A Crowdsourcing System for Integrated and Reproducible Evaluation in Scientific Visualization Appendix: Application Cases

Rickard Englund[1]* Sathish Kottravel[1]† Timo Ropinski[2]‡

[1] Interactive Visualization Group, Linköping University, Sweden
[2] Visual Computing Research Group, Ulm University, Germany

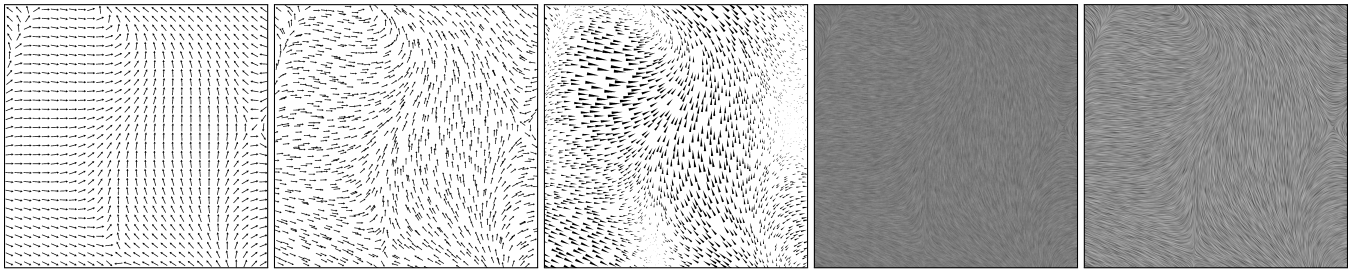


Figure 1: The five different techniques evaluated in the two dimensional Vector Field Visualization study. The techniques are, from left to right: GRID (Arrows on a regular grid), JIT (As GRID, but on a jittered grid), LIT (Triangles scaled according to velocity), LIC (Line-integral convolution) and Enhanced LIC

1 Appendix: Application Cases

In this appendix we present the three replicated studies [5, 6, 3] in detail that were conducted using the system described in the paper.

2 Study 1: two dimensional Vector Fields Visualization

This study is a replication of Laidlaw et al. study which compared six different techniques for visualization two dimensional vector fields [5]. In our study we have used a subset of the techniques and added one.

1. GRID : Arrows on a regular grid
2. JIT : As GRID, but on a jittered grid
3. LIT : Triangles scaled according to velocity [4]
4. LIC : Line-Integral Convolution [2]
5. Enhanced LIC: Enhanced LIC

Example renderings for each techniques are displayed in Figure 1. The enhanced LIC technique is implemented as a combination of two regular LIC, the output of the first LIC pass is filter using a Laplacian filter and then used as the noise picture to the second LIC pass. Two techniques are missing from the original study which are image-guided streamlines and streamlines seeded on a regular grid.

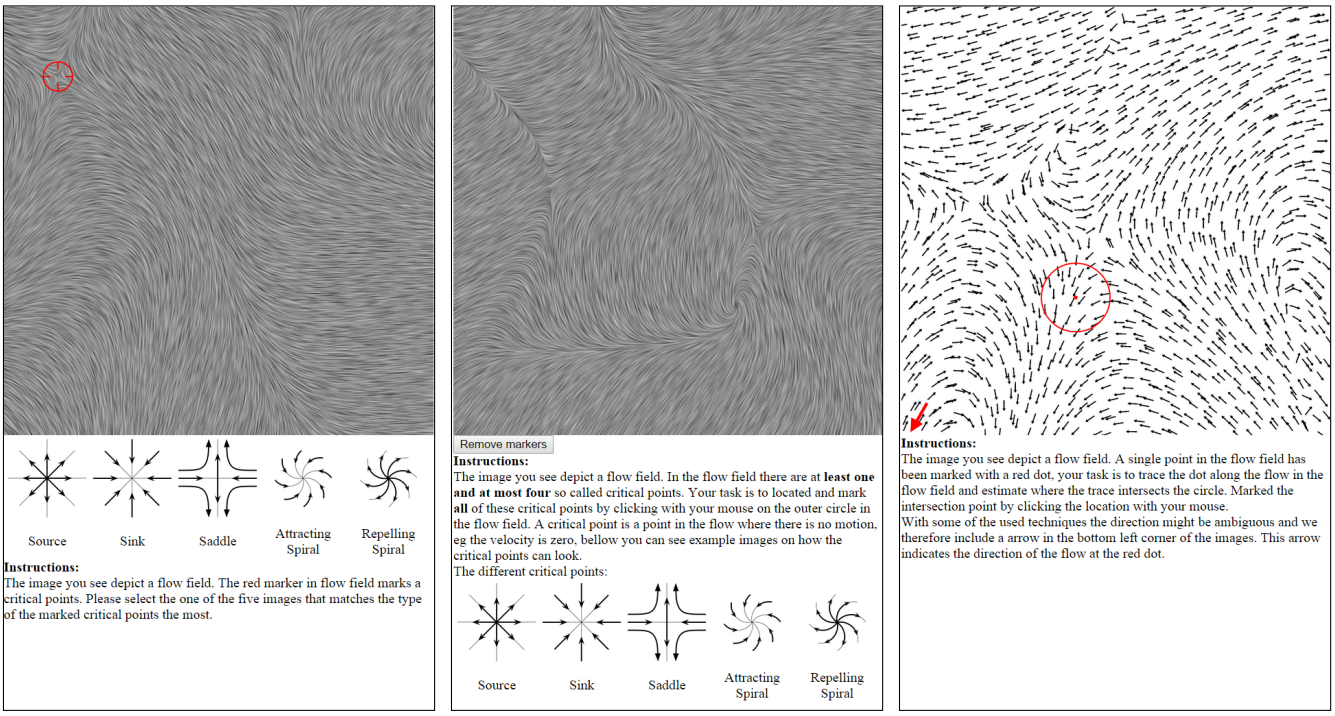
The study consists of three different task. The first task involves classifying a critical point marked with a red marker, the participants are shown images of the different types of critical points and are asked to click on the one that matches the marked point. In the second task the participant are asked to locate the critical points present in a vector field. When they find a critical point in the vector field they mark it by clicking on it with a mouse. In the final task evaluates the participants ability to trace a particle in the flow in the various techniques. This was done by marking a random point in the field and then draw a circle around it. The participants were asked to estimate where the trace would intersect the circle. For LIC and Enhanced LIC the direction of flow is not encoded in the visualization, to solve this ambiguity we draw an arrow in the bottom right part of the field to indicate the direction of flow under the starting position. Examples of all three task are shown in Figure 2.

The vector fields used in the study were created by generating a set of random two dimensional vectors at random locations in the range of $[0-1]$. A completed, continuous vector field was then approximated from these vectors using a Radial Basis Functions system [1]. Radial basis function systems is usually used for scalar values, so we use two sets of Radial basis function networks, one for the x-value of the velocities and one for the y-values. Further more, 3 additional vector field where generated using normal equations to be used in validation trials, these all had only one critical point located exactly in the middle, containing a sink, a source and a saddle respectively. Five trial list were generated with the trials organized in a counter-balanced order as described in the paper. At first we included 64 trials for each trial type for a total of 192 trials. To verify our study setup we first conducted a small prestudy consisting of 15 participants, this showed the 192 trials where far to many and people quickly lost focused. It also showed that some of the instructions needed to be made more clear, for example nearly no one had marked more than one critical point in the task to locate all critical points. Based on the results of the prestudy we decreased the amount of trials to 33 for each task type, for a total of 99 trials.

2.1 Results (auto-generated)

The study was conducted by recruiting a total of 73 participants. Each participant were payed $0.75 to complete a total 100 trials,

*e-mail: rickard.englund@liu.se
†e-mail:sathish.kottravel@liu.se
‡e-mail:timo.ropinski@uni-ulm.de

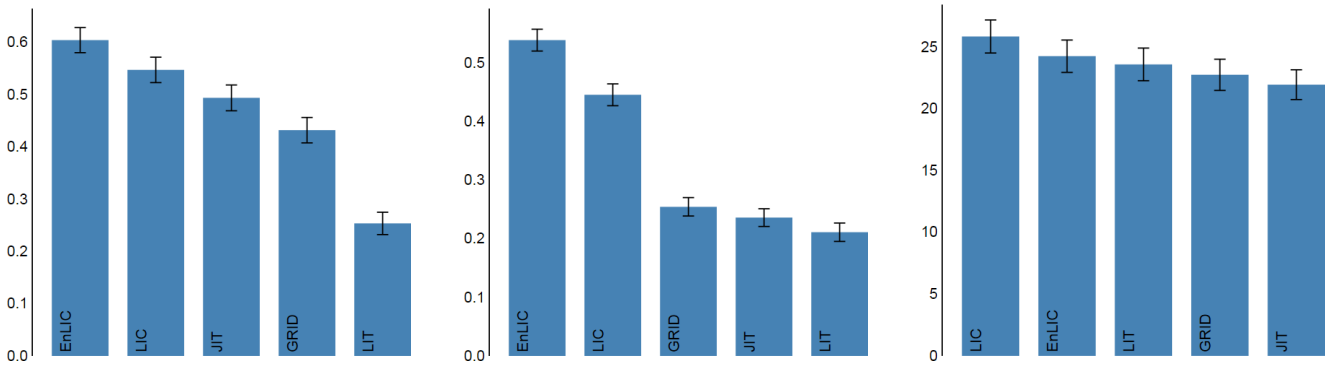


(a) An example trial using the **classify critical point** questionnaire design

(b) An example trial using the **locating critical points** questionnaire design

(c) An example trial using the **Tracing a particle** questionnaire design

Figure 2: Example trials using the designs of the three different tasks used in the two dimensional vector field visualization study.



(a) Results for the **Classifying critical points** task. Bars height represent correctness, higher is better.

(b) Results for the **Counting critical points** task. Bars height represent correctness, higher is better.

(c) Results for the **Tracing a particle** task. Bars height represent angular error, lower is better.

Figure 3: Bar charts showing the mean performance of users in the two dimensional vector field visualization study. Error bars shows one standard deviation.

total cost after CrowdFlower commission (~33%) was $81.0. Out of the 73 participants 4 were discarded due to failing validation trials.

2.1.1 User Performance

Part 1: Classifying critical points The Classify Critical Points part consisted of 3 validation trials and 30 regular trials. The rANOVA showed that there is a significant difference between the means ($F(4,1997) = 35.886, p < 0.001$). The Tukey HSD showed 100 groups which is summaries in 1.

Factor	Mean					
EnLIC	0.6043165	a				
LIC	0.5473934	a	b			
JIT	0.4938875		b	c		
GRID	0.4320388			c		
LIT	0.2536585					d

Table 1: Groups from the Tukey HSD analysis of the Flow Visualization Study : Classify Critical Points. Factors not sharing a letter are not significant difference.

Part 2: Locating critical points The Locating Critical Points part consisted of 3 validation trials and 30 regular trials. The rANOVA showed that there is a significant difference between the means ($F(4,1997) = 106.12, p < 0.001$). The Tukey HSD showed 99 groups which is summaries in 2.

Factor	Mean			
EnLIC	0.5389688	a		
LIC	0.4456951		b	
GRID	0.2544498			c
JIT	0.2361451			c
LIT	0.2111789			c

Table 2: Groups from the Tukey HSD analysis of the Flow Visualization Study : Locating Critical Points. Factors not sharing a letter are not significant difference.

Part 3: Tracing a particle The Tracing a particle part consisted of 3 validation trials and 30 regular trials. The rANOVA fails to find any significant differences between the means ($F(4,1997) = 1.788, p = 0.129$) so no further analysis is conducted.

Factor	Mean	
LIC	25.8485	a
EnLIC	24.25809	a
LIT	23.58869	a
GRID	22.75303	a
JIT	21.94942	a

Table 3: Groups from the Tukey HSD analysis of the Flow Visualization Study : Tracing a particle. Factors not sharing a letter are not significant difference.

2.2 Discussion

This study is divided into three parts. In the first part, classifying critical points, the rANOVA shows that there is a significant difference between the means. Enhanced LIC performed significantly better than all technique except against regular LIC. In the original study [5], Laidlaw's et al. fail to find any significant results, except between their best technique, GSTR and worst technique, LIC. This may be due to the fact that their participant pool was much smaller than ours, (17 vs 73).

In the second part, locating critical points, the rANOVA shows that there is a significant difference between the means. Enhanced LIC outperformed all other techniques followed by LIC. This is consistent with the results from Laidlaw's et al. original study [5] where LIC were better than GRID, JIT and LIT. It was expected that Enhanced LIC and LIC would perform well in this study since they are both dense visualization techniques, meaning they can depict the flow in every pixels, while the other techniques are sparse and values between glyphs has to be visually interpolated. In our study GRID and JIT does not perform significantly different, this is expected since they are quite similar, both uses arrows to show the direction of flow at a single point. Furthermore, LIT performed significantly worse than the other techniques, this is slightly different from Laidlaw's et al. results, where they found no significant difference between LIT, GRID and JIT. This might be due to the parameters used when rendering our images. It was not completely clear what parameters was used for the original study,which may lead to slightly different results. Regarding measured time there were no significant difference between the various techniques.

For the third part of the study, tracing a particle, the rANOVA fails to show any significant difference in user performance. In Laidlaw's et al. original study the analysis for this task shows two groupings, with OSTR and GSTR in one group performing significantly better than GRID, JIT, LIT and LIC. Since we only included GRID, JIT, LIT and LIC in our study our results agree with the results of Laidlaw's et al.

3 Study 2: Volume Illumination

In this subsection we describe the made findings when replicating a study on volume illumination models conducted by Lindemann and Ropinski [6]. The study investigates the impact on a user's depth perception of seven different volumetric illumination techniques. To determine this impact, questionnaires with tasks related to relative depth, absolute depth and beauty are used in the conducted study. We have replicated this study with our system, by generating the used questionnaires, and using them with the original study images as well as original marker positions.

3.1 Results

The study was conducted by recruiting a total of 51 participants. Each participant were payed $1.3 to complete a total 108 trials, total cost after CrowdFlower commission (~33%) was $93. Out of the 51 participant 13 was discarded due to failing validation trials.

Part 1 - Absolute Depth. The absolute depth part consisted of 1 validation trial and 39 regular trials. The rANOVA showed that there is a significant difference between the means ($F(6,1400) = 4.087, p < 0.001$). The Tukey HSD showed two overlapping groups with Directional Occlusion Shading (19.74%), Half Angle Slicing (20.91%), Shadow Volume Propagation (22.08%), Spherical Harmonic Lighting (23.33%) and Dynamic Ambient Occlusion (23.90%) in the first group. The second group consist of all techniques from group one, except Directional Occlusion Shading, plus Multidirectional Occlusion Shading (25.99%) and Phong Lighting (26.02%).

Part 2 - Ordinal Depth. The ordinal depth part consisted of 1 validation trial and 40 regular trials. The rANOVA showed that there is a significant difference between the means (F(6,1437)=3.715, p=0.001). The Tukey HSD showed two overlapping groups with Dynamic Ambient Occlusion (64.86%), Multidirectional Occlusion Shading (56.22%), Shadow Volume Propagation (55.68%), Half Angle Slicing (54.50%) and Directional Occlusion Shading (54.05%) in the first group and techniques except Dynamic Ambient Occlusion in the second group, with Phong

Lighting (46.85%) and Spherical Harmonic Lighting (45.50%).

Part 3 - Beauty Comparison The beauty comparison part consisted of 1 validation trial and 26 regular trials. All 21 possible combination of the illumination techniques were included in the list of trials. The preferred method was Phong Lighting (74.90%) followed by Half Angle Slicing (64.86%), Spherical Harmonic Lighting (61.54%), Shadow Volume Propagation (55.98%), Directional Occlusion Shading (46.92%), Dynamic Ambient Occlusion (33.22%) and finally Multidirectional Occlusion Shading (17.23%)

3.2 Discussion

This study consist of three parts, the first parts evaluates absolute depth perception. Both our study and the original study has significant strength according to the rANOVA, the distribution of the results are similar, with just a slight difference in the ordering. Their results showed that Half Angle Slicing and Shadow Volume Propagation where the top two techniques for absolute depth with 18.5% and 20.3% discrepancy, in our study they are on second and third place with 20.9% and 22.1% discrepancy and Directional Occlusion Shading performed best with 19.4% discrepancy. Similarly in the ordinal depth part we have similar results, our top three techniques was placed in their top four, after Directional Occlusion Shading. While our results is similar there are still some differences, this may be because of having too few validation trials, Lindemann and Ropinski's had one validation trials per task type which might not be enough when crowdsourcing is used for recruiting participants. For the beauty comparison study, the subjective preference on which method was the most beautiful we have very similar results. The only place were the results do not completely agree is place 3 an 4, were the order has been swapped.

4 Study 3: Multiclass Scatter Plots

In this subsection we describe the findings made when replicating a study on perception of averages in multiclass scatter plots, which has been presented by Gleicher et al. [3]. The authors analyze the perceptual averaging of the y-coordinate of multiclass scatter plots, whereby the classes are separated based on different visual attributes, such as size and shape. In their study they did two user evaluations. The first evaluation used a between-subjects design, this design did not provide enough evidence for all of their hypothesis so they constructed a follow up study with a blockwise within-subject design. We have replicated the second user evaluation study by reusing stimuli from the original study and recreated the questionnaire design.

4.1 Results

The study was conducted by recruiting a total of 159 participants. Each participant were payed $0.8 to complete a total 72 trials, the total cost after CrowdFlower commission (~33%) was $172.37. We have discarded 75 participants who failed on more than 50% of the validation trials. In this evaluation we investigated two different factors. The first factor is the absolute difference in pixels between the scatterplots averages, where the hypothesis is that the smaller the distance between the averages the harder the task becomes. According to the rANOVA ($F(5,5950) = 50.937, p < 0.001$) there is significant differences between the accuracies and the Tukey HSD shows 4 distinct groupings (Table 4). The grouping with best performance is the set of scatterplots with the largest difference between the averages, 80 pixels, with a accuracy of 83.4%. Followed by a grouping with the second largest difference, 36 pixels (0.74%). The third grouping, containing the scatter plots with 28 pixels (67.0%), 30 pixels (66.8%) and 12 pixels (63.3%) differences and lastly we have the final group with group containing the scatterplots with 4 pixels difference with a 54.6% accuracy. In Figure 5 the mean accuracy of each groups is visualized using bar charts. The second factor to investigate is the impact of the different encodings. This is done by running 9 different rANOVAs, one for each list of trials. Unfortunately, our analysis fails to find any significant difference in any of these 9 list of trials.

Absolute difference in pixels	Mean				
80	0.8343254	a			
36	0.7408143		b		
28	0.67			c	
20	0.6686508			c	
12	0.6329365			c	
4	0.5456349				d

Table 4: Groups from the Tukey HSD analysis of the impact of absolute difference in pixels in the multiclass scatter plots.

4.2 Discussion

The analysis of how the distance between the means affect the difficulty of the task shows similar finding as Gleicher et al., both with p-values very close to zero. In our study we had to discarded close to 50% of the participants due to the fact they failed 50% or more of the validation trials. Gleicher et al. [3] also had to discard some users, though not as many. After discarding participants, Gleicher et al. recruited new participants, which made their participant count remain high. We could do the same but at a monetary cost. Gleicher et al. used the Amazon's Mechanical Turk platform to recruit participants and did not pay for rejected users, we used the CrowdFlower platform, while they allow for rejecting participants, the cost of the rejected participants will still be charged and if new participants is to be recruited more money has to be spent. Furthermore, CrowdFlower uses various platforms on the web to distribute the tasks, while they assume all participant speaks English, it is a high probability the many participants of the participants does not have English as native language. CrowdFlower gives us a report on what country the participants who performed our study was located and only around 11% of the participants who completed the study were from a country were English is the native language. We believe that this might affect the ability to fully understand the instructions and therefore some participants may have completed the study incorrectly. We like to redo the study in the future, where we will limit the recruitment to include only countries where English is the native language.

References

[1] D. S. Broomhead and D. Lowe. Radial basis functions, multi-variable functional interpolation and adaptive networks. Technical report, DTIC Document, 1988.

[2] B. Cabral and L. C. Leedom. Imaging vector fields using line integral convolution. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, pages 263–270. ACM, 1993.

[3] M. Gleicher, M. Correll, C. Nothelfer, and S. Franconeri. Perception of average value in multiclass scatterplots. *IEEE TVCG*, 19(12):2316–2325, 2013.

[4] R. M. Kirby, H. Marmanis, and D. H. Laidlaw. Visualizing multivalued data from 2d incompressible flows using concepts from painting. In *Visualization'99. Proceedings*, pages 333–540. IEEE, 1999.

[5] D. H. Laidlaw, R. M. Kirby, C. D. Jackson, J. S. Davidson, T. S. Miller, M. Da Silva, W. H. Warren, and M. J. Tarr. Comparing 2d vector field visualization methods: A user study. *Visualization and Computer Graphics, IEEE Transactions on*, 11(1):59–70, 2005.

[6] F. Lindemann and T. Ropinski. About the influence of illumination models on image comprehension in direct volume rendering. *IEEE TVCG*, 17(12):1922–1931, 2011.

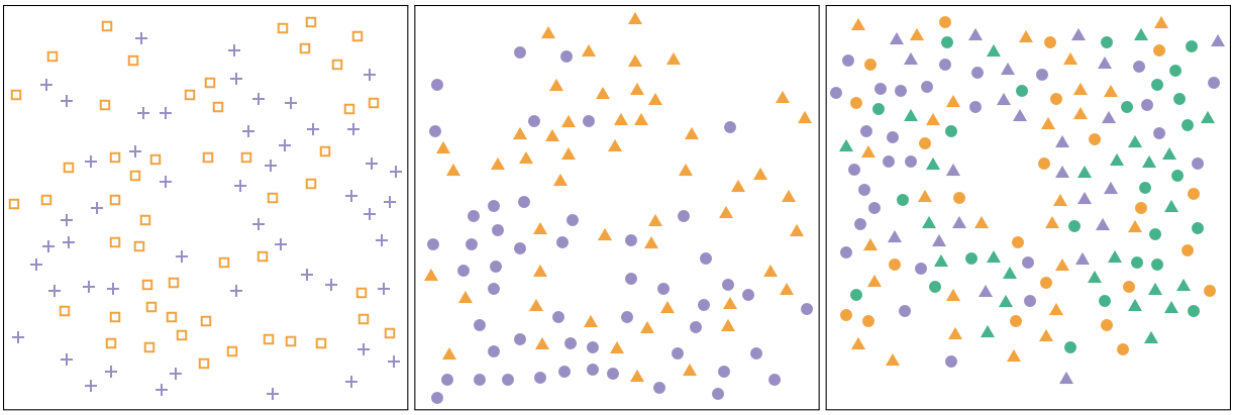


Figure 4: Example of some of the visual stumuli that where used to replicate the study originally conducted by Gleicher et al.[3]. The stimuli used where available from the original study.

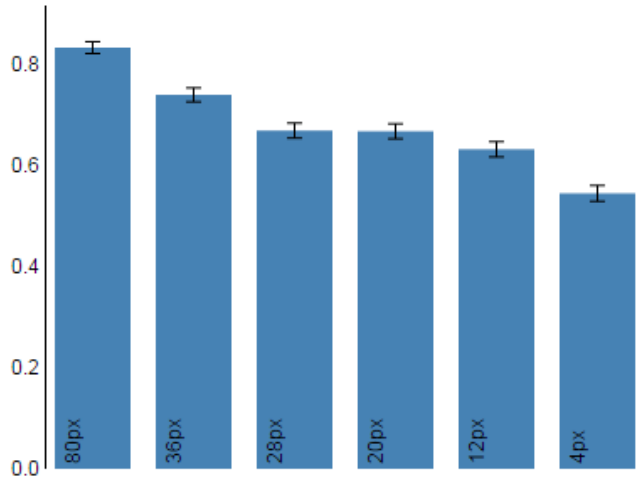


Figure 5: Bar charts showing the mean accuracy of the participants grouped by absolute difference in pixels from the analysis of the multiclass scatter plot study. (Chart is generated by the proposed system)

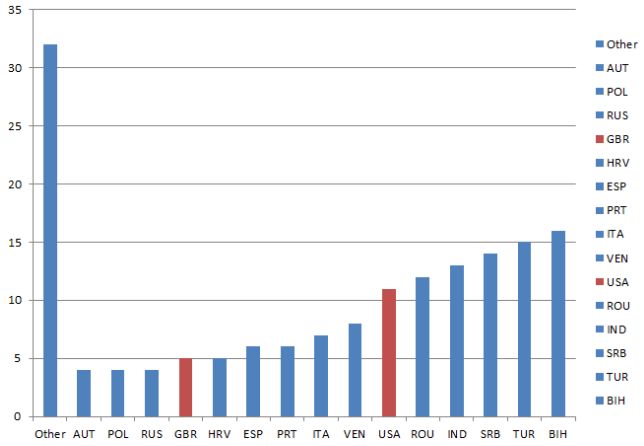


Figure 6: The distribution of participants from the multiclass scatterplot study per country. Countries where English is the native language is marked in red.