Notes after MTurk Abstract Study

Background

For the clouds category, we were able to recruit 33 valid subjects, among which we have

- 12 from Roy's England visit. We surveyed 13 people, and one person misread the instruction.
- 21 subjects from Mechanical Turk We initially planned for 50, and some people were not being reliable in their answer, and we were also experiencing some technical trouble with the server.

We were able to see some uniform agreement on complexities, and would like to carry this on to other categories.

Next, we tried running our experiment on the abstract category. We picked the next category based on how likely our subjects will form a uniform opinion on the complexity, or to put it in another way, how likely would our subjects agree on which image is more complex given any one pair of images.

Step

1. Post on Amazon mechanical Turk for 500 subjects
2. Rejected 2 subjects because their answers were tangential.
3. Ended up getting 504 log files (there might be people who refreshed the page and did it twice?)
4. Roy collected 13 more subjects from England to see if there are any differences between supervised v. unsupervised experiment condition.
5. Run thurstone scaling (TS) on all 517 (=504+13) log files.
    1. TS with subjects combined into one single matrix, result in an overall score vector.
    2. TS on each single subjects, result in individual score vectors.
6. Import the resulting vectors on Google spreadsheet.
7. Normalize the score (stretch to a scale of 0 - 100) and shade the cells (darker cell = more complex) as a method of data visualization.

Result

For the detailed result, see the Google spreadsheet.

Overall Score

The following is the overall score vector, normalized to a scale of 0 - 100. The higher the TS score, the more complex the image is. As we can see, overall, the subjects consider 4.jpg to be the most complex image, and 1.jpg the least complex.

Image Name	Thurstone Scaling Score (Normalized)
1.jpg	0
2.jpg	8.111726021
3.jpg	49.47083768
4.jpg	100
5.jpg	58.15664642
6.jpg	47.73151975
7.jpg	61.52387969
8.jpg	80.63661192
9.jpg	90.188486
10.jpg	55.420994

Individual Score

Unlike the clouds category, there is no clear definitive agreement among subjects. For example, if we do a majority vote on which subjects think which image is the most and the least complex, we get:

Number of people who believe this image is...	the most complex	the least complex
1.jpg	17	**183**
2.jpg	19	99
3.jpg	42	35
4.jpg	**130**	15
5.jpg	37	36
6.jpg	28	70
7.jpg	70	7
8.jpg	66	3
9.jpg	74	6
10.jpg	69	71

Although it is clear that 4.jpg and 1.jpg are the most and least complex images, consistent with the overall score, the votes are fairly uniformly distributed. This is also confirmed by the cell shading mentioned in step 6 described above.

The 13 England subjects mentioned in step 4 give different opinion on the most complex picture (7 out of 13 think 10.jpg is the most complex). And we observe the same chaotic shading.

Discussion

Although we are able to synthesize a overall TS score vector, looking into the individual score vectors, we learn that abstract might not be a good category.

- First, subjects do not agree with each other. Each subject seems to has their own philosophy, and they differ drastically.
- Second, subjects find some images really ambiguous to tell. We have a lot of subjects who are like "yeah we can tell between these four, but for the other half of this category, they all look the same complexity to me". We even have a subject who believes 5 images out of the 10 images are all "the most complex".
- Supervised study in England does not change the two points above.

We conclude that this is a bad category to use.

Next Step

- Get at least one more category. H proposes "office supply".
- Do small-scaled MTurk study on one more category (can be the additional category mentioned above) and see if the situation with the abstract category is unique.
- H will do a histogram to see which complexity judging criteria do people use.