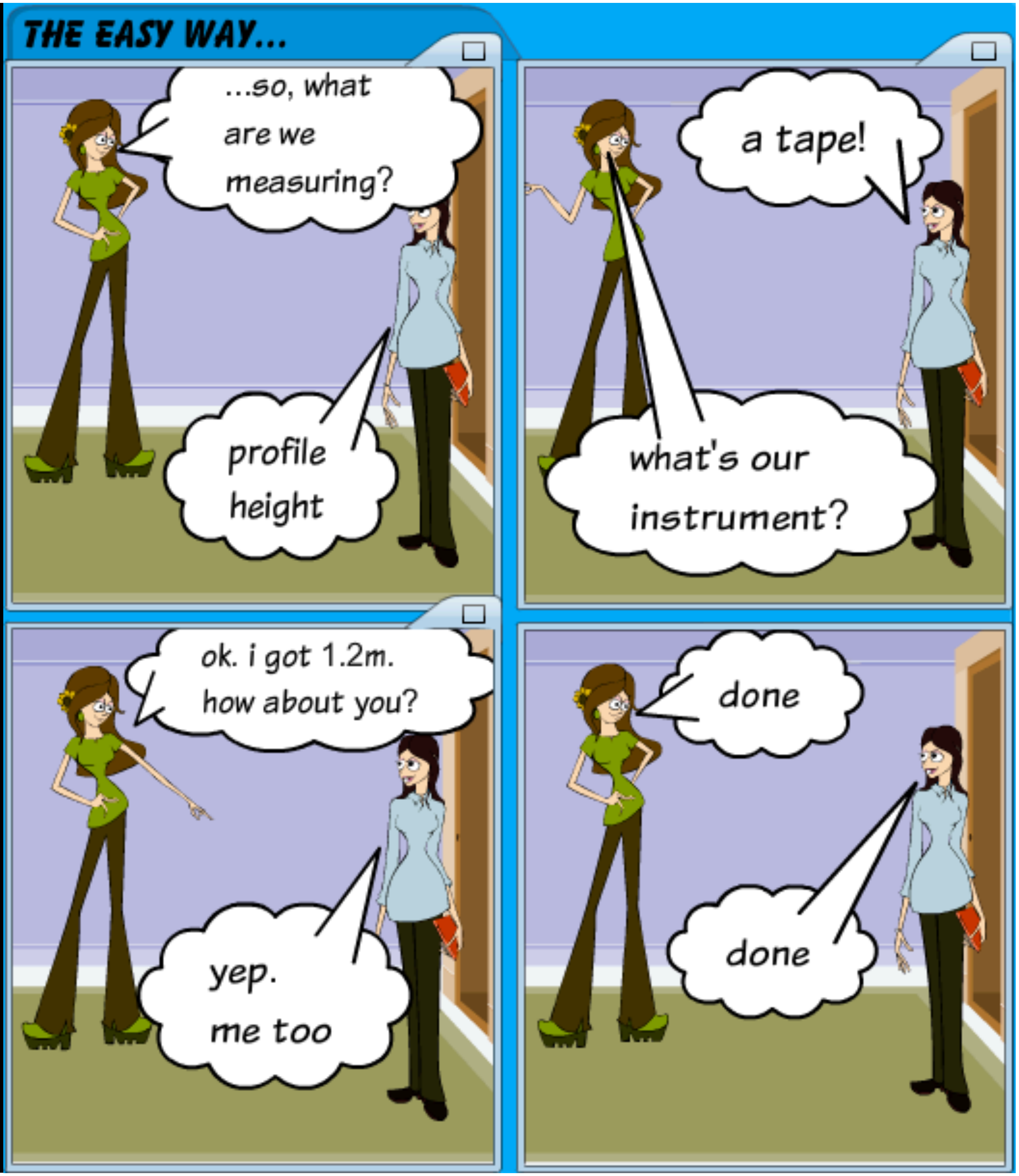
ICA Commission on
Cognitive Visualization

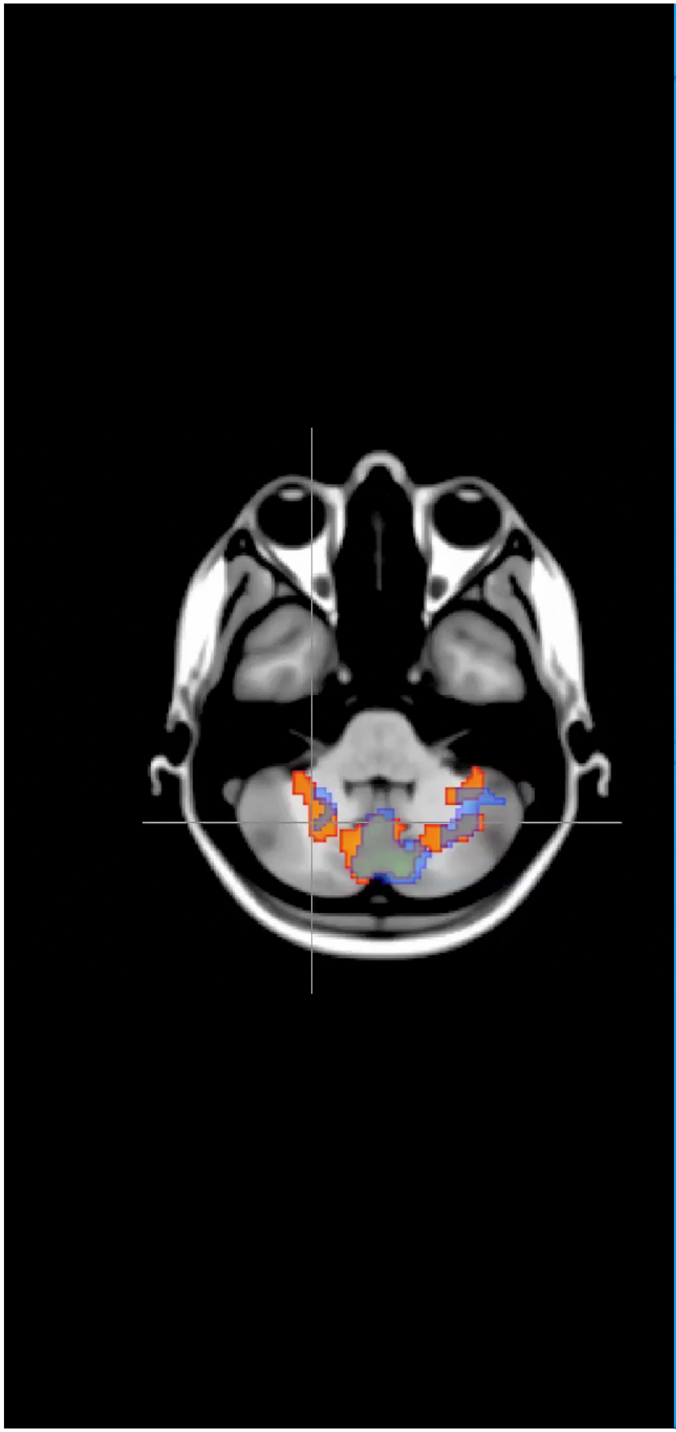
Instrument Integrity

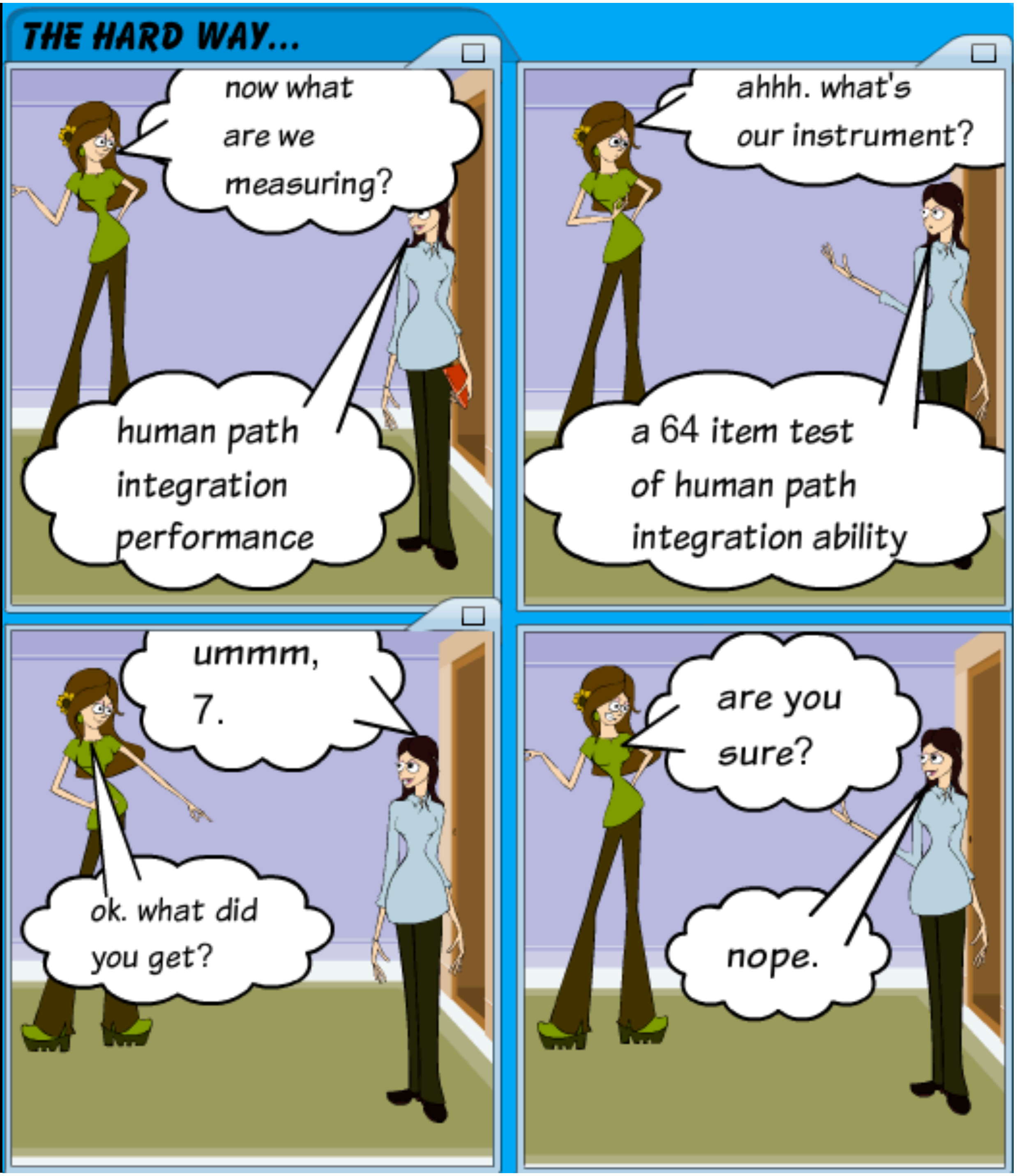
Amy Lobben

University of Oregon

20
40
60
80


THE EASY WAY...
...so, what are we measuring?
profile height
a tape!
what's our instrument?
ok. i got 1.2m. how about you?
yep. me too
done
done




THE HARD WAY...
now what are we measuring?
human path integration performance
ahhh. what's our instrument?
a 64 item test of human path integration ability
ummm, 7.
ok. what did you get?
are you sure?
nope.

the process…

what are we studying?

various human constructs of geospatial cognition

how do we measure?

psychological assessment – performance test

what are our methods?

computer testing, participant observation, eye-tracking, neuroimaging

the process…

what are we studying?

various human constructs of geospatial cognition

how do we measure?

Psychological assessment – performance test

what are our methods?

computer testing (performance), participant observation, eye-tracking, neuroimaging

the process…

what are we studying?

various human constructs of geospatial cognition

what are our methods?

computer testing, participant observation, eye-tracking, neuroimaging

how do we measure?

psychological assessment – performance test

let’s pose this one…

RQ: what environmental factors affect mental map encoding efficiency and effectiveness?

basically what we do…

- recruit participants
- balanced, randomized, true experimental design, …
- walk participants through an environment
- ask them to make sketch map of the environment
- maybe add talk aloud protocol just for fun
- develop systematic, robust post-hoc verbal analysis protocol
- all sounds good!!

voila! research question answered.

voila! research question answered.

but, are we sure?

did we really measure the mental map?

response #1 – yes. i know this because i’m well-trained and really smart.

response #2 – yes. i know this because i assessed the reliability of the instrument.

basically what we do…

- recruit participants
- balanced, randomized, true experimental design, …
- walk participants through an environment
- ask them to make sketch map of the environment
- maybe add talk aloud protocol just for fun
- develop systematic, robust post-hoc verbal analysis protocol
- all sounds good!!

ask them to make sketch map of the environment

maybe add talk aloud protocol just for fun

so, how do we assess the integrity of the measurement instrument (i.e. the reliability and validity of the sketch map and the talk-aloud protocol?

as designed…it’s impossible

a word about reliability and validity

- reliabilty – the consistency of a measure
- validity – the “truthiness” of a measure

reliability - 1

- consistency of scores obtained by the same person when examined with the same test on different occasions
- the interval, though, is important and should reflect test reliability and not behavior changes (i.e. 3 weeks versus 3 years)
- essentially reveals the extent to which differences between test scores are true differences or chance errors (not errors related to the test)
- several methods for assessing reliability
- the method you choose depends on the test and how it is designed and scored

reliability – 2 test/retest

- exact same test administered twice, with systematic interval applied between test takers
- systematic interval:
    - depends on age of test taker (usually shorter intervals for younger test takers, longer for older)
    - depends on complexity of test (shorter intervals for more complex longer for simpler)
    - should rarely exceed 6 months
- test scores between sessions compared
- advantage: if conducted appropriately, can give potentially most accurate measure of reliability
- disadvantage: learning effects, remembering questions

reliability – 3 test/retest

- in our example:
  - participants would perform the same walk through the environment and create the same mental map
  - probably a couple of months apart

reliability – 4 alternate form

- two forms of the test created
- administered in separate sessions over short interval
- direct comparison between scores
- higher correlation = better the reliability
- advantage: don’t have question memory issue, can administer over shorter interval
- disadvantage: not the same test

reliability – 5 alternate form

- in our example:
    - participants would perform a similar walk through a similar environment and create a similar mental map
    - shorter time interval – even same day

reliability – 6 split half

- one form of the test created
- administered in one session
- split test in sections
  - generally not good idea to split first half and second due to performance variation over the course of taking the test (fatigue…)
  - split by odd/even
  - but, must make sure that enough questions in each subject (i.e. if graphic is shown and questions relate to graphic, but have some odd and some even)
- direct comparison between scores
- higher correlation = better reliability
- advantage: one session, one test
- disadvantage: longer tests often better for this method

reliability – 7 split half

- in our example:
  - participants would perform walk through many similar environments and create several mental maps
  - at least 10
  - probably odd/even split half

reliability – 8 internal consistency

- measures homogeneity of test items, i.e. how closely related a group of questions are
- useful if the questions are designed to measure the same construct
- if a multiple construct test is assessed, treat each “section” as different test for reliability analysis
- internal consistency is indicated by Cronbach Alpha score, closer to 1 is higher reliability, above .8 is good
- advantage: one test, one testing session
- disadvantage: only measures test consistency, not necessarily between session consistency

reliability – 9   internal consistency

- in our example:
    - participants perform walk through at least 3 environments and create mental map
    - one test session

reliability – 10 interrater

- useful for both qualitative instrument and data analysis
- when open-ended questions are analyzed, a systematic scoring rubric should be developed
- multiple raters use the same rubric to analyze the same test taker's questions
- higher correlation between raters = higher reliability
- also useful for analyzing interviews; again, systematic coding sheet developed
- advantage: provides indication of post-hoc analysis reliability
- disadvantage: only provides indication of post-hoc analysis; not participant testing reliability

reliability – 11 interrater

- in our example:
  - systematic coding scheme for evaluating mental map construction
  - at least two raters apply the scheme

basically what we do…

- recruit participants
- balanced, randomized, true experimental design, …
- walk participants through an environment
- ask them to make sketch map of the environment
- maybe add talk aloud protocol just for fun
- develop systematic, robust post-hoc verbal analysis protocol
- all sounds good!!

ask them to make sketch map of the environment

maybe add talk aloud protocol just for fun

basically what we do…

- recruit participants
- balanced, randomized, true experimental design, …
- walk participants through an environment
- ask them to make sketch map of the environment
- maybe add talk aloud protocol just for fun
- develop systematic, robust post-hoc verbal analysis protocol
- all sounds good!!

ask them to make sketch map of the environment

maybe add talk aloud protocol just for fun

it can be done, but is convoluted

reliability

- which do we choose in our example:
    - test/retest
    - alternate form
    - split half
    - internal consistency
    - interrater

reliability

- which do we choose in our example:
  - test/retest
  - alternate form
  - split half
  - internal consistency
  - interrater

validity – 1

- the extent to which a test actually measures what it is intended to measure
- as with reliability, validity can be measured and is reported with most available tests
- types of validity:
  - face validity
  - content validity
  - criterion validity
  - construct validity

validity – 2 face

- test taker's perception of what the test actually measures
- a judgment of the relevancy of the test
- example: a test that says it measures map use, but contains no maps may not be perceived as a true measure of map use by the test taker
- face validity can be measured:
    - focus group
    - questionnaire
    - interview

validity – 3 face

- in our example:
    - simple structured or semi-structured interview with each participant
    - “what do you think we were measuring”

validity – 4 content

- how well a test samples knowledge or behavior its designed to measure
- commonly associated with achievement tests
  - example 1: course final exam – how well does a cumulative exam represent what was actually taught through the term?
  - example 2: employment test – considered content valid if the test represents job-related skills required for employment

validity – 5 content

- Measuring content validity
  - common approach: use raters to evaluate each question:
  - “is the skill or knowledge measured by this item…”
    - Essential
    - Useful but not essential
    - Not essential
  - develop acceptable threshold
    - example – if more than half of the raters judge the question as essential, the question passes the content validity test.

validity – 6 content

- in our example:
    - ask experts to participate in our experiment
    - semi-structured interview
    - focus group
    - “how well does our experiment capture participants’ mental maps”

validity – 7 criterion

- how well a test score can be used to infer an individual's standing on some measure of interest (the criterion)
- criterion – standard in which a judgment or decision may be based
- the process of establishing criterion validity involves comparing test results against a known criterion (either field measured behavior/activity/ability) or measured/diagnosed from another source
- validity coefficient – correlation coefficient that provides measure of the relationship between test scores and scores on the criterion measure

validity – 8 criterion

- 2 types of criterion-related validity
  - concurrent validity
    - the test and validating the criterion measured (or available) at the same time
    - example 1: test A is explored relative to criterion B, where B is existing measure or some other indicator of criterion
    - example 2: field validation
  - predictive validity
    - test scores taken at one point in time and criterion measured later – maybe after an intervention
    - example: comparison of Freshman admission test to end-of-year Freshman GPA (where GPA indicates academic success)

validity – 9 criterion

- in our example:
  - concurrent validity:
    - known test of mental mapping?
    - if looking at performance, maybe correlate with neuroanatomy (i.e. hippocampus tail or similar)
  - predictive validity:
    - compare mental maps to following field study in which participants are asked to conduct tasks related to mental map exercise

validity – 10 construct

- a judgment about the appropriateness of inferences drawn from test scores for a variable (the construct)
- construct – scientific idea that describes or explains a behavior
    - Example: Self-Location, intelligence, anxiety,…
- construct is intangible, researchers must formulate hypotheses about high and low scores on a test designed to measure the construct(s)
    - Example of contrasted (but somewhat related) constructs and test-based hypotheses: survey knowledge, route knowledge
- scientific activity and research is about finding evidence to support constructs

validity – 11 construct

- how do you find evidence of construct validity – 1?
  - depends on the research question and experimental design, but several approaches can be considered:
  - evidence of homogeneity
    - appropriate approach if the assumption is that the test measures the same construct
    - the extent to which test items correlate with each other

validity – 12 construct

- how do you find evidence of construct validity – 2?
  - – evidence of changes with age
    - appropriate if the assumption is that performance on the construct changes with age
    - example: increase, then later decrease in spatial abilities throughout your life
    - can be measured either longitudinally (using same subject group) or by using different age groups then comparing scores; results should follow hypothesized patterns

validity – 13 construct

- how do you find evidence of construct validity – 3?
  - evidence of pretest-posttest changes
    - should see measured, significantly different change as a result of an intervention
    - intervention can include: training, education, therapy, experience, medication
    - pretest, intervention, and posttest must be administered to each subject
    - direct comparison of scores

validity – 14 construct

- how do you find evidence of construct validity – 4?
  - convergent evidence
    - evidence that test results correlate with results from other known tests that are theorized to be related
  - discriminant evidence
    - evidence that test results are not statistically related to construct theorized not to be related
  - factor analysis
    - can be used to conduct an internal test of convergent and discriminant evidence

validity – 15 construct

- in our example:
  - convergent evidence: sketch maps correlate with field performance and also indicate environmental effects

basically what we do…

- recruit participants
- balanced, randomized, true experimental design, …
- walk participants through an environment
- ask them to make sketch map of the environment
- maybe add talk aloud protocol just for fun
- develop systematic, robust post-hoc verbal analysis protocol
- all sounds good!!

ask them to make sketch map of the environment

maybe add talk aloud protocol just for fun

basically what we do…

- recruit participants
- balanced, randomized, true experimental design, …
- walk participants through an environment
- ask them to make sketch map of the environment
- maybe add talk aloud protocol just for fun
- develop systematic, robust post-hoc verbal analysis protocol
- all sounds good!!

ask them to make sketch map of the environment

maybe add talk aloud protocol just for fun

it can be done, but is convoluted

validity

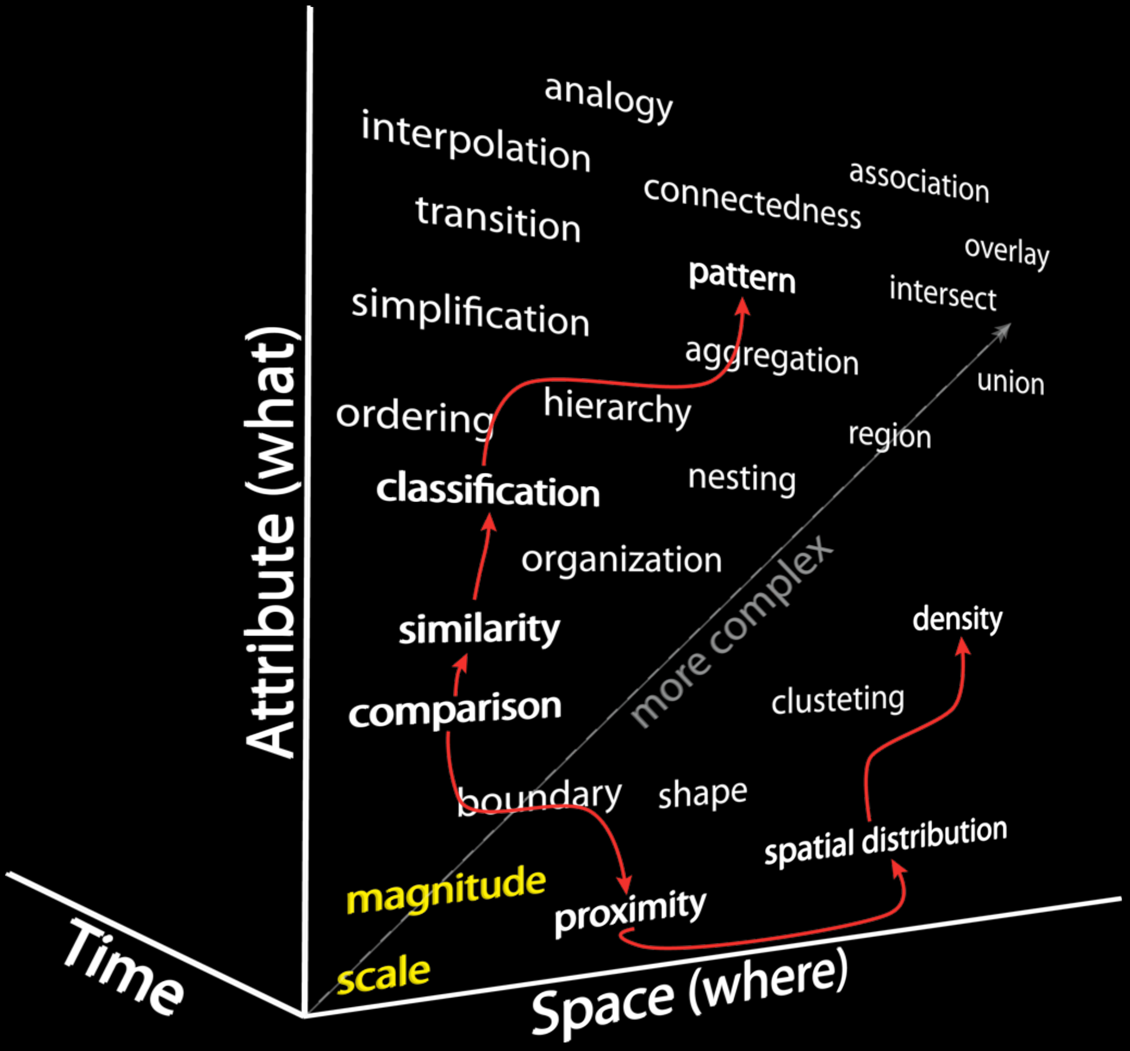
- which do we choose in our example:
    - face validity
    - content validity
    - criterion validity
    - construct validity

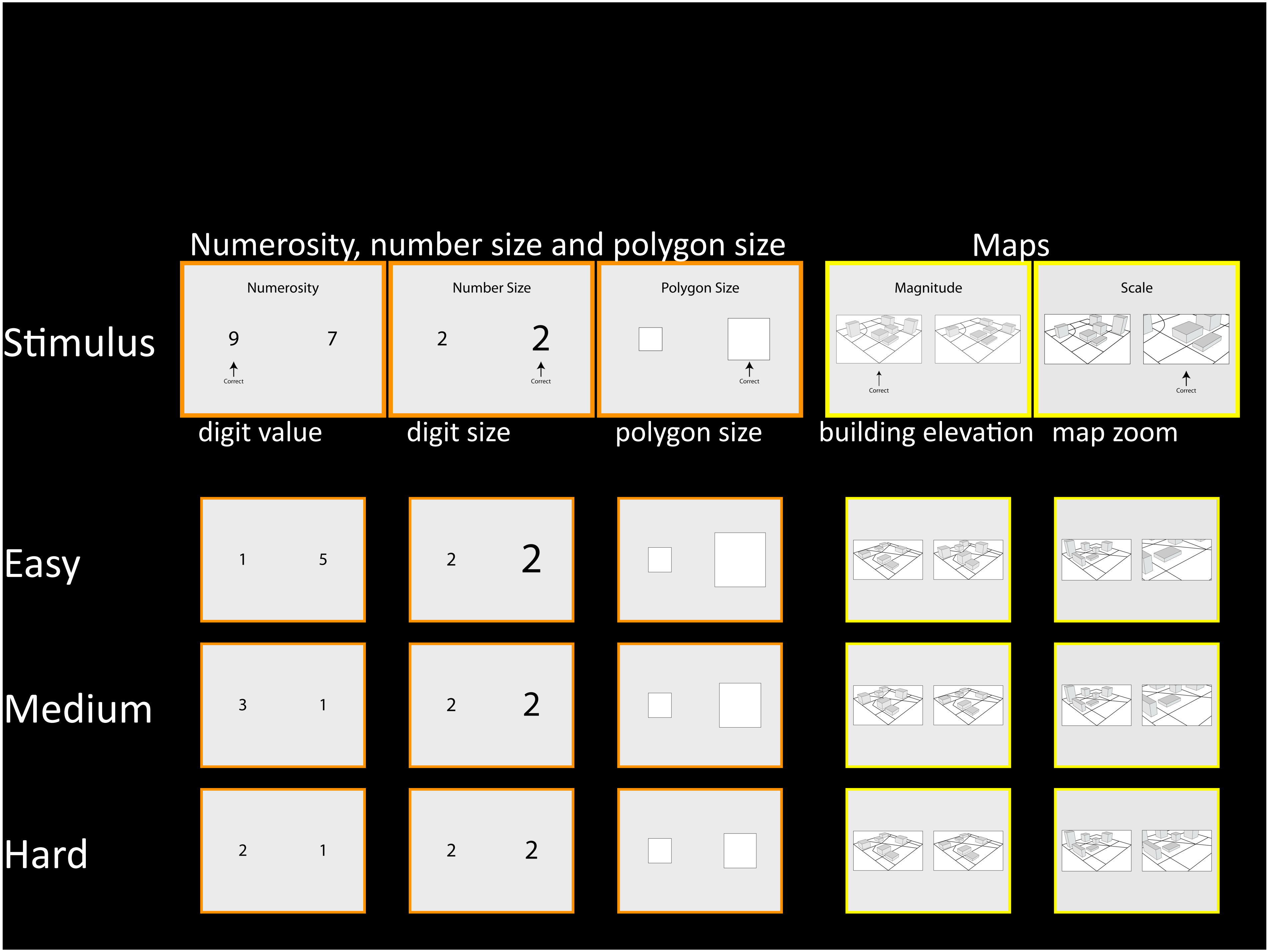
validity

- which do we choose in our example:
  - face validity
  - content validity
  - criterion validity
  - construct validity

**AND control for confounding variables

another example:





Numerosity, number size and polygon size

Maps

Stimulus

Easy

Medium

Hard

digit value

digit size

polygon size

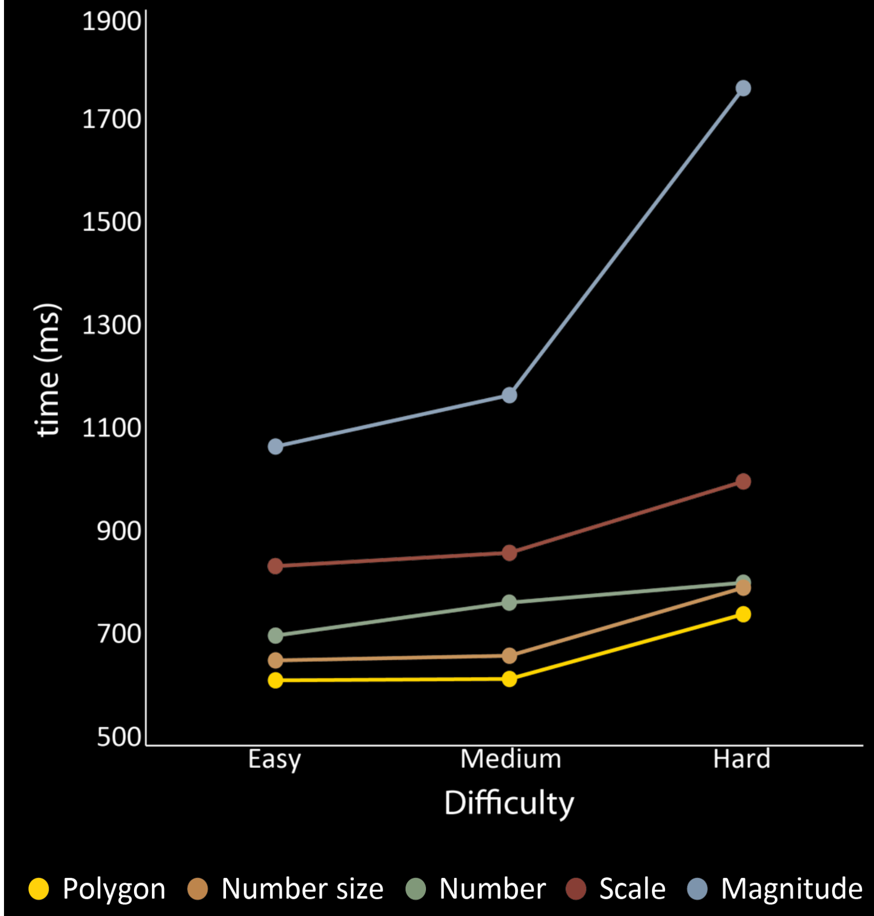
building elevation

map zoom

reliability analysis:

- internal consistency
    - computer-administered behavioral testing
    - 5 test sections, 5 measures of internal consistency

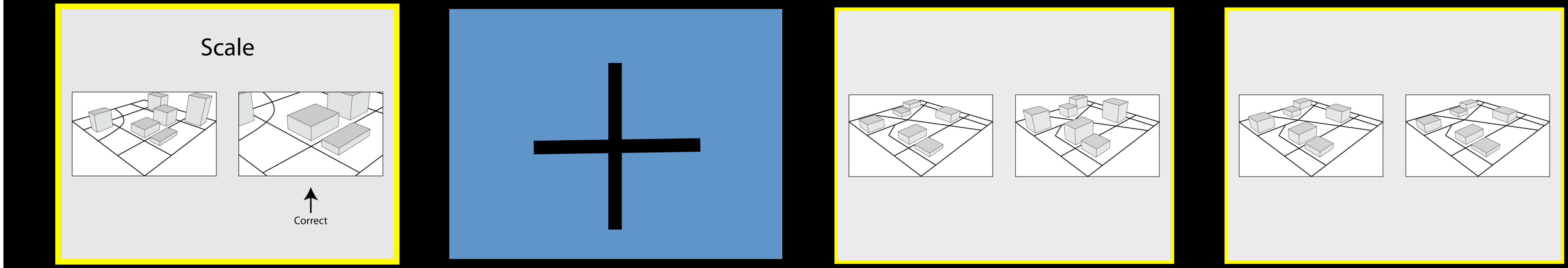
validity analysis (behavioral):



- Significant differences in reaction time by difficulty level within tasks (all p's < 0.001)
- Faster response times when differences in numeric and cartographic scale & magnitude are larger
- Results are consistent with previous research that identified task and difficulty level differences (Kadosh et al. 2005)
  - concurrent criterion validity

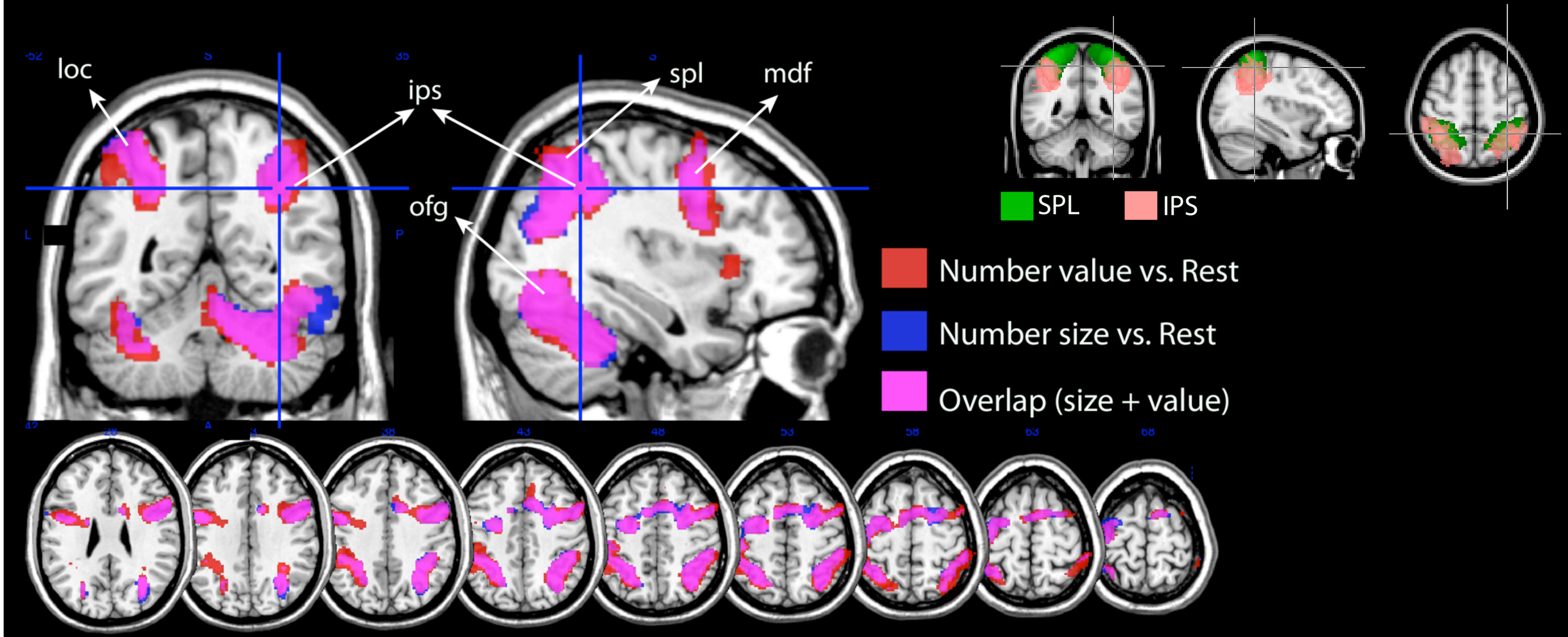
validity analysis (neuro):

- A total of 240 images were shown over 5 runs (48 images per run)
  - 24 images for each condition (12 for each difficulty)
  - Eliminated middle difficulty level (focused on easy vs. hard)
- Participants viewed a stimuli pair and reported which of the two images was larger
- Differences in BOLD were used to indicate encoding of scale and magnitude
  - **Data from numerosity & number size used as localizer for map data – again, looking for concurrent criterion validity**



6000ms 4000ms 6000ms 6000ms

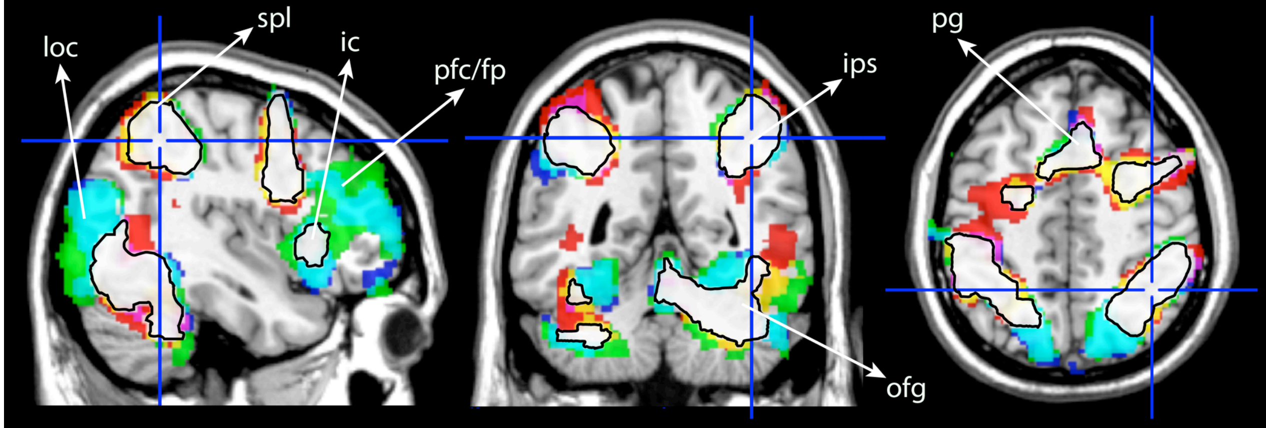
validity analysis (neuro): criterion



(all contrasts: $p < 0.05$ corrected)

- Previous research suggests that the IPS and the SPL are involved in numerical and physical magnitude comparisons (Dehaene et al., 2003; Kadosh et al., 2005)
- We have replicated previous findings by showing that magnitude comparisons of number value and number size activate the IPS & SPL

validity analysis (neuro): criterion

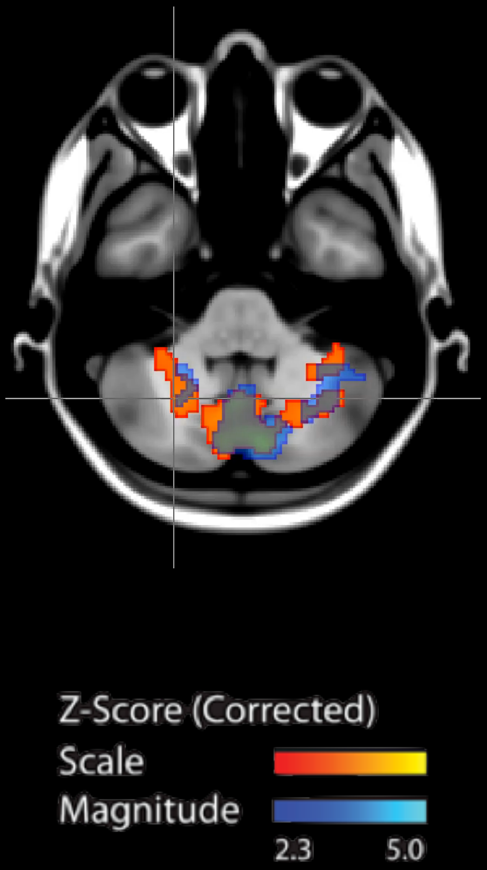


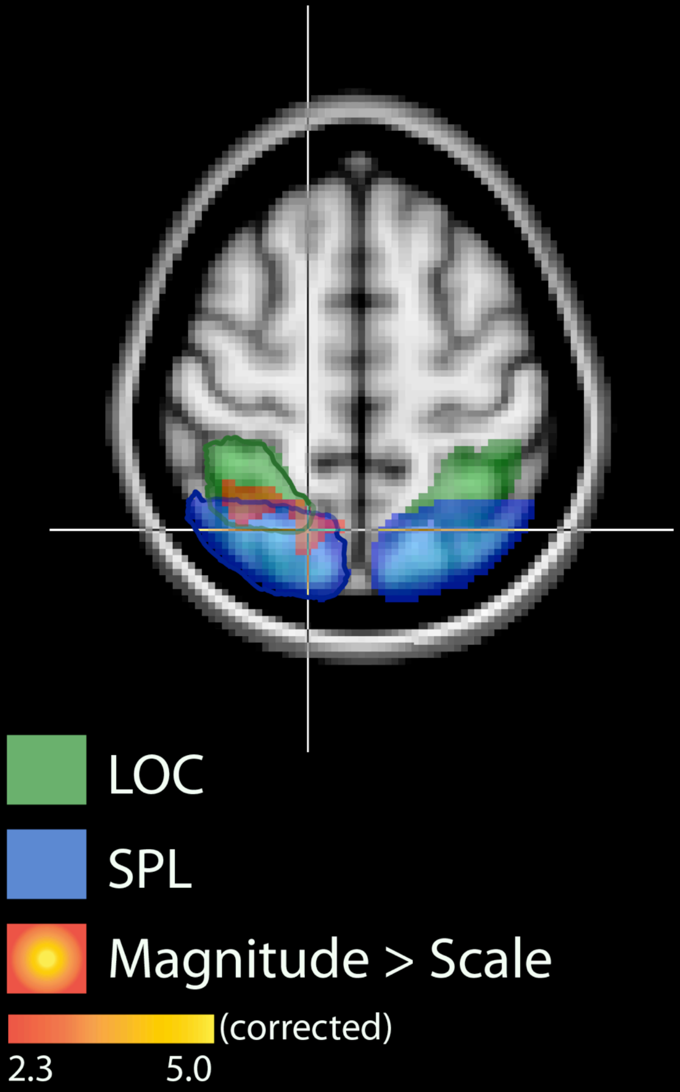
🟥 Number vs. Rest
🟦 Scale vs. Rest
🟩 Magnitude vs. Rest
🟪 Overlap (number + scale)
🟨 Overlap (number + Magnitude)
🟦 Overlap (scale + magnitude)
⬜ Overlap (number + Scale + Magnitude)

(all contrasts: $p < 0.05$ corrected)

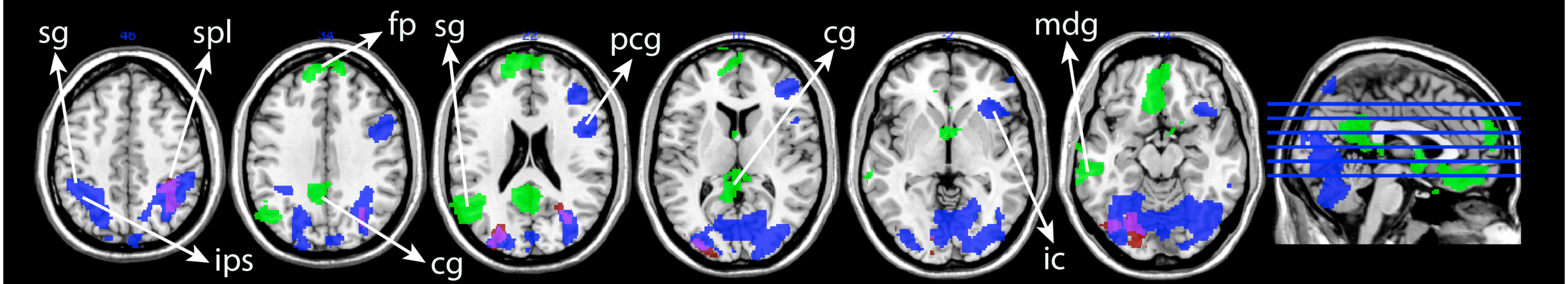
- Considerable overlap in the neural substrate between numerical, scale and magnitude comparison tasks Large overlap in the IPS & SPL between the three tasks Scale and magnitude tasks differentially activate a region in the LOC and PFC/FP Regions have previously been implicated in object recognition cognitive branching

validity analysis (neuro): construct





validity analysis (neuro): construct



🟩 Magnitude (easy vs. hard) 🟦 Magnitude (hard vs. easy) 🟥 Scale (hard vs. easy) 🟪 Overlap (hard vs. easy)

(all contrasts: $p < 0.05$ corrected)

- A whole brain analysis that contrasted task difficulty for maps revealed distinct networks for the magnitude condition with some overlap between scale and magnitude tasks

your challenge…

- design a protocol that does the following:
    - identifies the most effective substrate for tactile map symbols
    - 15 map symbols
    - 3 substrates
    - many facilitators