

# Psychology 360: Personality Research

## Pitfalls in Research

William Revelle

Department of Psychology  
Northwestern University  
Evanston, Illinois USA

November, 2022

## Outline

### Overview

### Investigator Paradigm Effects

#### Investigator design effects

### Investigator Loose Procedure Effects

#### Investigator Data Analysis Effects

### Investigator “Fudging” Effects and Scientific Fraud

### Experimenter effects

#### Experimenter Fudging and Expectancy Effects

### Replicability

### Recommendations

## The changing nature of science

1. The classic image: Little Science
  - (Mad) Scientist in white lab coat working by self
2. Reality: Medium to Big Science
  - Research Teams
  - Research Labs
  - Cross university-cross national research groups
3. Investigators and Experimenters
4. Sources of Errors due (primarily) to Investigators and Experimenters

(originally inspired and adapted from [Barber \(1976\)](#) Pitfalls in Human Research: Ten Pivotal Points.)

## Paradigms help and hurt research

1. Paradigm: a logical or conceptual structure serving us as a form of thought within a given area of experience
2. Kuhn and the philosophy of science: each period of normal science in the development of a scientific discipline corresponds to one and only one methodological framework or paradigm.
3. In a nut-shell, paradigms are 'universally recognized scientific achievements that for a time provide model problems and solutions to a community of practitioners'.

Kuhn (1970)



## Paradigmatic thinking as a potential source of error

1. Tenacity of paradigms and resistance to new discoveries  
theories aren't disproved, old theorists die
2. Failing to see events that do not fit within paradigm
3. Continental drift and plate tectonics
  - Alfred Wegner and theory of continental drift (1912)
  - Early suggestions by Snider-Pellegrini (1858) and Seuss (1885)
  - But no mechanism to explain it
  - Harry Hess (1962) and sea floor spreading to account for oceanographic findings of ridges, trenches, magnetic striping
4. Glacial dams and the Grand Coulee
  - Harlan Bretz (1922) described the channeled scablands of Montana and Eastern Washington
  - His catastrophic theory was rejected until 1965 by geologists holding to uniformitarianism

## Paradigmatic thinking as a potential source of error

1. Seeing non “events” that fit within a paradigm
  - Blondlot and N-Rays
  - Original claims of N-rays
  - Robert Wood and the critical experiments
2. Cold Fusion
  - Fleischman and Pons
3. Lysenko and the transmission of environmental effects
  - Although the rejection of Lysenkoism probably delayed the acceptance of epigenetics

## Paradigmatic thinking as a potential source of error

1. Seeing non “events” that fit within a paradigm
  - Blondlot and N-Rays
  - Original claims of N-rays
  - Robert Wood and the critical experiments
2. Cold Fusion
  - Fleischman and Pons
3. Lysenko and the transmission of environmental effects
  - Although the rejection of Lysenkoism probably delayed the acceptance of epigenetics
4. “They laughed at Galileo, they laughed at Columbus, they also laughed at Bozo the clown.” (Sagan, 1980)

## Paradigms in Psychology

1. Behaviorism
  - Forces operationalizations, ignores constructs
2. Cognitivism
  - cognitive psychology
  - cognitive social
  - cognitive learning
3. Tabula Rasa paradigm of development
  - Ignores evidence for genetic causes
4. Neuroscience reductionism
  - Ignores evidence for social causes

## Paradigms in Personality

1. Freudian dynamics
  - Importance of childhood in adult behavior
  - Unconscious processes driving thought
2. Trait theories of consistency
  - Assumes that people reflect underlying trait differences
3. Social learning theories of variability
  - Everything is learned, no trans-situational consistency
4. Biological bases of personality
  - Everything is genetic
  - Everything is biological

## Investigator Design Effects

1. Confirmatory designs
  - If A then B
  - Frequently test by doing A and observing B
  - But what about observing Not B?
2. Demonstrations versus tests
  - Experiments that are consistent with theoretical predictions, but do not test the theory
3. Failure to pit theory against theory

## Reasoning in Research

1. Karl Popper and the testability of theory
  - The hallmark of science is the testability of theory
  - Non-testable theories are not science
  - “it must be possible for all empirical scientific system to be refuted by experience”
  - Theories are not shown to be correct, they are shown to be incorrect
2. Science is the process of asking questions that have answers (Rep. Rush Holt)
3. All models (theories) are wrong, but some are useful

## J. Platt and Strong Inference? (Science, 1964)

### 4 signs of strong science

1. Devising alternative hypotheses;
2. Devising a crucial experiment (or several of them), with alternative possible outcomes, each of which will, as nearly is possible, exclude one or more of the hypotheses;
3. Carrying out the experiment so as to get a clean result;
4. Recycling the procedure, making subhypotheses or sequential hypotheses to refine the possibilities that remain, and so on.

## Strong inference

1. A theory which cannot be mortally endangered cannot be alive.
  - (Rushton, as cited by Platt)
2. “The problems of how enzymes are induced, of how proteins are synthesized, of how antibodies are formed, are closer to solution than is generally believed. If you do stupid experiments, and finish one a year, it can take 50 years. But if you stop doing experiments for a little while and think how proteins can possibly be synthesized, there are only about 5 different ways, not 50! And it will take only a few experiments to distinguish these.”
  - (Szilard, as cited by Platt)

## Platt and Strong Inference

*I will mention one severe but useful private test - a touchstone of strong inference - that removes the necessity for third-person criticism, because it is a test that anyone can learn to carry with him for use as needed. It is our old friend the Baconian 'exclusion,' but I call it 'The Question.' Obviously it should be applied as much to one's own thinking as to others'. It consists of asking in your own mind, on hearing any scientific explanation or theory put forward, 'But sir, what experiment could disprove your hypothesis?' or, on hearing a scientific experiment described, 'But sir, what hypothesis does your experiment disprove?'*

Platt, Science, 1964

## Confirmatory designs and weak inference

1. If Introverts are more aroused than extraverts
2. If arousal varies throughout the day
3. If arousal is increased by caffeine
4. If cognitive performance is curvilinearly related to arousal
5. Then introverts should be hurt (helped less) by caffeine than extraverts (in the morning) and extraverts should be hurt (helped less) by caffeine than introverts in the evening.

We observe 5, can we infer 1-4?

## Confirmatory designs and problems of inference

1. If A implies B and we observe B, does this imply A?
  - No. Not B implies not A.
2. If A and B and C and D imply E, and we observe E, does that imply A, B, C, & D?
  - No.
  - But Not E implies one of A, B, C, & D is not true, but which one?

## Correlation and inverse probabilities

1. Does observing that B almost always happens when we do A imply that doing A almost always leads to B?
2. Examples:

Table: Examples of inverse probability problem

Observe	Cause ?
Auto Accidents	Drinking alcohol
Lung Cancer	Smoking
Pregnancy	Intercourse

3. Although strong association in one direction, how strong is the association in the other direction?
4. We need to know the base rates as well as the one cell

## Correlation and inverse probabilities

1. If one has disease B, then one tests A+ with  $p=.99$
2. If one tests A-, then one has disease B with probability .01
3. 99% of people do not have the disease
4. If one tests A+, what the probability that they have disease B?

## Correlation and inverse probabilities

**Table:** The problem of inverse probability

	Then test A+	Then test A-
If Disease	.99	.01
if Healthy	.01	.99

## Correlation and inverse probabilities

But the base rates of the outcomes are very different

**Table:** The problem of inverse probability

	Then test A+	Then test A-	Frequency
If Disease	.99	.01	100
if Healthy	.01	.99	9900

## Correlation and inverse probabilities

We need to know the base rates of the two outcomes to understand the accuracy of the prediction

**Table:** The problem of inverse probability

	Then test A+	Then test A-	Frequency
If Disease	99	1	100
if Healthy	99	9801	9900
Total	198	9802	10,000

Probability of Disease if Test A+ is just .5!

## Correlation and inverse probabilities: A more accurate test

1. If one has disease B, then one tests A+ with  $p=.999$
2. If one tests A-, then one has disease B with probability .001
3. 99% of people do not have the disease
4. If one tests A+, what the probability that they have disease B?

## Correlation and inverse probabilities

**Table:** The problem of inverse probability

	Then test A+	Then test A-
If Disease	.999	.001
if Healthy	.001	.999

## Correlation and inverse probabilities

But the base rates of the outcomes are very different

**Table:** The problem of inverse probability

	Then test A+	Then test A-	Frequency
If Disease	.999	.001	100
if Healthy	.001	.999	9900

## Correlation and inverse probabilities

We need to know the base rates of the two outcomes to understand the accuracy of the prediction

Table: The problem of inverse probability

	Then test A+	Then test A-	Frequency
If Disease	99.9	.1	100
if Healthy	9.9	9891.1	9900
Total	108.8	9891.2	10,000

Probability of Disease if Test A+ is 92%!

## Investigator Loose Procedure Effects

1. Poor specification of how to conduct experiment
  - “make the subject relaxed”
  - Record memories of childhood
2. Badly defined specification of manipulation
  - Threaten with fear of public speaking

## Investigator Data Analysis Effects

1. “Pilot Studies ”and the filing cabinet syndrome
  - How many studies are done before “the experiment”
  - The meaning of significance tests when many studies are done but only one is reported
  - Is this the fault of the journal editors, or of the investigator?
2. Stopping rules, data “purification”, deletion of “outliers”
3. Hypothesis development and “confirmation”
  - Hypotheses developed after the data were collected are almost always correct!
  - Need to explore data, but call this hypothesis generation, not confirmation
4. Over interpretation of findings
5. Meaning of significance tests versus experiment-wise error rates

## Investigator Data Analysis Effects

1. Significance tests and experiment wide significance
2. Interpretability of randomness
3. Meta analytic technique for averaging results across studies
4. emphasis upon effect sizes and confidence estimates
5. The problem of testing “unlikely” hypotheses

## The problem of “sexy” (unlikely ) hypotheses

Table: 50% Power and likelihood of result

	Hypothesis false	Hypothesis is True
% of cases	90%	10%
Reject	95%	50%
Accept	5%	50%

## The problem of “sexy” (unlikely ) hypotheses

Table: Power and likelihood of result

	Hypothesis false	Hypothesis is True	
Number of cases	900	100	1,000
Reject	855	50	905
Accept	45	50	95

Probability that a significant result is a type 1 error is  $45/95 = 47\%$ !

## The problem of “sexy” (unlikely ) hypotheses

Table: 80% Power and likelihood of result

	Hypothesis false	Hypothesis is True
% of cases	90%	10%
Reject	95%	20%
Accept	5%	80%

## The problem of “sexy” (unlikely ) hypotheses

Table: Power and likelihood of result

	Hypothesis false	Hypothesis is True	
Number of cases	900	100	1,000
Reject	855	20	905
Accept	45	80	125

Probability that a significant result is a type 1 error is  $45/125 = 36\%$ !

## The problem of “sexy” (unlikely ) hypotheses

Table: 90% Power and likelihood of result

	Hypothesis false	Hypothesis is True
% of cases	90%	10%
Reject	95%	10%
Accept	5%	90%

## The problem of “sexy” (unlikely ) hypotheses

Table: Power and likelihood of result

	Hypothesis false	Hypothesis is True	
Number of cases	900	100	1,000
Reject	855	10	865
Accept	45	90	135

Probability that a significant result is a type 1 error is  $45/125 = 33\%$ !



## The problem of “sexy” (unlikely ) hypotheses

Table: Power and likelihood of result

	Hypothesis false	Hypothesis is True	
Number of cases	900	100	1,000
Reject	855	1	856
Accept	45	99	144

Probability that a significant result is a type 1 error is  $45/144 = 31\%$ !

## The problem of “sexy” (unlikely ) hypotheses: Increase power, lower alpha

Table: 99% Power and likelihood of result

	Hypothesis false	Hypothesis is True
% of cases	90%	10%
Reject	99%	1%
Accept	1%	99%

## The problem of “sexy” (unlikely ) hypotheses: power of .99, alpha of .01

Table: Power and likelihood of result

	Hypothesis false	Hypothesis is True	
Number of cases	900	100	1,000
Reject	891	1	892
Accept	9	99	108

Probability that a significant result is a type 1 error is  $9/108 = 8\%$ !

Unusual claims require unusual evidence.

## Investigator “Fudging” Effects and Scientific Fraud

1. Trimming, Cooking, idealizing
  - Galileo, Newton, Dalton, Mendel, Millikan
2. Generation of new “confirming cases”
  - Burt?
3. Outright Fraud
  - [Piltdown Man](#)
  - Summerlin, Schon, Darsee, Hwang, Stapel (2011), Förster (2014)
  - Nonexistent data, nonexistent subjects
4. Unknown prevalence assumed (hoped) to be small

## Examples of fraud in science

1. William T. Summerlin (chief of transplantation immunology at Sloan-Kettering) claimed he could transplant onto animals corneas, glands, and skin that would normally be rejected – sometimes even across species. He was discovered only after three years of this when a lab assistant noticed that the black “skin graphs” were drawn on with a marker (all the rest of his work turned out to be fake as well).
2. John Long (a resident) studied Hodgkins’s cell lines at Mass General in collaboration with MIT. A year later, a junior colleague charge fraud and it was discovered that the cell lines were from monkeys and healthy people.
3. Elias A. K. Alsabti (a researcher at Boston University) had published sixty papers by his mid-twenties, when it turned out that most of them were papers published in obscure foreign journals with only slight changes (like a new title)

from Judson, The Great Betrayal cited by swartz

<http://www.aaronsw.com/weblog/001616>

## Examples of fraud in science

1. Vijay Soman, an assistant professor at Yale, was asked to peer review a paper by Helena Wachslicht-Rodbard. He sent back a negative review, delaying publication, then turned around and submitted the same paper to another journal. He was found out when, in an amazing twist of fate, Helena Wachslicht-Rodbard was asked to peer review Soman's paper and recognized it as her own.
2. John Darsee had published dozens of papers with completely made up data—and done an incredibly bad job making up the data. (One paper claimed a father had four children – conceived when he was 8, 9, 11, and 12 years old, respectively.) To cover up this fact, Darsee had practiced “gift authorship” – adding people as co-authors even when they didn't do any work. Darsee had been at Harvard for three years before he was discovered by some postdocs, even then it took the university five months to admit the fraud.

from Judson, The Great Betrayal cited by swartz

<http://www.aaronsw.com/weblog/001616>

## The Darsee case

*Dr. John Darsee was regarded a brilliant student and medical researcher at the University of Notre Dame (1966-70), Indiana University (1970-74), Emory University (1974-9), and Harvard University (1979-1981). He was regarded by faculty at all four institutions as a potential "all-star" with a great research future ahead of him. At Harvard he reportedly often worked more than 90 hours a week as a Research Fellow in the Cardiac Research Laboratory headed by Dr. Eugene Braunwald. In less than two years at Harvard he was first author of seven publications in very good scientific journals. His special area of research concerned the testing of heart drugs on dogs.*

*In May 1981, three colleagues in the Cardiac Research Laboratory observed Darsee labeling data recordings "24 seconds," "72 hours," "one week," and "two weeks." In reality, only minutes had transpired. Confronted by his mentor Braunwald, Darsee admitted the fabrication; but he insisted that this was the only time he had done this, and that he had been under intense pressure to complete the study quickly. Shocked, Braunwald and Darsee's immediate supervisor, Dr. Robert Kroner, spent the next several months checking other research conducted by Darsee in their lab. Darsee's research fellowships were terminated, and an offer of a faculty position was withdrawn. However, he was allowed to continue his research projects at Harvard for the next several months (during which time Braunwald and Kroner observed his work very closely).*

*Hopeful that this was an isolated incident, Braunwald and Kroner were shocked again in October. A comparison of results from four different laboratories in a National Heart, Lung and Blood Institute (NHLBI) Models Study revealed an implausibly low degree of invariability in data provided by Darsee. In short, his data looked "too good." Since these data had been submitted in April, there was strong suspicion that Darsee had been fabricating or falsifying data for some time. Subsequent investigations seemed to indicate questionable research practices dating back as far as his undergraduate days.*

## The Diederik Stapel case

1. Data too good to be true?
2. More that 150 papers are thought to be completely fabricated
3. “Stapel’s eye-catching studies on aspects of social behaviour such as power and stereo-typing garnered wide press coverage. For example, in a recent Science paper (which the investigation has not identified as fraudulent), Stapel reported that untidy environments encouraged discrimination ( Science 332, 251-253; 2011).
4. “Somebody used the word ‘wunderkind’ says Miles Hewstone, a social psychologist at the University of Oxford, UK. “He was one of the bright thrusting young stars of Dutch social psychology – highly published, highly cited, prize-winning, worked with lots of people, and very well thought of in the field.” (From Nature , November 1, 2011.

## And it continues

1. The Jens Förster case
2. Several years ago, the news broke that the University of Amsterdam is recommending the retraction of a 2012 paper by one of its professors, social psychologist Prof Jens Förster , due to suspected data manipulation. The next day, Förster denied any wrongdoing.
3. Shortly afterwards, the Retraction Watch blog posted a (leaked?) copy of an internal report that set out the accusations against Förster.
4. The report, titled Suspicion of scientific misconduct by Dr. Jens Förster, is anonymous and dated September 2012. Reportedly it came from a statistician(s) at Förster's own university. It relates to three of Förster's papers, including the one that the University says should be retracted, plus two others.
5. A vigorous discussion of the allegations has been taking place in this Retraction Watch comment thread. The identity and motives of the unknown accuser(s) are one main topic of debate; another is whether Förster inability to produce raw data and records relating the studies is suspicious or not.
6. (See Science magazine) and [replicability index article](#)

## Reasons for Fraud; Situation or Trait?

1. Pressure for success
  - Grant funding
  - Tenure review
2. Hubris/Psychopathology
  - I know I am right, they are so stupid not to recognize it
3. Poor supervision – too large a lab
  - Schon's data not analyzed by supervisors
4. Lack of replication, pressure to be new
5. Field is starting to take open science seriously, recognizing the need to replicate and explain methods
6. Statistical sleuths are reanalyzing unusual studies

## Experimenter personal attributes

1. Stable
  - Sex, age, ethnicity, height, weight
  - Sex of experimenter affects pain tolerance in mice!
2. Modifiable
  - Prestige, anxiety, friendliness, warmth
3. Possible interactions of experimenter characteristics and subject characteristics
  - Sex of experimenter by sex of subject
  - Prestige of experimenter by anxiety
  - Etc.
4. Partial control for these by making every experimenter run all conditions (avoids experimenter by condition effects)

## Experimenter failing to follow procedures effects

1. Impossible, difficult and inconvenient procedures
  - Have they been pilot tested
2. Poorly specified procedures
  - What actually is supposed to be said or done?
3. Poor training, lack of practice, lack of supervision
  - Does the experimenter know how to do it?

## Experimenter Mis-recording Effect

### 1. Mistakes of data collection

- What condition was the subject supposed to be
- What manipulation was actually done

### 2. Mistakes at data entry

- Data entry and data checking
- What protection against mistakes in entry
- Double entry (two independent sources)
- Automatic entry (but what about programming errors)

## Experimenter Fudging Effects

1. Lack of interest in the outcome
  - Who cares, why bother
  - It is just a job
2. Too much interest in the outcome
  - Grade depends upon significance! (no!)
  - A true result that shows nothing is worth much more than a fake result that shows an effect.

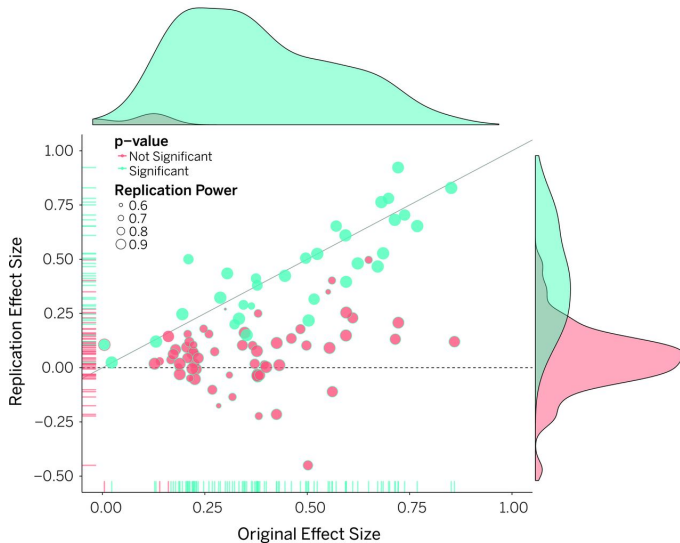
## Experimenter Expectancy Effects

1. Pygmalion in the classroom
  - An supposed example of expectancy effects
  - An actual example of sloppy procedures and misanalyzed data
  - See [Thorndike \(1968\)](#) for a very critical review
2. But a real effect when done properly
  - Younger sibs do better with same teacher if older sib did better, no difference with different teacher
3. Experimenter demand characteristics lead to results unintentionally
4. See the special issue celebrating 50 years of the “Pygmalion Effect” [Timmermans, Rubie-Davies & Rjosk \(2018\)](#)

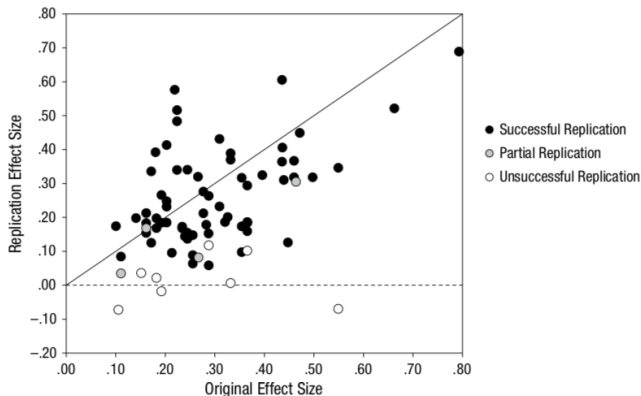
## Sample size, power, and replicability

1. How replicable is research?
2. What is the meaning of a replication
3. The 100 studies project
4. Estimating the reproducibility of psychological science  
(Science, 28 August, 2015)
5. But, [Soto & John \(2019\)](#) shows that personality studies are much more replicable
6. Personality studies, no flair, but replicable ([Atherton, Chung, Harris, Rohrer, Condon, Cheung, Vazire, Lucas, Donnellan, Mroczek, Soto, Antonoplis, Damian, Funder, Srivastava, Fraley, Jach, Roberts, Smillie, Sun, Tackett, Weston, Harden & Corker, 2021](#))

## Estimating the reproducibility of psychological science



## Estimating the reproducibility of personality research



Soto & John (2019)

## Recommendations

1. Be aware of one's own and alternative paradigms
  - Alternative procedures
  - Alternative explanations
2. Tighten protocols
3. Replicate
  - If it is worth doing and worth reporting, it is worth replicating
  - If you can't replicate your findings, who can?
  - can not replicate longitudinal studies, so do them right the first time.

## Recommendations-2

1. Delicate balance between theory testing and theory development
2. Methodological rigor versus theoretical speculation
3. Significance tests based upon prior hypotheses
4. Speculation based upon fortuitous findings
5. Issues of type I and Type II error
  - Type I: Finding something which is not there
  - Type II: Not finding something that is there
  - Type III: Asking the wrong question, not looking in the right place

## Recommendations - 3

1. Statistical Significance versus real significance
  - How likely is result to happen by chance (alpha level)
  - How important is this effect in real world
2. Point estimates and confidence intervals
  - What is the value of the effect and what is the range
3. Effect sizes
  - differences in terms of within cell error
  - Size of correlation coefficient
4. Power
  - How likely can you find a result if it is there?
  - Expressed in probability of significance given effect size of X
  - Expressed in precision of estimate of an effect

## Scepticism

1. Be skeptical of your own work
  - Be aware of your biases
  - Would you be willing to share the data and analyses with your strongest critic
2. Be skeptical of others as well
3. But help them improve the quality of your and their work

Atherton, O. E., Chung, J. M., Harris, K., Rohrer, J. M., Condon, D. M., Cheung, F., Vazire, S., Lucas, R. E., Donnellan, M. B., Mroczek, D. K., Soto, C. J., Antonoplis, S., Damian, R. I., Funder, D. C., Srivastava, S., Fraley, R. C., Jach, H., Roberts, B. W., Smillie, L. D., Sun, J., Tackett, J. L., Weston, S. J., Harden, K. P., & Corker, K. S. (2021). [Why Has Personality Psychology Played an Outsized Role in the Credibility Revolution?](#) *Personality Science*, 2(1), 1–21.

Barber, T. X. (1976). *Pitfalls in human research: Ten pivotal points*, volume 67. Elsevier.

Kuhn, T. (1970). *The Structure of Scientific Revolution* (2nd ed.). Chicago: University of Chicago Press.

Soto, C. J. & John, O. P. (2019). Optimizing the length, width, and balance of a personality scale: How do internal characteristics affect external validity? *Psychological Assessment*, 31(4), 444.

Thorndike, R. L. (1968). Reviews: Rosenthal, robert, and

jacobson, lenore. pygmalion in the classroom. new york: Holt, rinehart and winston, 1968. 240+ xi pp. *American Educational Research Journal*, 5(4), 708–711.

Timmermans, A. C., Rubie-Davies, C. M., & Rjosk, C. (2018). Pygmalion's 50th anniversary: the state of the art in teacher expectation research. *Educational Research and Evaluation*, 24(3-5), 91–98.