Discussion summary: Interpreting experiment results

(1) Many psycholinguistics articles include the following sections:

(a) A *theoretical overview*
- Gives a summary of the linguistic or psychological models that will be compared or tested
- May summarize or critique previous psycholinguistics papers

(b) A section outlining the *experiment design and methodology*
- Either here, or in the theoretical overview, there should be a discussion of exactly what predictions are made by the theories or frameworks under consideration

(c) One or more sections *reporting the results* and *discussing their implications*
- Today's discussion focuses on this part of the paper

(d) A general discussion
- This section will likely come back to the models or proposals from the theoretical overview and assess how they perform in light of the new experimental results
- This section may also discuss problems with the experiments, open questions, and directions for future research

(2) Experiment results: Comparing numerical values

(a) Most experiments are designed so that you are comparing different conditions to see if their outcomes are different

(b) In practice, this often means that you are collecting numerical data and you want to know if the numerical values obtained for the different categories are different
- Example: Average daily time spent on Facebook for different age brackets
- Example: Average reaction time (RT) in a lexical-decision task for high-frequency vs. low-frequency words

(c) This is not as straightforward as it sounds!

(3) Suppose you wanted to see if a coin was fair or not by flipping it 100 times.

(a) If you got 50 H and 50 T, would you conclude that the coin was fair?

(b) What about 10 H and 90 T?

(c) What about 48 H and 52 T?

(d) What about 40 H and 60 T?

(4) There are statistical tests that have been developed in order to determine how likely these various outcomes are. (For coin flipping, we can use a binomial distribution; other tests are more likely in psycholinguistics papers, but this is a simple example to look at.)

For 100 coin tosses, if the coin is fair (probability of H is 0.5):

	number of H	probability of *at most* this many heads occurring	
(a)	50	0.54 (54% chance)	coin is fair
(b)	10	0.00000000000000000137	coin is not fair
(c)	48	0.38 (38% chance)	???
(d)	40	0.03 (3% chance)	???

(5) As you can see from the last two coin-toss examples, we need a way to decide what probability range is acceptable

- If there is a 38% chance of 48 H when the coin is fair, should we conclude that the coin is fair or not fair when we get 48 H?
- If there is a 3% chance of 40 H when the coin is fair, should we conclude that the coin is fair or not fair when we get 40 H?

(6) This is what the ***p* value** is for when a result is reported: it tells you the probability that the result you got would have arisen by chance | remember: small p value is good!

(a) $p < 0.01$ (99% confidence that result is not due to chance) is generally considered highly significant

(b) $p < 0.05$ (95% confidence that result is not due to chance) is generally considered significant

(c) $p \geq 0.05$ is generally taken to show that the result is not significant

- Running more trials might let you get to significance if you are close!

(7) Something else we see in the Alegre & Gordon paper is an examination of significance by items and by subjects

- For a strong result, we want to see significance for both
- Significance for only one is indicative of a weaker result
- Warning: This is a slightly old-fashioned way of evaluating results; newer statistical models don't require this type of comparison