HUMANS AT THE CENTER OF MIR:

HUMAN-SUBJECTS RESEARCH BEST PRACTICES

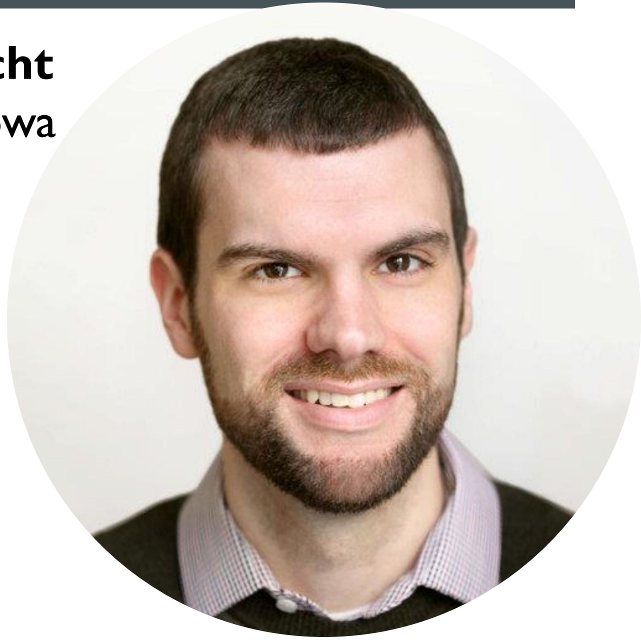
Not too late to do the experiment!

TUTORIAL OUTLINE

2:00 - 2:15	**INTRODUCTION**	Overview Use cases in MIR
2:15 - 3:00	**PART 1**	Experimental Design Validity and Reliability Sampling
3:00 - 3:30	**PART 2a**	Measurement & Randomness Statistical Evaluation
3:30 – 4:00		COFFEE BREAK
4:00 – 4:15	**PART 2a** (continued)	
4:15 - 4:45	**PART 2b**	Psychometrics
4:45 - 5:30	**PART 3**	Reporting in APA Ethics and Approval New Publication standards

WHO ARE WE?

Joshua Albrecht
University of Iowa



Claire Arthur
Georgia Institute of Technology

David Sears
Texas Tech University

Nat Condit-Schultz
Georgia Institute of Technology

J. Ashley Burgoyne
University of Amsterdam

INTRODUCTION

In one form or another, most MIR research depends on the judgment of humans.

Humans provide our ground-truth data, evaluate our results, and consume our tools, and models, and music. Will users like it? Will customers buy it? Does it sound good? These are all critical questions for MIR researchers which can only be answered by asking people.

USE CASES IN MIR

INTRODUCTION

SCENARIO 1: Source Separation

You have made a source separation model. You need to do subjective testing with human participants to validate that the separation isolates tracks better than other models, while minimizing artifacts and preserving audio fidelity. (Ultimately, perceptual problems).

INTRODUCTION

SCENARIO 2: Music Emotion Recognition

You have designed a model for automatically detecting musical emotion from audio (MER), a highly subjective (and human-dependent) task. You plan to use human participants both for subjective evaluation and, possibly, human feedback to refine model.

WHAT YOU WILL LEARN IN THIS TUTORIAL

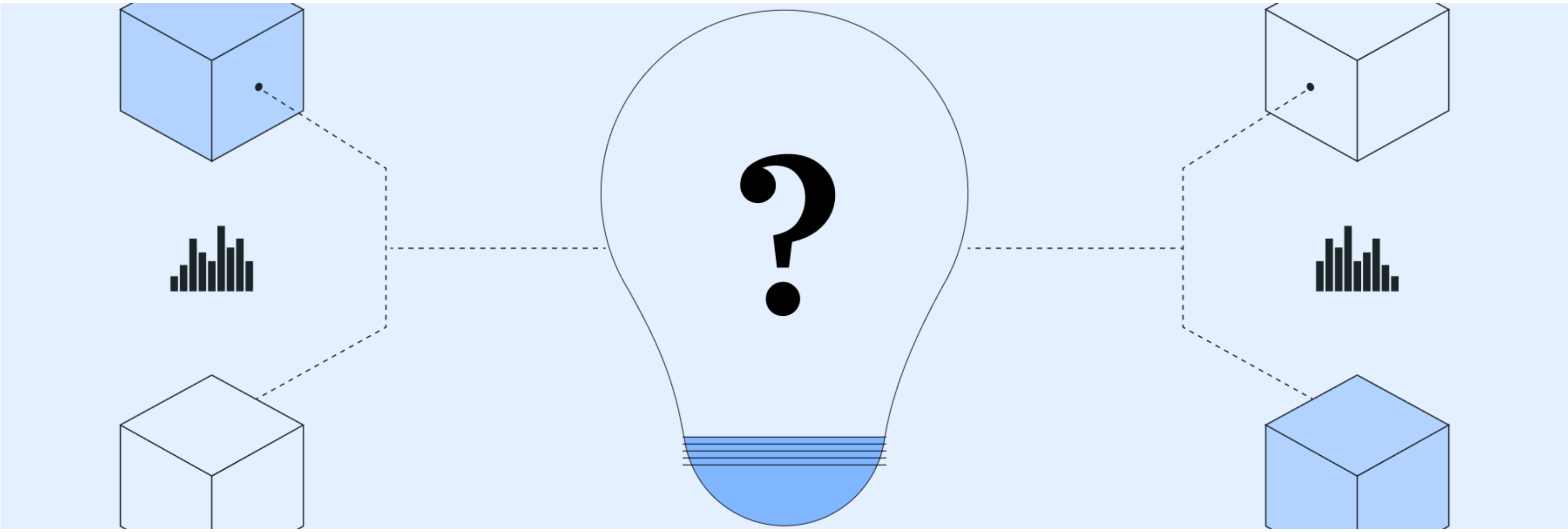
- Common pitfalls and considerations in designing User Studies and Human Subjects research
- Critical examination of experimental design to maximize results potential
- Intro/overview of statistical techniques and evaluation
- Dealing with human subjectivity (ignoring it doesn't make it go away!)
- Standards for methods, reporting and publishing

PART 1: EXPERIMENTAL DESIGN

PART 1 OUTLINE

- **Experimental Design**
  - Manipulation and causation
  - Variables & Groups
  - Controls
  - Measurement
  - Survey Design
  - Stimuli Design
  - Validity and Reliability
  - Demand characteristics
- **Sampling**
  - Sampling music
  - Sampling humans

DESIGNING YOUR EXPERIMENT



What is the goal?

- Capturing data that will *best* allow for the evaluation of your hypothesis.
- Ensuring the hypothesis has a chance to *fail*.

DESIGNING YOUR EXPERIMENT

Capturing data that will *best* allow for the evaluation of your hypothesis:

- Types of questions
- Type of task (rate; rank; choose; etc.)
- Clarity of instructions
- Appropriate stimuli (selection, length)
- Order effects
- Appropriate controls and conditions
- Participant pool & generalizability
- Understanding and avoiding bias

VARIABLES

Dependent Variable: The thing you are measuring. A.k.a. the response variable; the outcome variable.

Independent Variable: A variable that is manipulated or changed in the experiment/study to observe its effect on a dependent variable. A.k.a. the predictor variable; the condition.

MANIPULATION & CAUSATION

“No causation without manipulation”

- Subjective evaluation frequently in the form of surveys, which are often correlational

Internal Validity

- Manipulating one variable (independent variable) to examine impact on a response variable (dependent variable) in order to infer causality (and not just correlation).

EXAMPLES:

SCENARIO 1:

SOURCE SEPARATION EVALUATION

- Stimuli created by (or grouped by) condition (i.e., different models)
- Are there changes in the stimuli that could arise from causes *other* than the main condition (or IV) of interest?

EXAMPLES:

SCENARIO 2:

EMOTION RECOGNITION EVALUATION

- People rate perceived emotion of samples/stimuli, and:
  - Compare against ground truth labels from your model?
  - Compare against ground truth labels from multiple models?

CONDITIONS, CONTROLS, COVARIATES

To be confident claiming X *causes* Y, we have to be cautious that there are not other explanations.

Conditions: Different setups or variations in an experiment used to test effects on the outcome (i.e., IV).

Controls: Elements held constant to prevent them from influencing the outcome, ensuring that only the variable of interest affects the results.

Covariates: Additional variables that may impact the outcome and are measured to account for their influence in the analysis.

CONDITIONS, CONTROLS, COVARIATES

Conditions: Different setups or variations in an experiment used to test effects on the outcome (i.e., IV).

CONDITIONS, CONTROLS, COVARIATES

Conditions: Different setups or variations in an experiment used to test effects on the outcome (i.e., IV).

A.K.A. "Levels" of the Independent Variable(s)

E.g., Model A *vs.* Model B *vs.* Model C

CONDITIONS, CONTROLS, COVARIATES

Conditions: Different setups or variations in an experiment used to test effects on the outcome.

A.K.A. "Levels" of the Independent Variable(s)

(E.g., Medicine *vs.* Placebo)
(Fast *vs.* Slow Tempo)

CONDITIONS, CONTROLS, COVARIATES

Controls: Elements held constant to prevent them from influencing the outcome, ensuring that only the variable of interest affects the results.

E.g., Source Separation:

- Number of instruments
- Type of instruments/voice
- Presence of effects
- Source file!

E.g., MER

- Genre
- Loudness
- Presence of lyrics

CONDITIONS, CONTROLS, COVARIATES

Covariates: Additional variables that may impact the outcome and are measured to account for their influence in the analysis.

E.g., Source Separation:

- Musical expertise
- Listening environment
- Song familiarity

E.g., MER

- Person's current mood
- Musical preferences
- Time of day

RANDOMIZATION

Stimuli controls

- When not every confounding variable can possibly be controlled…

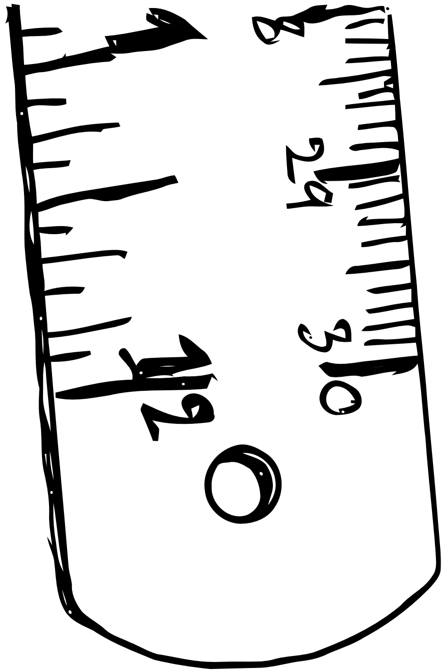
Order effects

- Hearing a sad song after an uplifting song?

- Getting better at a task over time

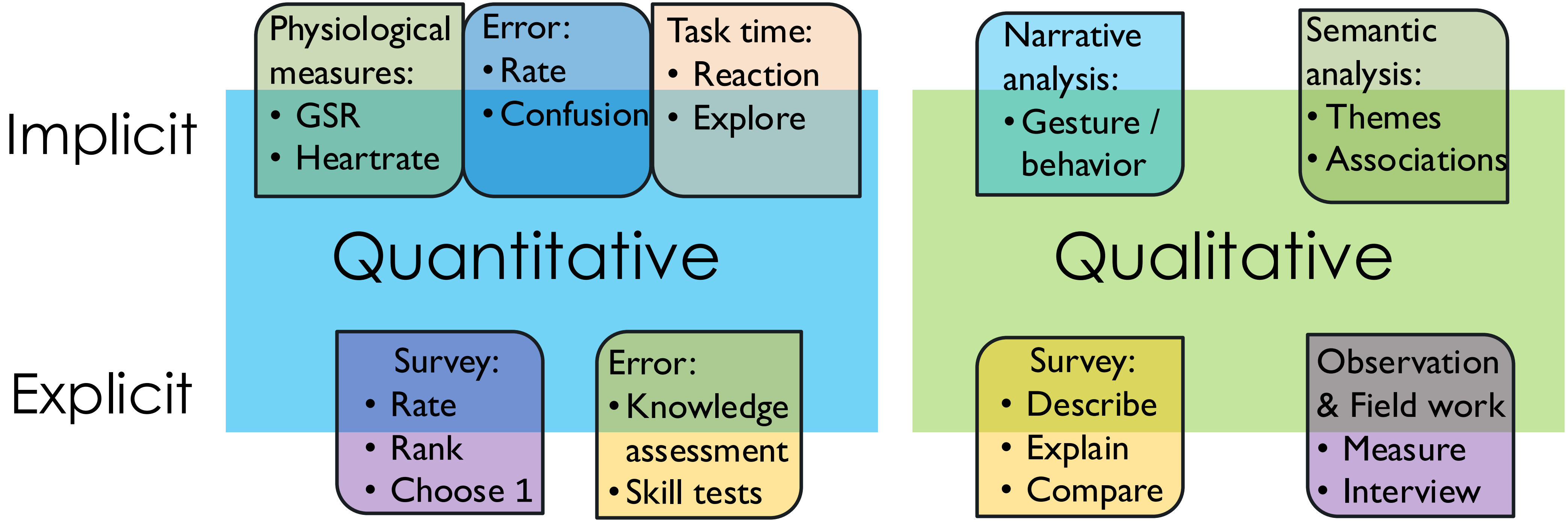
WHAT ARE WE MEASURING?

- Clarity of Instructions
- Type of task / question
- Scale and unit of measure
- Risks of bias



WHAT ARE WE MEASURING?

- Types of data collection:



Implicit

Explicit

Quantitative

Physiological measures:
- GSR
- Heartrate

Error:
- Rate
- Confusion

Task time:
- Reaction
- Explore

Survey:
- Rate
- Rank
- Choose 1

Error:
- Knowledge assessment
- Skill tests

Qualitative

Narrative analysis:
- Gesture / behavior

Semantic analysis:
- Themes
- Associations

Survey:
- Describe
- Explain
- Compare

Observation & Field work
- Measure
- Interview

WHAT ARE WE MEASURING?

SOURCE SEPARATION

MUSIC EMOTION RECOGNITION

WHAT ARE WE MEASURING?

SOURCE SEPARATION

WHAT ARE WE MEASURING?

SOURCE SEPARATION

Types of Questions:	Pros	Cons
Choose the file that…	Faster than rating; rank-order data	Lots of comparing = slow
Rate each file on a 0-100 scale for …	data for each file; (likely) normally-distributed data, specific	Less intuitive; more variation in responses; takes longer
Compare two files for …	Easy, fast, intuitive (more reliable)	Inefficient: Lots of pairs in order to be exhaustive; more difficult to analyze

WHAT ARE WE MEASURING?

SOURCE SEPARATION

Clarity of Instructions:

- What is meant by "overall quality"?
  - success of isolation?
  - fidelity of the audio?
  - timbral purity/accuracy?
  - smoothness and continuity of the isolated track?

WHAT ARE WE MEASURING?

MUSIC EMOTION RECOGNITION

WHAT ARE WE MEASURING?

MUSIC EMOTION RECOGNITION

Types of Questions:	Pros	Cons
Rate each file on a 0 - 7 scale for [various discrete emotions]…	data for each emotion; (likely) normal-ish data; detailed	Challenging; more variation in responses; takes longer
Choose one/all the emotion words that apply [select task]	Faster and easier than other options; binary	Experimenter-selected vocab (bias?)
Name the emotion term [free response]	Individualized (inclusive)	Messy to analyze
Rate valence & arousal in 2D space	Simplicity, common benchmark	Less intuitive; more variation; potentially limiting

WHAT ARE WE MEASURING?

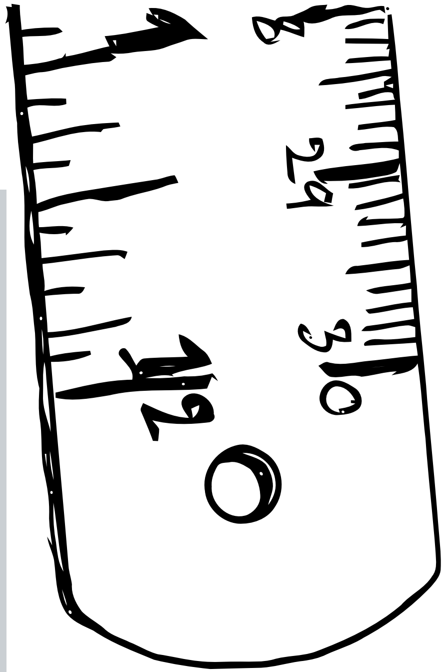
MUSIC EMOTION RECOGNITION

Clarity of Instructions:

- What is meant by "emotion"?
  - **Felt versus perceived**
  - Relation to mood
  - Categorical vs dimensional

WHAT ARE WE MEASURING?

- Scale and unit of measure



Measurement Scales

Nominal: categories with no order (e.g., genres)

Ordinal: rank order (e.g., disagree, neutral, agree)

Interval: equal intervals but no zero (e.g., dates, IQ

Ratio: same as interval but meaningful zero (e.g., time taken)

Continuous vs. Discrete data

Continuous: any unbounded (or bounded) value; highly precise (e.g., time, heart rate)

Discrete: distinct, separate values; often counts of items (e.g., key signatures, instrument labels, distributions)

Unit of Measurement

The magnitude of the measure

Examples:

- Decibels
- BPM for heartrate
- Milliseconds (ms) for reaction time.

FACTORS INFLUENCING RESPONSES

- Time of day
- Mood of participant
- Socioeconomic status/cultural background
- Order effects
- Maturation/fatigue effects
- Question asked
- Wording

RELIABILITY

Reliability

- The consistency of the measure
- Observing the same value over time in similar circumstances

RELIABILITY

Intrasubjective reliability

- Consistent responses from the same participant
  - Test/retest reliability
  - Re-engaging participant after some time

RELIABILITY

Intersubjective reliability

- Consistent responses from the different participants under similar conditions
- Dependent on effect size

RELIABILITY

MER

- Perceived emotion/affect more reliable than felt/induced emotion/affect
- “Please listen to the entire excerpt and rate how much of the following categories you feel it expresses.”

VALIDITY

Validity

- The degree to which the test measures what it claims to measure
- "Grade the Basic Audio Quality of the isolated part of each stimuli with respect to the reference mix."

VALIDITY

Internal Validity

- When the measures of the sample are correct
- The results of individual responses generalize to the sample studied

VALIDITY

External Validity

- When the sample is truly representative of the population of interest
- The results of your sample generalize to the broader population

VALIDITY

External Validity Strategies

- Compare your results to other studies
- Conduct replications
- Pre-registration
- Use more than one competing (independent) operationalizations
- Look for converging evidence
- Ecological validity

OPERATIONALIZATION

- Study how the timbres used in heavy metal music affect the perceptions of conveyed emotion
  - Heavy metal music
  - Timbre
  - Timbres used
  - Emotion
  - Conveyed emotion
  - Perceptions of conveyed emotion
  - Affects

OPERATIONALIZATION

- Approximation or estimation of concept that can be operationally used
- Perception of affect:
- "Selections on a 7-point scale in response to a question about perceived affect"

DEMAND CHARACTERISTICS

- Any aspect of the experiment that causes a change in a participant's behavior
- Participant perceptions of the experiment affecting their responses
  - e.g. placebo effect

TYPES OF DEMAND CHARACTERISTICS

- Cooperation bias
- Contrarian bias
- Acquiescence bias
- Social desirability bias

MINIMIZING DEMAND CHARACTERISTICS

- Blind procedure
- Deception
  - Explicitly lying
  - Implicitly suggesting
- Implicit measures
- Debriefing

SAMPLING

SAMPLING

- In academic environments (and even in industry), we rarely achieve 'N = all'.
- In MIR, we are almost always generalising from a model trained on a just a subset of musical fragments of interest.
- Sampling is the way we choose that subset.
- When human participants are involved, we not only need to sample music, but also our participants!

One Million Songs

1. Get the most ‘familiar’ artists according to The Echo Nest, then download as many songs as possible from each of them.
2. Get the 200 top terms from The Echo Nest, then using each term as a descriptor to find 100 artists, download as many of their songs as possible.
3. Get the songs and artists from the CAL500 dataset.
4. Get ‘extreme’ songs from The Echo Nest search parameters, e.g., songs with highest energy, lowest energy, tempo, song hotness, etc.
5. Take random walks along similar artist links, starting from the 100 most familiar artists.

Bertin-Mahieux et al. · 2011 · The Million Song Dataset

One Thousand Songs

1. Select a random week between 4 August 1958 (the first week the Billboard Hot 100 was published) and 23 November 1991 (the last week the Hot 100 was compiled using surveys instead of Nielsen SoundScan).
2. Select a random rank between 1 and 100.
3. If the track on the Hot 100 for the week selected in Step 1 and the rank selected in Step 2 is commercially available, download it.
   - If not, try the tracks either one rank above and one rank below the desired rank, in random order.
   - If neither of these are available, try the tracks two ranks above and two ranks below the desired rank, in random order.
   - If none of these five tracks are available, download nothing.
4. Repeat Steps 1–3 until the desired number of tracks is reached.

Burgoyne, Wild, et al. · 2011 · An Expert Ground Truth Set

POST-STRATIFICATION

- Some songs in the McGill Billboard corpus are sampled more than once (e.g., songs that charted for multiple weeks).
- An average listener would have been expected to have more exposure to these songs than those sampled only once.
- If we want our models and predictions to be based on an average listener's exposure, we need to weight these songs more heavily during training, testing, and validation.
- Post-stratification is the term used for re-weighting model during or after training to match population characteristics.

QUESTIONS TO ASK YOURSELF

- What kinds of music and people are relevant for my research?
- Are there subgroups (strata) of music or people that are relevant for my research?
- Should each stratum have equal weight in my corpus or participant pool, or should some be weighted more heavily than others?
- Within each stratum, do I have a random or near-random selection, or is there likely to be some sort of bias (e.g., toward more popular songs or highly-educated participants)? Is the bias a problem?

MISSING DATA CAN CAUSE BIAS

- Data missing completely at random (MCAR) are missing for reasons that have nothing to do with parameters of interest (e.g., an unexpected Internet outage). These are OK…but rare.
- Data missing at random (MAR) are missing for reasons that can be fully explained by observations (e.g., people who report that they do not like metal music are less likely to finish a study on metal music). These are often OK, but it can be difficult to prove that data are truly MAR.
- Data missing not at random (MNAR) will lead to bias.

HOW MANY PARTICIPANTS?

- Power analysis uses formulas to ensure that traditional statistical tests will be unlikely to miss a true result (i.e., make a Type II error). The traditional threshold is a 20% chance that the test will reject a true result (80% power).

- Traditional power analysis requires you to specify your expected effect size and population characteristics in advance. Online calculators like WebPower can help.

  - For example, when comparing the average ratings for two equally-sized groups, each of which have a standard deviation of σ across their ratings, traditional power analysis requires $(5.6\ \sigma / \Delta)^2$ participants.

- If you are optimistic in estimating effect size, actual power can easily drop to just 10%!

- A more modern approach is to use simulated data to assess how many participants are likely to be necessary.

HOW TO RECRUIT

- The best sampling strategies assume full access to your population of interest (e.g., choosing which users to A/B test).
- More often, researchers have to rely on non-probabilistic convenience samples of whomever they can get.
  - Spamming the ISMIR list
  - Mechanical Turk or Prolific
- Convenience samples may not generalise well!

STRATA TO CONSIDER AND BALANCE

- Gender
- Age
- Education level
- Cultural background
- Linguistic background

- Geographic location

- Socio-economic status
- Musical sophistication
- Musical taste
- Instruments played