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

Paul 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 converting SDs to percentile points involves a calculation that is not transparent to educational stakeholders. We show 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 three-fifths of a normal distribution, this rule of thumb is always accurate to within 1.6 percentile points for effect sizes of up to 0.8 SD. Two examples show that the rule can be just as accurate for empirical effects from real studies. Applying the rule to Kraft's empirical benchmarks, we find that the least effective third of educational interventions raise scores by 0 to 2 percentile points; the middle third raise scores by 2 to 7 percentile points; and the most effective third raise scores by more than 7 percentile points.

Multiply by 37 (Approximately): 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, a study might report that treatment raised some students' test scores by 0.2 SD—meaning that those students scored 0.2 SD higher, on average, than they would have without treatment.

Yet effects reported in SD units are unintuitive. When told that a program raised test scores by 0.2 SD, a nonprofit leader or principal will often ask, "Is that good or bad?" Different researchers answer this question differently. Some describe effects of nearly 0.2 SD as "quite large" (e.g., Kane & Staiger, 2008), while others describe them as "trivial to small" (e.g., Cheng et al., 2019).

Not only do researchers disagree on which SD effects are large or small, many researchers also have biased and unreliable intuitions about what effects expressed in SD units actually look like. When asked to illustrate an effect size of 0.2 SD, most researchers produce a misleading graph that at least doubles the requested effect, showing two normal distributions that are separated by 0.4 SD or more (Schuetze & Yan, 2023). Conversely, when shown two normal distributions that are actually separated by 0.2 SD, most researchers underestimate the illustrated effect, guessing that it is 0.05 SD c. Impressions of visually presented effect sizes vary dramatically across researchers, suggesting that we do not share a common or realistic understanding of what SD effects really represent (Schuetze & Yan, 2023).

One way to describe effects more intuitively is to express them in percentile points. Because standardized test scores are commonly reported as percentiles, many students, parents, teachers, and administrators understand that scoring at, say, the 50th percentile means scoring higher than half of students. They can also appreciate that improving a score by 10 percentile points—say from the 50th to

the 60th—means scoring higher than an additional 10 percent of students. So compared to SD effects, percentile point effects are relatively clear and accessible, not just to researchers but also to education stakeholders at all levels. The What Works Clearinghouse (2022) calls percentile point effects an "improvement index" and argues that they can "help readers judge the practical importance of the magnitude of intervention effects."

It is possible to calculate percentile point effects directly, for example by comparing the median percentile rank of treated and control students at the end of a study. But few studies do this, and reporting effects in SD units remains the norm. So people who communicate or consume the results of education research need a quick and convenient way to translate reported SD effects into equivalent percentile point effects.

In carrying out the translation, it is common to assume that the outcome variable—often a test score—is normally distributed. For a student who would score at the median of a normal distribution without treatment, an effect of 0.2 SD would raise their score by 8 percentile points—from the 50th percentile to the 58th. Translating an effect from 0.2 SD to 8 percentile points can convert a frustrating and abstract technical conversation about what an effect of 0.2 SD really *means* into a concrete policy conversation about whether an improvement of 8 percentile points is worth the intervention'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 is widely thought 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 many of us can do in our heads, and it is far from transparent to an education leader with limited statistical training. In fact, explaining the cumulative normal distribution to someone who is unfamiliar with it is every bit as hard as explaining an SD effect.

The second concern is that conversion from SDs to percentile points sometimes depends on where the student lies in the distribution. For a student who would score at the median without treatment, an effect of 0.20 SD will raise their percentile rank by 8 points (from the 50th percentile to the 58th). But for a student who would score at the 10th percentile without treatment, an effect of 0.20 SD will only raise their percentile rank by half as much—by 4 percentile points (from the 10th percentile to the 14th).

While both these concerns are valid, they only apply to unusually large effect sizes and students in the tails of the distribution. The following rule of thumb works surprisingly well for most students and the vast majority of effects reported in educational research:

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

Unlike calculations involving the cumulative standard normal distribution, 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 who have limited statistical training. Yet for most students and the vast majority of effect sizes, multiplying by 37 usually comes within 1 percentile point of the result obtained by using the cumulative normal distribution.

The calculation can also be reversed. For example, suppose you are planning a trial of a novel intervention. To clarify stakeholders' hopes, or calculate the sample size required to detect the effect, you would like a practitioner to tell you how large an effect they expect, or how small an effect would disappoint them. The practitioner may have trouble expressing their hopes in SD units, but they may more readily express themselves in percentile points. To translate their percentile point guess into SD units, you can simply divide it by 37.

In the rest of this paper, we'll demonstrate multiplying by 37 and explain its rationale, uses, and limitations.

Using the rule in data from a normal distribution

To start, let's assume that the outcome variable has a normal distribution. Then we can convert an effect from SD units to percentile points in two different ways:

1. an exact formula (given later) which uses the cumulative standard normal distribution, vs.
2. an approximation, which simply multiplies the SD effect by 37.

The top of Figure 1 compares these approaches for a student who, if untreated, would score at the median. For such a student, the two approaches agree closely. For any effect size up to 0.8 SD—that is, for over 90 percent of effects reported in educational research (Kraft, 2020)—the approximation of multiplying the SD effect by 37 comes within less than 1 percentile point of the exact answer obtained by using the cumulative normal distribution. For example, an effect of 0.8 SD translates to 28.8 percentile point if we use the cumulative normal distribution, or 29.6 percentile point if we multiply by 37. The approximation error is just 0.8 percentile points.

Percentile point effects are often reported as though they were only valid for students near the 50th percentile. But as the bottom of Figure 1 shows, multiplying by 37 works just as well for a student who, if untreated, would score at the 25th percentile of a normal distribution. For such a student, an effect of 0.80 SD translates to 30.0 percentile point if we use the cumulative normal distribution, or 29.6 percentile point if we multiply by 37—an approximation error of just 0.4 percentile points. More generally, for a student who if untreated would score at the 25th percentile of a normal distribution, multiplying by 37 comes within 1 percentile point of the exact answer (obtained using the cumulative normal distribution) for *any* effect size of between 0 and 0.8 SD (and even larger).

In general, if the outcome is normally distributed, multiplying by 37 usually comes within 1 percentile point of the exact answer—and always comes within 1.6 percentile points—for any student whose score stays approximately in the middle three-fifths of the distribution.

Can the rule work in real data?

Figure 2 shows that multiplying by 37 can also work well for effects from real empirical studies. To check this, we used results from two studies that reported effects in both SDs and percentile points.

- One study was a randomized controlled trial of a "personalized learning" intervention. The trial report estimated 24 effects of personalized learning—one effect for each of two subjects (reading, math) in each of 12 grades (kindergarten through 11th) (Baird & Pane, 2019; Pane et al., 2017).
- The other study was a regression discontinuity evaluation of the federally funded Reading First program (Gamse et al., 2008). The study report estimated 11 effects of Reading First; each estimate represented the program's effect on some skill (reading comprehension or the ability to decode letters into sound) at some grade level (1st, 2nd, or 3rd). Effects were also broken out by years of exposure (1, 2, 3, or all exposures together) and by when schools received Reading First funding (early, late, or all schools together).

Among 35 effects ranging from -0.20 SD to +0.43 SD (or -8 to +16 percentile points), multiplying the reported SD effect by 37 came within 1 point of the reported percentile point effect in every case but one.

This example shows that multiplying by 37 can be useful even if the outcome variable does not have a perfectly normal distribution. In the personalized learning study, the outcomes were math and reading scores from the Measures of Academic Progress (MAP) test published by NWEA. The

distribution of MAP scores can deviate noticeably from normality in the left tail.[1] In the Reading First Impact Study, one outcome was reading comprehension scores from the Stanford Achievement Test, 10th edition (SAT10), whose distribution can be bimodal, platykurtic, skewed left or skewed right (Shanley et al., 2019). Another outcome was decoding scores from the Test of Silent Word Reading Fluency, which counts the words that a student can read in 3 minutes; note that counts are often skewed to the right.

The empirical accuracy of multiplying these study estimates by 37 is also encouraging because each study calculated the percentile point effect somewhat differently. Neither study assumed that scores were normally distributed. Instead, in the Reading First Impact Study, the mean test scores of the treatment and control groups were converted to corresponding percentiles in a national reference distribution, and then the control percentile was subtracted from the treatment percentile. Mean scores in the control condition were between the 33rd and 46th percentiles, showing again that percentile gains (and the multiply-by-37 approximation) are not valid only at the median. Both the treatment and control percentiles were rounded to the nearest whole number before subtracting, so the estimated difference suffers from rounding error and may differ by 1 percentile point from the true difference; this may explain some slight discrepancies between the reported percentile effects and the approximation of multiplying by 37.

The personalized learning study estimated percentile point effects in a more complicated way—estimating the distribution of the control group's scores and then calculating the percentile that would be achieved if they added the treatment effect to the control median. The investigators noted that they estimated the control distribution in a nonparametric fashion that did not assume normality but

[1] Since neither the personalized learning study nor the Reading First Impact Study reported the distribution of scores, this paragraph reports the score distributions observed in other sources. We thank Megan Kuhfeld, Senior Research Scientist at NWEA, for sharing density plots of MAP scores.

"allow[ed] the distribution of scores to take any shape, such as skewed, bimodal, or highly kurtotic" (Baird & Pane, 2019, p. 223).

Despite these encouraging results, we would not claim that multiplying by 37 works well for every educational outcome. Surely, someone could find a highly non-normal distribution where it worked poorly—but then the usual cumulative normal calculation would work poorly as well! Our point is not that multiplying by 37 is always accurate, but that when the cumulative normal calculation is accurate, multiplying by 37 usually works practically as well.

Why does it work?

Figure 3 helps to illustrate why multiplying by 37 works for a normally distributed outcome. We will explain it in three different ways, starting with a simple explanation suitable for someone with limited exposure to statistics, and finishing with a more sophisticated explanation that uses a little calculus and optimization. We are brief in the article but provide more detail in the Appendix.

Explanation using the bell curve

The top of Figure 3 shows the standard normal probability density function (PDF)—the familiar bell curve taught in practically every introductory statistics course. A familiar fact about the normal distribution, also taught in practically every course, 68 percent of the probability lies within 1 SD of the mean. So an effect that raises a student's score from 1 SD below to 1 SD above the mean would raise their score by 68 percentile points. In this example, a 2 SD effect is equivalent to 68 percentile points, suggesting a rule that multiplies the SD effect by 34—not too far from 37.

Another fact about the normal distribution, not taught as often, is that 38 percent of the probability lies within half a SD of the mean. So an effect that raises a student's score from 0.5 SD

below to 0.5 SD above the mean would raise their score by 38 percentile points. In this example, a 1 SD effect is equivalent to 38 percentile points, suggesting a rule that multiplies the SD effect by 38—again not too far from 37.

In general, the best multiplier depends on where a student starts in the distribution and how much the intervention increases or decreases their score. But for students who stay in the middle three-fifths of the distribution, multipliers between 34 and 40 work reasonably well, and a multiplier of 37 works best on average.

Explanation using the cumulative normal distribution

The next two sections are more technical and some readers may wish to skip ahead to the section titled "When does multiplying by 37 work?"

Figure 3 shows the standard normal cumulative distribution function (CDF), which represents the relationship between a student's percentile rank *P* and how many standard deviations *Z* they are from the mean of a normal distribution. The standard normal CDF is often represented by the Greek letter Φ:[2]

$$P = \Phi(Z)$$

For example, a student who is 0 standard deviations from the mean is at the 50th percentile ($\Phi(0) = 50$), a student who is Z=0.8 standard deviations above the mean is at the 79th percentile ($\Phi(0.8) = 79$), and a student who is Z=0.8 standard deviations below the mean is at the 21st percentile ($\Phi(-0.8) = 21$).

Over the full range of the distribution, the CDF has an S shape, but between the *P*=21st and 79th percentiles—that is between Z=0.8 SD above and Z=-0.8 SD below the mean—the CDF is approximately linear with an intercept of 50 and a slope not too far from 37.

[2] The standard normal CDF is commonly defined as returning a fractile between 0 and 1. Here we want to return a percentile between 0 and 100, so our definition multiplies the common definition by 100.

$$\hat{P} = \hat{\Phi}(Z) = 50 + 37\,Z$$

The linear approximation $\hat{P}$ comes within 0.8 percentile points of the true value of *P* for every value of Z from -0.8 to 0.8.

This means that multiplying by 37 translates an SD effect into an approximate percentile point effect, provided the student stays between the 21st and 79th percentile. To see that explicitly, notice that the SD effect on an individual student (ΔZ) is the difference between how many standard deviations from the mean they would score under treatment (Z_t) vs. how many standard deviations from the mean they would score under the control (Z_c):

$$\Delta Z = Z_t - Z_c$$

Likewise the percentile point effect on an individual student (ΔP) is the difference between the student's percentile rank under treatment (Z_t) and their percentile rank under control (Z_c):

$$\Delta P = P_t - P_c$$

Over the full range of the normal distribution, the exact relationship between ΔP and ΔZ is not linear:

$$\begin{aligned} \Delta P &= P_t - P_c \\ &= \Phi(Z_t) - \Phi(Z_c) \qquad \text{(exact formula)} \end{aligned}$$

But in the range where $\Phi(Z)$ can be approximated by 50+37 *P*, the percentile point effect can be approximated by:

$$\begin{aligned} \Delta\hat{P} &= \hat{\Phi}(Z_t) - \hat{\Phi}(Z_c) \\ &\approx (50 + 37Z_t) - (50 + 37Z_c) \\ &= 37\,\Delta Z \qquad \text{(approximation)} \end{aligned}$$

That is, multiplying the SD effect by 37 can approximate the percentile point effect.

Explanation relating the bell curve to the cumulative distribution function

The previous explanations are compatible, because the probability density function (PDF) in the first explanation is related to the cumulative distribution function (CDF) in the second. Specifically, the PDF is the *derivative* or slope of the CDF, as students are taught in every mathematical statistics course (e.g., Hogg et al., 2012).[3] Knowing this, you can look at the PDF (top of Figure 3) to get the slope of the CDF (bottom of Figure 3) at each value of *Z* (SDs from the mean).

Specifically, at the mean (Z=0), the slope if the CDF is 40; at 0.4 SD above or below the mean (Z=0.4 or -0.4), the slope is 37; and at 0.8 SD above or below the mean (Z=0.8 or -0.8), the slope is 29. This is why our discussion of the bell curve found that larger multipliers were appropriate for students closer to the mean. Although the optimal slope is not exactly 37 through the whole range from Z=+0.8 to -0.8, it is close enough for 37 to be a good approximation for students who stay within this range.

When does multiplying by 37 work?

We have stated that multiplying by 37 works for students who "stay" between the 21st and 79th percentile (or, equivalently, between Z=+0.8 and -0.8 SDs from the mean). We should be explicit about what we mean by "stay." Multiplying the SD effect by 37 works for students who would score between the 21st and 79th percentile without treatment, *and would still score between the 21st and 79th percentile if they were treated*.

If treatment takes a student outside of that range, or if they would score outside that range if they were untreated, then multiplying the SD effect by 37 does not work as well, because the student will have strayed outside the range where the relationship between percentiles and SDs is approximately

[3] Like the CDF, the PDF has been multiplied by 100 here so that values can be interpreted on a percentile point scale.

linear. More specifically, if the treated or untreated score is in the top or bottom 20 percent of a normal distribution, then multiplying the SD effect by 37 will always *overestimate* the percentile point effect.

But knowing when we are overestimating an effect is useful, too, because it means that, in the tails, we can treat multiplying by 37 as an *upper bound*. It can be useful to place an upper bound on effect size, particularly if the effect is small. For example, if an SD effect is 0.05 SD, we can say that it is only 2 percentile points for students in the middle three-fifths of the distribution—and even less for students in the tails.

We tailored the multiply-by-37 rule to work for normally distributed outcomes, but Figure 2 showed that it can also work reasonably well for real test scores, which do not have a perfectly normal distribution. While would not claim that multiplying by 37 works for any data regardless of how it is distributed, we should point out that, since the rule is only meant to in the middle three-fifths of the distribution, non-normality in tails may not matter as much. For example, histograms shared with us by the publisher of the MAP tests (NWEA) suggest that MAP scores depart from normality primarily in the bottom 20 percent of the distribution. This may not have affected results from the Reading First Impact Study (Figure 2), which estimated the effect of Reading First on the MAP scores of students who would score in the 33rd to 50th percentiles if untreated.

Why 37?

Multiplying by 37 is just an approximation, and other multipliers are certainly possible. Any multiplier between 35 and 40 would give fairly similar estimates. A case can be made for multiplying by 40 because it is an easier mental calculation and gives very serviceable estimates, especially for small effects near the median.

That said, multiplying by 37 is optimal in the sense that it ensures the approximation error is never larger than 1.6 percentile points (and usually smaller than 1 percentile point) for students who stay between the 21st and 79th percentile of a normal distribution.

A different multiplier would be optimal if we focused on a different range of the normal distribution or tried to minimize a different function of the approximation error (such as the mean squared error). But the range between the 21st and 79th percentile seems to be the widest range over which any multiplier can work well, and it seemed desirable to minimize the largest error that can occur, rather than just ensuring that, say, squared errors are small on average.

The Appendix gives more detail on the derivation and properties of multiplying by 37.

Benchmarks for Percentile Point Effects

Although converting SD effects to percentile points makes them more interpretable, it can still be helpful to interpret percentile point effects with respect to some benchmark.

Table 1 uses the multiply-by-37 rule to convert benchmarks for effect size from SD units to percentile points. For example, Kraft (2020) derived empirical benchmarks from an inventory of rigorous studies of educational interventions. Converted from SDs to percentile points, Kraft's benchmarks suggest that the least effective third of interventions raise most students' scores by 0 to 2 percentile points, the middle third raise scores by 2 to 7 percentile points, and the most effective third raise scores by more than 7 percentile points.

Kraft's benchmarks contrast with Cohen's (1970) older benchmarks, which when converted to percentile points suggested that "small" effects raised most scores by approximately 7 percentile points, "medium" effects raised scores by approximately 19 percentile points, and "large" effects raised scores by approximately 30 percentile points. Cohen's benchmarks were "somewhat arbitrary" (Cohen, 1962)

and based on his experience with laboratory experiments in psychology. Cohen (1970) cautioned that they should not be generalized to other fields, but they have been widely used in education, where they now seem quite optimistic. According to Kraft's inventory, 70 percent of education effects would be "small" by Cohen's standards, and only about 4 percent of effects (less than 1 percent in large studies) would be "large" by Cohen's standards (Kraft, 2020, tbl. 1)

It can also be helpful to compare the effect of an intervention to its cost. As several authors have pointed out, if two interventions achieve similar effects, we should prefer the intervention with lower cost; likewise if two interventions have similar costs, we should prefer the intervention with the larger effect (Kraft, 2020; Levin et al., 2017). However, if interventions differ in both cost and effect, then comparison is more difficult. For example, suppose one intervention produces an effect of 0.06 SD (2 percentile points) at a cost of $100 per student, while an alternative can produce an effect 10 times larger (0.6 SD, 22 percentile points) at a cost that is 20 times higher ($2,000 per student). Measured in SDs or percentile points per dollar, the first intervention appears twice as cost-effective (Harris, 2009), but it only raises scores by 2 percentile points. The 22 percentile point improvement of the second intervention could be worth paying for, even if the cost per percentile point is twice as high.

Another popular option is to translate the effect into months or years or learning. This attractive metric is easy to apply in longitudinal studies where the benefit to the treatment group can be compared to the amount learned by the control group. But it is more problematic to compare the effect of an intervention to the amount learned by different students in a different study. The challenge is that learning rates are not constant but vary by age, subject, test, and other factors. Age is the most important factor; in general, young children gain reading and math skills much faster than older children, so that an effect of 0.2 SD is equivalent to about a month of learning in kindergarten but a year of learning in 9th grade (Bloom et al., 2008). When grade is held constant, annual gains still vary across tests; 9th grade

reading gains, for example, can be as large as 0.32 SD or as small as 0.04 SD, depending on what test is used to measure them (Bloom et al., 2008). Annual gains can be averaged across different tests , but the multi-test average will not necessarily be an appropriate benchmark for an intervention effect obtained on one test in particular. Annual gains vary across subjects (Bloom et al., 2008), and vary from one time and place to another (Matheny et al., 2023). On some tests, but not others, annual gains are different over 12 months than over a 9-month academic year that excludes summer vacation (Workman et al., 2023). For some topics, such as probability, Latin, or cartography, there may be no benchmarks for annual gains at all. In short, while it can be helpful to keep annual gains in mind as a rough comparison, all these sources of variation make it difficult to use gains as a stable benchmark for effect size. The challenge is fundamental, and switching from SDs to percentile points does not solve it.

Another option is to compare an effect to a gap in test scores between advantaged and disadvantaged children. This comparison can be appropriate in some settings, for example in a study that asked whether certain charter schools shrank the gap in reading and math scores between black children in Harlem and white children elsewhere in New York City (Dobbie & Fryer, 2011). But the comparison is not always pertinent and, if used habitually, can give the misleading and undesirable impression that score gaps are so constant and immutable that they can be engraved on a ruler (Quinn, 2020). In fact, score gaps between groups of children vary substantially across places, times, subjects, tests, grade levels, and the group characteristics used (e.g., family income, parental education, school-level poverty, or race and ethnicity) (Reardon et al., 2019; US Department of Education, 2023). The variable nature of score gaps can make it difficult to treat them as a fixed ruler to measure effect size.

Conclusion

Careful education researchers take pains to estimate treatment effects precisely and without bias. Yet we then describe those effects in SD units that many practitioners find unintelligible, and even trained researchers interpret in biased and inconsistent ways. Translating effects into percentile points puts effects on a shared and intelligible scale—and it can usually be accomplished simply by multiplying the SD effect by 37.

Even on an intelligible percentile point scale, educators and scholars may have different ideas about whether an effect is large enough to matter. Some may think that any effect smaller than 10 percentile points (0.27 SD) is negligible, while others may argue that even effects of 2 percentile points (0.05 SD) can be important in some settings. Translating effects into percentile points will not end all debate. But it is a good start to a conversation.

References

Angrist, J. D., Lavy, V., Leder-Luis, J., & Shany, A. (2019). Maimonides' rule redux. *American Economic Review: Insights*, *1*(3), 309–324. https://doi.org/10.1257/aeri.20180120

Baird, M. D., & Pane, J. F. (2019). Translating Standardized Effects of Education Programs Into More Interpretable Metrics. *Educational Researcher*, *48*(4), 217–228. https://doi.org/10.3102/0013189X19848729

Bloom, H. S., Hill, C. J., Black, A. B., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*(4), 289–328. https://doi.org/10.1080/19345740802400072

Cheng, L., Ritzhaupt, A. D., & Antonenko, P. (2019). Effects of the flipped classroom instructional strategy on students' learning outcomes: A meta-analysis. *Educational Technology Research and Development*, *67*(4), 793–824. https://doi.org/10.1007/s11423-018-9633-7

Chingos, M. M. (2013). Class Size and Student Outcomes: Research and Policy Implications. *Journal of Policy Analysis and Management*, *32*(2), 411–438. https://doi.org/10.1002/pam.21677

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145–153. https://doi.org/10.1037/h0045186

Cohen, J. (1970). Approximate Power and Sample Size Determination for Common one-Sample and two-Sample Hypothesis Tests. *Educational and Psychological Measurement*, *30*(4), 811–831. https://doi.org/10.1177/001316447003000404

Dobbie, W., & Fryer, R. G. (2011). Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone. *American Economic Journal: Applied Economics*, *3*(3), 158–187.

Gamse, B. C., Jacob, R. T., Horst, M., Boulay, B., & Unlu, F. (2008). Reading First Impact Study. Final Report. NCEE 2009-4038. In *National Center for Education Evaluation and Regional Assistance*. National Center for Education Evaluation and Regional Assistance. https://eric.ed.gov/?id=ED503344

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

Hogg, R., McKean, J., & Craig, A. (2012). *Introduction to Mathematical Statistics* (7th edition). Pearson.

Kane, T., & Staiger, D. O. (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. *National Bureau of Economic Research*, 14607.

Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Kraft, M. A. (2023). The Effect-Size Benchmark That Matters Most: Education Interventions Often Fail. *Educational Researcher*, *52*(3), 183–187. https://doi.org/10.3102/0013189X231155154

Lee, J., Finn, J., & Liu, X. (2019). Time-Indexed Effect Size for Educational Research and Evaluation: Reinterpreting Program Effects and Achievement Gaps in K–12 Reading and Math. *The Journal of Experimental Education*, *87*(2), 193–213. https://doi.org/10.1080/00220973.2017.1409183

Levin, H. M., McEwan, P. J., Belfield, C., Bowden, A. B., & Shand, R. (2017). *Economic Evaluation in Education: Cost-Effectiveness and Benefit-Cost Analysis*. SAGE Publications.

Matheny, K. T., Thompson, M. E., Townley-Flores, C., & reardon, sean f. (2023). Uneven Progress: Recent Trends in Academic Performance Among U.S. School Districts. *American Educational Research Journal*, *60*(3), 447–485. https://doi.org/10.3102/00028312221134769

NWEA. (updated regularly). *MAP Growth Norms*.

Pane, J. F., Steiner, E. D., Baird, M. D., Hamilton, L. S., & Pane, J. D. (2017). *Informing Progress: Insights on Personalized Learning Implementation and Effects*. RAND Corporation. https://www.rand.org/pubs/research_reports/RR2042.html

Quinn, D. M. (2020). Experimental Effects of "Achievement Gap" News Reporting on Viewers' Racial Stereotypes, Inequality Explanations, and Inequality Prioritization. *Educational Researcher*, *49*(7), 482–492. https://doi.org/10.3102/0013189X20932469

Reardon, S. F., Kalogrides, D., & Shores, K. (2019). The Geography of Racial/Ethnic Test Score Gaps. *American Journal of Sociology*, *124*(4), 1164–1221. https://doi.org/10.1086/700678

Schanzenbach, D. W. (2011). *Review of Class Size: What Research Says and What It Means for State Policy*. National Education Policy Center, University of Colorado. http://nepc.colorado.edu/thinktank/review-class-size-what-research-says-and-what-it-means

Schuetze, B. A., & Yan, V. X. (2023). Psychology Faculty Overestimate the Magnitude of Cohen's d Effect Sizes by Half a Standard Deviation. *Collabra: Psychology*, *9*(1), 74020. https://doi.org/10.1525/collabra.74020

Shanley, L., Clarke, B., Anderson, D. A., Turtura, J., Doabler, C. T., Kurtz-Nelson, E., & Fien, H. (2019). Exploring the utility of assessing early mathematics intervention response via embedded assessment. *School Psychology*, *34*(5), 541–554. https://doi.org/10.1037/spq0000326

US Department of Education. (2023). *NAEP Dashboards—Achievement Gaps*. https://www.nationsreportcard.gov/dashboards/achievement_gaps.aspx

von Hippel, P. T., Workman, J., & Downey, D. B. (2018). Inequality in Reading and Math Skills Forms Mainly before Kindergarten: A Replication, and Partial Correction, of "Are Schools the Great Equalizer?" *Sociology of Education*, *91*(4), 323–357. https://doi.org/10.1177/0038040718801760

What Works Clearinghouse. (2022). *WWC Version 5.0 Procedures and Standards Handbook* (WWC 2022008REV). US Department of Education. https://ies.ed.gov/ncee/wwc/handbooks

Workman, J., von Hippel, P. T., & Merry, J. (2023). Findings on Summer Learning Loss Often Fail to Replicate, Even in Recent Data. *Sociological Science*, *in press*.

Figures

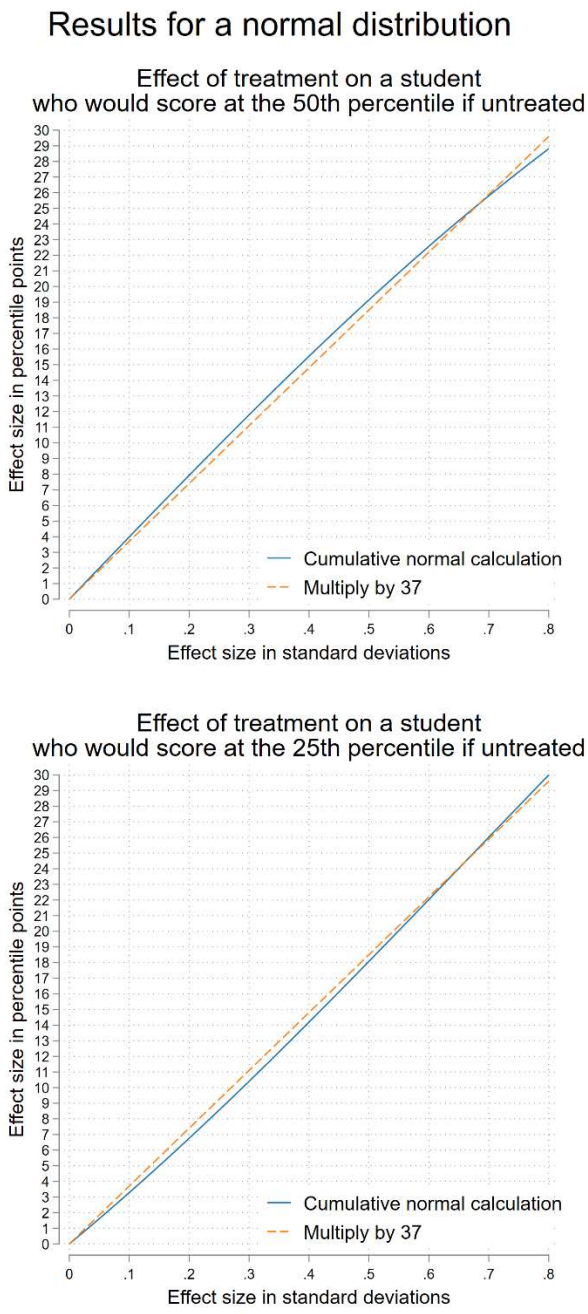

Figure 1. Theoretical accuracy of the multiply-by-37 rule for translating effect sizes from standard deviations to percentile points in a normal distribution. The top panel gives results for a student who, if untreated, would score at the 50th percentile; the bottom panel gives results for a student who, if untreated, would score at the 25th percentile.

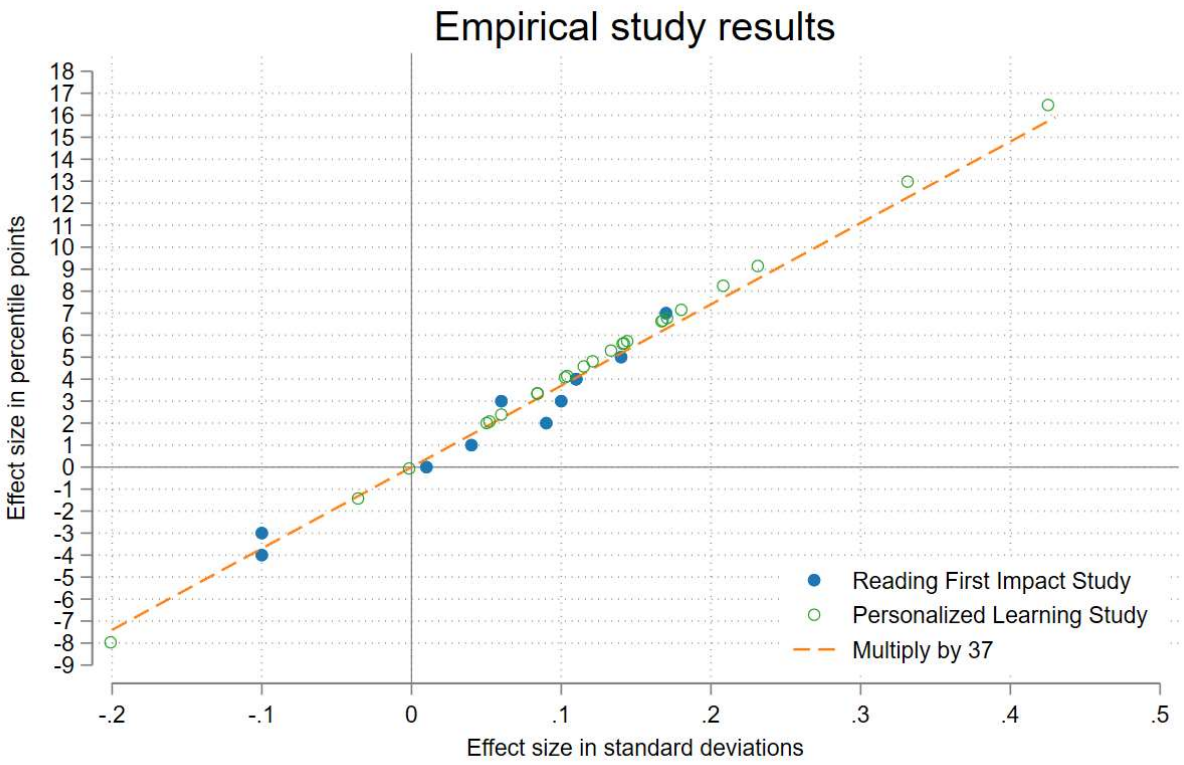

Figure 2. Empirical accuracy of the multiply-by-37 rule for converting effects observed in empirical studies from standard deviations to percentile points.

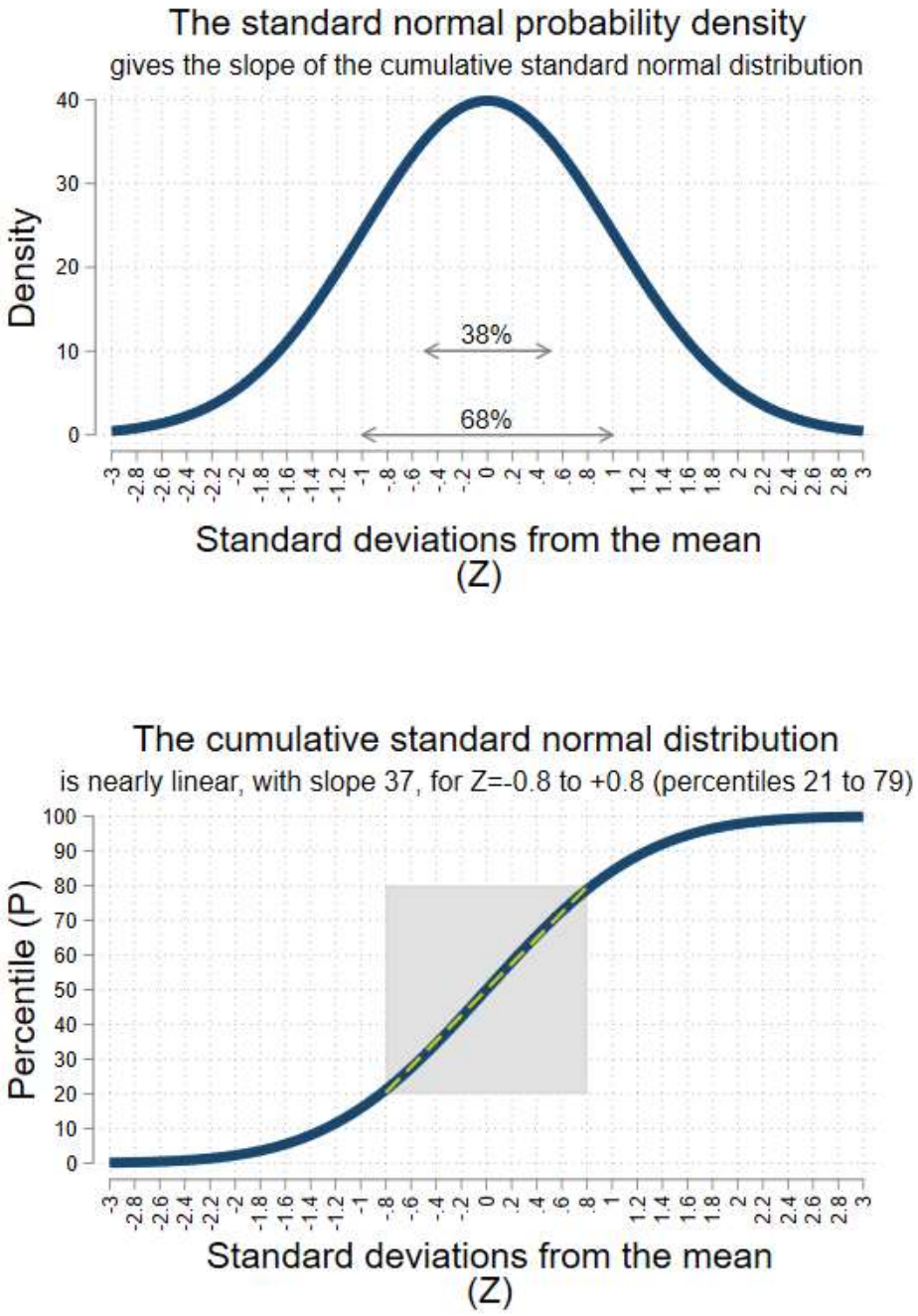

Figure 3. The slope of the cumulative standard normal distribution (CDF) is given by the standard normal probability density function. Between the 20th and 80th percentile the CDF is well approximated by a straight line with a slope of 37. (*Note*. To put both curves on a percentile point scale, the density function has been multiplied by 100.)

Table 1. Benchmarks for effect size

		Effect size		
	Units	Small	Medium	Large
Cohen (1969)	Standard deviations	0.2	0.5	0.8
	Percentile points	7	19	30
Kraft (2020)	Standard deviations	0.05	0.05 to 0.2	0.2
	Percentile points	Less than 2	2 to 7	More than 7

Appendix

The notion of multiplying the SD effect by a constant is attractive, and a variety of plausible values for the constant can be proposed. This Appendix shows how we settled on multiplying by 37 and shows when the approximation errors are large or small.

We simulated data representing students' potential outcomes under treatment and under control. For each student we kept 4 variables: P_c and Z_c represented their percentile and standardized outcome under the control condition, and P_t and $Z_t = Z_c$ represented their percentile and standardized outcome under the treatment condition. The control percentiles P_c took integer values from 1 to 99, and the SD effect $\Delta Z = Z_t - Z_c$ took values from 0.01 to 0.80 by 0.01. Note that the treatment outcomes P_t and Z_t represent where a treated student would fall in the control distribution, and we simulated the control condition to be standard normal. So $P_c = \Phi(Z_c)$ and $P_t = \Phi(Z_t)$, where Φ is the standard normal CDF.

From the simulated data we calculated the exact percentile point effect that corresponded to a given SD effect:

$$\Delta P = P_t - P_c \qquad \text{(exact formula)}$$

Our goal was to find the best approximation to the percentile point effect that could be achieved by simply multiplying the SD effect by a constant c, which is also the slope of the approximating line.

$$\Delta \hat{P} = c\, \Delta Z \qquad \text{(approximation)}$$

The "best" value of the slope c is one that, over some range of SD effects ΔZ and untreated percentiles P_c, minimizes some function of the approximation error e:

$$e = \Delta \hat{P} - \Delta P$$

The most common error function to minimize is the root mean squared error, $\sqrt{E(e^2)}$, but we thought it would be better to minimize the maximum of the absolute errors, $max\ (|e|)$. Minimizing root mean squared error can produce a line that fits well on average but may have large errors in some parts of the range; by contrast, minimizing the maximum absolute error produces a line that never exceeds a stated absolute error.

We calculated results for students with untreated percentiles between 21 and 50, subject to treatment effects between 0 and 0.8 SD. These are students who stay in the range where a linear approximation works best—between the 21st and 79th percentile—whether they are treated or not.

Figure A 1 gives the results. The max absolute error is minimized at a slope of approximately c=36.95, which we rounded to 37. At a slope of 37, the error of approximation never exceeds 1.6 percentile points in absolute value. By contrast, the root mean squared error is minimized at a slope of 38; at a slope of 38, the root mean squared error is 0.65 percentile points, but the absolute error can be as large as 2 percentile points under some conditions. While the difference between a slope of 37 or 38 is not large, we have a slight preference for minimizing the maximum absolute error—and that means choosing a slope of 37.

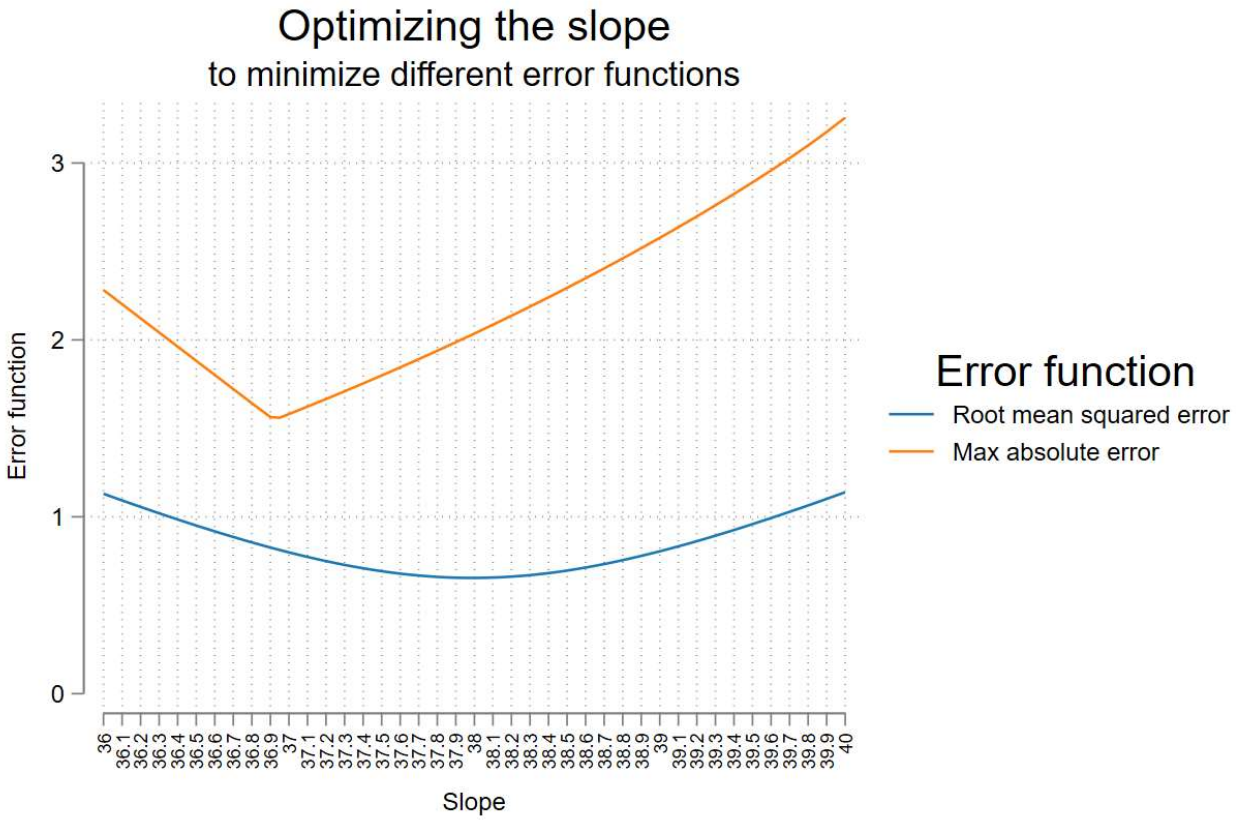

Figure A 1. The max absolute error is minimized at a slope of 37. The root mean squared error is minimized at a slope of 38. The error functions are calculated for students with untreated percentiles between 21 and 50, subject to treatment effects between 0 and 0.8 SD.

Now that we have settled on a slope of 37, we should ask under what circumstances the approximation works best and worst. Figure A 3 shows that the absolute approximation error is smallest if both the treated are untreated percentile are between the 20th and 80th percentile. For those students, the absolute approximation error is always below 1.6 percentile points and usually below 1 percentile point. The approximation error gets worse rapidly if the treated or untreated percentile goes below the 20th or above the 80th—which is why we recommend multiplying by 37 only for students who stay within that range.

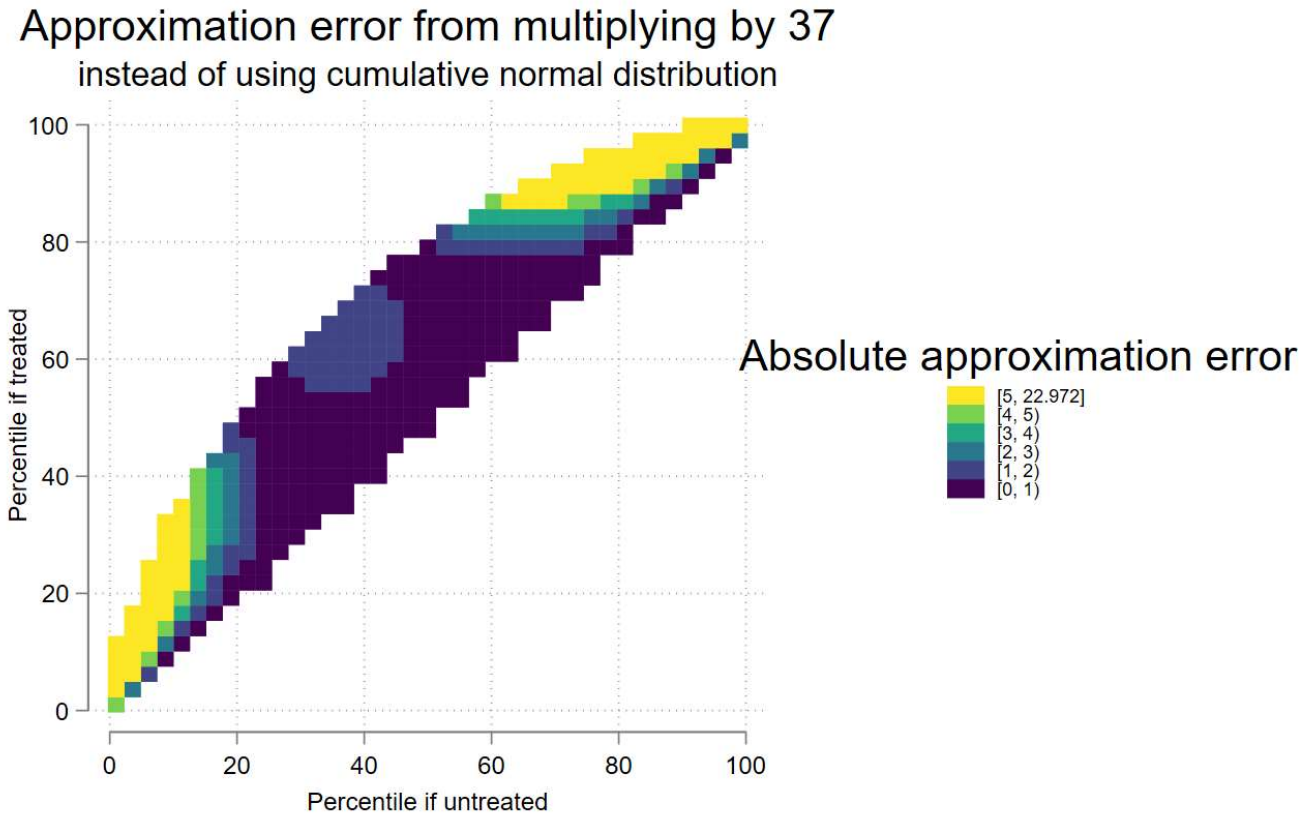

Figure A 2. If both the treated and untreated percentile are between approximately 20 and 80, multiplying the SD effect by 37 always comes within 1.6 percentile points of the percentile point effect, and usually comes within 1 percentile point.

While the limitation to students between the 20th and 80th percentile is clear enough, it is not useful in the usual situation where we don't have the treated and untreated percentiles and just have the SD effect. Figure A 3 addresses that situation by showing the approximation error of multiplying by 37 as a function of the untreated percentile and the treatment effect in SD units. The approximation works well for effect sizes of up to 0.8 SD for students who if untreated would score between the 21st and 52nd percentiles. Outside of that range, the approximation rapidly gets worse unless the effect size is quite small.

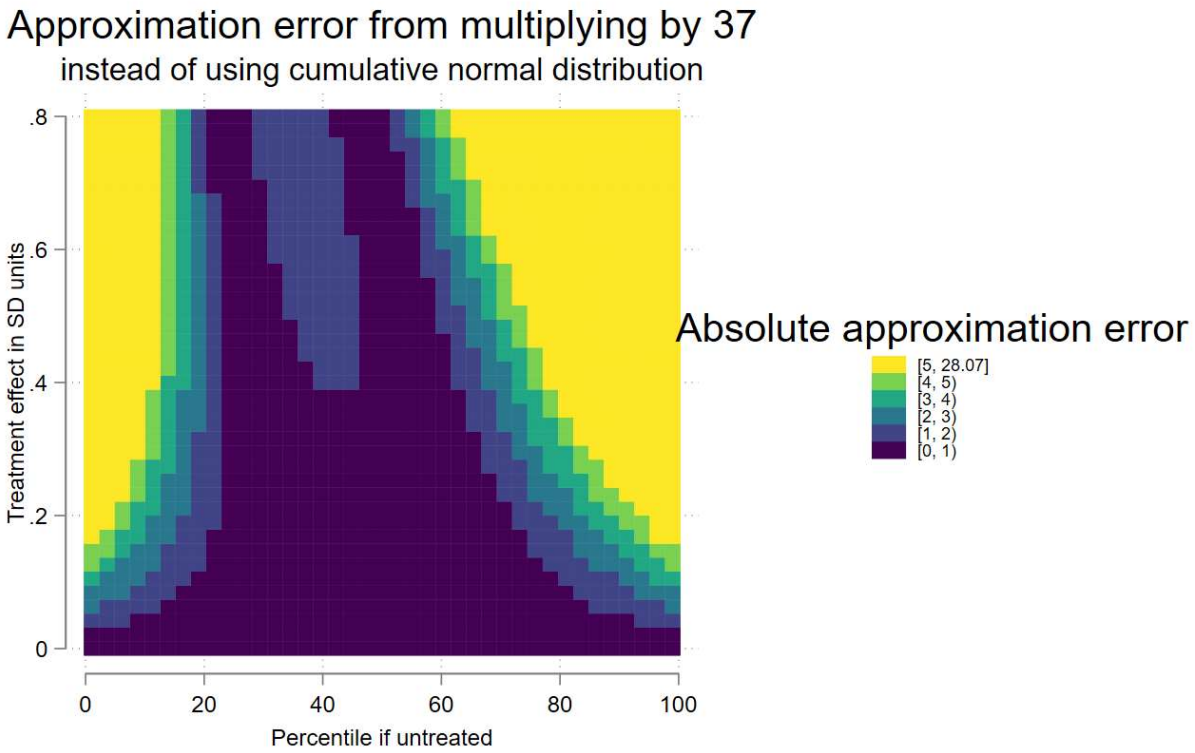

Figure A 3. The max absolute error is minimized at a slope of 37. The root mean squared error is minimized at a slope of 38. The error functions are calculated for students with untreated percentiles between 21 and 50, subject to treatment effects between 0 and 0.8 SD.