *log* and *SplitVectors*. *Text* had the highest confidence since data could be read directly; *linear* had the lowest. In RATIO, *log*, *SplitVectors*, and *text* had about the same confidence level, matching their performance measure in *relative error*.

It is interesting to note that the rank order and time-accuracy tradeoffs vary between detection and discrimination tasks. For the MAG discrimination tasks, linear performed the worst in completion time and accuracy (Figs. 8c and 8a); but for detection, linear was fastest but did not lead to greater accuracy (Figs. 10c and 9a).

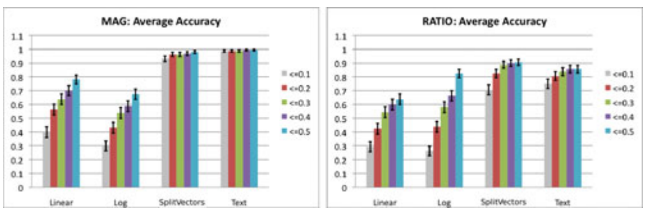


Fig. 11. Error sensitivity measure: uncertainty in relative error versus encoding. Error bars show standard error.

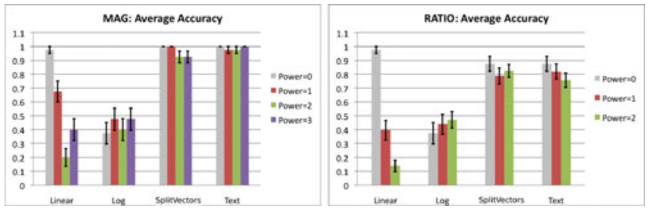


Fig. 12. Error sensitivity measure: exponent of data versus encoding. Error bars represent standard error.

6.4 Stereoscopy versus Time, Relative Error, and Confidence

The stereo effect was significant only in the detection task COMP of determining which one is larger (Table 2 and Figs. 8 bdf, 9 bdf and 10 bdf). Stereo also reduced relative errors on average by one, meaning doubled accuracy in the absolute values, for the relatively difficult RATIO task (Fig. 9b). Also note that the standard error in the mono condition is much larger than in the stereo condition, meaning that participant performance was more consistent in stereo in RATIO (Fig. 9b). Stereo and mono worked about equally well in the first "magnitude" task, with stereo marginally better than mono (Fig. 8b).

6.5 Sensitivity Analyses in Visual Encoding

We focus on the first two discrimination tasks, MAG and RATIO. We aggregate stereo and mono conditions, since we saw no significant main effect of stereo in accuracy measures.

Fig. 11 shows the results of our sensitivity analysis, varying the relative error uncertainty from 10 to 50 percent in increments of 10 percent. Log and linear are most sensitive to errors in both discrimination tasks.

Fig. 12 shows the sensitivity of these methods to the exponents in the data, since we would expect them to work about equally well for numerical values within a certain range but not equally well for large values. We can observe that linear worked very well for smaller values and then decayed quickly as the data values increased. Log was about equally poor and SplitVectors and Text were about equally insensitive to data input. This trend holds for both tasks.

6.6 Subjective Preferences and Comments

Preferences and comments were collected in the post-questionnaire and preferences were also rated from 1 to 7. We asked for preferences related to effectiveness and explained that effectiveness means how well a method helped accomplish a task. These results generally correlate with the performance data.

In the MAG tasks, participants preferred SplitVectors (mean score of 6) and text (6.45), compared to mean scores of 2.7 and 3.5 for linear and log approaches Fig. 13.

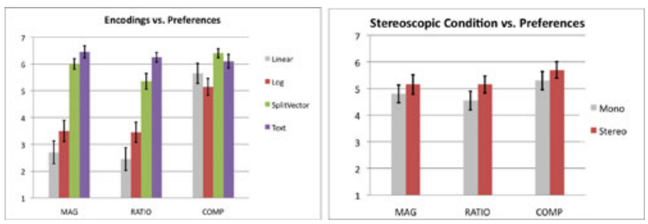


Fig. 13. Subjective preferences. Error bars show standard error.

Participants slightly preferred stereo (4.8 mono versus 5.15 stereo). Similar orders and levels were reported for the RATIO tasks (5.35 SplitVectors, 6.25 text, 2.45 linear, and 3.45 log; 4.55 mono and 5.15 stereo). For the final detection task, COMP, participant preferences were above five for all encodings and stereo displays.

Participants commented that for linear methods, they relied more on interaction and stereo helped more as well. Most participants still preferred SplitVectors, though they found linear more intuitive, commenting that if the vector chosen is lined up with the view or with others, the tasks became easier. Many took a stepwise approach in SplitVectors interpretation, though they had not been trained to do so: they first compared the power and then if needed looked into the inner cylinder for the digits term. Participants liked using log for detection since the scene was not very cluttered. Many commented that they disliked the text since it quickly cluttered the screen.

7 Discussion

This section discusses the design knowledge we can glean from our hypothesis testing and factors influencing our design and ability to address these factors. Most importantly, we also address reuse of our technique in other applications and tasks.

7.1 Visual Marks for Quantitative Visualization

H1. For large-range data in discrimination tasks, SplitVectors can lead to more accurate readings than linear and log approaches, especially for large-range values in discrimination tasks. [supported].

H2. For small dynamic range data, SplitVectors and linear methods are likely to show similar accuracy, while log will lead to the least accurate answers due to its nonlinear nature. [supported].

H7. For detection tasks, all encoding approaches will perform equally well. [not supported].

H1 is supported (Table 2 and Fig. 8a). We found that this SplitVectors design allowed about 10 times more precise numerical readings in about half of the task completion time compared to the traditional linear approach, and can achieve about the same precision as reading text.

H2 is also supported (Fig. 12) The same figures also show that SplitVectors is least sensitive to data range and is thus scalable to any range datasets.

H7 is not supported (Fig. 10c.) Linear has the advantage over all other approaches in task completion time. None of other factors were significant.

H1 and H2 show that our SplitVectors design is quite suitable to discrimination tasks. Results from all three hypotheses about the encoding, H1, H2, and H7, suggest that interpreting SplitVectors also needs some mental effort to get the correct answer. Compared to log, we should still use SplitVectors since the mental effort of interpreting log does not necessarily produce more correct answers.

Our technique is along the same lines as showing novel glyphs and our experimental validation shows that our technique worked well, at least for the discrimination tasks in large range data. Our research was inspired by recent developments in scientific visualization that have embraced novel encodings, such as "flow radar glyphs" [23], in which both uncertainty and flow directions are emphasized in *low-fidelity* glyphs, i.e., glyphs that were not direct plots of the original flow field. These results demonstrate that there might be value in considering presenting vector fields to address task scenarios rather than a direct visualization of the vectors.

Visual encoding methods can be studied more extensively in VE settings where more immersion is involved. Our qualitative evaluation in stereo conditions offers new insights that focus not only on high-fidelity length mapping but also on data-centric encoding to support new usage scenarios, in our case discrimination tasks in immersive metrology. We anticipate our technique will be useful in other fields such as biomedical data visualization for accurate diagnosis to immersive metrology in which tasks are not just about perceiving large patterns but also concern quantitative discrimination.

7.2 Stereoscopic Effects

H3. In both stereo and mono conditions, SplitVectors will lead to good task performance in task completion time and accuracy. [supported].

H6. For discrimination tasks, participants will prefer stereo over mono, especially for the highly cluttered linear approach. [Partially supported].

H8. For detection tasks, both mono and stereo will perform equally well. [Partially supported].

H3 is supported given that we saw no differences in mono and stereoscopic conditions as shown in Figs. 8d, 9d, and 10d. The aggregated analysis of error sensitivity provided evidences that our new approach worked equally well regardless of data range conditions (Figs. 11 and 12).

H6 is only partially supported: the stereo led to higher preference ratings but the effect was not significant. This matches the results on H5 in that preferences follow users' confidence about their answers.

H8 is also supported as shown in Figs. 10b, 10d, and 10f. The significant effect of stereoscopy was observed in the detection tasks for comparing which vector is larger (Fig. 10b).

These results on stereoscopic conditions need to be interpreted carefully on a task basis. A first and direct explanation for the lack of significance in MAG might be that that task did not need stereo as participants could see both points clearly. In other words, stereo and mono worked equally well. A second reason might be that the sample size was relatively small. The third reason is related to consistency in participants' results. We think these three reasons are applicable to different task conditions.

In RATIO, both the second and the third explanations apply, as we did observe that participants' errors were reduced by half in stereo and participant performance was more consistent in the stereo condition. The lack of significance arose mainly because participants were less consistent in the mono condition. Since RATIO requires more analytical steps than the two other tasks, we suggest that stereoscopy would be of more benefit in more difficult and cognitive demanding task conditions.

For the detection tasks of COMP, stereo improved accuracy. The strong effect of stereoscopy can be explained by the third reason in terms of consistency in participant performance. Both mono and stereo worked equally well. An interesting next step would be to compare these techniques with other display factors, such as those related to immersion and tracking.

7.3 Confidence

H4. For discrimination tasks, users' confidence level will be greatest in SplitVectors and text. [supported].

H5. For discrimination tasks, users' confidence level will be greater in stereo than mono. [Not supported].

H4 is supported as shown in Figs. 8e, 9e, and 10e. It is worth noting that log also led to higher confidence levels despite leading to higher error rate. This result suggests that designers should avoid log at all cost if possible.

H5 is not supported. Figs. 8f, 9f, and 10f show that all mono and stereo conditions led to about the same levels of confidence. The explanation could be that our tasks are quite complicated that the advantages of stereo become diminished.

Confidence can have at least two meanings. The confidence in our experiment was measured in terms of the fidelity of the participants' answers to the true answers, since we asked them to be as precise as possible. Another type of confidence would be confidence with a certain range of uncertainty. For example, in discrimination tasks of MAG and RATIO, a physicist might have high uncertainty about where the data would be in the range of 10 to 14, and thus might give a final answer of 12. Her or his uncertainty about the answer would be 100 percent in this case. We suspect this is why the study in Trafton et al. [16] suggested that scientists infer (precise) quantitative values based on (imprecise) qualitative information: partially because quantitative visualization was not available and also because scientists had an already established range of uncertainty.

The second type of uncertainty in confidence is interesting and yet is rarely considered in empirical studies. Such uncertainty is omnipresent in sciences and scientists often make judgments with this uncertainty in mind. This is true in medical sciences as well, since diagnostic accuracy reaches 80 percent, partially due to inaccurate human perception. This might be among the reasons that the data-driven sciences need more accurate interpretations of results needed.

7.4 Expanding Information-rich VEs for Knowledge Discovery in Sciences

IRVEs are used to integrate research in information visualization, scientific visualization, and VEs in order to help users visualize numbers and extract scientific insights. Interestingly, many information visualization papers lack a strong tie to application sciences, reporting instead completely new representations of datasets with certain characteristics. In scientific visualization, completely new representations are rare and efforts to improve existing ones in combination with application sciences are much more common [19]. IRVE research has been unable to use novel information visualization research results and is also rarely tied to application sciences. This study may suggest a new research agenda for IRVEs that stresses tight integration to applications to assist quantitative tasks to help scientists develop a more precise understanding of their models through the use of encodings.

One important issue is maximizing reuse of our technique in other domains. Reuse is possible partially due to overlapping task spaces in many application domains. This point is evidenced by the task formation approaches of task consolidation [42] so that low-level task study results are applicable to high-level cognitive tasks [43]. The other way is to consider specificity by considering *when*, *who*, and *what* [44], to add nuances to encoding design to improve affordances [45] suitable to application domains. For example, one might want to use arrow-sphere glyphs instead of cylinder-cylinder ones to better differentiate the digit and the exponent terms in the SplitVectors magnitude encoding. One can also emphasize the exponent term to show the magnitude range distribution.

7.5 Learning Effects

A measure that would be interesting to obtain from this study is the learning effect: how easily can one learn novel encoding approaches? There are two learning phases, the first occurs in the training, while the second occurs the empirical study itself to see how accuracy and efficacy improve over time. Our study could not measure the second effect due to the experimental design, but we did make some observations during training.

Participant training took from 40 to 50 minutes, including the time to fill out an entrance questionnaire and background survey and to take two short exams testing math to confirm understanding of logarithmic and SplitVectors design. The training was longer than most empirical studies in VEs but comparable to many measures in VIS field the authors have previously conducted. The overall time is still less than 2.5 hours in the study by Forsberg et al [15]. One reason for the longer training time was the many empirical conditions in this study and also that users needed to get familiar with the stereoscopic glasses. Another reason could be that participants needed to learn the tasks. Our goal was to measure experts' performance to ensure the validation of our empirical study. A desirable characteristic for visual marks would be that users can simply walk up and start using a visualization. The linear approach is apparently more direct and required the least training to interpret the results. For our collaborators, learning new techniques is never an issue and our study did demonstrate that, once familiar with the tasks, participants were able to accomplish them successfully. Most participants found the experiment quite enjoyable and instructive; some of them returned to run more studies with us.

7.6 When to Use Quantitative Measures?

The workflow in the physics domain of study includes both quantitative and qualitative tasks, and thus VE techniques fitting both scenarios will be most appropriate. Many new task scenarios and questions can be addressed with our new SplitVectors design, for example, whether simulation results differ in different experiments, the range of the differences, and the cause of those differences. Another use would be hypothesis testing, where the hypothesis is a tentative explanation that accounts for a set of explicit numerical discrimination and can be tested by further visual exploration. Our method can support scientific explanation with great accuracy.

7.7 Future Work

There are several ways to extend this research.

Encoding orientation. Our goal in this article is to design and validate a new glyph encoding to show large-range magnitudes. The logical next step is to study whether or not such a technique will be effective for tasks that require users to interpret both orientation and magnitude. Since physicists are interested in orientation similarity, a useful approach is to apply an orientation clustering algorithm to show orientation similarity and treat the orientation as categorical information. In that study, it will be important to consider how different glyph properties may interact with each other, i.e., by taking into account design principles relevant to integral and separable visual dimension and channel composition.

Tracking and immersion. Another research direction is to take into account kinetic cues and study the effects of tracking on the results. Immersion may also help data understanding through embodiment which can be further studied as well.

7.8 Environmental Factors

The study preparation involved a task parameter tuning phase whose objective was to select appropriate testing conditions. Through iterative testing we ultimately selected datasets and points for the tasks where applicable using the criteria stated in the previous sessions.

Dataset condition. We ran a series of pilot studies to determine parameters for the independent variables in visualization method, data complexity, and tasks, as well as the stereo display setup. In pilot studies we used both the real physical simulation data and data used in the Forsberg study, to balance two conditions of user expertise and regularity of the field; we found no differences and thus used the real data. Data density would have an impact on task performance, and generally more complex data would worsen the task performance. We chose to follow Forsberg et al. and use one density.

Interaction. A highly interactive system allows many usage patterns, more training, and the development of different strategies by users. These factors can be usefully studied only after we understand the effectiveness of new visualization design. Because our main interest here was visual performance on our new design, we minimize user interaction and suggest this direction as future work.

8 Conclusion

This paper describes new tasks in metrology and a new encoding design, as well as results in terms of design guidelines for understanding inferences in visual marks and stereoscopy. Our work extends the literature by designing visual encoding approaches that facilitate analytical tasks in VEs beyond just seeing patterns. This type of change discrimination is important in many other scientific visualization uses in VEs such as in-situ simulations. Our design recommendations are:

- Stereo is preferred and is more important for relatively complicated tasks or tasks requiring spatial understanding. For real-world applications where mixed task scenarios arise, stereo is also recommended.
- For discrimination tasks of reading large-range data, use SplitVectors if possible, especially for very large data. The linear method leads to large errors.
- Avoid using log for discrimination tasks, as it leads to large error rates coupled with high user confidence.
- For detection tasks, use linear if the task needs speed. Almost all measured techniques are equally accurate.

Acknowledgments

The authors would like to thank all participants and Katrina Avery for their contributions. We thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions. This work was supported in part by grants from NIST 70NANB13H181 and NSF IIS-1302755. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Certain commercial products are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the products identified are necessarily the best available for the purpose.

References

[1] R. D. Artuso and G. W. Bryant, "Quantum dot–quantum dot interactions mediated by a metal nanoparticle: Towards a fully quantum model," *Phys. Rev. B*, vol. 87, no. 12, p. 125423, 2013.

[2] D. A. Bowman, C. North, J. Chen, N. F. Polys, P. S. Pyla, and U. Yilmaz, "Information-rich virtual environments: Theory, tools, and research agenda," in *Proc. ACM Symp. Virtual Reality Softw. Technol.*, 2003, pp. 81–90.

[3] W. Müller and H. Schumann, "Visualization methods for time-dependent data-an overview," in *Proc. Winter Simul. Conf.*, 2003, vol. 1, pp. 737–745.

[4] A. M. MacEachren, *How Maps Work: Representation, Visualization, and Design*. New York, NY, USA: Guilford Press, 2004.

[5] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proc. IEEE Symp. Visual Languages*, 1996, pp. 336–343.

[6] N. F. Polys, D. A. Bowman, and C. North, "The role of depth and gestalt cues in information-rich virtual environments," *Int. J. Human-Comput. Stud.*, vol. 69, no. 1, pp. 30–51, 2011.

[7] T. Ni, D. A. Bowman, and J. Chen, "Increased display size and resolution improve task performance in information-rich virtual environments," in *Proc. Graph. Interface*, 2006, pp. 139–146.

[8] J. Chen, P. S. Pyla, and D. A. Bowman, "Testbed evaluation of navigation and text display techniques in an information-rich virtual environment," in *Proc. IEEE Virtual Reality*, 2004, pp. 181–289.

[9] J. Bertin, *Semiology of Graphics: Diagrams, Networks, Maps*. Madison, WI, USA: Univ. Wisconsin Press, 1967.

[10] R. Borgo, J. Dearden, and M. W. Jones, "Order of magnitude markers: An empirical study on large magnitude number detection," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 2261–2270, 2014.

[11] R. Borgo, J. Kehrer, D. H. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen, "Glyph-based visualization: Foundations, design guidelines, techniques and applications," *Eurographics State Art Rep.*, 2013, pp. 39–63.

[12] D. H. Pilar and C. Ware, "Representing flow patterns by using streamlines with glyphs," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 8, pp. 1331–1341, Aug. 2013.

[13] D. J. Bennett and W. Warren, "Size scaling: Retinal or environmental frame of reference?" *Perception Psychophysics*, vol. 64, no. 3, pp. 462–477, 2002.

[14] H. K. Distler, K. R. Gegenfurtner, H. Van Veen, and M. J. Hawken, "Velocity constancy in a virtual reality environment," *Perception*, vol. 29, no. 12, pp. 1423–1436, 2000.
[15] A. Forsberg, J. Chen, and D. H. Laidlaw, "Comparing 3D vector field visualization methods: A user study," *IEEE Trans. Vis. Comput. Graph.*, vol. 15, no. 6, pp. 1219–1226, Nov./Dec. 2009.
[16] J. Gregory Trafton, S. S. Kirschenbaum, T. L. Tsui, R. T. Miyamoto, J. A. Ballas, and P. D. Raymond, "Turning pictures into numbers: Extracting and generating information from complex visualizations," *Int. J. Human-Comput. Stud.*, vol. 53, no. 5, pp. 827–850, 2000.
[17] J. Terrill, W. George, T. Griffin, J. Hagedorn, J. Kelso, M. Olano, A. Peskin, S. Satterfield, J. Sims, J. Bullard, et al., "Extending measurement science to interactive visualization environments," *in Trends in Interactive Visualization*. New York, NY, USA: Springer, pp. 287–302, 2009.
[18] N. F. Polys, D. A. Bowman, C. North, R. Laubenbacher, and K. Duca, "Pathsim visualizer: An information-rich virtual environment framework for systems biology," in *Proc. Int. Conf. 3D Web Technol.*, 2004, pp. 7–14.
[19] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller, "A systematic review on the practice of evaluating visualization," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2818–2827, Dec. 2013.
[20] T. Munzner, "A nested model for visualization design and validation," *IEEE Trans. Vis. Comput. Graph.*, vol. 15, no. 6, pp. 921–928, Nov. 2009.
[21] R. Borgo, M. Chen, E. Maguire, and S. McDougall, "Glyph-based visualization," *IEEE VIS Tutorial*, 2014.
[22] T. Ropinski, S. Oeltze, and B. Preim, "Survey of glyph-based visualization techniques for spatial multivariate medical data," *Comput. Graph.*, vol. 35, no. 2, pp. 392–401, 2011.
[23] M. Hlawatsch, P. Leube, W. Nowak, and D. Weiskopf, "Flow radar Glyphs: Static visualization of unsteady flow with uncertainty," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 1949–1958, Dec. 2011.
[24] V. Interrante, B. Ries, and L. Anderson, "Distance perception in immersive virtual environments, revisited," in *Proc. IEEE Virtual Reality*, 2006, pp. 3–10.
[25] J. A. Jones, J. E. Swan II, G. Singh, E. Kolstad, and S. R. Ellis, "The effects of virtual reality, augmented reality, and motion parallax on egocentric depth perception," in *Proc. 5th Symp. Appl. Perception Graph. Vis.*, 2008, pp. 9–14.
[26] L. Phillips, B. Ries, V. Interrante, M. Kaeding, and L. Anderson, "Distance perception in NPR immersive virtual environments, revisited," in *Proc. 6th Symp. Appl. Perception Graph. Vis.*, 2009, pp. 11–14.
[27] C. Ware and G. Franck, "Evaluating stereo and motion cues for visualizing information nets in three dimensions," *ACM Trans. Graph.*, vol. 15, no. 2, pp. 121–140, 1996.
[28] C. Ware, "Designing with a 2 1/2-D attitude," *Inf. Des. J.*, vol. 10, no. 3, p. 258, 2001.
[29] B. Alper, T. Hollerer, J. Kuchera-Morin, and A. Forbes, "Stereoscopic highlighting: 2D graph visualization on stereo displays," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2325–2333, Dec. 2011.
[30] N. Elmqvist and P. Tsigas, "A taxonomy of 3D occlusion management for visualization," *IEEE Trans. Vis. Comput. Graph.*, vol. 14, no. 5, pp. 1095–1109, Sep./Oct. 2008.
[31] J. Chen, H. Cai, A. P. Auchus, and D. H. Laidlaw, "Effects of stereo and screen size on the legibility of three-dimensional streamtube visualization," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2130–2139, Dec. 2012.
[32] D. A. Bowman, E. Kruijff, J. J. LaViola Jr, and I. Poupyrev, *3D User Interfaces: Theory and Practice*. Reading, MA, USA: Addison-Wesley, 2004.
[33] J. S. Sims, W. L. George, T. J. Griffin, J. G. Hagedorn, M. Olano, A. P. Peskin, S. G. Satterfield, T. J. D., G. W. Bryant, and J. G. Diaz, "Accelerating scientific discovery through computation and visualization iii. tight-binding wave functions for quantum dots," *J. Res. Nat. Inst. Standards Technol.*, vol. 113, no. 3, pp. 223–245, 2008.
[34] M. B. Rosson and J. M. Carroll, *Usability Engineering: Scenario-Based Development of Human-Computer Interaction*. San Mateo, CA, USA: Morgan Kaufmann, 2002.
[35] N. Dell, V. Vaidyanathan, I. Medhi, E. Cutrell, and W. Thies, "Yours is better!: Participant response bias in HCI," in *Proc. ACM SIGCHI*, 2012, pp. 1321–1330.
[36] R. Rosenthal and R. L. e. Rosnow, Eds., *Artifacts in Behavioral Research*. New York, NY, USA: Academic Press, vol. 121, 1969.
[37] E. R. Tufte and P. Graves-Morris, *The Visual Display of Quantitative Information*. Cheshire, CT, USA: Graphics Press, vol. 2, 1983.
[38] J. Ahrens, B. Geveci, and C. Law, "Paraview: An end-user tool for large-data visualization," *Vis. Handbook*, vol. 836, p. 717, 2005.
[39] M. Harrower and C. A. Brewer, "Colorbrewer. org: An online tool for selecting colour schemes for maps," *Cartographic J.*, vol. 40, no. 1, pp. 27–37, 2003.
[40] J. R. Taylor, *An Introduction to Error Analysis: the study of Uncertainties in Physical Measurements*. Herndon, VA, USA: Univ. Sci. Books, 1996.
[41] B. Bell, S. Feiner, and T. Höllerer, "View management for virtual and augmented reality," in *Proc. 14th Annu. ACM Symp. User Interface Softw. Technol.*, 2001, pp. 101–110.
[42] H.-J. Schulz, T. Nocke, M. Heitzler, and H. Schumann, "A design space of visualization tasks," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2366–2375, Dec. 2013.
[43] M. Tory and T. Möller, "Human factors in visualization research," *IEEE Trans. Vis. Comput. Graph.*, vol. 10, no. 1, pp. 72–84, Jan./Feb. 2004.
[44] M. Brehmer and T. Munzner, "A multi-level typology of abstract visualization tasks," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2376–2385, Dec. 2013.
[45] R. Amar and J. Stasko, "A knowledge task-based framework for design and evaluation of information visualizations," in *Proc. IEEE Symp. Inf. Vis.*, 2004, pp. 143–150.

Henan Zhao received the BE degree in computer science and information security from Nankai University, China. She is currently working toward the PhD degree in computer science and electrical engineering, University of Maryland, Baltimore County. Her research interests include design and evaluation of visualization techniques. He is a student member of IEEE.

Garnett Bryant received the PhD degree from Indiana University, Bloomington, IN, USA, in theoretical condensed matter physics. After research positions at Washington State University, the National Bureau of Standards, McDonnell Research Labs and the Army Research Laboratory, he has worked at the National Institute of Standards and Technology (NIST) since 1994. He is now group leader of the Quantum Processes and Metrology Group at NIST with experimental and theoretical programs on nanoscale, condensed matter systems for quantum information science and metrology. His theoretical research program focuses on nanosystems, nanooptics, and quantum science. He is a fellow of the Joint Quantum Institute of NIST/University of Maryland, a fellow of the American Physical Society and a member of the IEEE.

Wesley Griffin received the MS degree in computer science from the University of Maryland, Baltimore County, in 2010. He is currently working toward the PhD degree in computer science from UMBC, Baltimore, MD, USA. He is a computer scientist in the High Performance Computing and Visualization Group of the Applied and Computational Mathematics Division of the Information Technology Laboratory at the National Institute of Standards and Technology (NIST). He is working in the Visualization, Animation, Nonphotorealistic Graphics, Object modeling, and Graphics Hardware (VANGOGH) Lab. His research interests include real-time graphics and graphics hardware. He is a member of ACM SIGGRAPH, student member of the IEEE and the IEEE computer society.

Judith E. Terrill is a computer scientist and the leader of the High Performance Computing and Visualization Group at the National Institute of Standards and Technology. She is a member of the IEEE Computer Society, the association for Computing machinery, and the association for the advancement of artificial intelligence.

Jian Chen received the PhD degree in computer science from Virginia Polytechnic Institute and State University (Virginia Tech), Blacksburg, VA, USA. She did her post-doctoral work in computer science and BioMed at Brown University, Providence, RI,USA. She is an assistant professor in computer science and electrical engineering at University of Maryland, Baltimore County where she directs the Interactive Visual Computing Laboratory. Her research interests include design and evaluation of visualization techniques, 3D interface, and visual analytics. She is a member of the IEEE and the IEEE Computer Society.