Exam 1 Topics.

This list is \*\*\*not\*\*\* exclusive but meant to give you an impression of the kinds of topics you should be familiar with and help you steer your study of the readings, notes, and assignments.

**The Trapped Minds Hypothesis**

**The Therapeutic Benefit Hypothesis**

**Review Prisoner’s of Silence. Can you discuss some of the fallacious arguments put forth by Biklen and relate these to other examples we’ve discussed in the class?**

* The kids are savants
* Appeals to emotions
* Case studies as definitive proof and dismissal of empirical studies that falsified the “theory”
* Stereotype threat as reasons for lack of replication – But participants produce the Gibberish that facilitators are told (e.g. the Cardinal study)
* Creeping hypotheses – there are savants somewhere for whom this effect appears
* the missing/ignored **base rate** problem. what is the percentage of individuals **‘diagnosed’** as autistic, whose communication skills improve across development **without** the use of facilitated communication?

In addition:

* *Post-hoc ergo propter hoc* – communication improved after FC
* *Special pleading* – language savants
* *Ad hominem* – the skeptics were heartless and non-inclusive
* *Appeal to emotion* – accounts from parents
* *Augment from ignorance* – You can’t explain

**Can you fashion an account, based on tested cognitive processes, of why the FC phenomenon occurs at the facilitator-facilitatee level?**

* Ideomotor responses: As in the Chevreul Pendulum, participants, regardless of whether they are told to imagine or prevent the pendelum to move, make it move in the suggested direction. Basically, this can be thought of as the brain-motor systems being so intricately intertwined that there are implicit and unconscious motor centers activated just by thinking about a specific response (or actively not thinking about it). Then once it begins, the error is exaggerated (it’s working) and becomes amplified
* Expectancy effects: Related to the ideomotor responses, but the experimenter (not just the facilitator) may be able to influence the results (e.g. as in the Cardinal study when both were present)
* Demand Characteristics (external reward): The students in the Cardinal study and many other studies of FC are used to being rewarded for being “successfully” facilitated, as are the facilitators themselves.

**What are the major flaws of the best examples of research cited by Biklen and his center?**

* Based on case studies
* Documentaries
* Confirmatory bias – they dismiss the empirical evidence coming from the OD Heck studies and other sources

**Can you draw connections between these papers and Popper’s paper on falsification? For example, Popper stated that every good theory is a prohibition…how does this relate to the extremely varied etiology of subjects used in the Cardinal paper?**

Popper said that good theories should provide testable hypotheses that allow the theory to be false – that is, it specifies the boundary conditions under which we should see success. In this study, there participants with autism, MR, cerebral palsy, and Down Syndrome. The theory should specify ways in which there should be performance differences among these groups with “categorically” different standing on different diagnosable disorders. But the theory can basically be “bent” to fit any of these

**Even if one took the findings and statistical results of Cardinal et al. at face value (bad idea), would these findings support the trapped minds hypothesis given the description of the students enrolled in the study?**

* Selection effects were really strong – participants had to score badly in order to get in a likely regressed to the mean
* These students were supposed to have been successful in FC before starting and were supposed to be using FC to succeed in age appropriate school tasks, then they should be almost at ceiling for accuracy with these simple, one word tasks

**Are you comfortable with the Burgess et al paper that tests the possibility that ideomotor effects underlie demonstrations of facilitator control?**

**If I claimed that study was invalid because the student facilitators only received brief training before administering facilitation, could you fashion a counter argument?**

* There are no criteria for **what makes someone an expert at FC**. Also, if FC is being guided by the facilitatee, then it shouldn’t matter how much experience the facilitator has
* The number of responses provided by the facilitatee was correlated with the **number of words that the facilitator remembered**, which suggests that training had nothing to do with it, unless is about memorizing information.
* The facilitatee should have produced **gibberish** to everything but ht ename because it was **double blind**
* Susceptibility to the **ideomotor** effect

**A related concern often raised about critical tests of FC that demonstrate facilitator control, is that these studies suffer from stereotype threat and the negative features of testing situations. Does this critique make sense given the nature of the findings we have seen under controlled tests of FC authorship (e.g., O.D. Heck study)?**

In the OD Heck studies, facilitators were blind to the information (words and images) that the facilitatee saw. Moreover, the vision of the facilitator was blocked

**Howe does the Mook ‘In defense of external validity paper’ relate to the Burgess et al. paper? How about the ‘Drama of the Artificial’? Is the Burgess paper an ‘analogue’ or applied study? If not, then what was the goal?**

Sometimes we don’t want a study that generalizes to real life. That isn’t always the point / goal.

**Are you comfortable with the following R commands/structures and do you feel like you could anticipate the output of a few lines of simple R script? Can you figure out which elements of a data frame, matrix, or vector are being fetched using bracket notation (i.e., foo[,3])?**

read.table(), for loops, rnorm(), c(), rbinom(), replicate()

For example, what will the following print to the console?

for (i in seq(from = 1, to = 10, by = 2){

print(i<5)

}

**Can you describe some of the potential problems of the original Carney, Cuddy & Yap (2010) power posing study?**

* Experimenters placed the participants into positions
* There was no baseline cortisol / testosterone collected
* They didn’t test whether hormone levels were > 0
* So much over-claiming
* Small sample size
* No external validity – the measures weren’t behavioral

**What is likely the most serious problem with that report?**

The most glaring issue was the experimenter expectancy effects. The participants were placed into the positions by the experimenters which could have strongly influenced participants to feel a certain way in the poses

**Was the Ranehill et al (2015) replication attempt less or more powerful and how did the design of that report differ in terms of methods.**

The report was more powerful.

It was a conceptual replication, not an exact replication, with the following differences:

* Participants held the poses longer
* There was no deception in the study
* Participants put themselves into the positions following on-screen instructions
* Different filler task

The only thing they were able to replicate was the self-report, which is basically just that the manipulation works explicitly

**What is an experimenter expectancy effect and can you think of a couple of ways to control for this potential confound?**

The presence or knowledge of an experimenter implicitly impacts the results of a study.

This can be prevented by:

* Double blind studies
* Limiting contact between the participant and experimenter
* Using blind coders
* Using standardized procedures as much as possible

**In thinking about the Simmons and Somonsohn (2017) p-curve analysis of power posing studies can you describe the shape of the p-value distribution one would observe if no effect is actually present (i.e., under the NULL hypothesis)? How about if a significant and robust effect is present…will this distribution be right skewed, left skewed or uniform….why?**

* If there was no effect, then the shape of the distribution would be uniform – all p-values are more or less equally likely
* If there is an effect, then the shape of the distribution should right skewed – we should see that there should be lots of really small p-values
* If there is publication bias or p-hacking, then we should see left skewed results – lots of results should be around .05 with small p-values less likely

**What is publication bias?**

Publication bias occurs when journals tend to only publish significant results, resulting in an inflation in only seeing studies that tend to confirm theories / expectations.

**What is a journal impact factor?**

Journal impact factor is a measure of the # of citations for a journal / the number of articles in the last 2 years. Essentially this number will be higher for journals that produce papers that are more likely to be cited multiple times over two year periods.

**If I claimed that the Tuzzi (2009) psycholinguistic paper provided a strong confirmation of autistic authorship during facilitation because it used fine grained statistical comparisons of language could you offer counter arguments?**

**Say we are conducting a research study together as a team and we hypothesize that one condition (A) will increase the range of each subjects’ reaction times across trials versus another condition (B) (i.e., range of A reaction times > range of B reaction times). We want to use a paired t-test to test this hypothesis but are unsure if ranges are cool to use this way, particularly since we have 40 trials in condition A and 20 trials in condition B. Can you describe broadly, how would we use Monte Carlo Simulation to find out whether the t-test would yield valid inference in this situation?**

You would want to test two possible outcomes – one where there are differences in ranges and one where there are not. You would simulate data to these specifications and test pairwise differences in ranges across participants. Then you’d look at the distribution of the test-statistic in the case of range differences and in their absence. They should look different across approximately 1000 trials. You would also want to vary the number of trials in each condition (20-20, 40-20, and 40-40) to see if this changes the distribution. You’re likely to get larger ranges in the condition that uses more trials because of the possibility of more extreme values across more trials (ranges are *very* susceptible to outliers)

**Reliability and Validity. Can you distinguish them? Are you familiar with different types?**

Reliability estimates the consistency of a measure:

* *Internal consistency*: do different indicators appear to be measuring the same thing
* *Inter-rater reliability*: does the construct appear to operate the same across different measures
* *Split-half reliability*: another way of estimating internal consistency generally

Validity tests whether you are actually measuring the thing you think you’re measuring

* *Construct validity*: basically replicability across measures, contexts, tasks, etc
* *Generalizability*: Will the results generalize across different operationalization of the same
* *External Validity*: do the findings / operationalize apply outside of the research context
* *Structural Validity*: Using CFA, gets at divergence, unidimensionality, etc
* *Content Validity*: Are you measuring the construct you mean to measure
* *Substantive Validity*: basically does it mediate stuff it should impact

**How does one potentially constrain or limit the other?**

If an unreliable test is used to test the validity of a method, this could lead to the creation of a test that is neither reliable nor valid. However, you can have a reliable test that is not valid. An example of this would be that homosexuality used to be diagnosable via the DSM. The test was reliable at identifying whether someone was homosexual. However, it was not a valid test (homosexuality is \*not\* a disorder.

**Can you discuss the multi-trait multi-method approach to external validity at a broad level?**

Multi-trait, multi-method approaches mean to test the validity of an effect by separating method variance from true variance. In other words, can we test whether a specific construct is robust across different ways of testing it (e.g. behavioral task v. self-report v. observer report) and different operationalizations of it (e.g. if testing personality, related and unrelated constructs). If this works, we should see that divergent traits within methods shouldn’t be correlated than convergent traits across methods. Moreover, the reliability of any of these methods should exceed mulit-trait and multi-method correlations.

**Are you comfortable with the output of the alpha function from the psych package…for example, which column shows the item to test correlations with the item removed?**

Raw alpha is the alpha based on the covariances, while standardized alpha is the one based on the correlations (which are standardized)

Raw.r is the correlation between the item and total score when the item is included in the calculation of the total score

**If given a small inter-item correlation matrix for items from a test can you hand calculate Cronbach’s alpha?**

Cronbach’s alpha is the expected split half correlation across all possible split halves

Alpha = n\*mean(rij) / (n\*mean(rij) + (1- mean(rij)))

**Does the Spearman-Brown prophecy formula make sense and can you use it to decide whether adding items to a test would make a worthwhile improvement in reliability?**

Spearman-Brown prophecy formula: 

The SB formula estimates the relationship between the number of elements and the reliability of the overall composite when variances across indicators are similar and pairwise correlations are equivalent

Adding items to a test improves reliability when pairwise correlations are lower (.6)

**What are the limitations of Cronbach’s alpha…for example, what two things does it conflate in its calculation?**

Alpha conflates internal consistency and unidimensionality in its calculation because it is based on average inter-item correlations. You need to examine the specific correlations in order to assess whether the items seem to be tapping the same thing

**Can you think of a few of ways that the FC program and the Microaggression Research Program are similar? For example, how do they approach falsifiability or divergent validity? How about case evidence or testimonial evidence? How about implementation outside of academia, do they have something in common here?**

*Falsifiability*: FC and Microaggression both seem to be shifting their hypotheses around, similar to Popper’s discussion of Marxism.

*Divergent validity*: The boundary conditions surrounding both are vague. Lilienfield discusses how the definition of microaggressions is almost boundless, especially because the definition provided by Sue et al. argues that it is entirely in the eyes of the perceiver.

*Case evidence / testimonial evidence*: Sue et al.’s definition of microagressions were almost completely based on anecdotal evidence and discussion among the authors.

*Implementation outside academia*: Both were quickly and wildly accepted for implementation without rigorous scientific testing, with FC being used at both OD Heck and Biklen’s institutes and microaggressions being adopted at universities like in the UC system.

**What was Baumeister et al’s (2007) general concern when it comes to behavioral observation. Under what type of validity would this concern generally fall?**

Baumeister et al.,’s concern was that behavioral observation is almost completely absent from theories about *behavior.* This would fail under external validity.

**Is there a potential for conflict between the Baumeister paper and the Mook (1983) paper…why?**

The goal of theory testing (process, model, mechanism) (**basic science**) and the goal of analogue prediction (**applied science**) are not subject to the same types of validity concerns.

In contrast, the Baumeister et al., paper was arguing that behavior is all but missing in psychology and this places serious threats on the external validity of our results. Basically arguing that there is no good excuse for not measuring actual behavior when that is actually what we care about

**Cunningham, Preacher and Banaji considered the reliability and convergent validity of implicit attitude measures. What was the impetus for this research? In other words, what problem or concern was it designed to address?**

The question was whether the IAT is useful in terms of ranking individuals – that is, can it be used to predict other outcomes. There was also concern about the test-retest consistency of the IAT, which is necessary (but not sufficient) for establishing that the IAT could be used to rank people.

**Does the work suggest any practical limitations of the reliability of the considered implicit measures?**

The implicit measures correlate poorly among themselves, suggesting that they can’t stably rank people over time. The implicit measures also only correlate weakly to moderately with the explicit measure of racism

**What does it tells us about the composites; that is, should we generally expect at least modest gains in reliability when we administer the same test multiple times?**

This is going to be a given because the mean across administrations is always sgoing to be a better estimate than a single administration alone

**Do you think that the sampling of implicit measures was sufficiently broad to cover questions of the reliability and validity of implicit attitude measures in general?**

No, they used almost identical implicit measures of attitudes to test the reliability and validity of their measure, which per the MTMM approach, does not have varying enough methods to actually test the heteromethod-monotrait part of the matrix. They would need an alternative implicit measure.