ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 23-829

Multiply by 37: A Surprisingly Accurate Rule of Thumb for Converting Effect Sizes from Standard Deviations to Percentile Points

Paul T. von Hippel
University of Texas, Austin

Educational researchers often report effect sizes in standard deviation units (SD), but SD effects are hard to interpret. Effects are easier to interpret in percentile points, but conversion from SDs to percentile points involves a calculation that is not intuitive to educational stakeholders. We point out that, if the outcome variable is normally distributed, simply multiplying the SD effect by 37 usually gives an excellent approximation to the percentile-point effect. For students in the middle half of the distribution, the approximation is accurate to within 1 percentile point for effect sizes of up to 0.8 SD (or 29 to 30 percentile points).

VERSION: September 2023

Suggested citation: von Hippel, Paul T.. (2023). Multiply by 37: A Surprisingly Accurate Rule of Thumb for Converting Effect Sizes from Standard Deviations to Percentile Points. (EdWorkingPaper: 23-829). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/xk0b-ft25

Multiply by 37:
A Surprisingly Accurate Rule of Thumb for Converting Effect Sizes from Standard Deviations to Percentile Points

Paul T. von Hippel

University of Texas, Austin

Abstract

Educational researchers often report effect sizes in standard deviation units (SD), but SD effects are hard to interpret. Effects are easier to interpret in percentile points, but conversion from SDs to percentile points involves a calculation that is not intuitive to educational stakeholders. We point out that, if the outcome variable is normally distributed, simply multiplying the SD effect by 37 usually gives an excellent approximation to the percentile-point effect. For students in the middle half of the distribution, the approximation is accurate to within 1 percentile point for effect sizes of up to 0.8 SD (or 29 to 30 percentile points).

Multiply by 37:
A Surprisingly Accurate Rule of Thumb for Converting Effect Sizes from Standard Deviations to Percentile Points

Educational researchers often report effect sizes in standard deviations (SD). For example, we might say that a treatment raises students' test scores by 0.15 SD—or more precisely, that treated students score 0.15 SD higher than they would without treatment.

Yet effects reported in SDs can be unintuitive—and not just to educators and policy makers with limited training in statistics. Even among trained researchers, there is little agreement on how to interpret an effect expressed in SDs. An effect of 0.15 SD may be described as "trivial" in one study, but "substantial" in another. Authorities once classified all effects of less than 0.2 SD as "small" (Cohen, 1988), yet today it seems that most educational interventions have effects of 0.1 SD or less (Kraft, 2020).

There are several ways to make an effect expressed in SDs more interpretable (Lipsey et al, 2012). For example, we may compare the effect to the cost of treatment (Harris, 2009), to the effects of alternative treatments (Kraft, 2020), to the amount learned in a month of school (Lee et al. 2012), or to an achievement gap that we want to shrink (Lipsey et al., 2012).

Perhaps the simplest approach is to translate an effect from SDs to percentile points (Baird & Pane, 2019). If we assume that an educational outcome is normally distributed (as many test scores approximately are), then an effect of 0.15 SD means that the treatment would raise a median student's score by approximately 6 percentile points—from the 50th percentile to the 56th. Because educational stakeholders are familiar with percentile points, translating an effect into percentile points can convert an abstract technical conversation about what an effect of 0.15 SD really means into a concrete policy conversation about whether an improvement of 6 percentile points is worth a treatment's cost in time, trouble, or money.

There are two concerns that deter researchers from converting effects to percentile points as often as we might like. First, the calculation seems to require using the cumulative standard normal distribution. The calculation is not difficult for someone with a little statistical training and a spreadsheet, but it is not a calculation that we can do in our heads, and the cumulative normal distribution is as difficult to explain to stakeholders as the SD is.

The second concern is that conversion from SDs to percentile points depends on where the student lies in the distribution. For a student who would score at the median without treatment, an effect of 0.15 SD will raise their score by 6 percentile points (from the 50^{th} to the 56^{th}). But for a student who would score at the 10^{th} percentile without treatment, an effect of 0.15 SD will only raise their score by 3 percentile points (from the 10^{th} to the 13^{th}).

While both concerns are valid, they are only relevant for extremely large effect sizes and students in the extremes of the distribution. The following rule of thumb works surprisingly well for most students and the vast majority of effect sizes encountered in practice:

To convert an effect size to approximate percentile points, simply multiply the SD effect by 37.

Multiplying by 37 is a transparent calculation that many of us can approximate in our heads while reading a report, giving a presentation, or discussing results in a meeting. We do not need to consult a table or spreadsheet, and we can explain the calculation to stakeholders with limited statistical training. Although the rule is imperfect, it quite often comes within a percentile point of the correct answer.

The left half of Table 1 describes treatment effects on a student who would score at the median if untreated. For treatment effects up to 0.80 SD—which account for over 90 percent of the effects obtained in educational research (Kraft, 2020)—the approximation of multiplying the SD effect by 37 comes within a percentile point of the answer obtained by using the cumulative normal distribution.

The right half of Table 1 shows that the accuracy of the approximation is not limited to students near the median. Indeed, for a student who, if untreated, would score at the 25th percentile, the approximation comes within a percentile point of the cumulative normal calculation for every effect size up to 1.00 SD.

Although Table 1 describes the difference between the linear approximation and the cumulative normal calculation as approximation errors, we should not take the cumulative normal calculation as gospel. Most test scores are not quite normally distributed, so calculations based on the normal distribution are approximations as well. In fact, the true effect of treatment often differs by about a percentile point from the effect implied by the cumulative normal calculation (Baird & Pane, 2019). In other words, the approximation error that comes from multiplying the SD effect by 37 is typically no worse than the approximation error that comes from assuming test scores are normally distributed.

Figure 1 uses the standard normal cumulative distribution to illustrate why the rule of thumb works. The horizontal axis shows Z scores representing student outcomes (e.g., test scores) in SDs from the mean, and the vertical axis represents the same outcomes as percentile points. The relationship between SDs and percentiles is nonlinear over the full range of the graph, but between approximately the 20th and 80th percentiles (i.e., between Z scores of -0.8 and

+0.8) the relationship is approximately linear with a slope of approximately 37.[1] What this means is that multiplying the SD effect by 37 works reasonably well for students who score between the 20th and 80th percentiles.

Outside of that range, the curve is flatter, so multiplying the SD effect by 37 will overestimate the percentile point effect. That means that an SD effect will correspond to a smaller percentile point effect in the extremes of the distribution than in the middle. For example, we can say that an effect of 0.1 SD is 4 percentile points at most—about 4 percentile points for students near the middle of the distribution, and less for students near the extremes.

[1] Under different criteria for fit, slightly different lines fit the curve inside the box. For example, the least squares line has a slope of 37.5, and the secant line connecting the 20th and 80th percentiles has a slope of 36. We chose a slope of 37 because it ensured that the line was within 1 percentile point of the curve for every point inside the box.

References

1. Baird, M. D., & Pane, J. F. (2019). Translating standardized effects of education programs into more interpretable metrics. *Educational Researcher, 48*(4), 217-228.
2. Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum.
3. Harris, D. N. (2009). Toward Policy-Relevant Benchmarks for Interpreting Effect Sizes: Combining Effects With Costs. *Educational Evaluation and Policy Analysis*, *31*(1), 3–29. https://doi.org/10.3102/0162373708327524
4. Lee, J., Finn, J., & Liu, X. (2018). Time-Indexed Effect Size for Educational Research and Evaluation: Reinterpreting Program Effects and Achievement Gaps in K–12 Reading and Math. *Journal of Experimental Education*.
5. Lipsey, M.W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M.W., Roberts, M., Anthony, K.S. and Busick, M.D. (2012). *Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms*. (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Tables and Figures

Table 1. Effect sizes on two hypothetical students.

For a 50th percentile student				For a 25th percentile student			
Effect in SD	Effect in normal percentile points	Approximate percentile effect as 37 times SD effect	Approximation error	Effect in SD	Effect in normal percentile points	Approximate percentile effect as 37 times SD effect	Approximation error
0.05	2	2	-0.1	0.05	2	2	0.2
0.10	4	4	-0.3	0.10	4	3	0.4
0.15	6	6	-0.4	0.15	6	5	0.6
0.20	8	7	-0.5	0.20	7	7	0.6
0.25	10	9	-0.6	0.25	9	9	0.7
0.30	12	11	-0.7	0.30	11	10	0.7
0.35	14	13	-0.7	0.35	13	12	0.7
0.40	16	15	-0.7	0.40	15	14	0.6
0.45	17	17	-0.7	0.45	17	16	0.5
0.50	19	19	-0.6	0.50	19	18	0.4
0.55	21	20	-0.5	0.55	20	20	0.3
0.60	23	22	-0.4	0.60	22	22	0.2
0.65	24	24	-0.2	0.65	24	24	0.0
0.70	26	26	0.1	0.70	26	26	-0.1
0.75	27	28	0.4	0.75	28	28	-0.3
0.80	29	30	0.8	0.80	30	30	-0.4
0.85	30	31	1.2	0.85	31	32	-0.5
0.90	32	33	1.7	0.90	33	34	-0.6
0.95	33	35	2.3	0.95	35	36	-0.7
1.00	34	37	2.9	1.00	37	38	-0.8

Note. The effect in SD is $\Delta Z = Z_1 - Z_0$, where Z_1 and Z_0 represent the student's standardized scores with vs. without treatment. If scores are normally distributed, then the effect in percentile points is $\Delta P = \Phi(Z_1) - \Phi(Z_0)$, where $\Phi()$ is the cumulative standard normal distribution. We approximate the percentile point effect by $\Delta\tilde{P} = 37\Delta Z$, and the table shows the error of the approximation $e = \Delta\tilde{P} - \Delta P$.

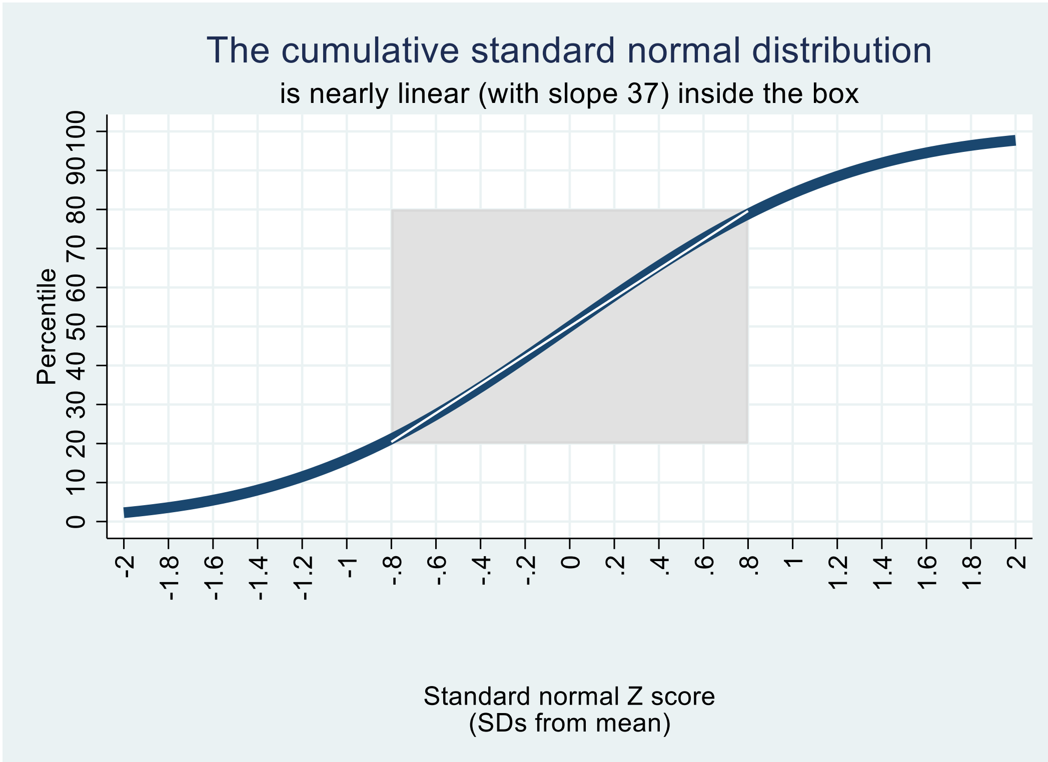

Figure 1. Why multiplying SD effects by 37 is a good approximation for scores inside the box.