**Overview of changes**

**The paper has been rearranged to make the major points clearer for the reader, as suggested by the reviewers. The paper has also been shortened, and additional material shifted to the appendix.**

**Author’s response to reviews**

**Associate Editor  
Comments to the Author:**Two referees and an AE carefully read this paper. The referees make some interesting comments that should be addressed, e.g., if the order of the plots was randomized (seems so but this was not clearly explained). There were also some references potentially missed by the authors, e.g.,  
  
J. Hannig, T. C. M Lee and C. Park (2013), Metrics for SiZer Map Comparison, STAT, The ISI’s Journal for the Rapid Dissemination of Statistics Research, 2, 49-60.  
  
Marron, J. S., & Tsybakov, A. B. (1995). Visual error criteria for qualitative smoothing. Journal of the American Statistical Association, 90(430), 499-507.

**Thank you for the additional references. We have pointed to them at appropriate points in the paper. Of course, the order of plots is randomized, which is described in the experimental design where the original data was collected. This paper utilizes the data collected in the other experiments in order to assess possible metrics. In the paper re-organization we have explained this more clearly.**

**Reviewer(s)' Comments to Author:  
  
Reviewer: 1  
  
Comments to the Author**One of the great challenges of our time is "what does the human mind do?".  This paper takes on that issue in an interesting and at least to me novel way.  It is a compelling attempt at integrating statistical methods into this process.  It does feel like there is very much more that could be done, but this is a reasonable start.  
  
I like the paper, and have just a few rather minor comments.  
  
Some earlier thinking in this type of direction was done in:  
  
Marron, J. S., & Tsybakov, A. B. (1995). Visual error criteria for qualitative smoothing. Journal of the American Statistical Association, 90(430), 499-507.

**Author: Thank you for this. We have added this as a reference. The ideas are a bit different from our paper. That article is concerned that the typical error metric for regression provides a model fit using a smoother, that does not match what we expect based on eyeballing. The benefit of the current paper is that the lineup protocol provides a rigorous approach to assessing structure in plots, and actually could be used to validate the metrics described in Marron and Tsybakov. The examples that we have included are more general than regression. The approach could be extended much more broadly but we specifically wanted to take advantage of the large number of Turk experiments, and study the way that subjects read the specific set of plots.**

Specific comments (page, line):  
  
**(4,10)** I am not sure this provides strong evidence of a significant difference.  In particular, if this is tried with no true steepest curve (all 20 are from the null), it seems that you could end up with almost everybody choosing the next steepest, which I guess is number 6.  It would be interesting to test this as well.

**We can calculate the probability that 66 out of 70 independent observers picked the true plot, under the assumption that there is no pattern in the population, using the methods described in Majumder et al (2013). It is very small. Yes, it is possible that a null plot will have more structure than the data, as is the case with classical testing. People do tend to pick the plot with the strongest structure, regardless of whether it is a null plot. However, when all the plots are null, there is not a lot of difference from one plot to another, and multiple observers typically have some differences in their selections, which reflects this closeness, and is reflected in higher p-values.**

**In practice, with actual data, multiple lineups are made with different sets of null plots.**

**In follow up work by Hofmann and Follett, multiple choices or no choices by the observer are possible, which involves a small adjustment to the probability calculations.**

**Roy Chowdhury et. al (2015) tested structure in plots using some lineups with null plots only and some lineups with actual data, for projections of high-dimensional data, where there is not a lot of difference between actual and null data. The combined results of multiple human subjects showed that they could collectively identify the true data plots from purely null plots.**

**We’ve restructured the section to make all of this clearer.  
  
(5,-13**)    OK, the above issue is resolved here, by showing different observers different realizations of the null plots.  Consider re-ordering the discussion to make this critical point clear earlier on.

**We’ve restructured the section to make all of this clearer.  
  
(11,9)** Good to control the aspect ratio, but why regression of Y on X?  That is sensible when thinking of X as predicting Y here.  One could also use Y to predict X, but better here would be to use the projected residuals, i.e. to replace the regression line with the 1st PCA line.

**Yes, that could be. This example was constructed as part of a validation study done on 2010. We are simply re-analyzing the data, collected during the large Turk studies.   
  
(14,11)** At this point, it would be interesting to study how the distribution in Fig. 5b varies across the 40 realizations, that the subjects looked at.  Is the biggest black line actually bigger than the orange?  Same comment applies to other examples.

**Yes, this is of interest. It is what we have done, and summarized in the latter part of the paper. Fig 7 shows the relationships for boxplots of Fig 6a,b.**

**(20,-6)** Why not scatterplots, instead of side by side boxplots?  The latter only shows marginal distributions, but wouldn't the full joint distributions be even more interesting?

**This example was also constructed as part of a validation study done on 2010, and we are simply re-analyzing the Turk study data.   
  
(23,2)** Now I am confused.  Is each subject looking at different realizations, or are all subjects looking at these same plots?  This needs to be clarified everywhere.

**Each subject is looking at one lineup with one particular true plot. None of the human subjects evaluates two lineups with the same true plot. We have expanded the experiment description at the start of the paper. A detailed description of the data collection of Amazon Turk experiment is provided in the supplementary material of the validation study paper (Majumder et al, 2013).  
  
  
  
Reviewer: 2  
  
Comments to the Author**In this paper, the authors proposed several metrics to quantify structures in data plots and how human subjects read the structures from the plots. My overall impression is that this manuscript is well written. The motivation and application are clear. The proposed method could be potentially useful for exploratory data analysis. Here are some comments. **1.** The title of this paper is quite misleading. The authors developed several metrics between plots; while, I am not sure if those developments will directly help to “examine what people read from data plots”.

**The title of the paper has been updated to better match the main message of the paper.   
  
2.** Throughout the paper, the authors intended to “test” if human subjects are able to read the structure or pattern in the plots. The null hypothesis is not clear. The authors compared the visual inference versus the traditional inference. What is visual inference? For visual inference, what are the error and the power of the “test”?

**Buja et al (2009) and Majumder et al. (2013) introduce the concept of visual inference, describe the validation experiments, and how to calculate p-values and the power of the visual inference tests. An overview of these is given in the revised intro, but in depth explanations require the reader visit the original papers.**

**3**.      In this paper, are human subjects or observers independent? Will their responses form a random sample? Are there any other factors, e.g., personalities, that might influence the individual response/choice? This seems to be rather important since the topic of the paper is “What people read from data plots?”

**Each human subject evaluated the lineups independently through Amazon Mechanical Turk website with responses from many different countries of the world. Each evaluation of a lineup can be considered to be an independent sample. They may not be identical, because individual skills may vary, and this is incorporated in the original experimental analysis by using mixed effects models.**

**Zhao et al. (2013) looked at how people look at lineups using an eye-tracking device. It was noticed that people indeed look at and read the lineups differently. However, their responses and their accuracy in detecting the true data plot was not different, whether they eyeballed from left to right or up and down.**

**In the experiments, filter plots were used to determine that observers were making an effort to evaluate lineups diligently.**

**In the analyses used in this paper results are aggregated across observers, which mitigates individual differences.  
  
4.** In the example depicted in Figure 1, how many covariates in the full model? The human subjects were asked to identify the plot with the steepest slope. Is this equivalent to the problem of hypothesis testing, H0: beta\_k=0? As a minor comment, the authors only discussed the generation of null plots in the caption of figure 1.

**The explanation of the way the lineup was generated is expanded in the introduction now, and a reference to the full experimental details is provided. It is the equivalent to testing H0: beta\_k = 0. That was the purpose of the validation study that collected the data that we have re-analyzed to study metrics.**

**5.** In Figure 2, panel (b) is not clear. What is the test statistic here? The authors stated that “plots are used as test statistics in visual inference …”

**The test statistic in visual inference is a plot. Figure 2 attempts to show the conceptual difference between classical inference and visual inference. In classical inference, the test statistic is a real number but in visual inference, the test statistic is a plot. In Figure 2b we have a distribution of the null plots. But in a lineup of size 20 we can only observe 19 null plots to compare the test statistic (true plot) unlike in classical inference, where we have the whole distribution to compare the test statistic. The revison of the introduction hopefully makes this clearer.**

**6.** In all lineup plots, will the order of those subplots matter? What happen if those plots are randomly permuted? Will the human subjects choose the same answer? The distance metric distribution in Section 2.2 seems to be permutation invariant. Is that reasonable?

**Roy Chowdhury et al (2015) studied the effect of position of the data plot in the lineup. The order of plots in the lineup does not have a significant effect on the response of the observers. Hence the distance metric does not take the permutation into consideration.   
  
7.** The authors discussed three null generating methods. I think the choice of the methods depends on the purpose of the study, which is not clear from the current manuscript. Will bootstrap be an option? In the null generating process, how about model misspecification?

**Bootstrap method is typically not an option for graphics because all plots will look almost identical, because the same data values, perhaps with some overplotting, are plotted. Permutation is the most common null generating mechanism that we have used, in practice, because it breaks association between variables while keeping marginal distributions the same. Other null generating mechanisms using simulation from known distributions, and rotation residuals, are available in the R package (nullabor). In the experiments underlying this paper, the cluster separation study utilized permutation, and the two linear regression studies (scatterplots and boxplots) used simulation from the model with beta\_k=0 to generate nulls.**

**We are not sure what the reviewer means about model misspecification. This paper is not about null generating mechanisms specifically. It is possible that an experimenter might not use an appropriate null generating mechanism for their problem. The data that we have re-analyzed to examine metrics comes from very controlled simulation studies, with clearly defined null generating mechanisms.**