KW thoughts on auto-encoding analyses
Last updated: 2021-02-05

High-level thoughts:

This is a very difficult design to analyze! When we were writing the 2015 paper I went down a long rabbit hole about it, involving extended conversations with the resident statistician at Stanford (Ewart Thomas). He convinced me that the original analysis (in the 2015 paper, based on others' previous work) was reasonable, and probably better than any of the trial- or block-level stuff I was attempting to invent from scratch. I've re-convinced myself of this from scratch at least 3 times since then, so my best advice is to go with the analysis plan we articulated in the original paper.

FAQs about analyses

Q: How do you calculate the adjusted error difference score (AES)?

To test the main research question for each condition - *did children make disproportionately more within-category than between-category errors?* - we calculated "adjusted error difference scores" (AESs) for each participant (collapsing across trial blocks), using the following equation:

AES = (number of "within-category" errors) - (0.5 * number of "between-category" errors)

So, e.g., if a kid made 2 within-category errors (e.g., confusing a girl with another girl, or a boy with another boy) and 2 between-category errors (confusing a boy with a girl or vice-versa), they would receive an adjusted error difference score of 2 - 0.5 * 2 = 1.

The possible range for AESs for each block was -1 to +4, and we had 4 trial blocks - so in the final dataset for analysis, each participant had one adjusted error difference score, ranging from -4 to +16.

Q: Is it really correct to compare adjusted error difference scores (AESs) to zero? This is not the theoretical midpoint of the range of AESs!

A: I have spent many hours worrying about this, and I always end up re-convincing myself that 0 is the correct comparison. This is because we want to compare the observed AESs to the expected value of AESs at chance, i.e., if children were responding randomly.

The reason we're not comparing to the midpoint of the scale (-1 to +2 for each block) is that different scores have different likelihoods, at chance:

1.	4 correct:	4.17%	AES = 0 – 0.5 * 0 = **0**
2.	2 correct, 2 within errors:	8.33%	AES = 2 – 0.5 * 0 = **2**
3.	2 correct, 2 between errors:	16.67%	AES = 0 – 0.5 * 2 = **-1**
4.	1 correct, 1 within, 2 between errors:	33.33%	AES = 1 – 0.5 * 1 = **0**
5.	4 within errors:	4.17%	AES = 4 – 0.5 * 0 = **4**
6.	2 within, 2 between errors:	16.67%	AES = 2 – 0.5 * 2 = **1**
7.	4 between errors:	16.67%	AES = 0 – 0.5 * 4 = **-2**

If we multiply these out, we can an expected value of zero:
0.0417*0 + 0.0833*2 + 0.1667*(-1) + 0.3333*0 + 0.0417*4 + 0.1667*1 + 0.1667*(-2) = 0.

Q: What's up with the Wilcoxon signed rank tests?

A: The Wilcoxon signed ranks tests were just a non-parametric alternative to the t-tests we reported as the primary tests of our main question for each condition. We did this because the adjusted error difference score violates a lot of the assumptions of t-tests (e.g., the distribution of possible scores isn't quite normal). So whereas the t-tests reported for each condition compared the mean AES (for that condition) to 0 using standard distributions and assuming that the scores were normally distributed, the Wilcoxon tests assessed whether the distribution of AESs (for that condition) is symmetric around 0 in a non-parametric way.

The main take-away is that these are two ways of testing the same thing, and I would recommend doing them both.

Q: How do you take into account children's own identity? (e.g., whether children made more within-gender errors for children of their own gender vs. another gender)?

We did a secondary analysis for this. For each kid we tallied up the number of "within-category" errors that happened with targets of their own gender, and the number of "within-category errors" that happened with targets of the other gender. Then we calculated difference scores (same gender - other gender). So, e.g., if a girl made 2 within-category errors involving girls and 4 within-category errors involving boys, she'd have a difference score of 2 - 4 = -2.

We used a t-test to compare these difference scores to 0. (This is equivalent to doing a paired t-test, which might have made more sense to say!) You could do the same Wilcoxon stuff to do a non-parametric version of this, but we didn't.

Q: Can we model responses at the level of individual "trials" (e.g., what animal did the child match Girl 1 to?)?

A: I don't think it makes sense to try to think of this as individual trials rather blocks, because for each blocks the "trials" are not independent – e.g., if you match Girl 1 to Animal A, you can't match any other kid to that animal, so your choices are constrained. To me, this rules out any alternative analysis that treats each response as an independent trial.

Q: Can we model responses at the level of "blocks" (e.g., what 4 animals did the child match these 4 children to?)?

A: I think this would bvery reasonable, but pretty complicated.

In principle, I'd love to see someone develop an analysis that is tailored to the block level (with repeated measures if participants completed more than 1 block). I once worked out all the possible patterns of responses you could get on a block (e.g., 4 corrrect; 2 correct and 2 within-gender errors; 2 correct and 2 between-gender errors; ...), and how often you'd expect each of these possible response patterns if kids were responding randomly. Basically, there are 7

possible response patterns, and by chance you'd expect them to occur at radically different rates:

8.	4 correct:	4.17%	AES = 0 – 0.5 * 0 = **0**
9.	2 correct, 2 within errors:	8.33%	AES = 2 – 0.5 * 0 = **2**
10.	2 correct, 2 between errors:	16.67%	AES = 0 – 0.5 * 2 = **-1**
11.	1 correct, 1 within, 2 between errors:	33.33%	AES = 1 – 0.5 * 1 = **0**
12.	4 within errors:	4.17%	AES = 4 – 0.5 * 0 = **4**
13.	2 within, 2 between errors:	16.67%	AES = 2 – 0.5 * 2 = **1**
14.	4 between errors:	16.67%	AES = 0 – 0.5 * 4 = **-2**

You could imagine doing something like a chi-squared test (or a multinomial regression) to test whether the observed distribution of response patterns differs from this expected distribution, and then doing follow-up tests to see which response patterns occur more/less often than you'd expect by chance. If kids are automatically encoding the category in question (e.g., gender), you'd expect patterns #2 and #5 to occur more likely than they would be chance, and patterns #3 and #7 to occur less likely than they would by chance. I don't think you'd expect pattern #1 to occur any more/less likely than it would by chance. But I'm not sure about patterns #4 or #6... which is somewhat disturbing, because together patterns #4 and #6 would be expected to make up 50% of the data! Maybe someone else has better intutions about these patterns and could build this out into something workable, but it feels difficult (and a little convoluted) to me.

What about correct responses? Could we come up with an alternative scoring approach that would take those into account?

Tara Mandalaywala proposed an alternative scoring approach that considers <u>correct</u> responses to be partial indicators of auto-encoding (worth half as much as within-category errors):

AES_2 = (number of "within-category" errors) - (0.5 * number of "between-category" errors) **+ (0.5 * number of correct respones)**

1.	4 correct:	4.17%	AES_2 = 0 – 0.5 * 0 + 0.5 * 4 = **2**
2.	2 correct, 2 within errors:	8.33%	AES_2 = 2 – 0.5 * 0 + 0.5 * 2 = **3**
3.	2 correct, 2 between errors:	16.67%	AES_2 = 0 – 0.5 * 2 + 0.5 * 2 = **0**
4.	1 correct, 1 within, 2 between errors:	33.33%	AES_2 = 1 – 0.5 * 1 + 0.5 * 1 = **0.5**
5.	4 within errors:	4.17%	AES_2 = 4 – 0.5 * 0 + 0.5 * 0 = **4**
6.	2 within, 2 between errors:	16.67%	AES_2 = 2 – 0.5 * 2 + 0.5 * 0 = **1**
7.	4 between errors:	16.67%	AES_2 = 0 – 0.5 * 4 + 0.5 * 0 = **-2**

If we multiply these out, we can an expected value of 0.5:
0.0417*2 + 0.0833*3 + 0.1667 *0 + 0.3333*0.5 + 0.0417*4 + 0.1667*1 + 0.1667*(-2) = 0.50.

I've worked this out here to demonstrate that you could definitely score these data in different ways and use the same overall analysis approach – you'd just have to re-calculate the expected value of the score to determine what to compare these scores. (And, of course, justify <u>why</u> a correct response is worth "half as much" as a within-category error – or whatever scoring scheme you might develop.)