8. Statistical power and effect size

LPO.8800: Statistical Methods in Education Research

Sean P. Corcoran

Last time

- Statistical hypothesis testing: tests about μ and π
- Null vs. alternative hypotheses; one-tailed vs. two-tailed tests
- Test statistics and *p*-values
- Significance levels (α)
- Using a $(1-\alpha)$% confidence interval to test a hypothesis with an α significance level
- Using `mean` or `ttest` in Stata for hypothesis testing

Type I and Type II errors

A hypothesis test can result in one of two types of incorrect decisions:

- **Type I error**: rejecting H_0 when it is actually true
- **Type II error**: not rejecting H_0 when it is false

Type I and Type II errors

	Reject H_0	**Do Not Reject H_0**
H_0 is true	Incorrect decision— ***Type I error*** (Pr = α)	*Correct decision* (Pr = 1-α)
H_0 is false	Correct decision (Pr = 1-β)	Incorrect decision— ***Type II error*** (Pr = β)

Type I and Type II errors

Example: criminal court

- H_0: not guilty
- **Type I error**: rejecting H_0 and convicting an innocent man
- **Type II error**: not rejecting H_0 and letting a guilty man go free
- Guilt "beyond a reasonable doubt"—implies a very low p-value is required, a very high confidence level (low significance level α).

Type I and Type II errors

Example: PSA (prostate specific antigen) screening for prostate cancer

- H_0: no prostate cancer
- **Type I error**: false positive—finding elevated levels of PSA and inferring a cancer growth when it does not exist
- **Type II error**: false negative—failing to detect an actual cancerous growth when it does exist

There are costly consequences for both types of errors:

- If a Type II error is made, growth exists and is untreated
- If a Type I error is made, detect tumor and perform unnecessary surgery

Probability of a Type II error

The probability of committing a Type II error is a bit more difficult to calculate than the probability of committing a Type I error (which is set by the researcher as α). This is because this probability depends on how far away the plausible alternative to μ_0 is.

- All else equal, we will be *more* likely to make a Type II error if the true μ is close—but not equal to—μ_0.
- All else equal, we will be *less* likely to make a Type II error if the true μ is far away from μ_0.

Probability of a Type II error

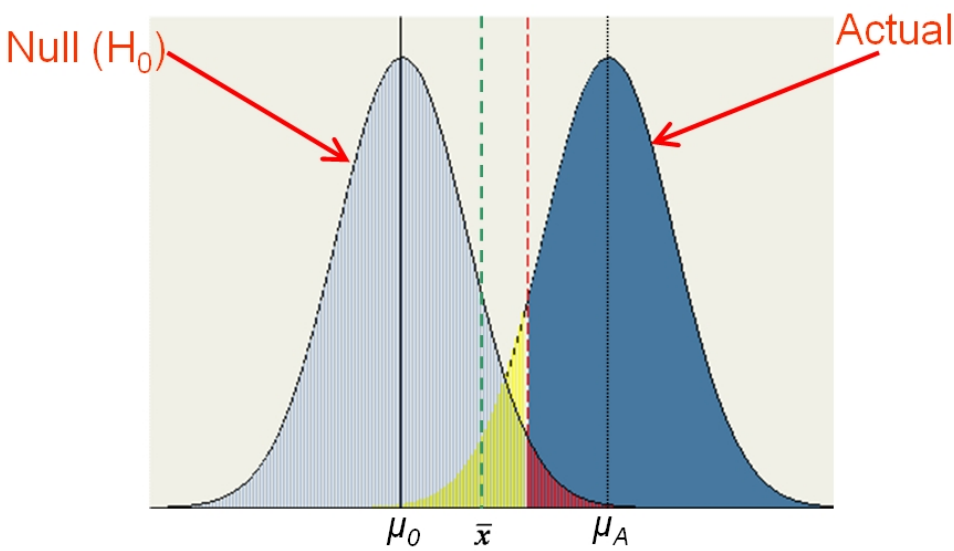


Figure: Distribution of $\bar{x}$ under H_0 and a specific alternative H_A

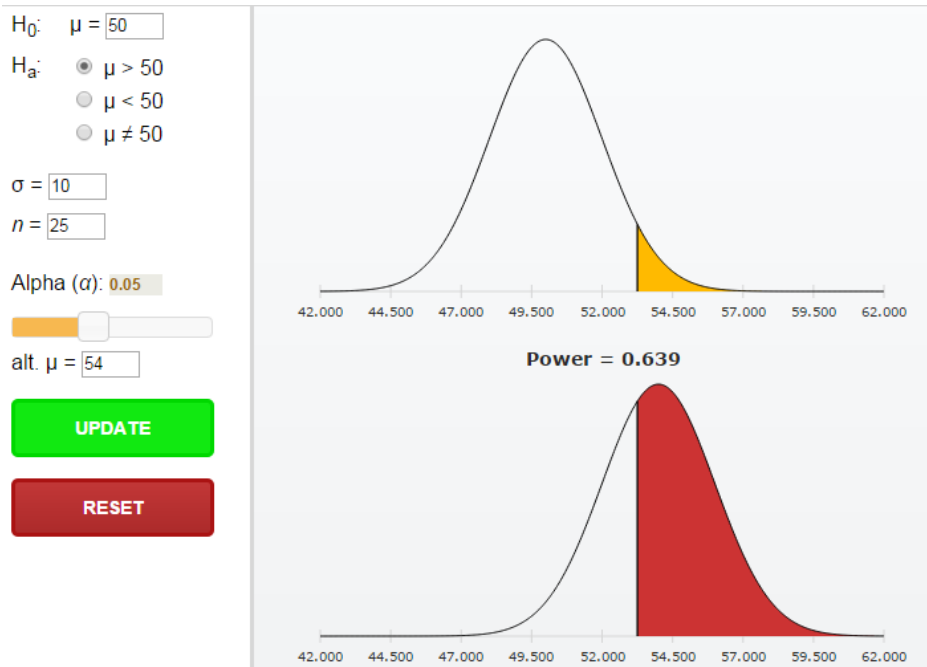
Types of errors

The probability of *correctly* rejecting H_0 when H_0 is false $(1 - \beta)$ is called the **power of the test**. The power of the test represents our ability to detect a difference between the null hypothesis and a particular alternative hypothesis. This is the dark blue region in the previous slide.

Note: the following figures were taken from an online applet linked on the class website:

`http://digitalfirst.bfwpub.com/stats_applet/stats_applet_9_power.html`

Power - 1



One-sided hypothesis test: $\mu_0 = 50$, $\sigma = 10$, $n = 25$, $\alpha = 0.05$. Find statistical power $(1 - \beta)$ when μ is actually 54.

Power - 1

If you were doing this manually, you would need to determine the value of $\bar{x}$ beyond which H_0 will be rejected (i.e., the yellow region above), find its z (or t) score in the *alternative* sampling distribution, and determine the probability of obtaining that score or something greater if H_a were true.

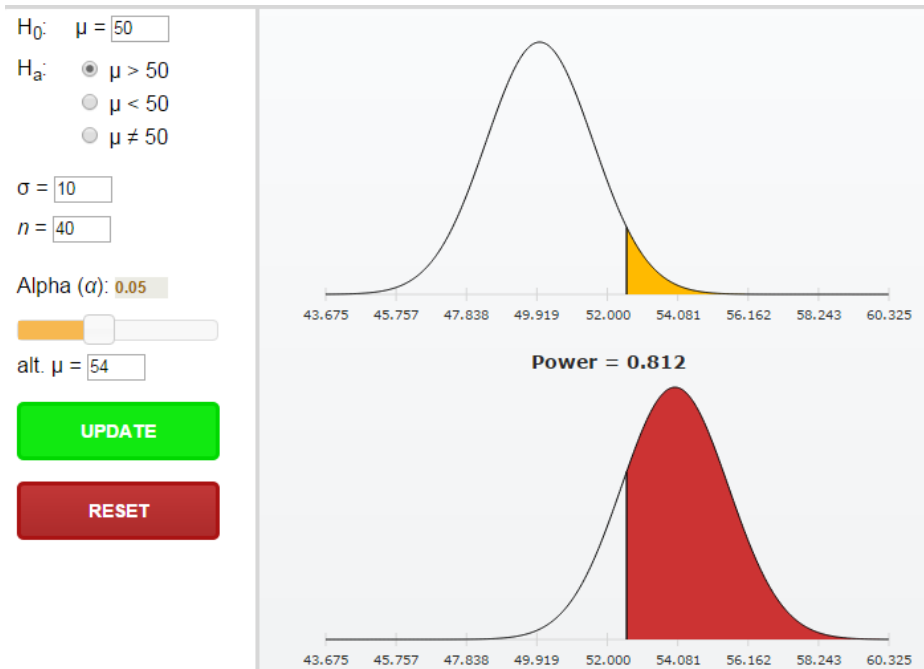
The $\bar{x}$ beyond which H_0 is rejected is: $50 + 1.645 * (10/\sqrt{25}) = 53.29$

The z-value in the *alternative* is: $(53.29 - 54)/(10/\sqrt{25}) = -0.355$

The probability of obtaining a $z > -0.355$ is **0.639**. `1-normal(-0.355)`

I have assumed normality for simplicity here. If σ were unknown this would affect the t value used above (it would be 1.677 rather than 1.645).

Power - 2

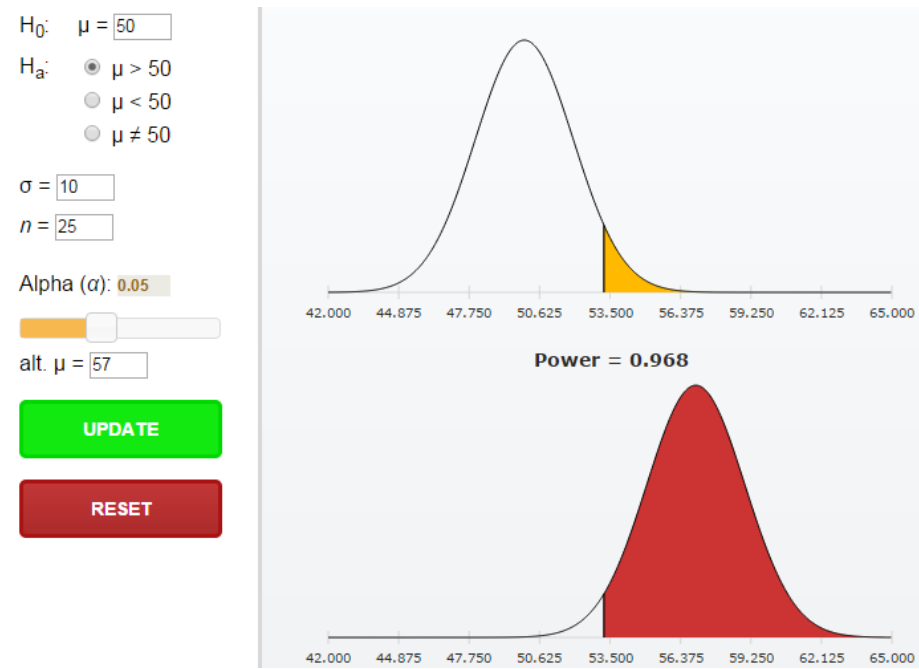


Consider what happens when n increases to 40.

Power - 3

Consider what happens when σ increases to 20 (keep $n = 25$).

Power - 4



Consider what happens when the alternative is further away (e.g. $\mu = 57$).

Power of the test

Things that affect the power of the test (our ability to discern our null hypothesis from an alternative):

- The effect size of interest: how far the alternative is away from the null. All else equal, the closer the alternative to the null, the lower the power.
- α, which determines when we reject. All else equal, a higher significance level the greater the power of the test.
- The standard error of the sample mean ($\sigma/\sqrt{n}$). All else equal, the smaller the standard error, the greater the power of the test. Because *sample size* decreases the standard error, a larger n (holding σ constant) will increase the power of the test.

Power of the test

Tools for calculating power

- Power applet: `http://digitalfirst.bfwpub.com/stats_applet/stats_applet_9_power.html`
- Stata power command: Statistics → Power and sample size → Means → One-sample → Test comparing one mean to a reference value. Select Compute: Power. Can calculate:
  - Power $(1 - \beta)$
  - Sample size requirements
- Stata can accept ranges of values (e.g., sample sizes, alternative hypotheses) and plot the results

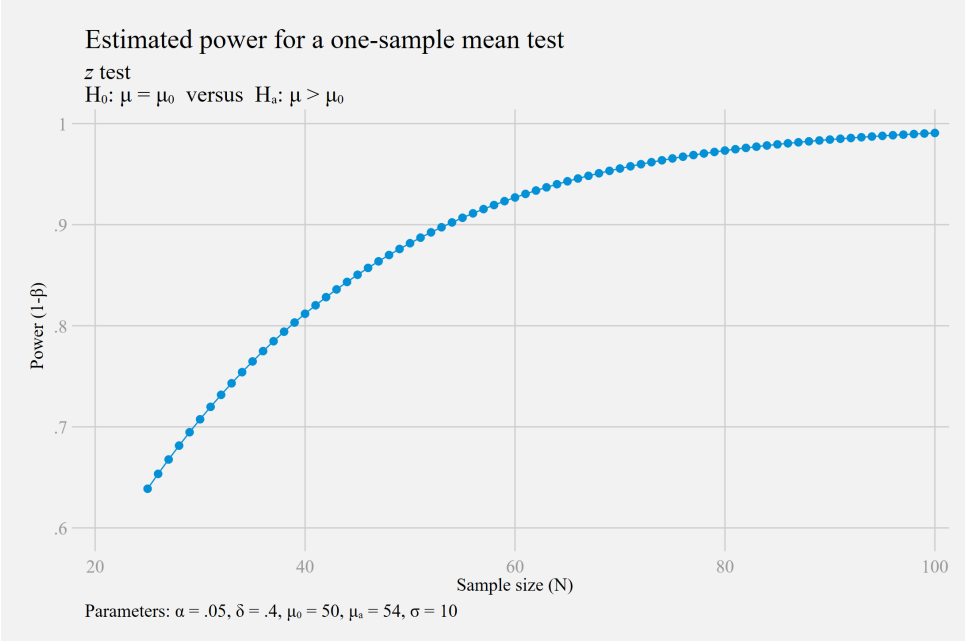
Power calculation in Stata - 1

Using "Power - 1" example above. $\mu_0 = 50, \sigma = 10, n = 25, \alpha = 0.05$. Find statistical power when μ is actually 54.

```
. power onemean 50 54, n(25) sd(10) knownsd onesided

Estimated power for a one-sample mean test
z test
Ho: m = m0  versus  Ha: m > m0

Study parameters:

        alpha =    0.0500
            N =        25
        delta =    0.4000
           m0 =   50.0000
           ma =   54.0000
           sd =   10.0000

Estimated power:

        power =    0.6388
```


Power calculation in Stata - 1

Using "Power - 1" example. $\mu_0 = 50, \mu_a = 54, \sigma = 10, \alpha = 0.05$. Find power for sample sizes ranging from $n = 25$ to $n = 100$.



Power calculation in Stata - 2

Your research group has developed an intervention designed to improve reading comprehension in 3rd grade. The typical (mean) gain on the 3rd grade reading test is 10 points, with a standard deviation of 6. Your intervention intends to improve on this. You randomly select n students to receive the intervention and calculate their mean gains ($\bar{x}$).

A standard significance test would be set up as:

$$H_0 : \mu = 10$$

$$H_1 : \mu > 10$$

The test statistic is: $t = (\bar{x} - 10)/(6/\sqrt{n})$, and you will reject if the probability of obtaining a t at least that large is < 0.05.

Power calculation in Stata - 2

In some circumstances, your test will fail to reject H_0 even when it is false (a Type II error). If your intervention has a positive effect, you'd like your test to reject H_0.

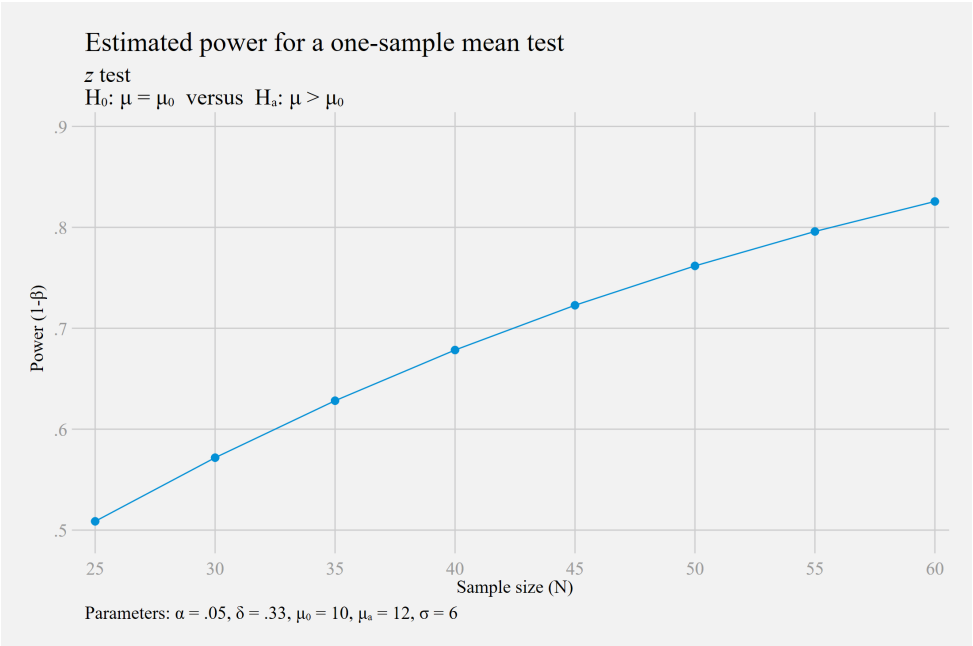
Your team believes the intervention will increase gains by 2 (from 10 to 12). What is the probability of a Type II error (and power) associated with various sample sizes (25-60)?

```
. power onemean 10 12, n(25(5)60) sd(6) knownsd onesided table(power beta N) graph

Estimated power for a one-sample mean test
z test
Ho: m = m0  versus  Ha: m > m0
```

power	beta	N
.5087	.4913	25
.5718	.4282	30
.6282	.3718	35
.6784	.3216	40
.7228	.2772	45
.7618	.2382	50
.7959	.2041	55
.8257	.1743	60

Power calculation in Stata - 2

Graphically:



Power - interpretation

Sticking with the previous example and $n = 45$:

- In 95% of random samples, this test will not reject H_0 if it is true (i.e., the mean gain in the study sample is 10, no different than the general population).
- In 5% of random samples, this test *will* reject H_0 when it is true—a Type I error.
- The above statements are by design, since we set $\alpha = 0.05$
- Suppose $\mu = 12$ in the study population—the study had a $+2$ point effect. If $n = 45$, we fail to reject H_0 in 27.7% of random samples. We don't detect the effect in these cases.
- In 72.3% of random samples, we properly reject H_0

What is a desirable power? A generally accepted value is **80%**.

Power - uses

The most common uses for power analysis are:

- Determining the **minimum required sample size**. How large of a sample n do I need in order to detect a given effect size $(\mu_a - \mu_0)$ $(1 - \beta)\%$ of the time?

- Determining the **minimum detectable effect size**. Given a sample size n, what is the smallest effect I will detect $(1 - \beta)\%$ of the time?

Practical significance

A *statistically significant* effect or difference is not necessarily a *practically important* one. In fact, with a large enough n, one can find statistically significant differences between the observed and hypothesized mean, even when the absolute difference between the two is quite small. For example:

- In our test of mean body temperature $H_0 : \mu = 98.6$, a large sample size could result in $\bar{x} = 98.59$ leading to the rejection of H_0. In other words, the difference could be statistically significant.

- But whether this difference is **practically significant** is a different matter, and depends on the context.

Practical significance is sometimes referred to as a "meaningfully large" effect, an "educationally significant" effect, or an "economically significant effect," depending on the context.

Effect size

An **effect size** is a measure of the degree to which the null hypothesis is false, in some meaningful unit (rather than in probabilistic terms). One measure of effect size is the number of *standard deviations* in the original distribution the observed sample mean is from the hypothesized one. This measure is sometimes called **Cohen's d**:

$$d = \frac{\bar{x} - \mu_0}{s}$$

Note we are using s from the original scale of x (not standard errors, which always decrease as n gets larger).

Note: in the Stata `power` output this is called *delta*.

Effect size

For example, suppose we were able to reject $H_0 : \mu = 98.6$ with $\bar{x} = 98.59$ (and assume $s = 0.60$). The effect size is:

$$d = \frac{\bar{x} - \mu_0}{s} = \frac{98.59 - 98.6}{0.60} = -0.017$$

The standard deviation of body temperature in the sample is 0.60. Our observed sample mean temperature is only 0.017 standard deviations below our hypothesized mean. Even if we can reject H_0, this small difference in temperature may not be practically significant.

Practical vs. statistical significance

	A	B	C	D
Sample size	10,000	10,000	9	1,000
Overall mean test score	200	200	200	200
Standard deviation	25	25	100	25
Girls' mean test score	175	199	175	199
Boys' mean test score	225	201	225	201
Δ = Difference (Boys - Girls)	**50**	**2**	**50**	**2**
Effect size (Δ/SD)	2	0.08	0.5	0.08
Practically significant?	Yes	No	Yes (if true)	No
Standard error (se) of difference in means	0.5	0.5	66.6	1.58
t-statistic for difference in means (Δ/se)	100	4	0.75	1.26
p-value	$p<0.0001$	$p<0.001$	$p>0.40$	$p>0.20$
Statistically significant?	Yes	Yes	No	No
Confidence interval	$50 \pm 1.96 * 0.5$ (49.02, 50.98)	$2 \pm 1.96 * 0.5$ (1.02, 2.98)	$50 \pm 2.31 * 66.6$ (-103.8, 203.85)	$2 \pm 1.96 * 1.58$ (-1.10, 5.10)

Source: Remler & Van Ryzin ch. 8. Note standard error for difference in means is $2 * (SD/\sqrt{n})$. Assumes the standard deviation is the same for boys and girls, and an equal number of boys and girls.

Practical vs. statistical significance

In column D, the results are neither practically nor statistically significant. But they still provide valuable information. Note the 95% confidence interval of (-1.1, 5.1). If we don't consider the *bounds* of this interval to be meaningful differences, then we can rule out practically meaningful effects.

When the confidence interval includes zero and rules out meaningful effects, it is sometimes called a "precise zero."

How to report results: advice

Most papers emphasize two characteristics of their findings, statistical significance and practical significance.

- Statistical significance tells us that the point estimate is statistically different from H_0 (often, zero).
- Practical significance assesses whether the point estimate is meaningful in size, given the context.

Some limitations to this approach:

- Zero and the point estimate itself are not the only values of interest. Sometimes zero is not an interesting null hypothesis.
- No information is provided about the *strength* of the evidence against the null.

How to report results: advice

Empirical papers in leading economics journals rarely discuss confidence intervals or the size of standard errors (Romer, 2020).

TABLE 1—INFORMATION REPORTED IN THE TEXT OF EMPIRICAL PAPERS IN THREE LEADING JOURNALS IN 2019

Discussed prominently	
Confidence intervals	14%
	(3)
Standard errors but not confidence intervals	10%
	(3)
Mentioned in passing	
Confidence intervals	6%
	(2)
Standard errors but not confidence intervals	7%
	(2)
Neither confidence intervals nor standard errors discussed	64%
	(5)

Notes: Standard errors are in parentheses. See text for details.

The upper end of a 2 SE confidence interval for papers that discuss confidence intervals is 20%. ;)

How to report results: advice

Romer (2020): "the tone is often that once it is known that estimates are statistically different from zero, the only aspect of the results that matters is the point estimates—almost as though when an estimate is significantly different from zero, it can be treated as exact."

Romer recommends reporting and discussing confidence intervals.

"Knowing significance is not enough to know what values of the parameter other than zero the data provide strong evidence against, and what values they provide little reason to object to."

Example

Consider two papers estimating the rate of return to an additional year of education (i.e., the % increase in annual earnings) . Both papers estimate $\bar{x} = 9.0$.

- Paper 1 has a standard error of 3.9
- Paper 2 has a standard error of 1.8

Both papers would claim statistical and practical significance. (For most people, the benefit to additional education would outweigh the costs).

- Paper 1 has a 95% CI of (1.4, 16.6)
- Paper 2 has a 95% CI of (5.5, 12,5)

Paper 1 cannot rule out effects that may be considered practically small.

How to report results: advice

The common interpretation of confidence intervals is that results provide strong evidence against parameter values outside of the CI, and equally supportive evidence for those inside the CI.

This exaggerates the uncertainty in the results. Ex post, one should view the point estimate as "more likely" while those further away "less likely".

Romer (2020) suggests presentation of narrower confidence intervals may be appropriate (e.g. ± 1 standard error):

"the natural, and roughly correct, shortcut interpretation would be that the results provide little information about the relative merits of different values within the 1 SE interval, moderate evidence against values in the 2 SE but not the 1 SE interval relative to the point estimate, and strong evidence against values outside the 2 SE band relative to the point estimate"

Benchmarking effect sizes

How do we know if an effect size is practically meaningful? Cohen (1969) proposed the following guidelines for interpreting d:

- 0.2 = small effect
- 0.5 = medium effect
- 0.8 = large effect

These benchmarks do not work well in all contexts, however, and the evidence suggests they are much too large for educational interventions (Kraft, 2020; Hill et al., 2008).

Benchmarking effect sizes

A better approach to interpreting effect size is to look to empirical benchmarks—that is, looking to existing evidence to tell us whether an effect is meaningful or not. Approaches in Hill et al. (2008):

- Normative expectations for growth in student achievement: "typical yearly growth"
- Policy-relevant gaps in student achievement by demographic group or school performance
- Effect sizes from past research for similar interventions and target populations.

Benchmarking effect sizes: typical student growth

Table 1

Average Annual Gain in Effect Size From Nationally Normed Tests

	Reading tests		Math tests	
Grade transition	Mean	Margin of error	Mean	Margin of error
Grade K–1	1.52	±0.21	1.14	±0.49
Grade 1–2	0.97	±0.10	1.03	±0.14
Grade 2–3	0.60	±0.10	0.89	±0.16
Grade 3–4	0.36	±0.12	0.52	±0.14
Grade 4–5	0.40	±0.06	0.56	±0.11
Grade 5–6	0.32	±0.11	0.41	±0.08
Grade 6–7	0.23	±0.11	0.30	±0.06
Grade 7–8	0.26	±0.03	0.32	±0.05
Grade 8–9	0.24	±0.10	0.22	±0.10
Grade 9–10	0.19	±0.08	0.25	±0.07
Grade 10–11	0.19	±0.17	0.14	±0.16
Grade 11–12	0.06	±0.11	0.01	±0.14

Sources. Annual gain for reading is calculated from seven nationally normed tests: California Achievement Test (CAT)–5th edition, Stanford Achievement Test (SAT)–9th edition, TerraNova–Comprehensive Test of Basic Skills (CTBS), Metropolitan Achievement Test (MAT8), TerraNova–CAT, SAT10, and Gates-MacGinitie. Annual gain for math is calculated from six nationally normed tests: CAT5, SAT9, TerraNova–CTBS, MAT8, Terra Nova–CAT, and SAT10. For further details, contact the authors (Bloom et al. 2006a, 2006b, 2007a, 2007b, in press; Lipsey et al., 2007).

Source: Hill et al. (2008)

Benchmarking effect sizes: typical achievement gaps

Table 2

Demographic Performance Gap in Mean NAEP Scores, by Grade (in Effect Size)

Subject and grade	Black–White	Hispanic–White	Eligible–ineligible for free/reduced-price lunch	Male–Female
Reading				
Grade 4	−0.83	−0.77	−0.74	−0.18
Grade 8	−0.80	−0.76	−0.66	−0.28
Grade 12	−0.67	−0.53	−0.45	−0.44
Math				
Grade 4	−0.99	−0.85	−0.85	0.08
Grade 8	−1.04	−0.82	−0.80	0.04
Grade 12	−0.94	−0.68	−0.72	0.09

Sources. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Reading Assessment, and 2000 Mathematics Assessment (Bloom et al., 2007a, 2007b, in press; Lipsey et al., 2007).

Source: Hill et al. (2008)

Benchmarking effect sizes

Based on a review of 747 randomized controlled trials in education, Kraft (2020) proposes the following benchmarks:

- < 0.05 = small effect
- 0.05 - 0.2 = medium effect
- > 0.20 = large effect

In determining practical significance in your context, ask "how large is the effect relative to other studies with broadly comparable features?"

Benchmarking effect sizes

Typical effect sizes vary by test subject (math or reading), scope of test, and sample size.

Table 1
Empirical Distributions of Effect Sizes From Randomized Control Trials of Education Interventions With Standardized Achievement Outcomes

		Subject		Sample Size					Scope of Test		DoE
	Overall	Math	Reading	≤100	101–250	251–500	501–2,000	>2,000	Broad	Narrow	Studies
Mean	0.16	0.11	0.17	0.30	0.16	0.16	0.10	0.05	0.14	0.25	0.03
Standard deviation	0.28	0.22	0.29	0.41	0.29	0.22	0.15	0.11	0.24	0.44	0.16
Mean (weighted)	0.04	0.03	0.05	0.29	0.15	0.16	0.10	0.02	0.04	0.08	0.02
P1	–0.38	–0.34	–0.38	–0.56	–0.42	–0.29	–0.23	–0.22	–0.38	–0.78	–0.38
P10	–0.08	–0.08	–0.08	–0.10	–0.14	–0.07	–0.05	–0.06	–0.08	–0.12	–0.14
P20	–0.01	–0.03	–0.01	0.02	–0.04	0.00	–0.01	–0.03	–0.03	0.00	–0.07
P30	0.02	0.01	0.03	0.10	0.02	0.06	0.03	0.00	0.02	0.05	–0.04
P40	0.06	0.04	0.08	0.16	0.07	0.10	0.06	0.01	0.06	0.11	–0.01
P50	0.10	0.07	0.12	0.24	0.12	0.15	0.09	0.03	0.10	0.17	0.03
P60	0.15	0.11	0.17	0.32	0.17	0.18	0.12	0.05	0.14	0.22	0.05
P70	0.21	0.16	0.23	0.43	0.25	0.22	0.15	0.08	0.20	0.34	0.09
P80	0.30	0.22	0.33	0.55	0.35	0.29	0.19	0.11	0.29	0.47	0.14
P90	0.47	0.37	0.50	0.77	0.49	0.40	0.27	0.17	0.43	0.70	0.23
P99	1.08	0.91	1.14	1.58	0.93	0.91	0.61	0.48	0.93	2.12	0.50
k (number of effect sizes)	1,942	588	1,260	408	452	328	395	327	1,352	243	139
n (number of studies)	747	314	495	202	169	173	181	124	527	91	49

Note. A majority of the standardized achievement outcomes (95%) are based on math and English language art test scores, with the remaining based on science, social studies, or general achievement. Weights are based on sample size for weighted mean estimates. For details about data sources, see Appendix A, available on the journal website. DoE = U.S. Department of Education.

Benchmarking effect sizes

How should we think about effect sizes? (Kraft, 2020)

- Effect sizes can be *descriptive* or *causal effects*. Correlational "effect sizes" are often much larger than causal ones.
- Effects on short-run outcomes are often larger than effects on long-run outcomes.
- Effects on specialized and researcher-designed instruments are often larger than those on broader instruments.
- Effect sizes are smaller when more measurement error is expected.

Benchmarking effect sizes

How should we think about effect sizes? (Kraft, 2020) - continued

- Studies with more targeted samples tend to have bigger effects than those with more inclusive samples.
- Effect sizes for an intervention tend to be larger when there is a greater treatment-control contrast.
- Treatment effects are larger if they are based on actual treatment, rather than a treatment *offer*.
- Cost matters: effects from lower-cost interventions are arguably more impressive than effects from higher-cost interventions.
- Effects of interventions are generally smaller when they are taken to scale.