**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.  
  
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.**

**For each true data, 3 to 5 lineups are made with different sets of null plots. More than one human subject observes each lineup and picks up the plot that has the steepest slope. We can calculate the probability that 66 out of 70 independent observers picked the true plot (which is very small).**

**Also, Roy Chowdhury et. al (2015) tested lineups with null plots only and lineups with one actual plot with clusters. It can be observed from the results that human subjects can identify the true plots in the lineup when the difference between the null plots and true plot are insignificant.  
  
(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.**

**Each observer can judge a lineup with one particular true data. Once the true plot is revealed (as we reveal to the observer if their response was correct or incorrect), we believe that their response will be biased if they view another lineup with the same true plot.   
  
(11,9)    Good to control the apsect 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.**

**In this case, X and Y are two variables, which can be interchanged or replaced by other variables. Surely, we can use the residuals instead of the Y variable in certain situations. But since we use scatterplot as our test statistic, we believe that using conventional names may be helpful for the viewers to understand.  
  
(14,11)    At this point, it would be interesting to study how the distribution in Fig. 5b varies accross the 40 realizations, that the subjects looked at.  Is the biggest black line actually bigger than the orange?  Same comment applies to other examples.**

**This is done later in the paper. We calculate the difference between the mean distance for the true plot and the maximum of the mean distances for the null plots. This difference helps us understand if the lineups are easy or difficult. Larger the difference, easier it is to pick the actual plot. However, there are some exceptions showing that this does not match with people actually detected. We did not look at the overall distribution of these metrics.  
  
(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?**

**Since the covariate here is discrete with two levels, we think it is more appropriate to look at the marginal distributions using a side-by-side boxplot. A scatterplot may be an option, but we wanted to see if subjects can pick up the vertical difference between the side-by-side boxplots.  
  
(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. A detailed description of the data collection of Amazon Turk experiment is provided in the Appendix.  
  
  
  
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 more suit 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”?**

**Majumder et al. (2013) describes in details the concept of visual inference and power of the visual inference. Since this paper was referenced, we believe that this discussion is not needed in details. However more details have been provided in the introduction section of the paper.  
  
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. The responses were received from different parts of the world. Since the human subjects did not observe the true plot before evaluating the actual lineup, we can assume that their responses represent an independent random sample.**

**Zhao et al. (2013) looked at how people look at lineups using an eye-tracking device. It was noticed that people indeed look at the plots differently. However, their responses are not influenced by how they look at the plot. Also, to make sure that the human subjects are picking up the plots honestly, various examples were provided on the website. In addition, there was one test lineup (a very easy one) in a set of lineups and the responses were disregarded if the human subject did not understand the test plot correctly.  
  
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.**

**In the lineup, human subjects are asked to identify the plot which has the strongest signal. So we can ask the question “which plot is different?” But in that case, human subjects can pick any plot which has an outlier or a plot with a steep slope or any other features. Since we specifically wanted to test H0: beta\_k = 0, we asked the human subjects to identify the plots which has the steepest slope.**

**Here the true plot is a scatterplot between the dependent variable Y and the independent variable X\_k with a regression line overlaid. We want to test H0: beta\_k = 0. Assuming that the null hypothesis is true, the nulls are generated from N(X\*betahat, sigmahat^2) using the null model. Hence if the human subjects can identify the true plot from the lineup, we can conclude that the null hypothesis is rejected.  
  
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.**

**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?**

**The order of subplots does not have a significant effect on the response of the observers. Also, a number of replicates are used for each true plot and the lineups are evaluated by different human subjects. 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 not an option as we may have issues with overplotting resulting in the null plots having fewer observations in the plot compared to the true plot.**

**Not sure how to answer the question on model misspecification  
  
  
Associate Editor  
Comments to the Author:  
Two referees and and 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.**