

# Chapter 1

---

## *Changing Times*

They always say time changes things, but you actually have to change them yourself.

—Andy Warhol

### Overview

- Historical developments
- Current context for change in data analysis practices
- Why change is inevitable and desirable

## Historical developments

---

- Development of the hybrid logic of statistical tests: 1920-1960
- Institutionalization of statistical tests in psychology: 1940-1960
- Increasing criticism of the use of statistical tests: 1940-present
- Failure of early “suggestions” to report effect sizes: 1994-present
- Rise of meta-analysis and meta-analytic thinking: 1976-present
- Report of the TFSI, 5th ed. *APA Publication Manual*: 1999-present

## Hybrid logic of statistical tests (1920-1960)

---

- Contemporary null hypothesis significance testing (NHST) is actually a hybrid of two different conceptual models:

- ✓ Fisher ( $p$ -value approach):

$p$  values under  $H_0$  only (1920s)

- ✓ Neyman-Pearson (fixed- $p$  approach):

$H_1$ , power analysis (1930s)

## Hybrid logic of statistical tests (1920-1960)

---

- Fisher and Neyman-Pearson models were integrated roughly between 1935-1950 by other statisticians
- The resulting hybrid (“Intro Stats”) model may have been rejected by Fisher and Neyman-Pearson, but for different reasons
- Composite nature of NHST may be a source of confusion today

## Institutionalization of statistical tests (1940-1960)

---

- Gigerenzer (1993): The *inference revolution* in psychology—roughly 1940-1960
- During this time, the “Intro Stats” method is adopted as basically *the* only way to test hypotheses
- Coincided with the *probabilistic revolution* in the natural sciences
- This “revolution” concerned the application of the concept of indeterminism as an explanatory construct to help explain the *subject* of study

Example: Chaos theory

## Institutionalization of statistical tests (1940-1960)

---

- However, in psychology the concept of indeterminism was applied to mechanize the *inference process*
- Since the 1970s, the use of statistical tests in psychology and other behavioral sciences is almost universal (see Figure 1.1)
  - ✓ Advantage: Made the administration of behavioral science research easier
  - ✓ Disadvantage: Gave the illusion of objectivity, has become dogma

## Increasing criticism of statistical tests (1940-present)

---

- Examples of early critical works:
  - ✓ Boring (1919)
  - ✓ Berkson (1942)
  - ✓ Morrison and Henkel (1970)
  - ✓ Rozeboom (1960)
  
- Examples of more recent critical works:
  - ✓ Carver (1978)
  - ✓ Cohen (1994)
  - ✓ Gigerenzer (1998)
  - ✓ Lykken (1991)

## Increasing criticism of statistical tests (1940-present)

---

- Exponential increase in the number of critical works since the 1940s (see Figure 1.2)
- Increasing numbers of respected behavioral scientists have called for a ban on statistical tests in psychology research journals (see Nix & Barnette, 1998)

## Failure of “suggestions” to report effect sizes (1994-present)

---

- Idea of effect size estimation is over a century old (i.e., it is not new)
- 4th ed. of the *Publication Manual of the American Psychological Association* (1994) encouraged, but did not require, the reporting of effect sizes
- A self-canceling mixed message?

## Failure of “suggestions” to report effect sizes (1994-present)

---

- Finch, Cumming, and Thomason (2001) found little evidence of reform in the reporting of results in journal articles over the period 1940-1999
- That there has been so little apparent change in methods of data analysis in the behavioral sciences is surprising
- Many researchers use today’s fast computers to conduct types of statistical analyses from the 1920s

## Meta-analysis and meta-analytic thinking (1976-present)

---

- Modern form of meta-analysis generally attributed to Gene Glass (1976) and Robert Rosenthal (1976)
- Has become an important tool for research synthesis in several disciplines, including psychology, education, and medicine
- Usually analyzes standardized effect sizes from a set of studies (published or not)

# Meta-analysis and meta-analytic thinking (1976-present)

---

- Meta-analytic thinking requires:
  - ✓ An accurate appreciation of the results of prior studies
  - ✓ A view of one's own study as making a modest contribution
  - ✓ Reporting of results so they can be included in future meta-analyses
  - ✓ Interpretation of new results via comparison with prior effect sizes
- This view is *incompatible* with the use of statistical tests as the sole way to test hypotheses

## Report of the TFSI (1999)

---

- APA's Board of Scientific Affairs convenes the Task Force on Statistical Inference (TFSI) in 1996
- Wilkinson and TFSI (1999) report—also available at:

<http://www.psycinfo.com/psycarticles/1999-03403-008.html>

## Report of the TFSI (1999)

---

- Some recommendations:
  - ✓ Minimally sufficient analyses (i.e., simpler is better)
  - ✓ Do not report statistics from computer output without knowing what they mean
  - ✓ Report observed effect sizes for primary outcomes
  - ✓ Report confidence intervals based on observed effect sizes
- However, the TFSI did *not* recommend a ban on statistical tests

## 5th ed. *Publication Manual of the APA* (2001)

---

- Effect sizes should “almost always” be reported
- Absence of effect sizes is cited as an example of a study defect
- Use of confidence intervals is “strongly recommended”

## 5th ed. *Publication Manual of the APA* (2001)

---

- Also did not call for ban on statistical tests
- Has been criticized for lack of detail about *how* to reform data analysis methods
- Not always possible to calculate effect sizes or confidence intervals, but it is in perhaps most studies

## Contexts for change in data analysis practices

---

- About 20+ research journals in the behavioral sciences now require the reporting of effect sizes
- This includes some flagship journals of large associations
- See B. Thompson's list of journals requiring effect sizes:

<http://www.coe.tamu.edu/~bthompson/journals.htm>

- This requirement is an effective ban on use of statistical tests alone

# Contexts for change in data analysis practices

---

- Merits of statistical tests are now debated in many different disciplines besides psychology

Examples: Nursing, medicine, and wildlife management (e.g., Anderson, Burnham, & W. Thompson, 2000)

- W. Thompson (2001) made a list of 402 citations of works in different disciplines that question the indiscriminant use of statistical tests:

<http://biology.uark.edu/Coop/Courses/thompson5.html>

- Kaufman (1998): The controversy over statistical tests is the major methodological issue of our generation

## Why change is inevitable and desirable

---

- > 50 years of debate about the technical merits of statistical tests has produced little real change in data-analysis practices
- Outcomes of statistical tests are widely misunderstood
- These false beliefs may result in cognitive distortions that hinder research progress in the behavioral sciences
- Results of statistical tests in perhaps most behavioral studies may not be very meaningful

# Why change is inevitable and desirable

---

- Statistical tests do not generally tell researchers what they really want to know
- Statistical significance says nothing directly about effect size or *substantive significance*, which includes:

Theoretical, practical, and clinical significance

- A specific set of methods is associated with clinical significance (chap. 4)

## Why change is inevitable and desirable

---

- Behavioral science research often has little impact—examples:
  - ✓ Miller (1999): Educational research seems to have little relevance for educational practices and policy
  - ✓ Beutler, Williams, Wakefield, and Entwistle (1995): Clinical psychology practitioners say that the clinical research literature is of limited value
  - ✓ Lykken (1991): Perhaps most published worked are never cited by another author, and many works have few readers (e.g., < 200)
- Perhaps part of the problem is the excessive or uncritical use of statistical tests in the behavioral sciences
- *If we don't make the future, others will do it for us*

# References

---

- American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: Author.
- American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.
- Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management*, *64*, 912-923.
- Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association*, *37*, 325-335.
- Beutler, L. E., Williams, R. E., Wakefield, P. J., & Entwistle, S. R. (1995). Bridging scientist and practitioner perspectives in clinical psychology. *American Psychologist*, *50*, 984-994.
- Boring, E. G. (1919). Mathematical vs. scientific importance. *Psychological Bulletin*, *16*, 335-338.
- Carver, R. P. (1978). The case against significance testing. *Harvard Educational Review*, *48*, 378-399.
- Cohen, J. (1994). The earth is round ( $p < .05$ ). *American Psychologist*, *49*, 997-1003.
- Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform. *Educational and Psychological Measurement*, *61*, 181-210.
- Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Vol. 1. Methodological issues* (pp. 311-339). Hillsdale, NJ: Erlbaum.
- Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, *21*, 199-200.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *10*, 3-8.
-

- Kaufman, A. S. (1998). Introduction to the special issue on statistical significance testing. *Research in the Schools*, 5, 1.
- Lykken, D. T. (1991). *What's wrong with psychology, anyway?* In D. Chicchetti & W. Grove (Eds.), *Thinking clearly about psychology* (Vol 1, pp. 3-39). Minneapolis, MN: University of Minnesota Press.
- Miller, D. W. (1999, August 6). The black hole of education research: Why do academic studies play such a minimal role in efforts to improve the schools? *Chronicle of Higher Education*, 45(48), A17-A18.
- Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy*. Chicago: Aldine.
- Nix, T. W., & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the Schools*, 5, 3-14.
- Rosenthal, R. (1976). *Experimenter effects in behavioral research*. New York: Halstead Press.
- Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, 57, 416-428.
- Thompson, W. L. (2001). 402 citations questioning the indiscriminate use of null hypothesis significance tests in observational studies. Retrieved November 11, 2001, from <http://biology.uark.edu/Coop/Courses/thompson5.html>
- Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.