

# CRISSP Lecture Series



A program for experimental syntax:  
data, theory, and biology

Jon Sprouse  
University of Connecticut

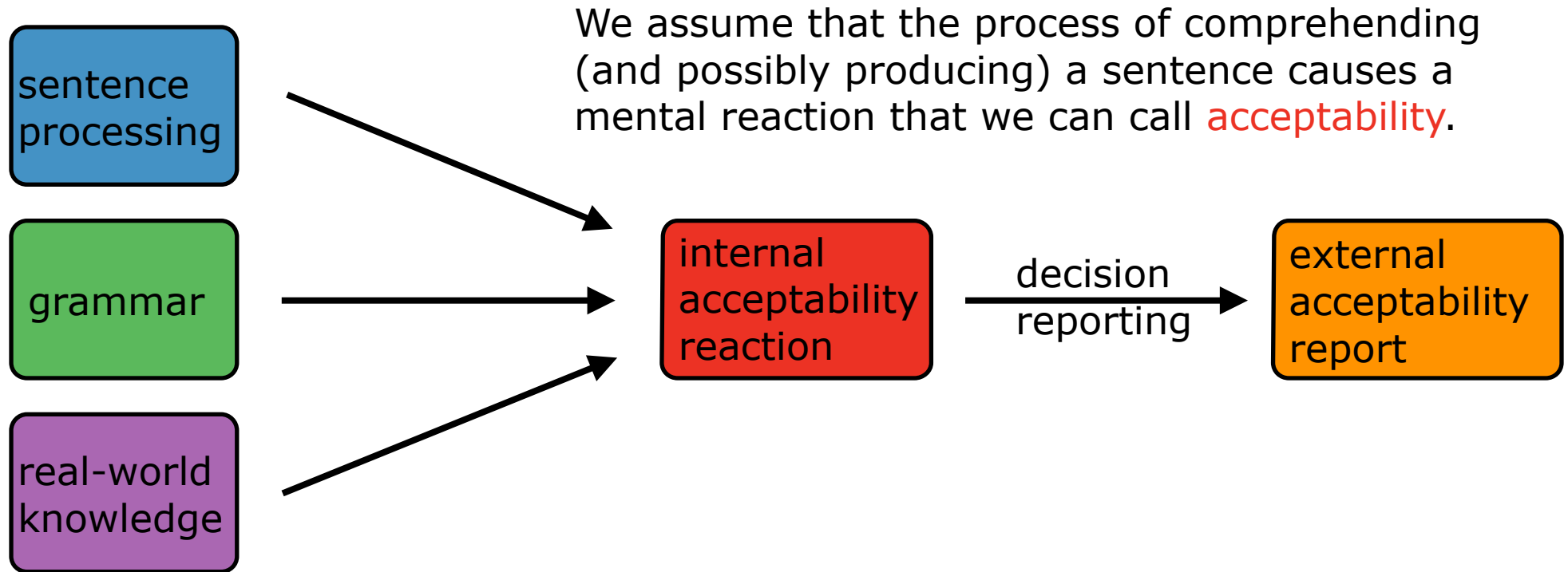
Data 1:

Experimental Syntax and the **validity** of  
acceptability judgments

KU Leuven - Brussels 03.16.15

# What are acceptability judgments?

To be honest, I don't think we have a full theory of what acceptability judgments are. But we do have some rough ideas:



This reaction is not a reaction to a single property (but it does appear to be a monolithic reaction). It is a consequence of **all of the cognitive systems** that contribute to sentence comprehension (and possibly production).

We then ask participants to **report this internal reaction**, which we call an **acceptability judgment**.

# How are judgments collected?

construct conditions that isolate the phenomenon of interest

create a **handful** of lexicalizations (<5)

ask **linguists** to give you ratings (<10)

using **forced-choice** or **yes-no** tasks

report an **intuitive analysis** using a diacritic

What do you think **that John bought** \_\_\_ ?

What do you wonder **whether John bought** \_\_\_ ?

What do you think that Amy cooked?

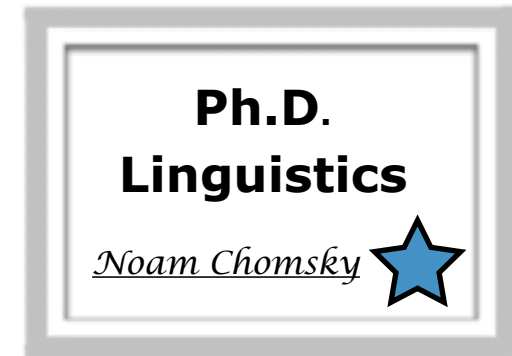
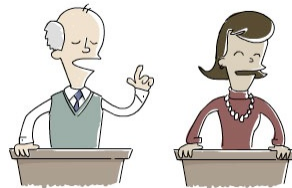
What do you wonder whether Amy cooked?

Who do you think that the dog bit?

Who do you wonder whether the dog bit?

What do you think that he read?

What do you wonder whether he read?



What do you think **that John bought** \_\_\_ ?

\*What do you wonder **whether John bought** \_\_\_ ?

# The concern prior to 2013

*But in recent developments in linguistics the intuitions have become **more and more subtle**, and more difficult for non linguists to intuit themselves or to accept. This disturbing development has led to the question of whether or not linguists' intuitions can be uncritically accepted as being valid and basic to the speech community.*

*Spencer 1973*

*When linguistics began, it made a great deal of sense that the primary data would be intuitions about whether sentences were grammatical or ungrammatical. The field needed to get off the ground, and the techniques used in other areas of cognitive science were hardly more sophisticated. Moreover the **contrasts were extremely clear**... Other areas of cognitive science have moved on to far more powerful methodologies... In addition, in formal syntax, **the intuitions are no longer uncontroversial**.*

*Ferreira 2005*

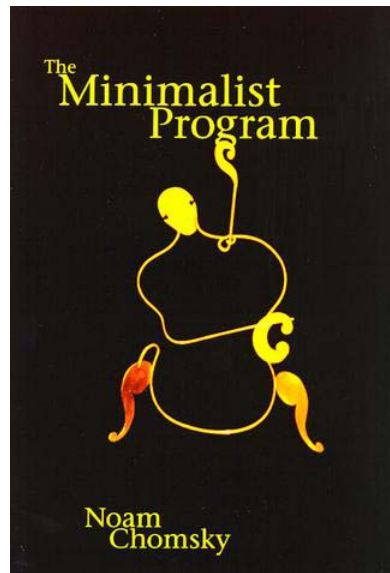
*The **lack of validity** of the standard linguistic methodology has led to many cases in the literature where questionable judgments have led to **incorrect generalizations and unsound theorizing**.*

*Gibson and Fedorenko 2012*

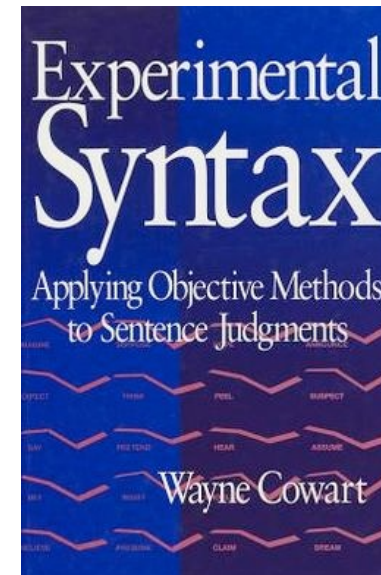
# So, we want a more formal method...

The concerns that have been raised about acceptability judgments set up a dichotomy between two methods. Let's call them the **informal method** that is traditionally used, and the **formal method** of experimental syntax.

Let's compare these methods to get a better sense of potential differences.



VS



# Here is what it looks like

What do you think **that John bought** \_\_\_ ?

What do you wonder **whether John bought** \_\_\_ ?

## Magnitude Estimation

reference sentence 100

target sentence \_\_\_\_\_

target sentence \_\_\_\_\_

target sentence \_\_\_\_\_

## Likert Scales

target sentence 1 - 7

target sentence 1 - 7

target sentence 1 - 7

target sentence 1 - 7

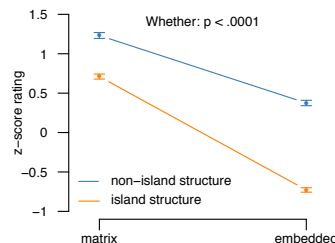
construct conditions that isolate the phenomenon of interest

create a **multiple** lexicalizations (**8+**)

ask **students (20+)** to give you ratings

using **numerical scaling** tasks

analyze the results using **inferential statistics**



# What is the nature of the differences?

construct conditions that isolate the phenomenon of interest

=

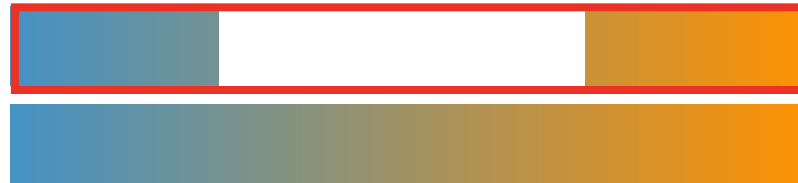
construct conditions that isolate the phenomenon of interest

create a **handful** of lexicalizations (<5)



create a **multiple** lexicalizations (8+)

ask **linguists** to give you ratings (<10)



ask **students (20+)** to give you ratings

using **forced-choice** or **yes-no** tasks



using **numerical scaling** tasks

report an **intuitive analysis** using a diacritic



analyze the results using **inferential statistics**

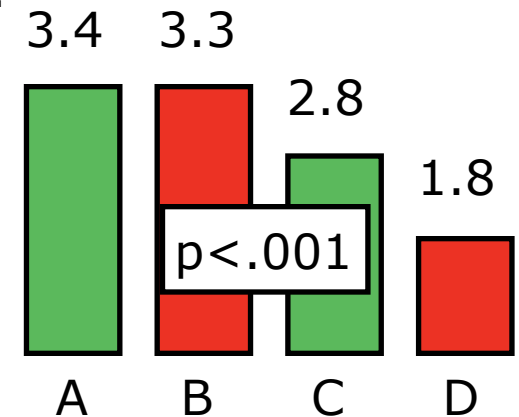
# The **evidence** prior to 2013

Gibson and Fedorenko 2012 (re)present five examples of judgments in the literature that do not appear to hold up to formal experimentation.

Two come from previously published papers by other authors:

**Wasow and Arnold 2005** (claim from Chomsky 1955):

- A. The children took **in** [NP all our instructions].
- B. The children took **in** [RC everything we said].
- C. The children took [NP all our instructions] **in**.
- D. The children took [RC everything we said] **in**.



This is exactly what W&A find!

**Langendoen, Kalish-London, and Dore 1973** (claim from Fillmore 1965):

- \*Who did you show \_\_\_ the woman?
- Who did you show the woman \_\_\_?

~22 responses

~87 responses

$p < .0000000018$

This is exactly what they find!



# The **evidence** prior to 2013



Gibson and Fedorenko 2012 (re)present five examples of judgments in the literature that do not appear to hold up to formal experimentation.

One is a phenomenon that was already tested and published, twice:

**Clifton and Frazier 2006:** Does a 3rd wh-word ameliorate a superiority violation (claim from Kayne 1983)?

*What can who do about it?	2.27	2.27	out of 5
?What can who do about it when?			

**Fedorenko and Gibson 2012:** same thing, but embedded questions, and with context

*Peter tried to remember what who carried.	3.72	3.76	out of 7
?Peter tried to remember what who carried when.			

Notice that nobody finds a result going the wrong direction. The two sentences are judged to be **equal**. As C&F 2006 point out, this is interesting given that adding wh-words typically lowers acceptability!

# The **evidence** prior to 2013

Gibson and Fedorenko 2012 (re)present five examples of judgments in the literature that do not appear to hold up to formal experimentation.

And two are new to this paper:

**Gibson 1991:** a parsing preference from his dissertation

\*The man that the woman the dog bit likes eats fish.

?I saw the man that the woman that the dog bit likes.

This is a parsing preference, not a grammatical claim!

**Chomsky 1986:** something about Superiority

\*I wonder what who saw

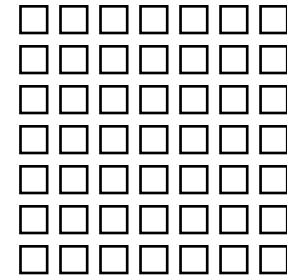
What do you wonder who saw?

These two sentences are not a minimal pair, nor are they presented as such in the text.

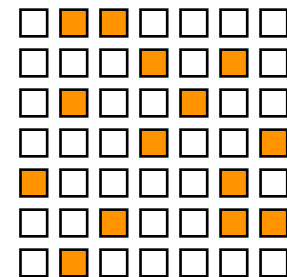
# The **problem** with the evidence prior to 2013

**First problem:** Populations, samples, sampling methods, and sampling bias

A **population** is the set of all items that meet some criterion. It could be something like “all humans on earth”, or something smaller like “all data points published in LI”.

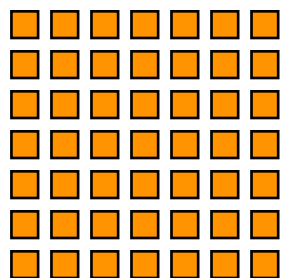


A **sample** is a finite subset of a population. It can be any size from 1 up to the size of the population itself.

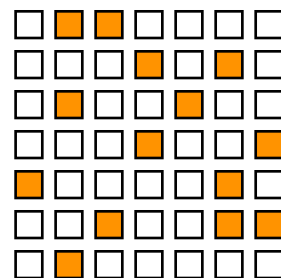


The **sampling method** is the method you use to select your sample:

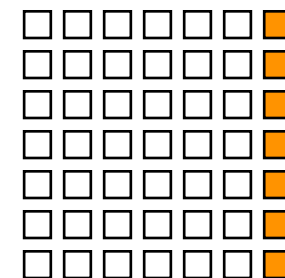
**exhaustive**



**random**



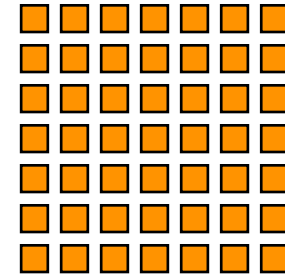
**biased**



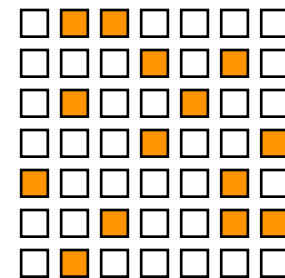
# The **problem** with the evidence prior to 2013

**First problem:** Populations, samples, sampling methods, and sampling bias

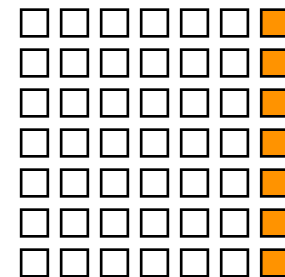
**Exhaustive sampling** gives you **perfect knowledge** about your population. It is often impractical.



**Random sampling** allows you to **estimate** the properties of your population. We say that you can statistically (mathematically) **generalize** from the sample to the population. This is what pollsters do during elections. The results have a margin of error.



**Biased sampling** doesn't allow you to make any statistical (mathematical) claims about the population.



Notice that no method lets you make claims about other populations, but that is ok, because the population is typically your object of interest.

# The **problem** with the evidence prior to 2013

**First problem:** Populations, samples, sampling methods, and sampling bias

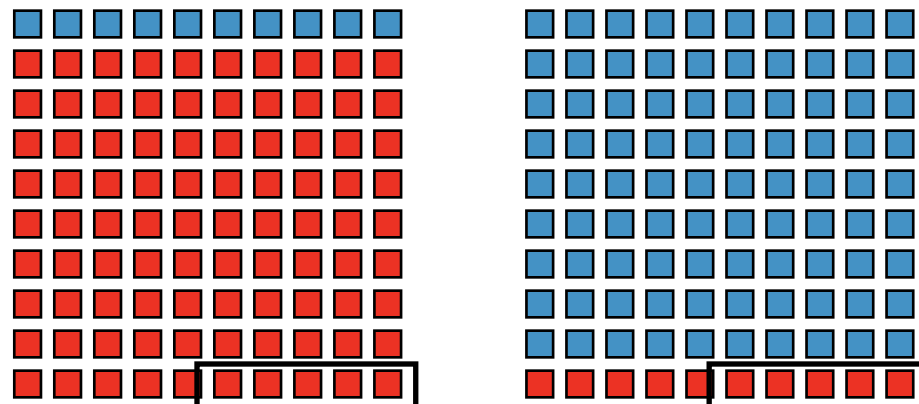
To my knowledge, all evidence presented prior to 2013 was based on **biased sampling**.

So, even if we assume that the 5 examples in G&F2012 are evidence of a problem with informal methods (which they aren't), we can't use these examples to make claims about the full population of evidence in linguistics.

Basically, those 5 examples could come from very different populations:

■ phenomena for which informal methods are **correct**

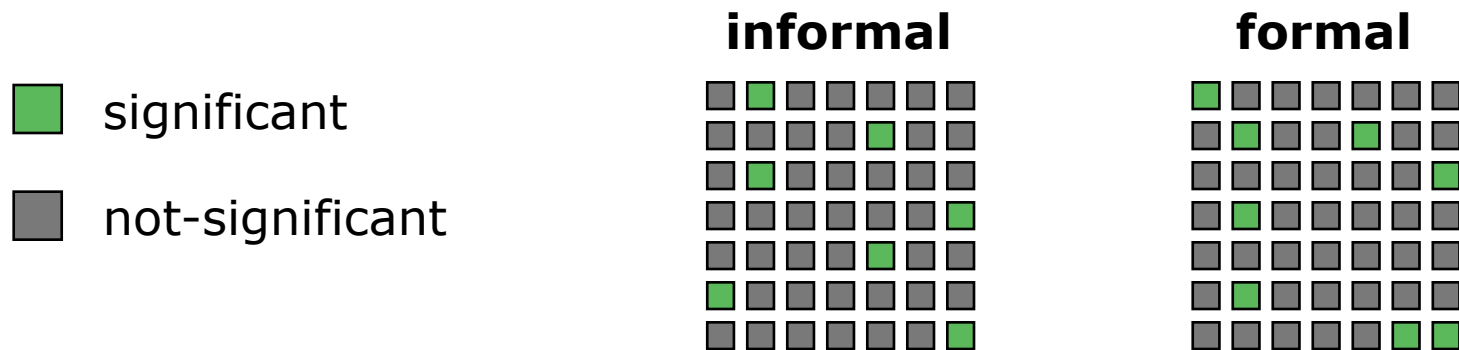
■ phenomena for which informal methods are **incorrect**



# The **problem** with the evidence prior to 2013

**Second problem:** Differences are just differences

Let's say that you have a sample of phenomena, and you run them twice: once with informal methods, and once with formal methods.



Which one is a **better reflection of the universe**? (Notice I didn't say correct/true, because we can never know that!)

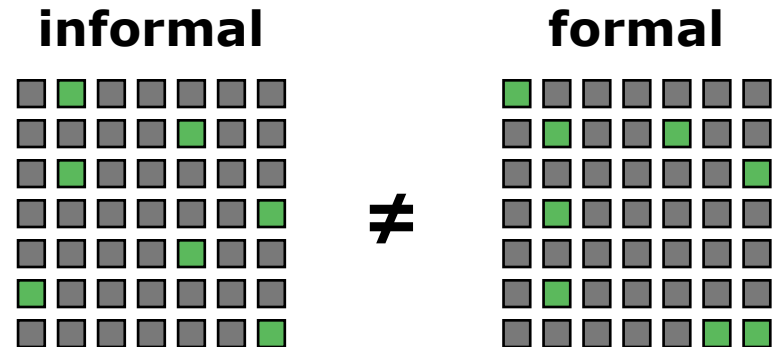
G&F2012 would say that the formal results are correct. But where is the **evidence** for this? Which aspect of the experiment demonstrated that the formal results are the correct ones?

Nothing in the experiment provided evidence for this. It is entirely based on their **prior belief** that formal experiments are superior to informal experiments.

# The **problem** with the evidence prior to 2013

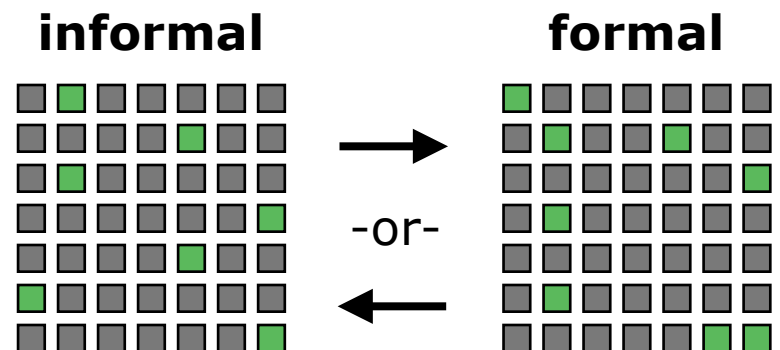
**Second problem:** Differences are just differences

Experiments that compare methods **only provide evidence that there is a difference**, they do not provide evidence about which method is a better reflection of reality.



If you really wanted to gather evidence for the superiority of one measure over another, you would have to do the following:

1. Identify the phenomena where the two methods diverge:
2. Postulate a hypothesis about the mechanism that **causes** one method to be superior.
3. Manipulate that mechanism and see if it makes the results converge.



# Correcting these problems

**Step 1:** Unbiased selection of phenomena

**Biased selection:**

Biased sampling is when the selection is not random. Biased selection allows **no statistical generalization** beyond the sample. So we know nothing about the population the sample came from. (In theory one could make non-statistical arguments for generalization, but that is rarely done, and is highly subjective.)

**Exhaustive selection:**

This means to select every phenomenon there is. This is probably impractical on a large scale, but potentially possible for smaller bodies of work (books or journal volumes). Exhaustive selection provides **perfect knowledge** of the population selected from. (It doesn't allow generalizability beyond the population, but then again, no method does.)

**Random selection:**

Randomly sampling phenomena from a body of work (a population) allows us to **statistically estimate** a convergence rate for the population. The estimate comes with a margin of error.



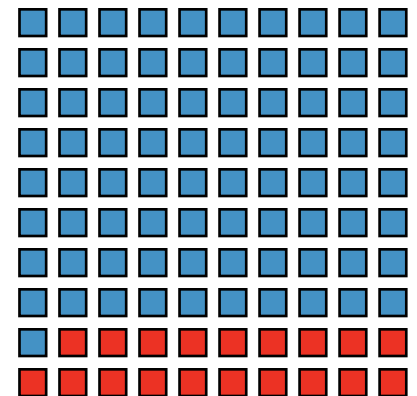
# Correcting these problems

**Step 2:** Focus on convergence/divergence rates

**Convergence rate:** The proportion (or percentage) of effects that are detected by both informal and formal methods.

**Divergence rate:** The proportion (or percentage) of effects that are detected in informal methods, but not formal methods.

The idea behind this is that it will tell us how many phenomena would be removed from the theory if we threw away informal methods, and replaced them with formal methods. It is an estimate of the size of the consequences of the debate.



I should note that this won't be a complete picture. We could also look at effects that weren't detected by informal methods, and ask if formal methods would detect an effect. We could also ask about effects that aren't detected by formal methods, and ask if informal methods would detect them. These are harder questions because it is hard to list undetected effects, so the first step is to focus on effects that have been detected and are in the literature.

# Two experiments

## Random selection:

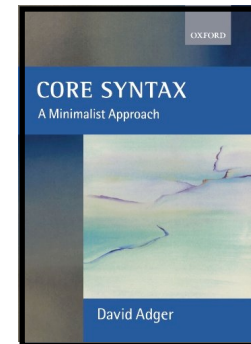
Randomly sampling phenomena from a body of work (a population) allows us to **statistically estimate** a convergence rate for the population. The estimate comes with a margin of error.



**2001-  
2010**

## Exhaustive selection:

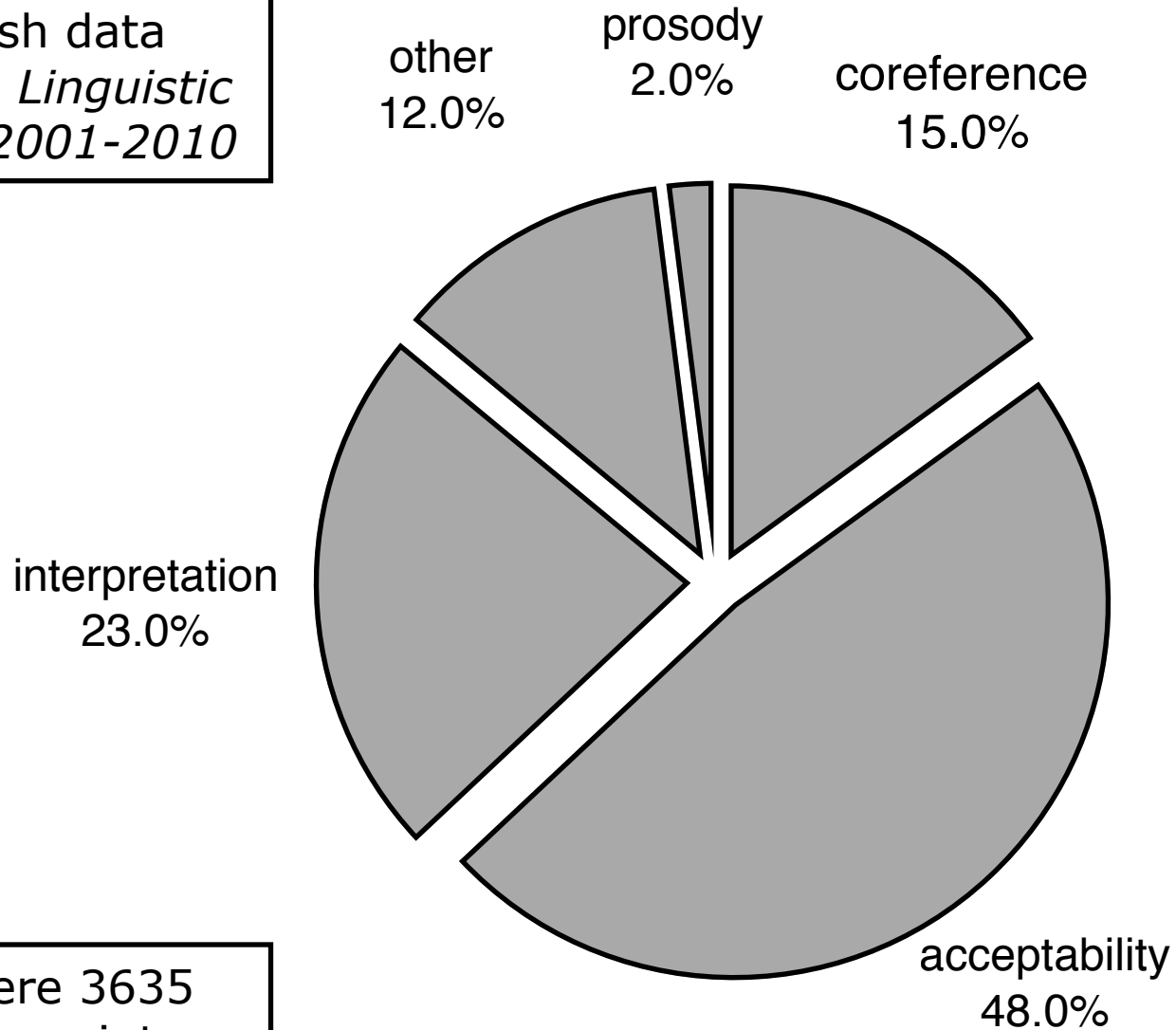
This means to select every phenomenon there is. This is probably impractical on a large scale, but potentially possible for smaller bodies of work (books or journal volumes). Exhaustive selection provides **perfect knowledge** of the population selected from.



**Adger  
2003**

# Step 1: Defining a population

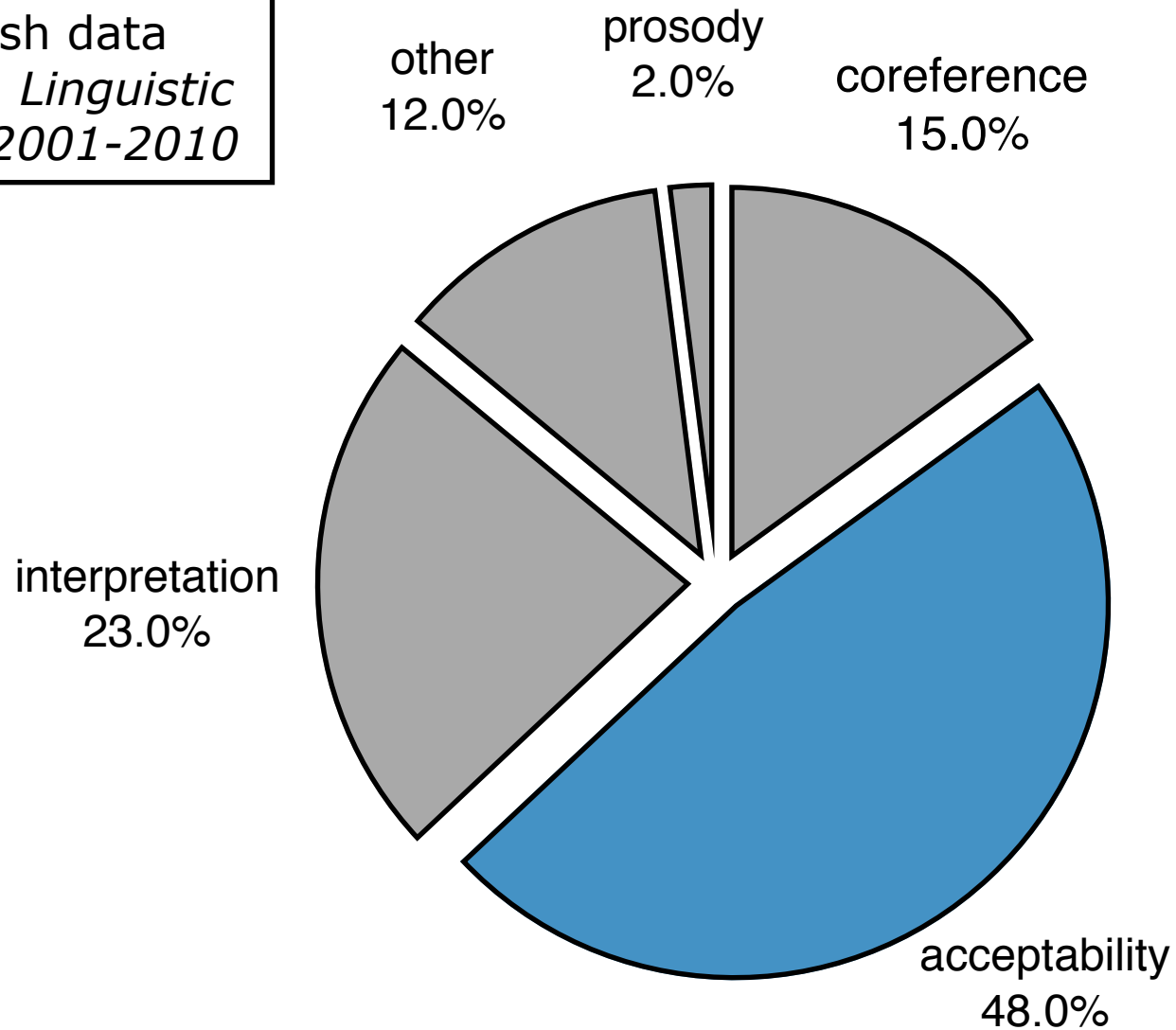
US English data  
points in *Linguistic  
Inquiry 2001-2010*



There were 3635  
total data points  
published over this  
10 year period

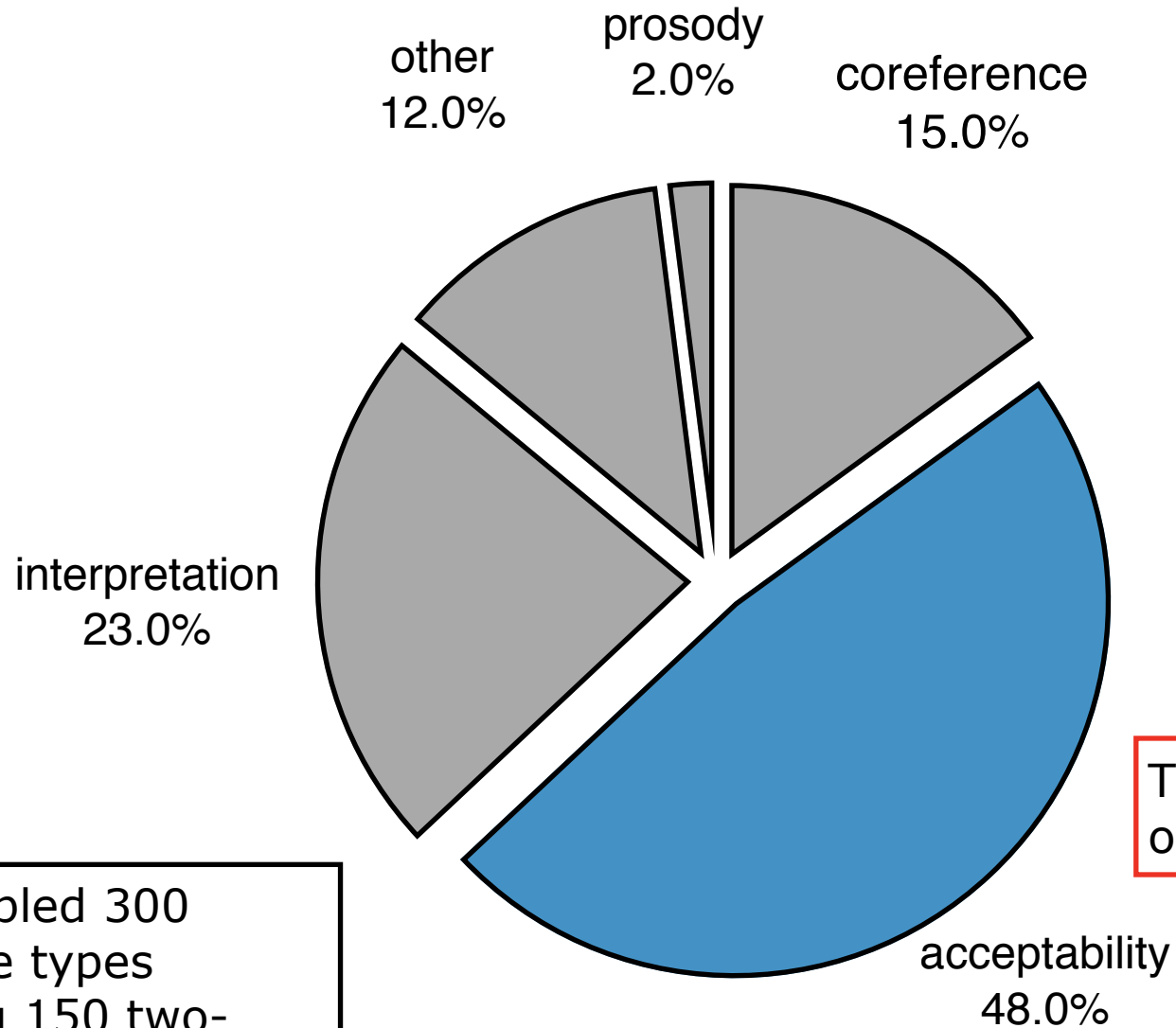
# Step 1: Defining a population

US English data  
points in *Linguistic  
Inquiry 2001-2010*



1743 data points

## Step 2: Choosing a sample size



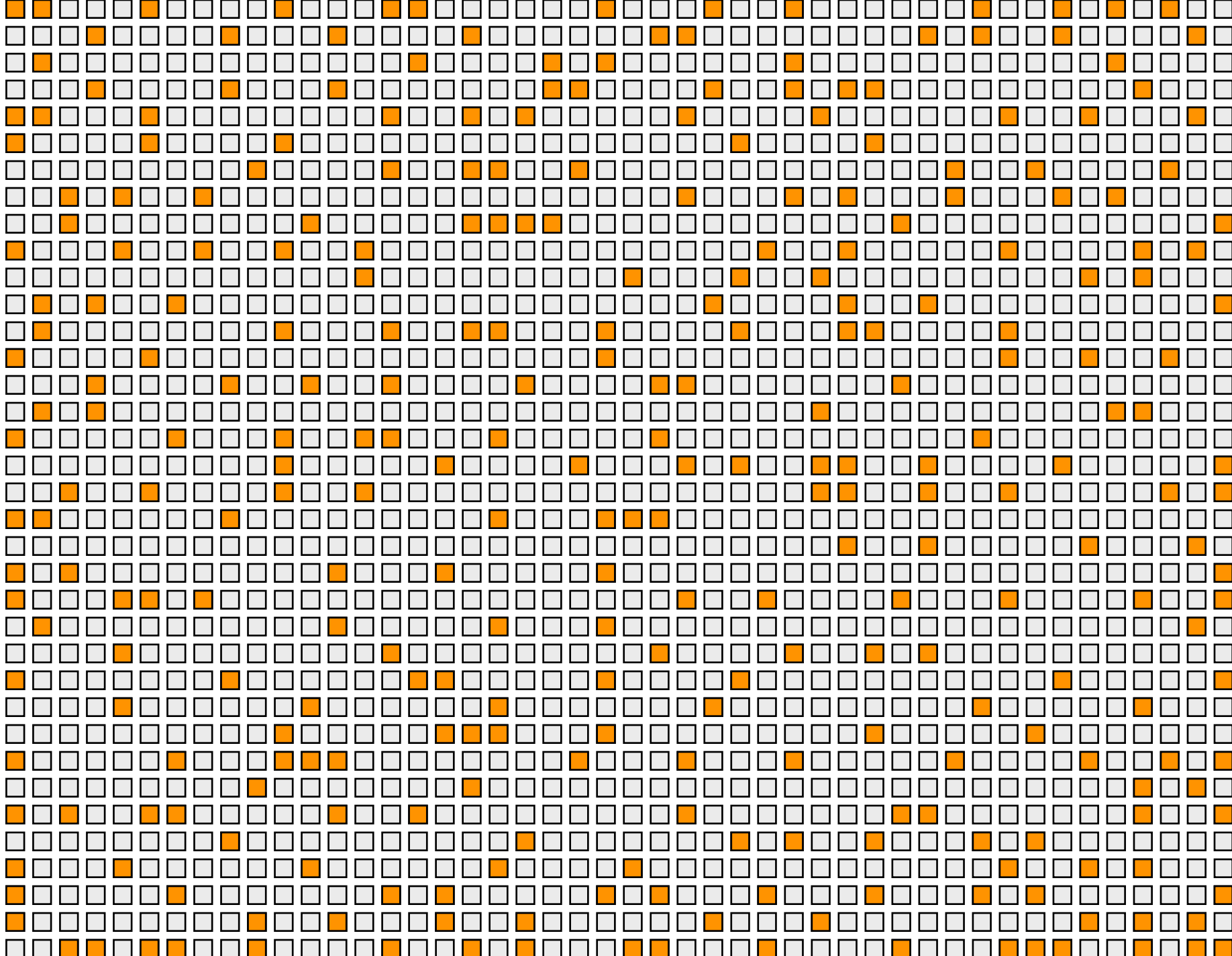
We sampled 300 sentence types (forming 150 two-condition phenomena)

This yields a margin of error of  $\pm 5$





**2001-  
2010**



# Step 3: Creating the experiment

**150 pairwise phenomena**

What do you think **that John bought** \_\_\_ ?

**x8 tokens**

What do you wonder **whether John bought** \_\_\_ ?

**2001-  
2010**

**Latin Square**

A	B	C
C	A	B
B	C	A

+

**Pseudo-  
randomization**

**Magnitude Estimation**

reference sentence 100

target sentence \_\_\_\_\_

**Likert Scales**

target sentence 1 - 7

target sentence 1 - 7

**Forced Choice**

sentence A 0

sentence B 0

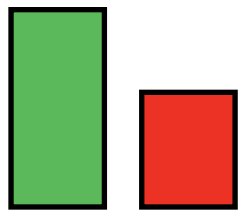


**x936**

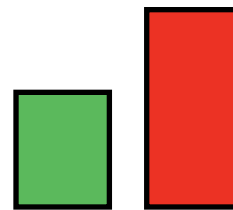


# Step 4: Defining Convergence

All of the phenomena tested were reported in the informal literature as having one condition more acceptable than the other. This means we can define (at least) 6 possible results in the formal experiments:



statistically significant  
in the correct direction



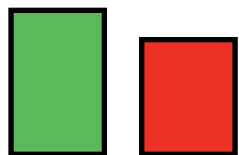
statistically significant  
in the opposite direction



marginally significant  
in the correct direction



marginally significant  
in the opposite direction



non-significant  
in the correct direction



non-significant  
in the opposite direction

The question is which of these 6 results are we going to count as converging with the informal results?

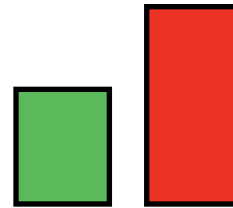
# Step 4: Defining Convergence

## Conservative Definition:

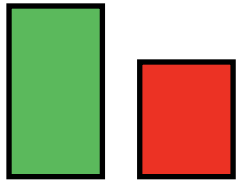
By conservative, I mean the definition that returns the smallest number of converging phenomena.



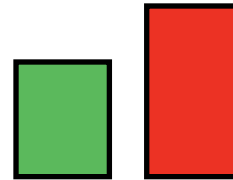
statistically significant  
in the correct direction



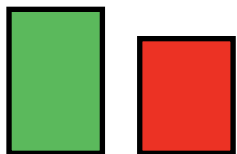
statistically significant  
in the opposite direction



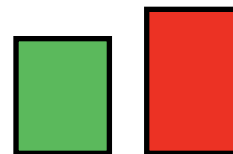
marginally significant  
in the correct direction



marginally significant  
in the opposite direction



non-significant  
in the correct direction



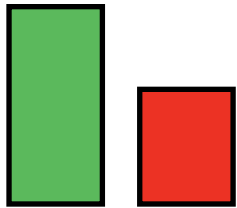
non-significant  
in the opposite direction

The most conservative measure would be to only count statistically significant in the correct direction; call everything else divergence.

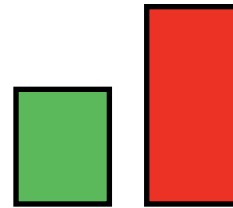
# Step 4: Defining Convergence

## **Liberal Definition:**

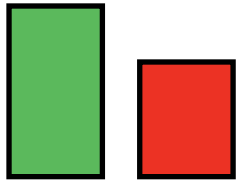
By liberal, I mean the definition that returns the largest number of converging phenomena.



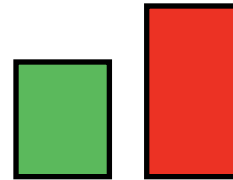
statistically significant  
in the correct direction



statistically significant  
in the opposite direction



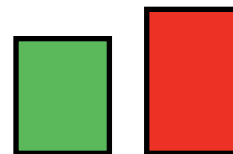
marginally significant  
in the correct direction



marginally significant  
in the opposite direction



non-significant  
in the correct direction



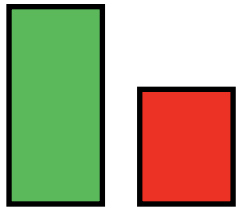
non-significant  
in the opposite direction

The most liberal measure would be to count everything except significant in the opposite direction as convergence.

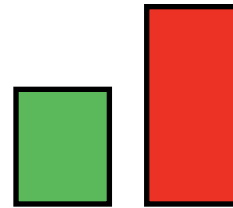
# Step 4: Defining Convergence

## **Trend-based Definition:**

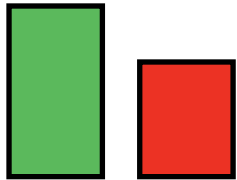
A middle ground is to count the trends - if the means are in the correct direction, it is convergence.



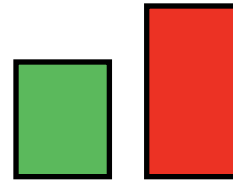
statistically significant  
in the correct direction



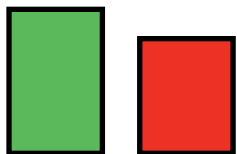
statistically significant  
in the opposite direction



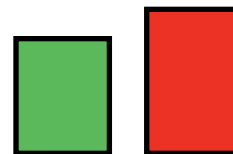
marginally significant  
in the correct direction



marginally significant  
in the opposite direction



non-significant  
in the correct direction



non-significant  
in the opposite direction

This method admits that statistical significance is just another way of saying large versus small effect (relative to spread).

# Step 5: Defining beliefs ahead of time

One thing that is absolutely critical in order to make valid inferences from data is to define the inference rules **before** we see the results of the experiment.

An inference rule for our current experiment would be something like:

If the convergence rate is above/below X%, conclude Y.

Most likely, we will only want to make one of two conclusions:

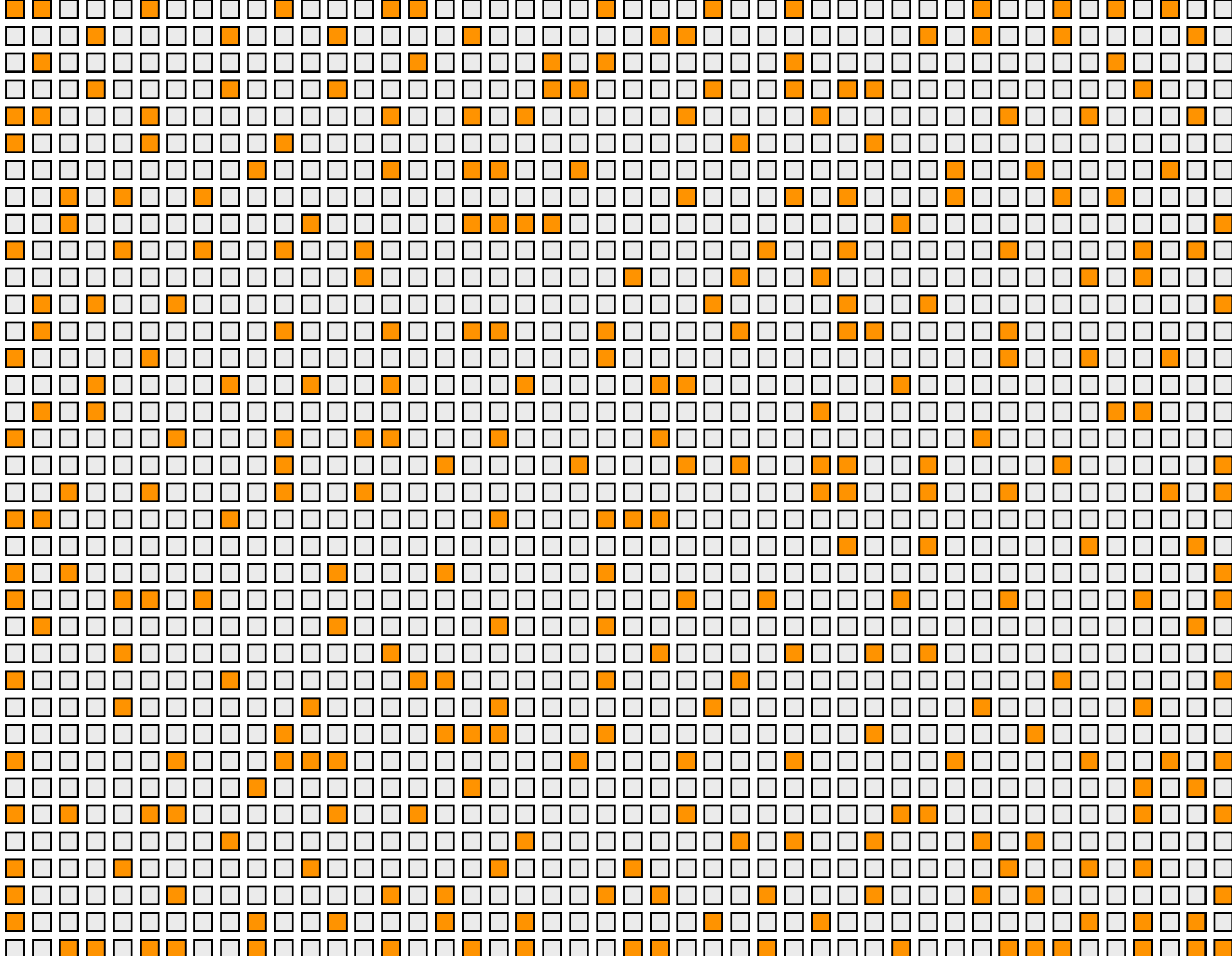
If the convergence rate is **above** X%, conclude that the two methods are close enough that they should both be considered valid (when there is no reason to be concerned about the results).

If the convergence rate is **below** X%, conclude that the two methods are sufficiently different that we need to run follow-up studies to determine which one provides a more accurate reflection of reality.

**So what are the values for X?**



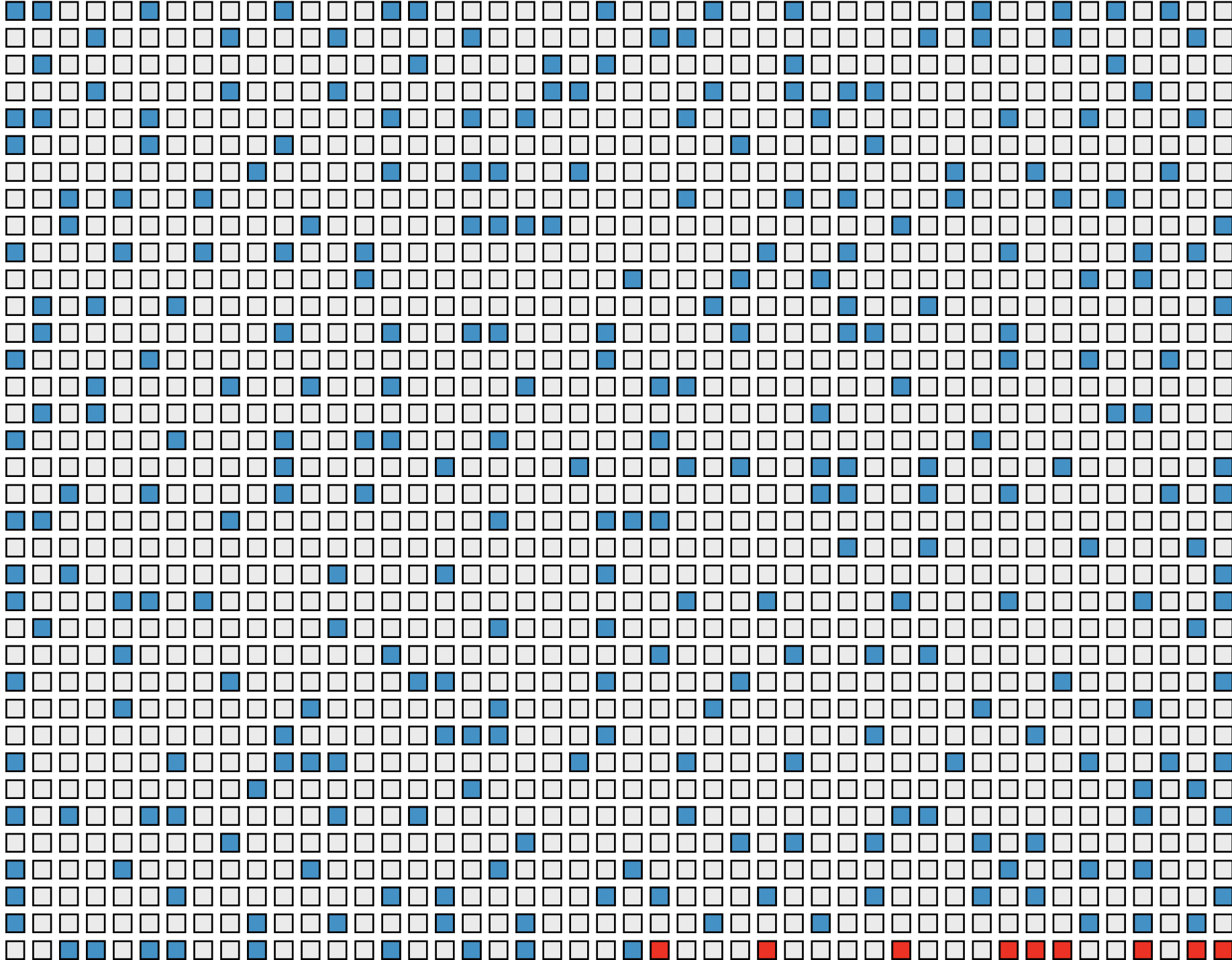
**2001-  
2010**



# Conservative Rate: 95% ± 5



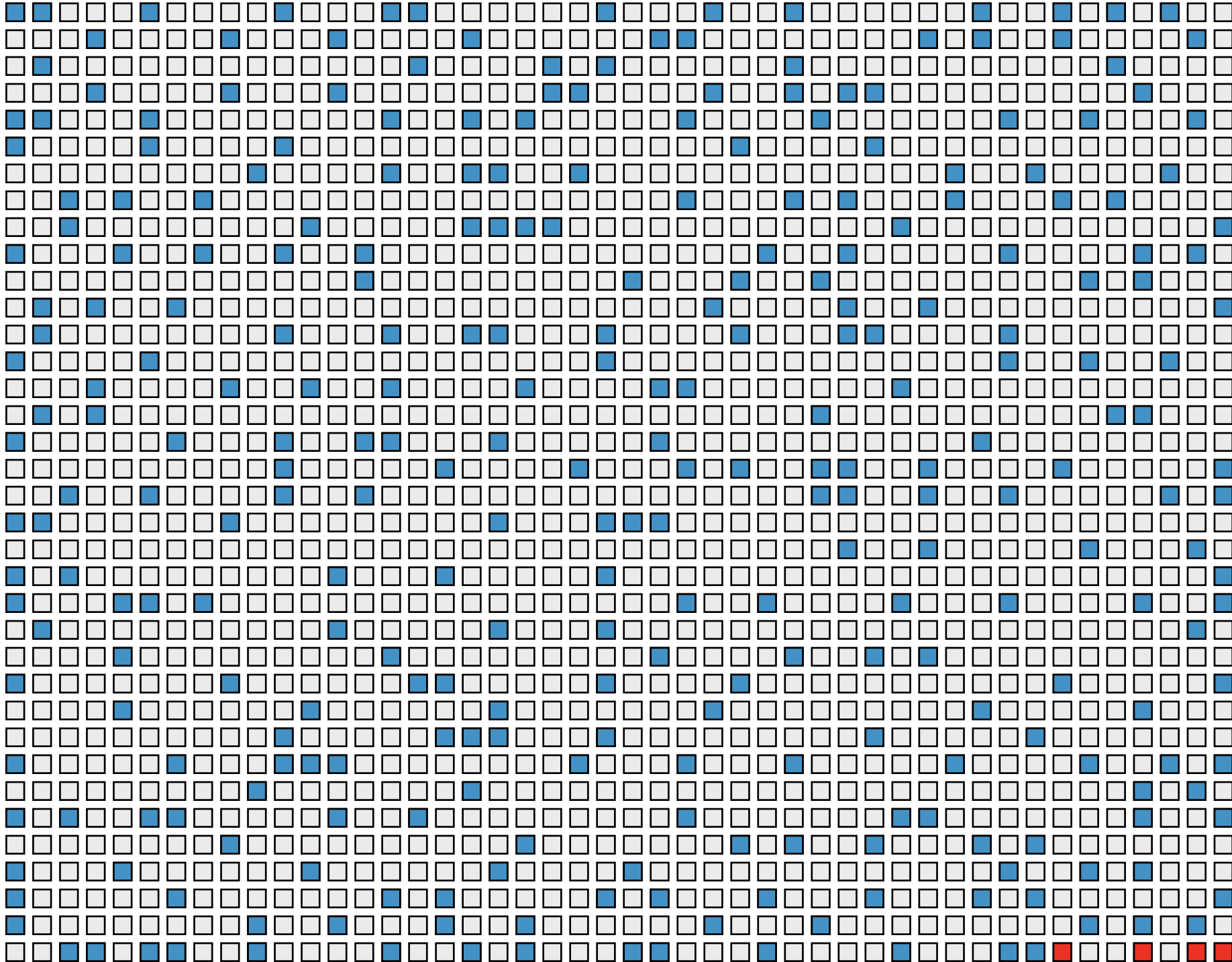
2001-  
2010



**Liberal Rate: 97% ± 5**



**2001-  
2010**

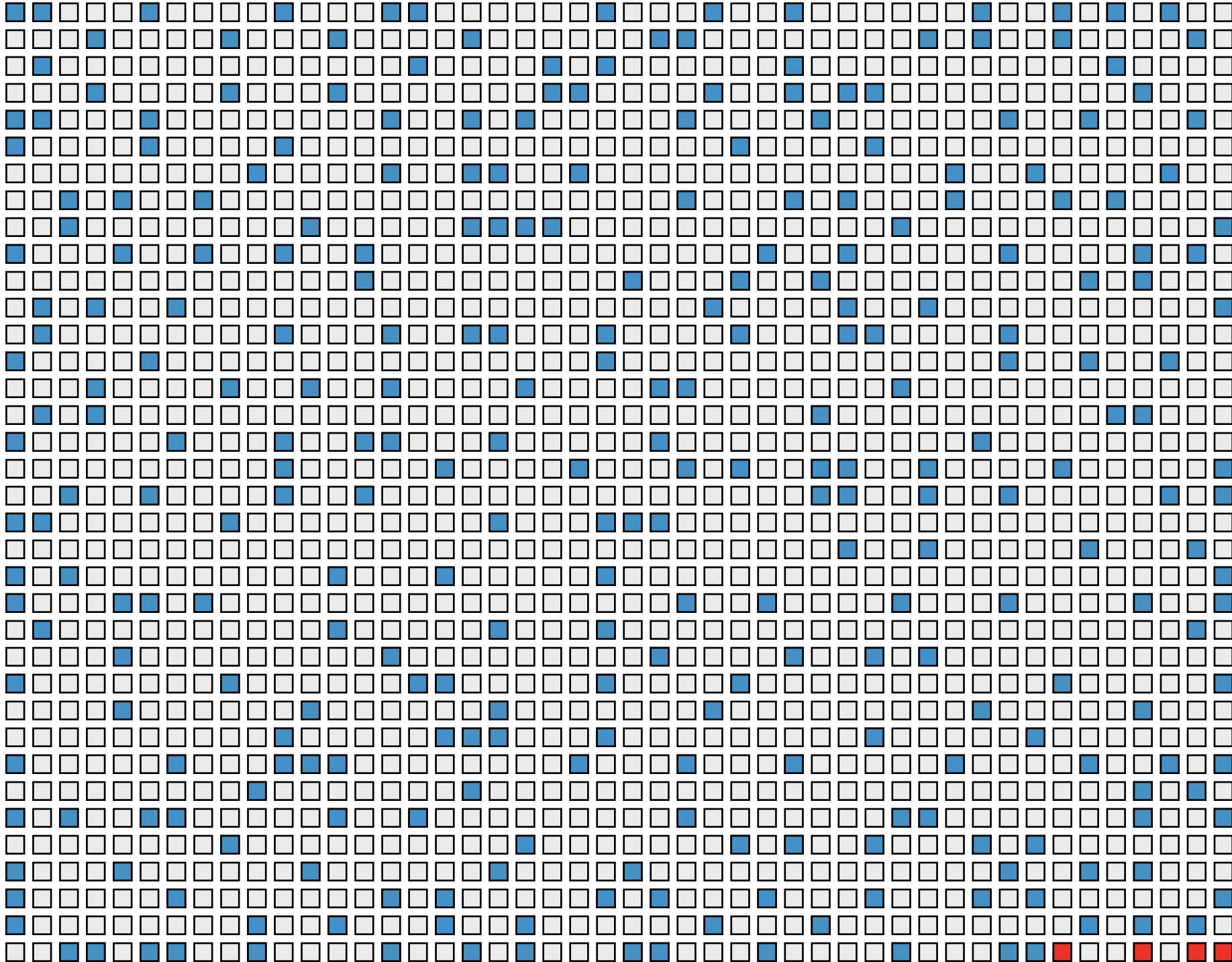




**Trend Rate: 97% ± 5**

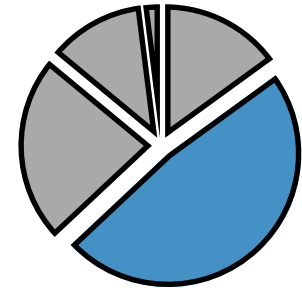


**2001-  
2010**



# What can we conclude from these rates?

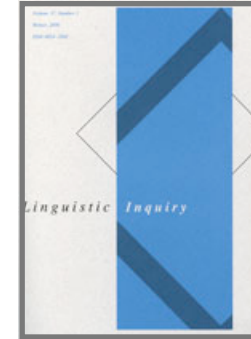
1. We can statistically generalize to the population within a margin of error. No previous study has ever been able to do that.
2. We can state exactly how big this debate is for the population of interest. It is 3%-5%, with a 5 point margin of error.
3. Is 3%-5% big enough to worry? That is a personal question.
4. We can't attach a probability to the data types outside our population. But we can use inductive (Bayesian) inference and say that the likelihood of rampant problems is smaller now that we've tested 48% of the data.
5. We can say something about theoretical bias. If linguists are using their theoretical knowledge to give their judgments, then we would expect a substantial number of direction-reversals. But we don't see many at all (at most 4). This suggests that theoretical bias is not a rampant problem.



# Two experiments

## Random selection:

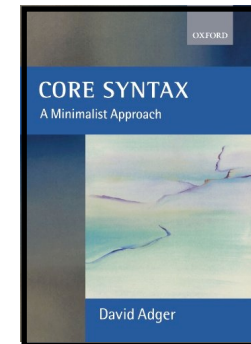
Randomly sampling phenomena from a body of work (a population) allows us to statistically estimate a convergence rate for the population. The estimate comes with a margin of error.



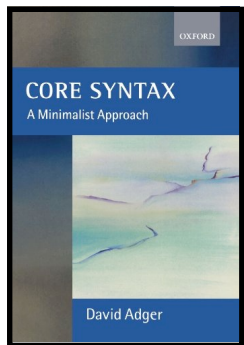
**2001-  
2010**

## Exhaustive selection:

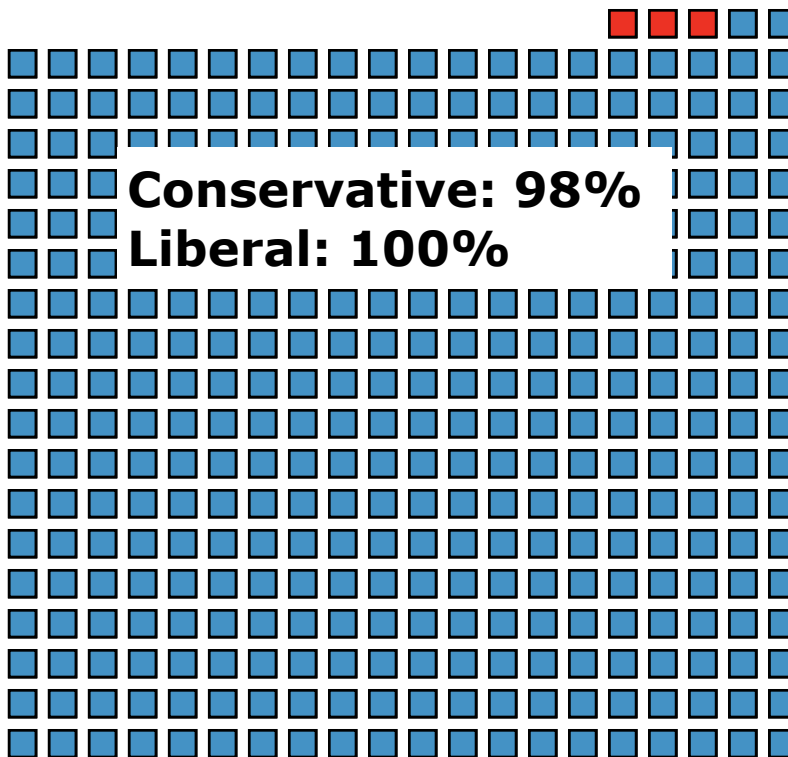
This means to select every phenomenon there is. This is probably impractical on a large scale, but potentially possible for smaller bodies of work (books or journal volumes). Exhaustive selection provides **perfect knowledge** of the population selected from.



**Adger  
2003**



**Adger  
2003**



**x8 tokens**

### Latin Square

A	B	C
C	A	B
B	C	A

+

**Pseudo-  
randomization**

<b>Magnitude Estimation</b>
reference sentence 100
target sentence _____

<b>Yes - No</b>
target sentence Y or N
target sentence Y or N



**x440**

# What can we conclude from these rates?

1. To the extent that the theory constructed in this textbook is predicated upon the data, we can say that the theory is well founded.
2. This doesn't mean the theory is correct. Any given data set is compatible with an infinite number of theories. This is just one theory that this data set is compatible with. But the theory does rest on a valid empirical foundation.
3. Some people have the opinion that this text book data will be more convergent than journal data, and it is, by a small amount. This is interesting because the only reason to believe that there should be a difference is the assumption that linguists police the data over time — identifying robust data and discarding less robust data, at least when it comes to constructing the core of the theory. That undermines the entire claim that formal methods are necessary to police the data.

# What is the current state of the topic?

## **The claim from perfection:**

Gibson (G&F2013, various pc) has made the argument that a 5% divergence rate is a **major problem**. The argument goes like this: if 5% of your data points are incorrect, you need to know which ones are incorrect.

For example, if you have a theory that explains 80 data points, and 4 of them are incorrect, then there are **1.6 million combinations** of 76 correct and 4 incorrect data points. Assuming 1 grammar per combination, that is 1.6 million possible grammars.

## **The problem with this claim:**

First, this argument assumes that we can know which data points are correct and which are incorrect. But we can't. We can increase our **confidence** in their correctness, but we can never know the truth.

Second, this argument assumes that the formal experiments reveal the (better) truth. It **conflates convergence with correctness**. No one has shown (experimentally) that formal experiments are superior to informal experiments.

# What is the current state of the topic?

## **The claim from utility:**

Gibson (G&F2013, various pc) has made the argument that formal experiments are superior to informal experiments because they yield various bits of **quantitative information** that informal experiments do not, such as effect sizes, distributions, and confidence intervals.

## **The problem with this claim:**

Everybody agrees with this! This is not the issue under debate.

Experimental syntacticians have been using formal experiments for years because of these reasons (and the rest of my lectures here will show various examples of this. But the original debate was about the necessity of formal experiments.

But the original debate was about the necessity of formal experiments:

*The **lack of validity** of the standard linguistic methodology has led to many cases in the literature where questionable judgments have led to **incorrect generalizations and unsound theorizing**. (G&F 2012)*

# What is the current state of the topic?

## **The claim from the shrinking corner:**

Gibson (G&F2013, various pc) has made the argument that we've only tested standard acceptability judgments, so the problem must be in other data types: interpretation judgments, coreference judgments, etc.

## **The problem with this claim:**

This shows that their **strong prior beliefs were not updated** on the new evidence. They are maintaining the same claim, but retreating into the untested portion of the data.

The problem with this is that those **prior beliefs were not based on valid evidence**. All previous evidence was based on biased sampling. These beliefs should not be strong enough to completely overwhelm the new evidence.

Also, even though we can't make a frequentist statistical argument for generalization beyond our population, we can make a **Bayesian argument**. Whatever the cause of bad informal judgments is (e.g., theoretical bias), it will most likely be operative for all data types. Therefore the 48% we tested provides **some amount of inductive evidence** that the remaining 52% are less likely to be contaminated.



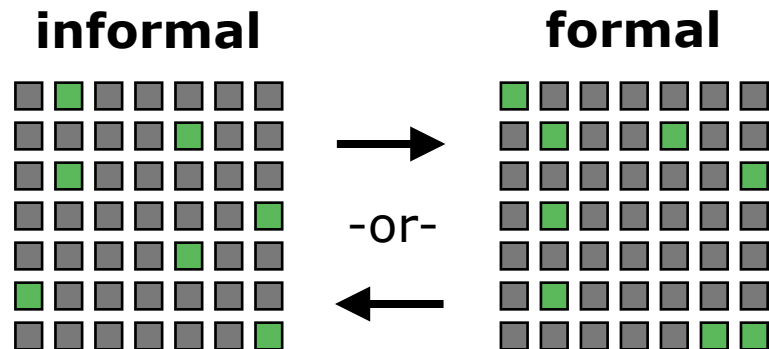
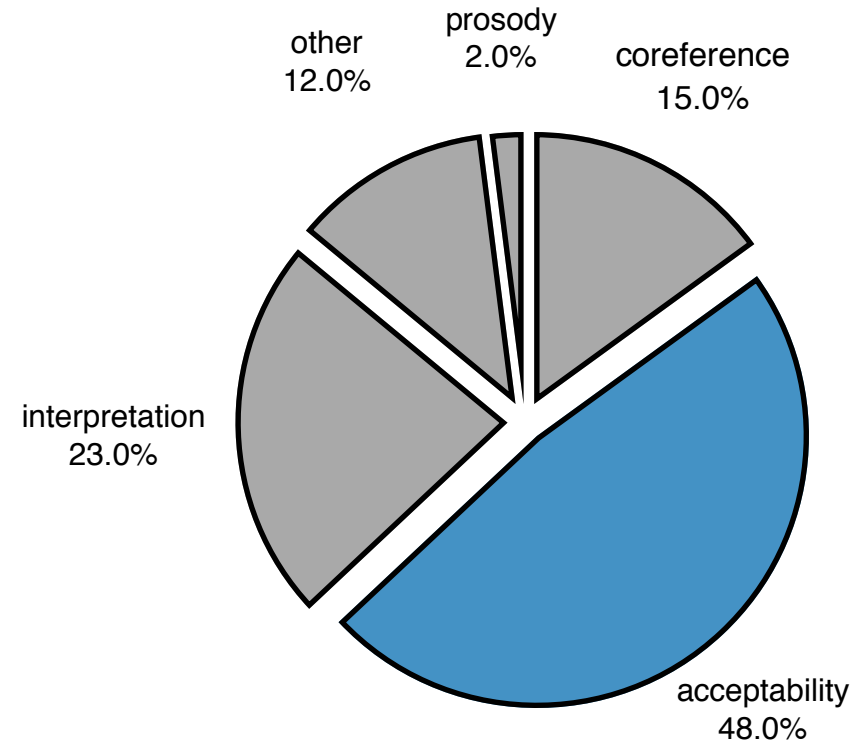
# What should we do next?

1. Test the other data types (other populations of data).

Right now we can't make any frequentist statistical claims about these other data types. All we can say is that we tested the largest chunk! And by induction, that is some amount of evidence.

2. Run follow-up studies on the divergent phenomena:

- (i) Identify the phenomena where the two methods diverge.
- (ii) Postulate a hypothesis about the mechanism that **causes** one method to be superior.
- (iii) Manipulate that mechanism and see if it makes the results converge.



# THANK YOU!

## and thank you to my generous collaborators!



**Diogo Almeida**  
NYU - Abu Dhabi



**Carson Schütze**  
UCLA

# Extra slides about Statistical Power

# There are two types of errors

If the concern is that informal methods are invalid, and therefore that the resulting syntactic theories are invalid, then what we really want to do is investigate **the number of errors** that have been made in the literature.

But there are actually (at least) two types of errors that can be committed. We can see this by contrasting states of the world with our behaviors relative to those states:



	No difference	Difference
Difference	Type I Error	Correct Action
No difference	Correct Action	Type II Error

## Type I error:

There is no difference between conditions, but we act as if there is a difference. A false positive.

## Type II error:

There is a difference between conditions, but we act as if there is no difference. A false negative.

# Side note: Fisher vs Neyman-Pearson



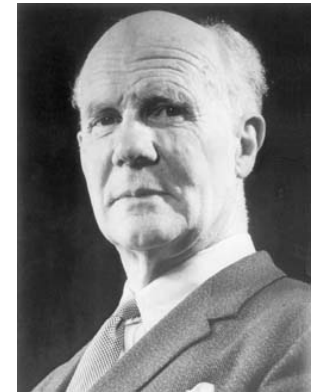
Ronald A. Fisher  
(1890-1962)

There are (at least) two schools of thought in null hypothesis significance testing. The first was Fisher's approach, which interprets p-values as the strength of evidence against the null hypothesis. This is probably the way most scientists think about p-values.

The second is the Neyman-Pearson approach, which interprets p-values as a number for making a decision. That decision is either to accept or reject the null hypothesis. The precise number of the p-value is irrelevant; all that matters is what decision you make (and therefore the behavior that you engage in). This is where the ideas of significance criteria come in to play. Although not particularly popular in science, this makes a lot of sense in quality control fields.



Jerzy Neyman  
(1894-1981)



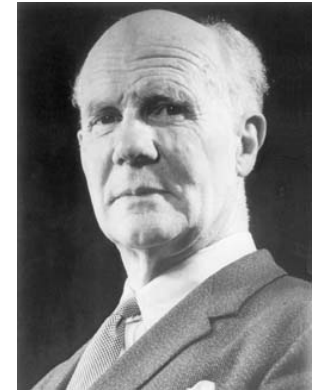
Egon Pearson  
(1895-1980)

# Side note: Fisher vs Neyman-Pearson

The terminology that I just used for our error types is the Neyman-Pearson terminology. This does not imply any particular endorsement of the N-P approach (I am personally more of a fan of Fisher's approach). But it is a nice framework for classifying errors, and the terminology is standardized.



Jerzy Neyman  
(1894-1981)



Egon Pearson  
(1895-1980)

		State of the World	
		$H_0$ True	$H_0$ False
Decision	Reject $H_0$	Type I Error	Correct Action
	Accept $H_0$	Correct Action	Type II Error

# We need to investigate two types of errors

Once again, we have (at least) two types of errors that invalid methods could lead to:



**Difference**  
**No difference**

**No difference**

**Difference**

Type I Error	Correct Action
Correct Action	Type II Error

Therefore we are going to need to run at least two types of studies to quantify the number of errors that have crept into the field.

# A method for counting Type I errors

**Type I error:** There is no difference between conditions, but we act as if there is a difference. A false positive.

In experimental syntax terms, Type I errors would be published acceptability differences that are not true differences. This means that we can count errors by working through the published differences in the literature, and determining if they are true differences or not.

But then the old problem rears its head: we can't know if a difference is real or not. All we can do is increase confidence in it. So here is a two step process:

**Step 1:** Re-test the informal differences using formal methods to see if the two methods converge.

**Step 2:** Investigate the source of any divergences between the two methods by manipulating potential sources of divergence.

[More details about these steps on the next two slides...]



# A method for Type II Errors

**Type II error:** There is a difference between conditions, but we act as if there is no difference. A false negative.

Type II errors are a bit more difficult to assess, because non-differences are rarely reported in the literature. So there is no set of phenomena that we can re-test. Instead, we need to try to assess the likelihood that syntacticians have missed detecting differences that are really there.

So what we need is a measure of **the probability of detecting an effect** when one is present. This is called **statistical power**:

**Statistical Power:** The probability of detecting an effect when one is truly present.

From the statistical power, we can calculate the Type II error rate. **The Type II error rate is simply  $1 - \text{statistical power}$ .**

If we can estimate the statistical power for informal judgment methods, we can estimate the likelihood of Type II errors in the field.

# Statistical Power

**Statistical Power** is dependent on several factors:

the task

the size of the difference (effect size) that you want to detect

the size of the sample that you are using

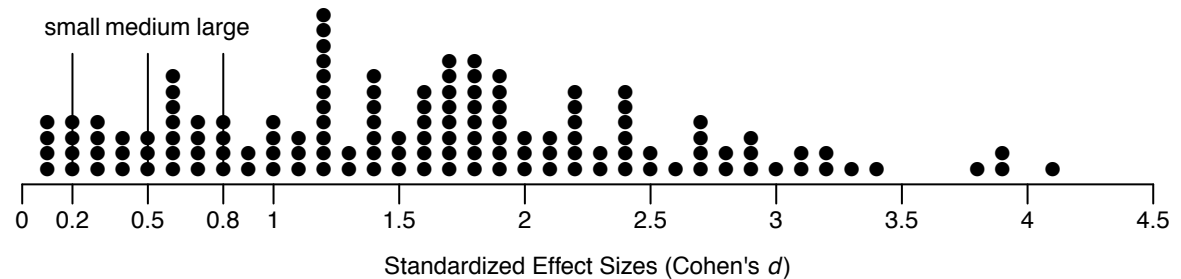
the Type I error rate that you are willing to tolerate (e.g., .05)

So if we want to estimate the statistical power for any possible judgment experiment, the first step would be to vary all of these factors and calculate statistical power. This would provide a (multi-dimensional) range of possible values for statistical power in judgment collection.

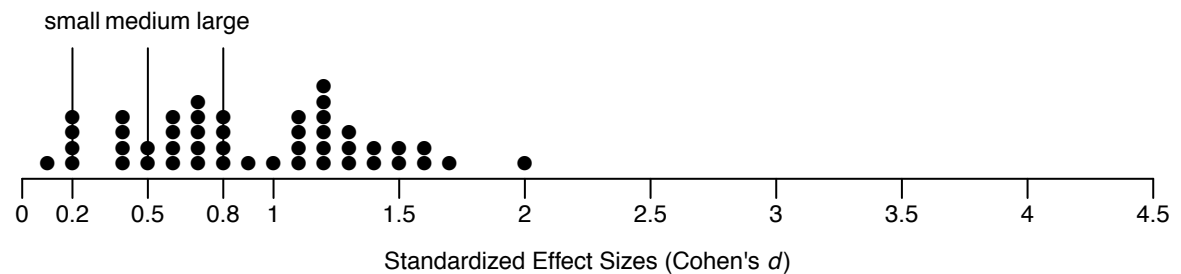
# A first investigation

We tested 50 two-condition (convergent) phenomena from the LI project. We chose the phenomena such that they span the range of effect sizes we saw in the full LI experiment (150 phenomena), with a slight focus on the smaller effect sizes to increase the informativity of the study:

The distribution of effect sizes in the full LI experiment.



The distribution of effect sizes in the 50 LI phenomena that were re-tested.

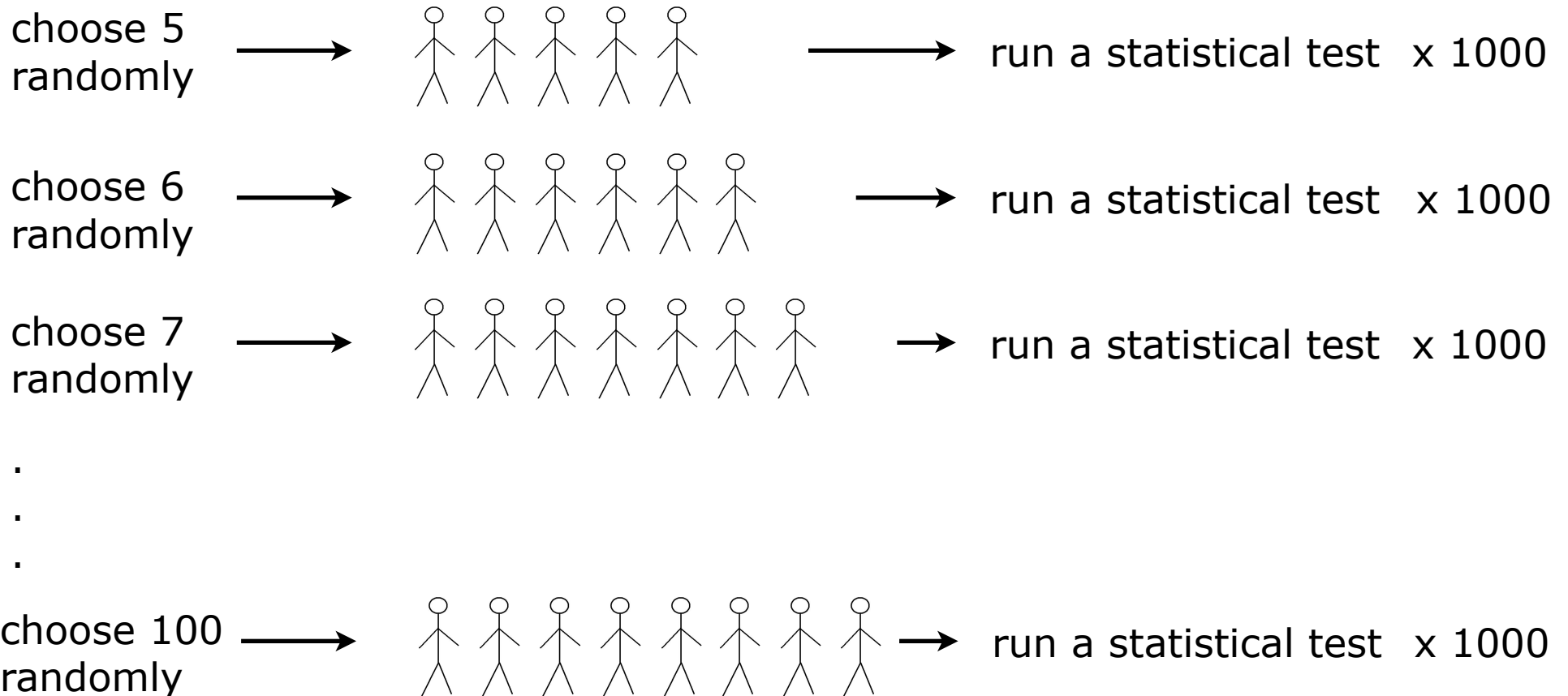


Because these phenomena were already tested in the LI project, and were part of the convergent results, we have high confidence that they are real effects. This means that our experiments should detect an effect. If they don't, it is a Type II error!

# A first investigation

We tested 100 participants (on Amazon Mechanical Turk) on all 50 phenomena.

This allows us to simulate the results for a range of effect sizes. All we have to do is sample from our 100 participants for each sample size that we want!



# A first investigation

We ran the experiment four times, each time with a different task.

This allows us to estimate the statistical power for each task. This is important given the range of possible tasks that could be used in judgment studies.

Magnitude Estimation	
reference sentence	100
target sentence	_____

Likert Scales	
target sentence	1 - 7
target sentence	1 - 7

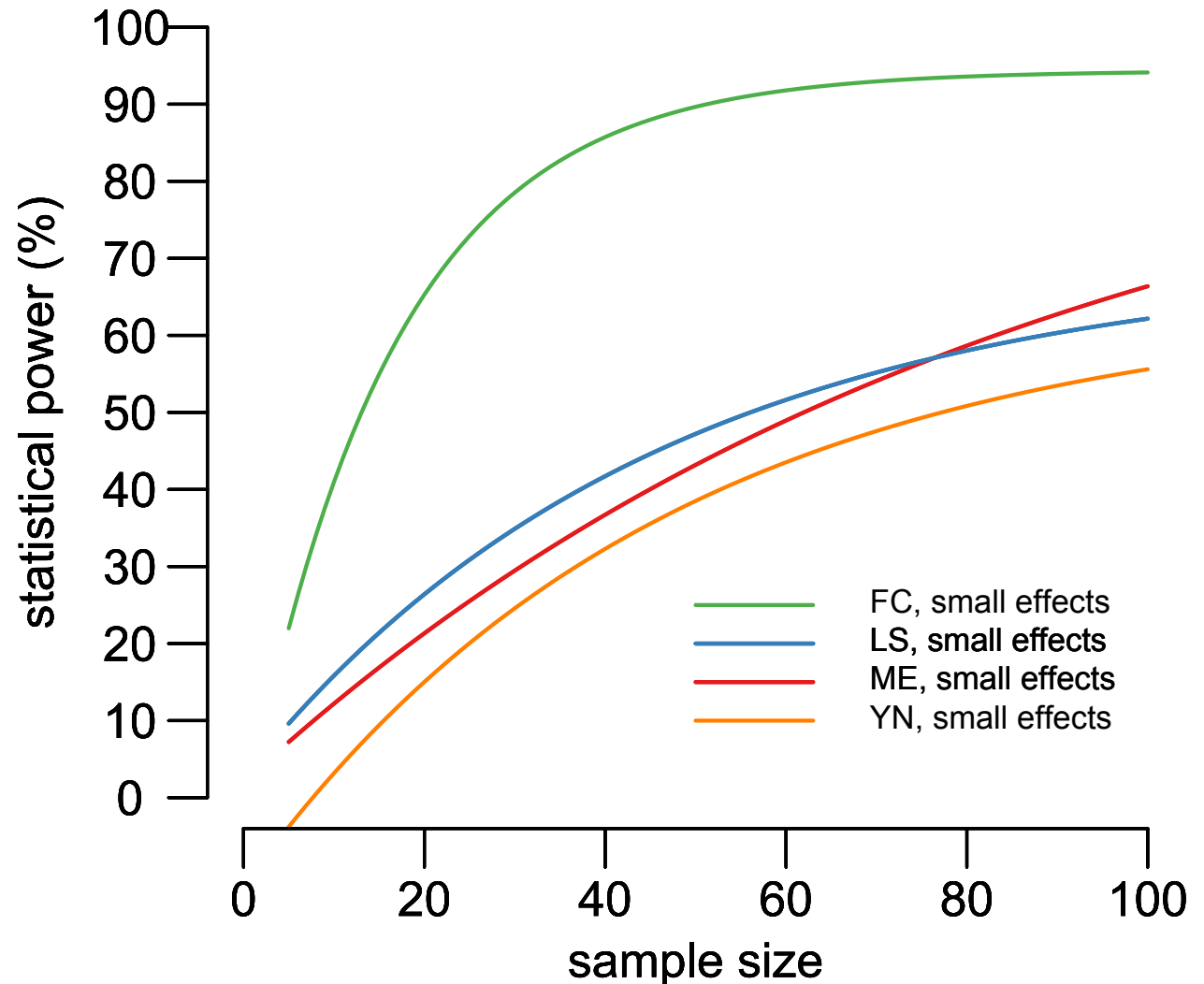
Forced Choice	
sentence A	<input type="radio"/>
sentence B	<input type="radio"/>

Yes - No		
sentence A	<input type="checkbox"/>	<input type="checkbox"/>
sentence B	<input type="checkbox"/>	<input type="checkbox"/>

# A first investigation: results

We analyze the results by counting the number of significant results (by t-test) in each 1000 simulations, at each sample size, for each effect size. This is an estimate of statistical power. We can plot these points into lines that show the change in power by sample size:

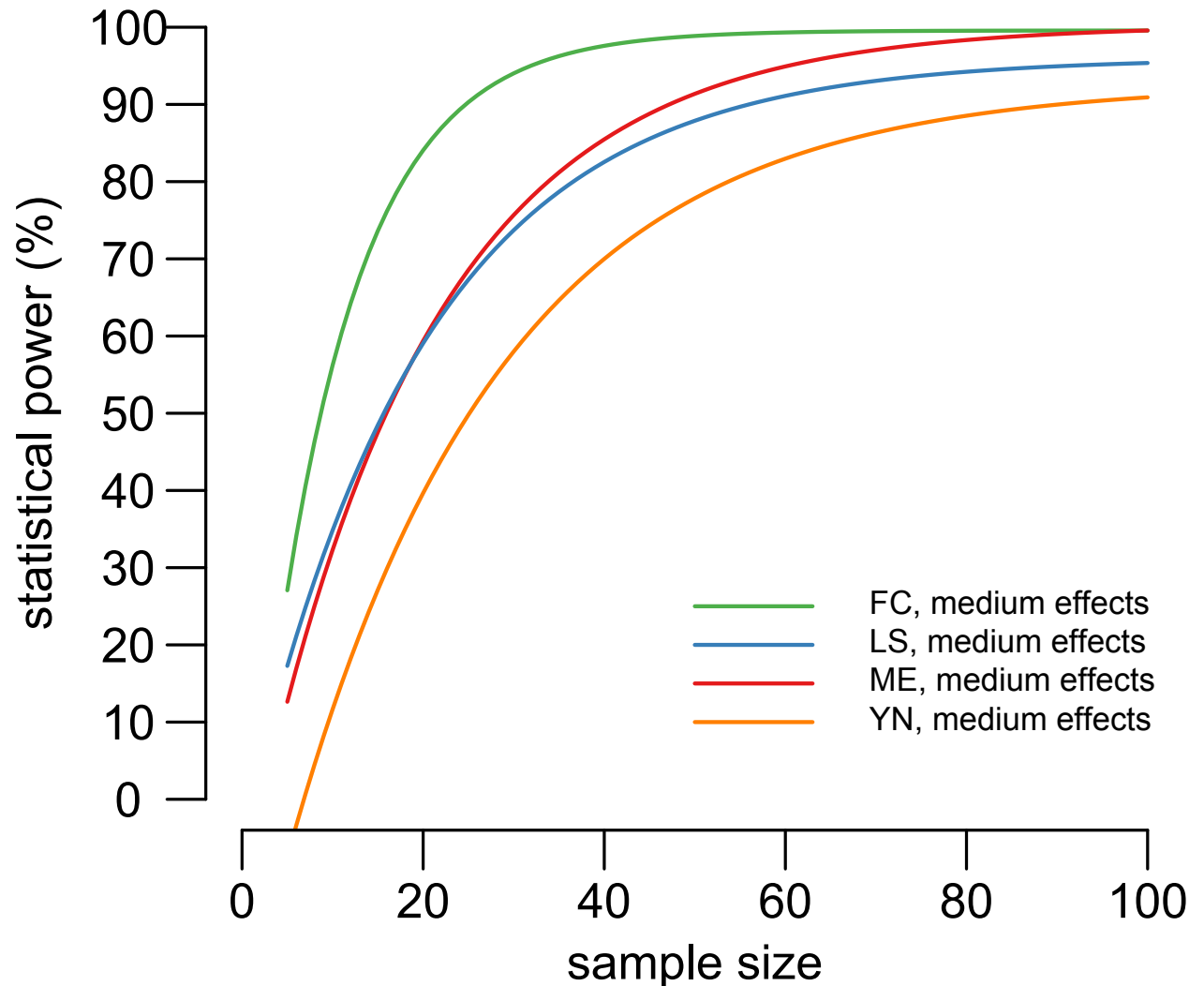
- 50 two-condition phenomena from LI split into 4 groups (small, medium, large, extra-large effect sizes)
- 4 tasks
- 100 participants per phenomenon per task
- 1000 re-sampling simulations per phenomenon per task to estimate detectability



# A first investigation: results

We analyze the results by counting the number of significant results (by t-test) in each 1000 simulations, at each sample size, for each effect size. This is an estimate of statistical power. We can plot these points into lines that show the change in power by sample size:

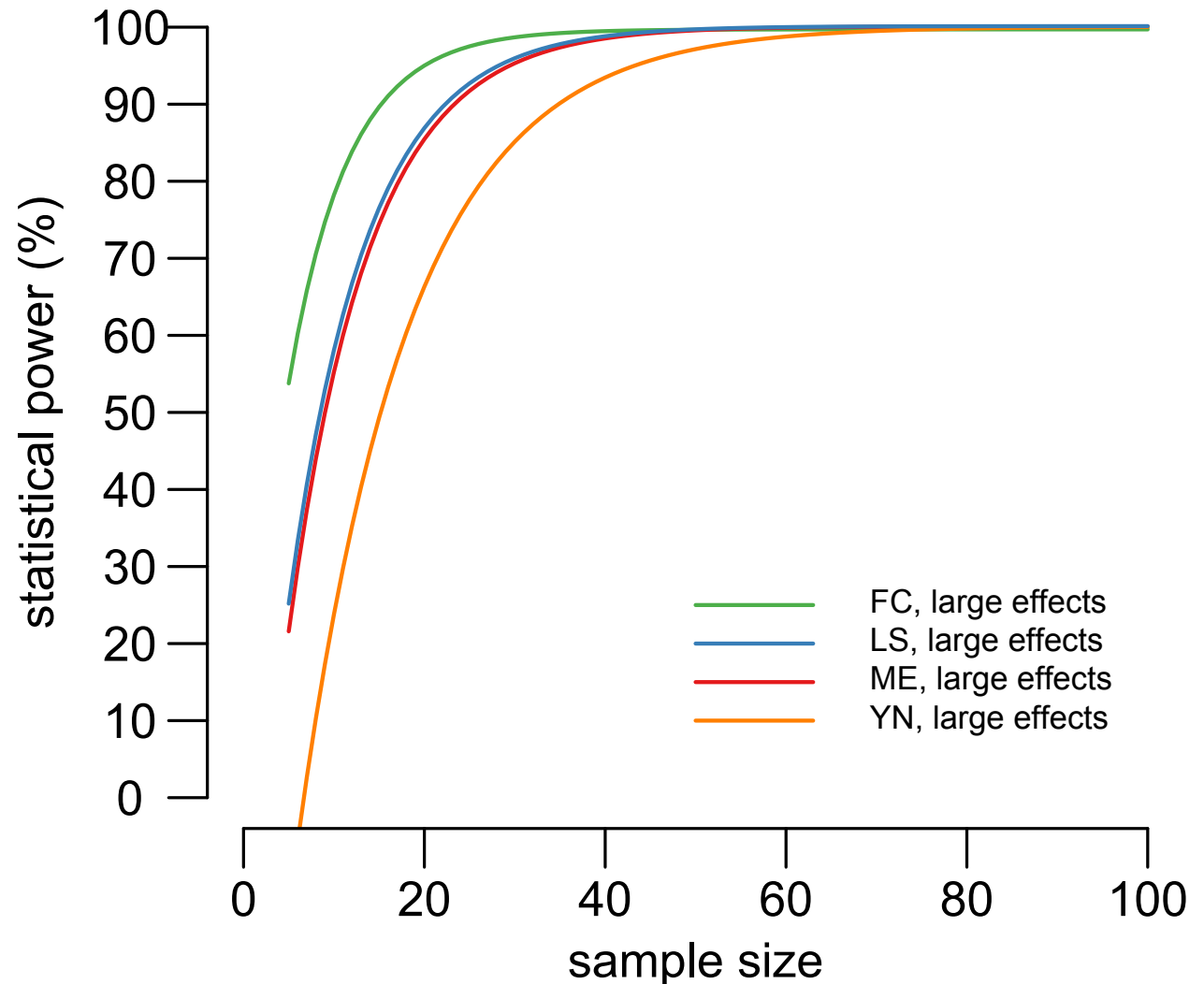
- 50 two-condition phenomena from LI split into 4 groups (small, medium, large, extra-large effect sizes)
- 4 tasks
- 100 participants per phenomenon per task
- 1000 re-sampling simulations per phenomenon per task to estimate detectability



# A first investigation: results

We analyze the results by counting the number of significant results (by t-test) in each 1000 simulations, at each sample size, for each effect size. This is an estimate of statistical power. We can plot these points into lines that show the change in power by sample size:

- 50 two-condition phenomena from LI split into 4 groups (small, medium, large, extra-large effect sizes)
- 4 tasks
- 100 participants per phenomenon per task
- 1000 re-sampling simulations per phenomenon per task to estimate detectability

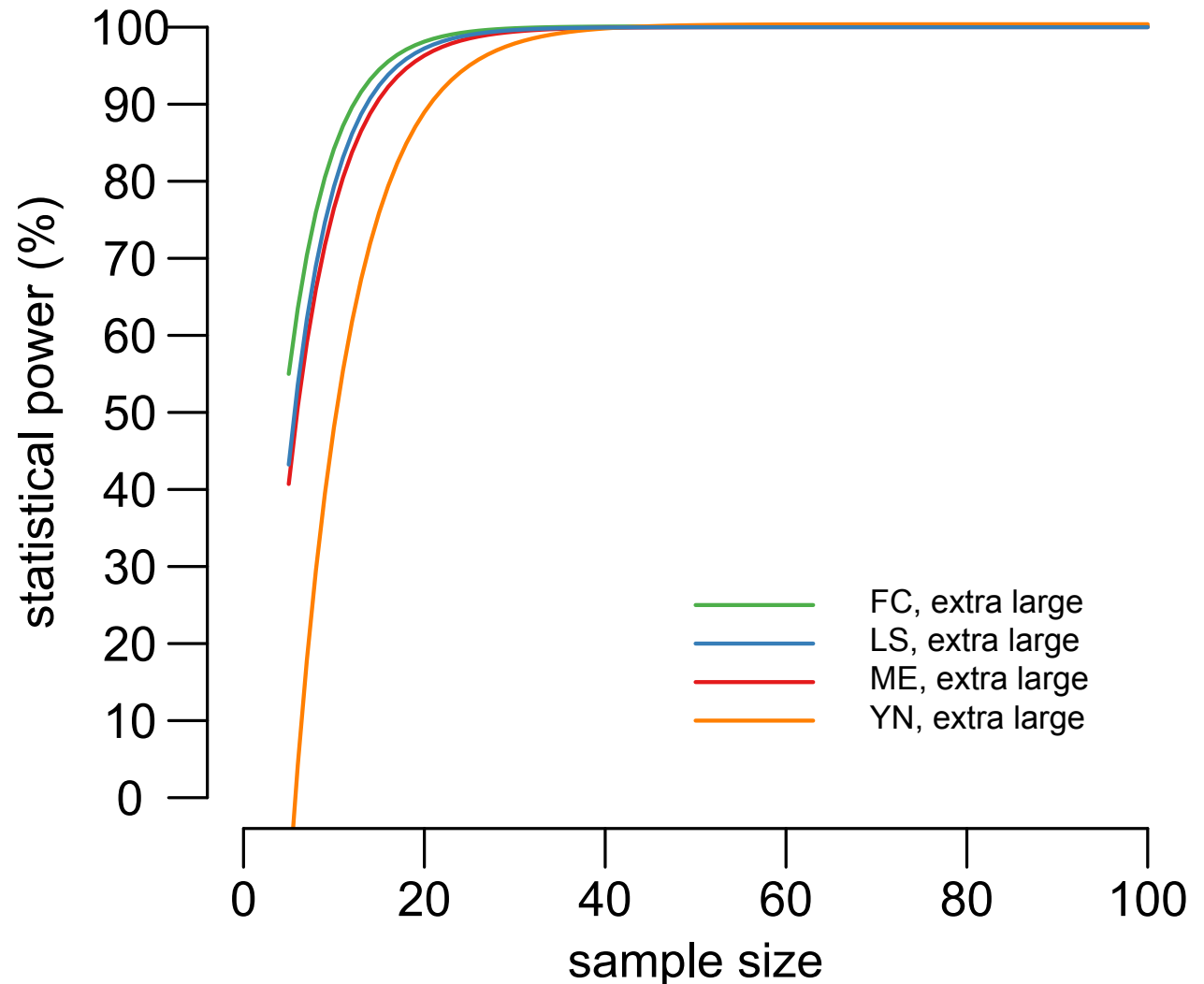




# A first investigation: results

We analyze the results by counting the number of significant results (by t-test) in each 1000 simulations, at each sample size, for each effect size. This is an estimate of statistical power. We can plot these points into lines that show the change in power by sample size:

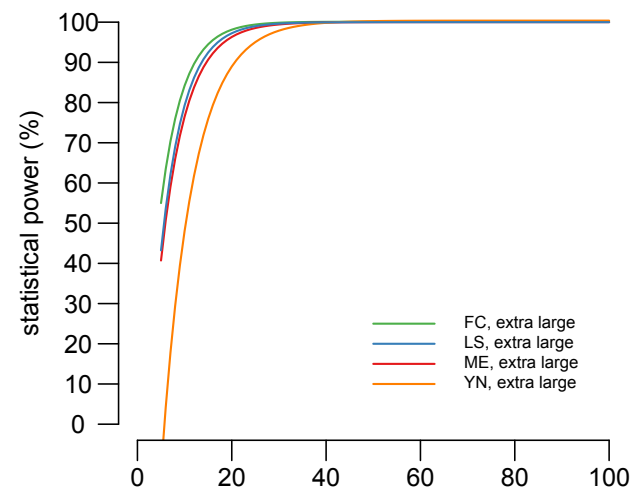
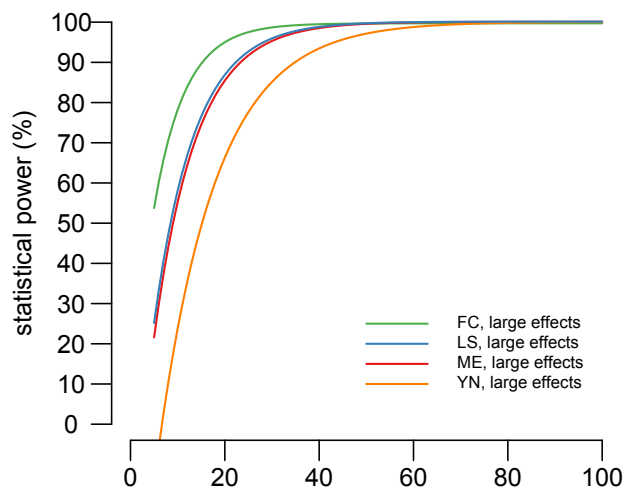
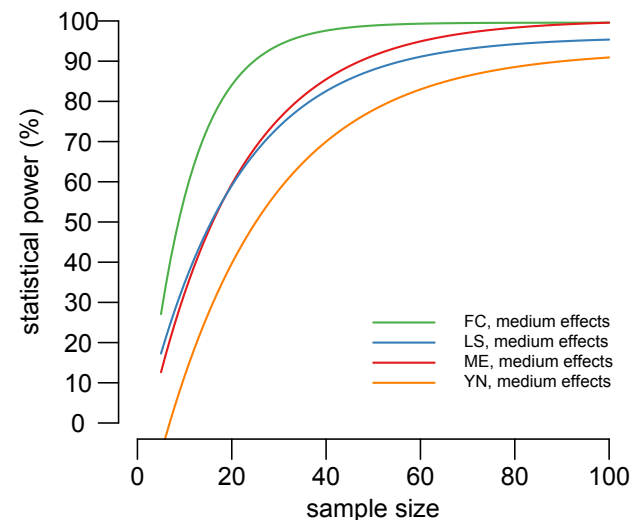
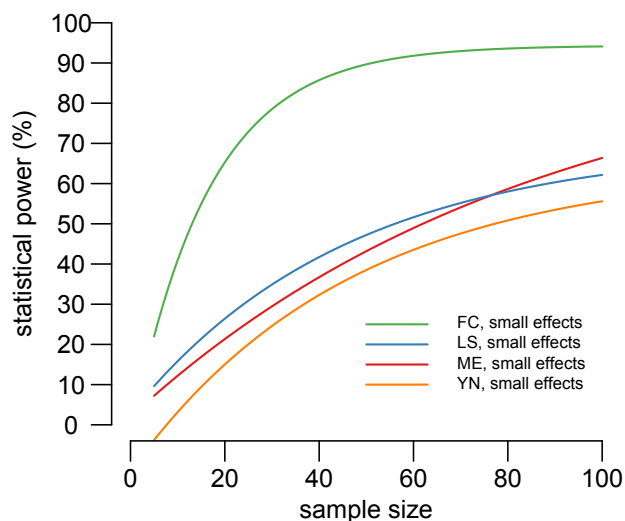
- 50 two-condition phenomena from LI split into 4 groups (small, medium, large, extra-large effect sizes)
- 4 tasks
- 100 participants per phenomenon per task
- 1000 re-sampling simulations per phenomenon per task to estimate detectability



# How do we interpret the results?

The problem is that informal methods by definition don't report things like task and sample size. So it is difficult to determine exactly what the statistical power has been for the tens of thousands of informal studies that have been conducted in syntax.

The best we can do is provide a range of statistical power estimates. Each syntactician can use their best judgment about the properties of their own studies to determine if they had a high or low Type II error rate.



# A side note about Type II error rates

Much like Type I error rates, what counts as large or small Type II error rates is open to interpretation.

The standard in experimental psychology seems to be the suggestion by Cohen (1962) that a good target for statistical power is 80%. Cohen arrived at this number with the following logic:

1. **Type I errors** are tolerated at a rate of .05 by convention.
2. **Type I errors** are about **4x more dangerous** than **Type II errors**, therefore the maximum **Type II error rate should be .2**.
3. Power is  $1 - \text{Type II error rate}$ , therefore **statistical power should be .8**.

To my knowledge there have been no published discussion of statistical power/ Type II errors in syntax. But I could imagine that syntacticians might want a lower Type II error rate, especially given that grammatical theories tend to place more than nominal importance on the ability to capture the absence of differences between sentence types.

# Validity and Reliability (and accuracy and precision)

# Building confidence in experimental results

Perhaps the fundamental problem with experiments is that **we cannot know whether our results are a true** reflection of the universe as it is.

That would require independent, true knowledge of the universe (which would also negate the need for measurements).

So what do we do? We **build confidence** in our results. And to build confidence in our results, we look for (at least) two properties:

**Validity:** A measurement method is valid if it measures the property it is intended to measure. Basically, we want a method that tests acceptability, not temperature or emotional states.

**Reliability:** A measurement method is reliable if it consistently produces the same output (under the same circumstances). Basically, we don't want a method that works on Mondays but not Tuesdays.

# Validity

**Validity:** A measurement method is valid if it measures the property it is intended to measure. Basically, we want a method that tests acceptability, not temperature or emotional states.

There are at least two ways to establish validity:

1. Look for correlations between the method of interest and other, established methods that measure the intended property.

This is not really possible for acceptability judgments (or really any cognitive rating task). We tend to use these tasks because there aren't any other measures that provide the same information.

2. Use predictions from an uncontroversial theory to evaluate how well the method behaves.

This is probably what most syntacticians do. Informal acceptability judgments collection works well for "clear" cases, so they believe the validity extends to unclear cases. Critics of informal judgments most likely don't believe the validity extends to unclear cases.

# A side note on validity through similarity

One interesting aspect of this conversation that is rarely discussed is the asymmetry of the investigation. Informal methods must establish their validity *bona fides*, but formal methods need not.

So this raises the question: **How was the validity of formal collection methods established?** It was not through the standard validation methods: both methods fail the correlation test (there is no measure to correlate with), and both pass the theoretical prediction test.

## **Validity through similarity:**

My impression is that formal methods are considered valid because they appear similar to validated methods in other domains (e.g. reaction times or ERPs). The properties of these validated methods exist for good reason (the measures in question require them), so some people assume that all measures should employ them.

*...the fact that this methodology is not valid has the unwelcome consequence that **researchers with higher methodological standards** will often ignore the current theories from the field of linguistics.*

*Gibson and Fedorenko 2010*

# Reliability

**Reliability:** A measurement method is reliable if it consistently produces the same output (under the same circumstances). Basically, we don't want a method that works on Mondays but not Tuesdays.

The primary way to establish reliability is through replication: the repeated deployment of an experiment under certain constraints.

The constraints on replication:

**Fixed factors:** The components of an experiment that must not change. These are the factors whose levels are directly chosen. Typically, the manipulation(s) of the independent variable(s).

**Random factors:** The components of an experiment that may change between replications. These are the factors whose levels are randomly chosen. Typically, the participants, the items, the time of day, the location.



# Validity seems to be the issue

Discussions of judgment collection methodology rarely use the technical terms validity and reliability, but it seems to me that the issue at stake for most is **validity**. In fact, Gibson and Fedorenko 2010 appear to be explicit about this:

*The **lack of validity** of the standard linguistic methodology has led to many cases in the literature where questionable judgments have led to **incorrect generalizations and unsound theorizing**.*

And in a different quote from the same paper, it appears as if there is at least a suggestion that **reliability is not a concern**:

*Although acceptability judgments are a **good dependent measure** of linguistic complexity (results from acceptability-judgment experiments are **highly systematic** across speakers...)...*

# A side note on accuracy and precision

When evaluating measurement methods, it is not uncommon to encounter two other properties:

**Accuracy:** Accuracy quantifies how close a measurement is to the true value of the property in question.

**Precision:** Precision quantifies the fineness of the scale of measurement, often operationalized as how well the measurement method can distinguish two measurements that in fact differ from one another.

A & P are useful for quantifying how well a measurement method is working (especially in statistics, where we care about the accuracy of our sample statistics, and the variability in our measurements). But in some ways they already assume validity and reliability, so they are secondary to our concerns here.

