



Morehouse
THE FORCE IN CALIBRATION SINCE 1925

Morehouse Instrument Company
1742 Sixth Ave.
York, PA 17403-2675
(717) 843-0081

Morehouse Instrument Company is ISO/IEC 17025 / ANSI/NCSLI Z540.3 Accredited

www.mhforce.com

The GUM in Plain English

This paper explains measurement uncertainty in plain language, with force-calibration examples drawn from real Morehouse uncertainty budgets. It is an independent Morehouse technical paper, a companion to JCGM 100:2008, not a reproduction or a translation of it. Section numbers in parentheses point to the official document, which you should keep on the shelf and read at the source.

One example threads the whole paper: a force-measuring instrument calibrated on a Morehouse machine at 5000 lbf (22.241 kN), following ASTM E74. We build its uncertainty budget one piece at a time and add it all up at the end. Every number in that budget is real. It comes straight from a published Morehouse single-point analysis, so you can check our work against your own.



Contents

1. Introduction.....	3
2. Error versus uncertainty.....	6
2.1 The measurement model.....	7
3. Type A evaluation of uncertainty	9
4. Type B evaluation of uncertainty.....	12
4.1 Why the GUM retired “random” and “systematic”	15
5. Combining uncertainty components.....	17
5.1 Correlation between inputs	18
6. Expanded uncertainty and the coverage factor	20
6.1 Effective degrees of freedom	21
7. Reporting the result	24
7.1 Worked example: the 5000 lbf budget.....	25
Appendix A. Summary of formulas	29
References	30

1. Introduction

Pull any calibration certificate off the shelf, and you will find a line like $U = 1.71 \text{ lbf}$, $k = 2.57$, approximately 95 % confidence. Most of the people who read that line, including some who sign it, could not tell you where the 2.57 came from, what the 95 % is really claiming, or what would make the number bigger.

One thing to settle before we go further, because it trips up careful readers. That 2.57 is not a 99 % coverage factor. It is the Student's t factor for 95 % confidence on a budget with only 5 effective degrees of freedom, and it is the honest multiplier when the largest term in the budget rests on four readings.¹ If "effective degrees of freedom" means nothing to you yet, keep reading. Part 6 earns every word of it.

That is not their fault. Every honest uncertainty statement on Earth traces back to one document, the Guide to the Expression of Uncertainty in Measurement, JCGM 100:2008, known to everyone as the GUM. And the GUM is hard to read. It was written by a committee drawn from seven international organizations, in the careful language of people who could not afford to be wrong, and it shows.

This paper walks the same path the GUM walks, in the order it builds, and turns each step into plain language. We build one complete uncertainty budget for a force-measuring instrument calibrated at 5000 lbf (22.241 kN), following ASTM E74. By the final section you will be able to defend every number on it.

¹ Most certificates state k and a coverage probability and nothing else, so a reader who sees $k = 2.57$ reasonably assumes a 99 % interval with infinite degrees of freedom behind it, the same way $k = 1.96$, usually rounded to 2, is read as 95 %. The GUM asks that the effective degrees of freedom be reported when it is small enough to change k (GUM 7.2.7), and here it changes it from 2.00 to 2.57.



A MEASUREMENT IS A BEST ESTIMATE PLUS AN INTERVAL OF DOUBT

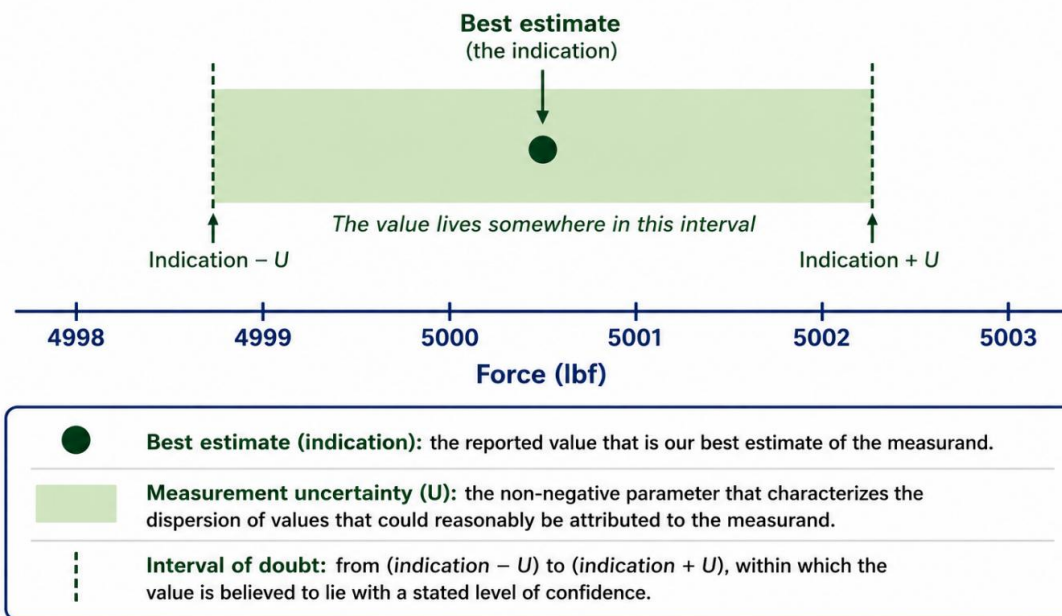


Figure 1. A measured value is an estimate sitting inside a neighborhood of doubt. Uncertainty is the width of that neighborhood.

A measurement without an uncertainty is an opinion with a decimal point.

Put a load cell on a machine and read 5000 lbf. What you have is not the force. It is your best estimate of the force. The real value sits somewhere in a small neighborhood around your reading, and the entire job of uncertainty is to say how big that neighborhood is. The quantity you set out to measure has a name: the measurand (JCGM 200:2012, the VIM). The result is only an estimate of it, and that estimate is complete only when you state how much you doubt it (GUM 3.1.2). No statement of doubt, no result. Just a number floating free.

Where does the doubt come from? The GUM lists ten sources (GUM 3.3.2). In a force lab, they collapse to a familiar handful: the reference standard is not perfect, and it drifts between calibrations; the load cell does not repeat exactly when you reload it; it reads a little differently when a different technician runs it; it reads differently in a machine that does not have the same characteristics of plumb, level, square, rigid, and low torsion as that on which it was calibrated; adapters and loading conditions change the stress concentrations, and therefore the output; the indicator has finite resolution; and temperature, alignment, and a few smaller effects move the reading.

Every one of those is a piece of doubt. The GUM's method is to size each piece, combine them honestly, and report the total. That is the whole arc of this paper. The rest is detail.



Morehouse
THE FORCE IN CALIBRATION SINCE 1925

Morehouse Instrument Company
1742 Sixth Ave.
York, PA 17403-2675
(717) 843-0081

Morehouse Instrument Company is ISO/IEC 17025 / ANSI/NCSLI Z540.3 Accredited

www.mhforce.com

There is a hard limit worth stating on the first page. You cannot claim an uncertainty smaller than the reference laboratory that calibrated your standard (ISO/IEC 17025:2017, Clause 7.6). If the lab above you works to 0.002 % of applied force, your number lives above that floor, not below it. Math does not work the other way.

Notice what is missing from the list: mistakes. Uncertainty does not cover blunders, a transposed digit, a misread dial, or a wrong constant (GUM 3.4.7). Those you catch by review, not by arithmetic. Uncertainty is the honest doubt that remains after you have done everything right.

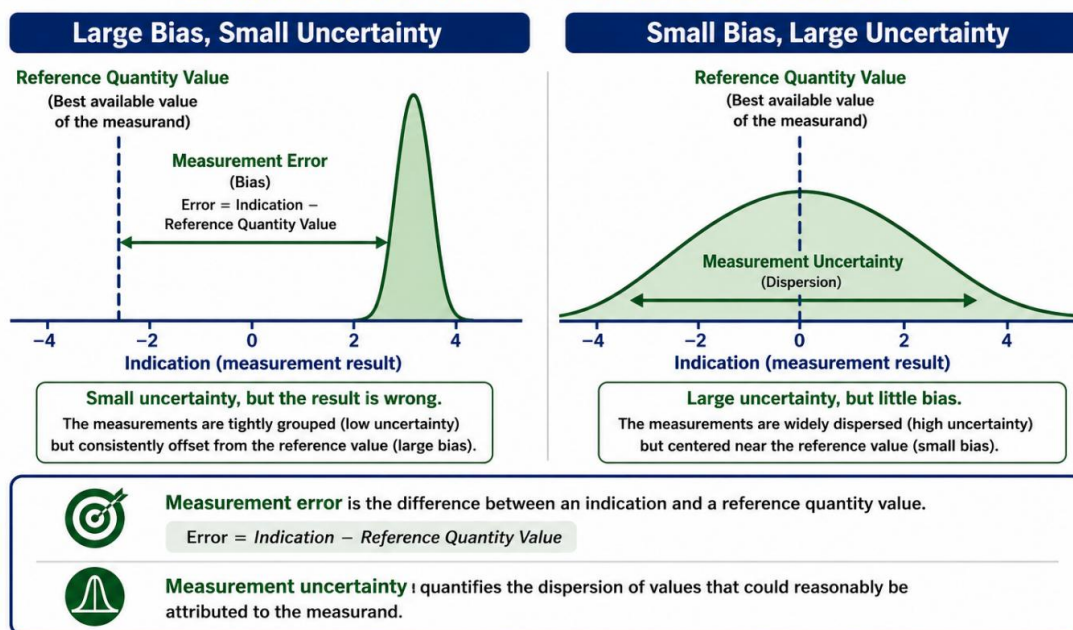
Since 2008: the VIM's third edition (JCGM 200:2012) turned the point of this part into a definition. A measurement result is now the measured value together with its uncertainty, not the value alone (VIM 2.9). That is the no doubt, no result rule stated above, made official. The same edition settled on metrological traceability (VIM 2.41) where older texts wrote only traceability.



2. Error versus uncertainty

In Section 1 we agreed that a measurement without an uncertainty is an opinion with a decimal point, and we listed where the doubt in a force lab comes from. Before we can size any of it, we have to clear up the confusion that trips up more working metrologists than any other, and then write down the one equation that makes a budget possible.

ERROR IS NOT UNCERTAINTY



Bias and uncertainty are different properties of a measurement result. | A result may have low uncertainty and still be wrong. | A result may have high uncertainty and little bias.

Terminology in accordance with the VIM (JCGM 200:2012) and the GUM (JCGM 100:2008).

Figure 2. Large error with small scatter on the left; small error with large scatter on the right. Different problems, different fixes.

Here is a distinction the GUM buried most of in Annex D, where few people look.

Picture two bathroom scales. The first reads 2 lb heavy every single time, dead consistent. The second is perfectly centered on average, but jumps around by a pound either way. The first scale has a large error and small scatter. The second has near-zero error and large scatter. They are not the same problem, and you do not fix them the same way.

Error is the difference between your measured value and a reference value (the VIM, 2.16). That reference is the best available value of the measurand: sometimes the true value, which you can never know, and sometimes a calibrated standard, which you can. Uncertainty quantifies the dispersion of the



values that could reasonably be attributed to the measurand. They are not synonyms. They are not even the same kind of thing.

Measured against the true value, error is a single number you can almost never know, because to know it, you would need the true value, and if you knew the true value, you would not be measuring.

Measured against a calibrated standard, error becomes a bias you can find and correct. Uncertainty is a number you can always estimate either way, because it is built from what you do know: your data, your standards, your specs.

Force calibration shows the difference in dollars. Load a 50 000 lbf tension link with the wrong size pin, a 1.75 in pin where the design calls for a 50 mm (1.97 in) pin, and it reads 49 140 lbf. That is an 860 lbf miss, a 1.72 % error, and no amount of uncertainty accounting repairs it. It is a bias you remove by using the right adapter, not a doubt you report. Confuse the two, and you will pad your uncertainty while the real error sits there untouched.

The GUM makes the point in a line worth taping to the wall. A corrected result can sit very close to the true value, and so carry a negligible error, and still have a large uncertainty (GUM 3.3.1, note). Small error, big uncertainty. The two move independently. The GUM's rule is direct: identify systematic effects, correct for them, then carry the uncertainty of the correction (GUM 3.2.3). You do not get to skip the correction and pad the uncertainty instead (GUM 6.3.1, note).

Keep them separate. Error is the known bias you remove. Uncertainty is the honest doubt you report.

2.1 The measurement model

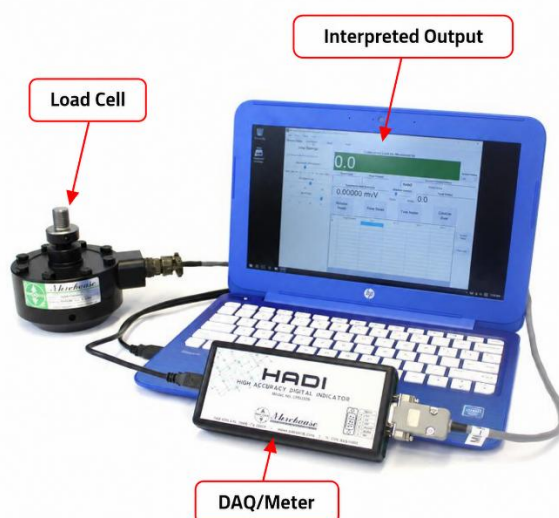


Figure 3. The measurement chain. Every box from the load cell to the interpreted output is an input with its own doubt.

You cannot budget for doubt you have not written down.

Before any uncertainty work, you need the equation that turns your raw readings into the answer. The GUM calls it the measurement model, and it is just this: the measurand is some function of input quantities (GUM 4.1.1).

In a force calibration, the model looks plain. The force on the load cell equals the reference force applied by the machine, adjusted for the few effects that move it: the cell's repeatability, the spread between technicians, the indicator resolution, temperature, and alignment. Write that out, and you have named every input. Name the inputs, and you have named the budget, because each input carries a standard uncertainty, and those are the pieces you will combine.

Why bother writing the obvious? Because the model is where the hidden inputs hide. The deadweight machine's force is not just the weights. It is the weights, the local value of gravity, the air buoyancy correction, repeatability studies with a great load cell, how level it is, how rigid it is, how plumb the weights hang, the travel of the stage beam, and the alignment, each an input with its own doubt (GUM 4.1.2). The reference standard itself splits into two inputs that behave differently: its calibration uncertainty, fixed on a certificate, and its stability, the drift you will see before the next calibration. Miss one in the model and you will miss it in the budget, and your final uncertainty will come out too small and too confident.

The Death Star had this exact problem. The plans were complete except for one thermal exhaust port, two meters wide, that nobody treated as worth modeling. One proton torpedo later, the whole station was gone. A missing input works the same way. It is small, it is easy to leave off the list, and it is the one that takes down the result. Build the model so the exhaust port is on it.

The GUM is blunt about the limits here. The model can be incomplete, so vary everything you can and let real data, not assumptions, drive the numbers (GUM 3.4.2). And no formula substitutes for judgment. Evaluating uncertainty, in the GUM's own words, "cannot substitute for critical thinking, intellectual honesty and professional skill" (GUM 3.4.8). Math is the easy part. Knowing what belongs in the model is the craft.

For the rest of this paper, our model is a Morehouse load cell at 5000 lbf (22.241 kN). The full budget breaks into nine inputs. We size each one, then combine them.

Since 2008: the step this part calls the craft now has its own guide. JCGM GUM-6:2020, Developing and using measurement models, once numbered JCGM 103, walks you from naming the measurand to a finished equation, exactly the work above. If the model is where you get stuck, that is where the JCGM put the help.



3. Type A evaluation of uncertainty

In Section 2 we wrote the measurement model and named every input in our 5000 lbf budget. Now we start filling in numbers, beginning with the ones we earn ourselves.

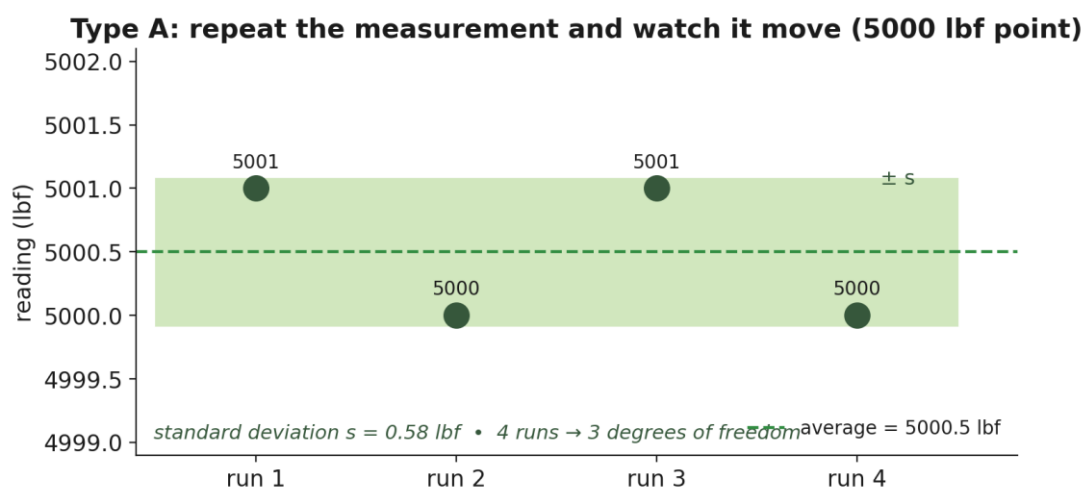


Figure 4. Four loadings at 5000 lbf: 5000, 5001, 5000, 5001. Mean 5000.5 lbf, standard deviation 0.58 lbf, 3 degrees of freedom.

The fastest way to see uncertainty is to measure the same thing twice. It will not agree with itself.

Load the cell to 5000 lbf, then unload it and reload it. Do it four times, and you get 5000, 5001, 5000, and 5001 lbf. The mean is 5000.5 lbf. The spread in those readings is real doubt, and you analyze it with ordinary statistics. The sample standard deviation of the four runs is 0.58 lbf (GUM 4.2.2). That is a Type A evaluation: the GUM's name for any uncertainty obtained by the statistical analysis of a series of observations (GUM 2.3.2).

None of this is special to force. The GUM writes the same recipe for any quantity you can measure more than once under the same conditions (GUM 4.2.1). Weigh the same mass ten times, time the same pendulum swing, read the same voltage on the same meter, and the readings will scatter. Take the arithmetic mean as your best estimate. Take the sample standard deviation of those readings as the size of the scatter (GUM 4.2.2). Any time you can repeat a measurement, the measured data hands you its own doubt.

Here is the subtlety, and it is not a flaw in the GUM. The same four runs give you two numbers that answer two different questions. The sample standard deviation, 0.58 lbf, describes the spread of single readings (GUM 4.2.2). Divide it by the square root of the number of runs, and you get the experimental



standard deviation of the mean, 0.29 lbf, which is how tightly those four runs pin down their own average (GUM 4.2.3). Neither is more correct; the right one matches what you report. A force instrument is not used to report an average. It is used to make a single reading at a time, so the figure that matters is the doubt in a single measurement determination at the point. So the repeatability we carry is the sample standard deviation itself, 0.58 lbf, not the experimental standard deviation of the mean. Know which question you are answering before you divide.

In summary: if four single repeated measurements are made, use the sample standard deviation. If four averages of multiple individual measurements of a known sample size (say 5) are used, use the sample standard deviation of the individual readings divided by $\sqrt{5}$. If four averages of unknown sample sizes are used, use the sample standard deviation of the four averages divided by $\sqrt{4}$, the number of averages.

That single contribution will turn out to dominate the entire budget. Hold that thought; it is the whole story of Part 6.

Repeatability between technicians is the same idea, one level up. Put a different operator on the same point and the reading shifts a little. The overall repeatability within technicians, when there is more than one technician, is the square root of the average variance of each technician's repeatability. The reproducibility between technicians is the sample standard deviation of each technician's average.

Reading	Technician 1	Technician 2
1	-4.11328	-4.11328
2	-4.11325	-4.11324
3	-4.11325	-4.11324
4	-4.11324	-4.11324
5	-4.11324	-4.11322
6	-4.1132	-4.1132
Mean (mV/V)	-4.113243333	-4.113236667
Sample Standard Deviation	25.82E-6	26.58E-6
Variance	666.67E-12	706.67E-12
Repeatability (mV/V)	26.20E-6	=SQRT(AVERAGE(C11:D11))
Reproducibility (mV/V)	4.71E-6	=STDEV(C9:D9)

Figure 5. The two-technician repeatability and reproducibility study at 5000 lbf: six readings from each technician in mV/V, with the within-technician repeatability and the between-technician reproducibility computed below.

Worked example: two technicians, six readings each at the 5000 lbf point, cell output in mV/V.

Repeatability between technicians is the square root of the average of the two variances:

$\sqrt{[(666.67 \times 10^{-12} + 706.67 \times 10^{-12}) \div 2]} = 26.20 \times 10^{-6}$ mV/V, at 10 degrees of freedom. Reproducibility

between technicians is the sample standard deviation of the two technician averages: 4.71×10^{-6} mV/V,

at 1 degree of freedom. The cell reads 4.11324 mV/V for 5000 lbf, so one mV/V equals 1215.6 lbf. In force units, the repeatability is 0.032 lbf and the reproducibility 0.006 lbf, the two values carried in the budget.

On a deadweight primary standard with good adapters (Morehouse recommends following ISO 376 recommendations), these often vanish into the noise. On a working machine, they can grow, so you test them rather than assume them.

One more Type A piece belongs here, because ASTM E74 hands it to you already computed: the lower limit factor (llf). It is a statistical estimate of the fit error in the calibration equation, sometimes called the reproducibility or expected performance of the device when used under similar conditions. ASTM reports it at a coverage factor of 2.4, so to use it, you divide by 2.4 to reach one standard deviation. A reported llf of 0.25 lbf becomes 0.10 lbf.

Table 1. The Type A contributors for the 5000 lbf example, with the value and degrees of freedom each carries.

Type A contributor	Standard uncertainty	Degrees of freedom	Notes
Single-reading repeatability	0.58 lbf	3	Dominates the budget; changes at each loading point
Between-technician repeatability	0.032 lbf	10	Square root of the average of each technician's variance
Between-technician reproducibility	0.006 lbf	1	Sample standard deviation of the technician averages
ASTM lower limit factor (llf)	0.10 lbf	32	Reported at $k = 2.4$; divide by 2.4 to reach one standard deviation

Look hard at that degrees-of-freedom column, because it hides a choice. The lower limit factor carries 32, and for good reason: ASTM E74 fits the calibration equation across the full range, roughly 35 points, which with a second-degree fit leaves 32 degrees of freedom, so its fit error rests on a lot of data. The single-reading repeatability carries only 3, because we sized it from four determinations at one point. We could have borrowed the E74 fit for repeatability too and inherited all 32 degrees of freedom. We did not. A force instrument is used to make one reading at one point, so the number we want is the honest spread of a single determination at that point, not a value pulled from a regression across the whole range. That choice keeps the repeatability term low in degrees of freedom, and Part 6 shows exactly what it costs.

Every number in Table 1 came from data we took ourselves, and every one carries its own degrees of freedom, which is why that column rides along in the budget. What none of them covers is the doubt we never re-measured: the reference, the resolution, the room.

That is the rest of the story, as Paul Harvey used to promise on the radio. The good part is always in the half you have not heard yet.

4. Type B evaluation of uncertainty

Part 3 earned four numbers by repeating measurements: our Type A pieces, led by a 0.58 lbf single-reading repeatability at 3 degrees of freedom. But not all doubt comes from repeating a measurement. Most of it, in a good lab, comes from things you already know and never re-measure.

You did not determine your reference standard's uncertainty by running it four times this morning. You read it off the calibration certificate. You did not characterize your indicator's resolution statistically. You read it off the front panel. That kind of evaluation, built from existing information rather than fresh repetition, is a Type B evaluation (GUM 4.3). It is not second-class. In this budget, the Type B pieces include the second-largest contributor of all.

The trick with Type B is turning what you know into a standard uncertainty, a one-sigma number you can combine with the rest. There are only a few moves, and they cover almost everything you will meet.

Divide a stated uncertainty by its coverage factor. When the information already comes as an uncertainty with a coverage factor, divide it back down. Our reference was calibrated by deadweight primary standards with a calibration and measurement capability (CMC) of 0.002 % of applied force, reported at $k = 2$. The CMC is the lab's best case under ideal conditions; the uncertainty you report on a customer's instrument is normally larger, because it also carries that instrument's own doubt. At 5000 lbf, that is a 0.1 lbf expanded value. Divide by 2, and the standard uncertainty of the reference is 0.05 lbf. Done.

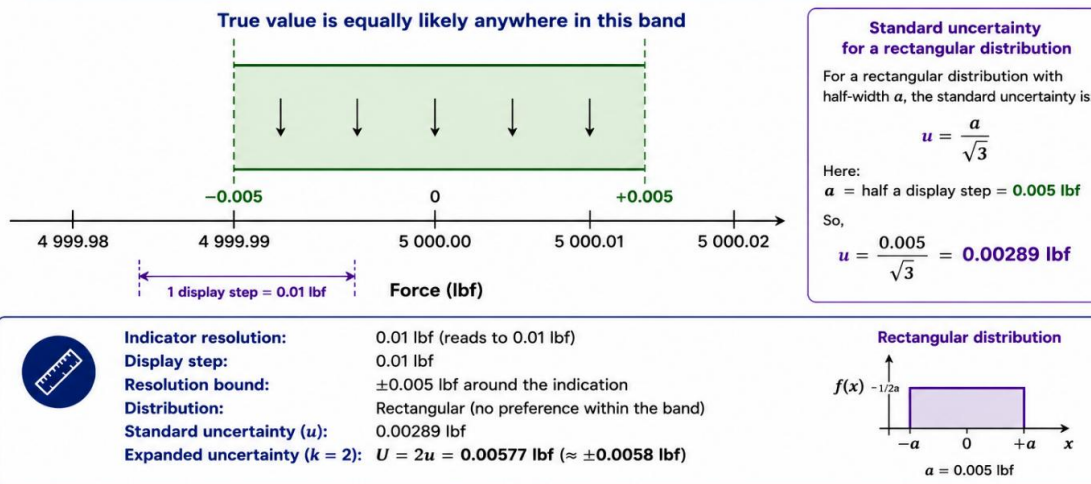
Turn a bound into a shape. When all you have is a bound, an "it cannot be off by more than this," you have to assume a shape for the doubt inside the bound. If you have no reason to favor the middle over the edges, you assume every value is equally likely, a rectangular distribution, and divide the half-width by the square root of 3. Three of our inputs are sized this way. Reference standard stability: the reference shifted 0.01 % between its last two calibrations, which at 5000 lbf is a 0.5 lbf bound; rectangular, divided by $\sqrt{3}$, gives 0.29 lbf, the second-largest line in the whole budget, and it is pure Type B. Environmental factors: the cell's temperature coefficient is about 0.0015 % per °C, and the lab holds the temperature to within 1 °C; that is 0.075 lbf at 5000 lbf; rectangular, divided by $\sqrt{3}$, gives 0.043 lbf. Other error sources: alignment is 1/16 in off the centerline, and the side-load sensitivity may be 0.05 % per inch, so the contribution is about 0.003 % of reading, or 0.15 lbf; rectangular, divided by $\sqrt{3}$, gives 0.087 lbf.



Resolution as a Rectangular Bound

(after UKAS M3003 §2.8)

An indicator that reads to 0.01 lbf can only show values in steps of 0.01 lbf.
When the display shows **5 000.00 lbf**, the true value could be anywhere within ± 0.005 lbf.
With no preference inside this band, the resolution uncertainty is **rectangular**.



i This follows UKAS M3003 §2.8: when an error is bounded by $\pm a$ with equal likelihood within the bounds (**rectangular**), the standard uncertainty is $u = \frac{a}{\sqrt{3}}$.

Figure 6. Resolution as a rectangular bound. An indicator reading to 0.01 lbf places the true value anywhere within one display step, so with no preference inside the band, the standard uncertainty is half a step divided by $\sqrt{3}$ (after UKAS M3003 §2.8).

Resolution is the same bound idea with a tighter step. The indicator resolves to 0.01 lbf at this point, the least increment of its display, so a reading can sit anywhere within that step. Divide by $\sqrt{12}$, which is just $2\sqrt{3}$ for a half-step, and the resolution contribution is 0.0029 lbf. This is the oldest move in the Type B toolkit, and UKAS M3003 states it plainly: whenever a value is read off a finite display, there is always an uncertainty of plus or minus half the last increment (M3003, §2.8). A reading of 5000.00 lbf stands for any value from 4999.995 to 5000.005 lbf.

Why these particular divisors? They are just the standard deviations of those shapes, worked out once, so you do not have to. Here is the part the GUM points out and most people miss: the divisors land surprisingly close together, even though the shapes behind them demand wildly different amounts of knowledge to justify (GUM 4.3.9, note). The shape matters less than people fear. Picking a sensible bound matters far more.



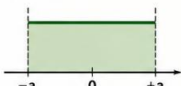
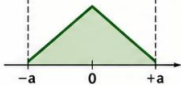
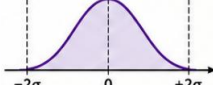
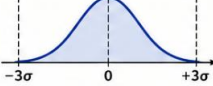
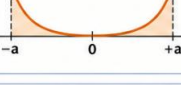
COMMON PROBABILITY DISTRIBUTIONS USED IN UNCERTAINTY ANALYSIS (TYPE B)				
Distribution	Shape (within $\pm a$ unless noted)	Assumption About Values	Standard Uncertainty (u)	Typical Use / Examples
Rectangular		All values within the bounds are equally likely.	$u = \frac{a}{\sqrt{3}}$	Specification limits, stability limits, environmental limits, resolution, alignment bounds
Triangular		Values near zero (center) are more likely than values near the limits.	$u = \frac{a}{\sqrt{6}}$	Drift between calibrations, wear or aging, temperature variation with center most likely
Normal (95.45 % Bounds) ($\pm 2\sigma$ bounds)		Values cluster around the center and become progressively less likely away from the mean.	$u = \frac{a}{2}$ or $u = \frac{U}{k}$ ($k = 2$) (95.45 % bounds)	Calibration certificate uncertainty, stated uncertainty with $k = 2$ (95.45 % bounds), previously evaluated uncertainty
Normal (99.73 % Bounds) ($\pm 3\sigma$ bounds)		Values cluster around the center and become progressively less likely away from the mean.	$u = \frac{a}{3}$ or $u = \frac{U}{k}$ ($k = 3$) (99.73 % bounds)	When bounds represent approximately 99.73 % coverage ($\pm 3\sigma$)
U-Shaped		Values near the limits are more likely than values near the center.	$u = \frac{a}{\sqrt{2}}$	Situations where extreme values are more plausible than central values (rare in calibration)
i The GUM (4.3.9, note) points out that the resulting standard uncertainties from these distributions are often surprisingly similar. Selecting a realistic bound is usually more important than selecting the exact distribution.				

Figure 7. Common Type B probability distributions and their divisors: rectangular ($\div \sqrt{3}$), triangular ($\div \sqrt{6}$), normal at 95.45 % ($\div 2$) and 99.73 % ($\div 3$), and U-shaped ($\div \sqrt{2}$), with the assumption behind each.

Table 2. The Type B contributors for the 5000 lbf example, with the distribution assumed and the divisor that turns each bound into a standard uncertainty.

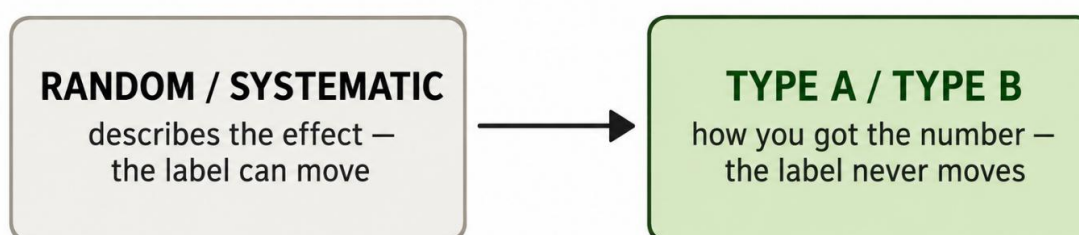
Type B contributor	Standard uncertainty	Distribution / divisor	Notes
Reference standard uncertainty	0.05 lbf	Normal, $\div 2$	From the certificate (CMC 0.002 %, $k = 2$); may change across the range
Reference standard stability	0.29 lbf	Rectangular, $\div \sqrt{3}$	Second-largest line; changes at each test point
Environmental factors	0.043 lbf	Rectangular, $\div \sqrt{3}$	Temperature coefficient times the lab's temperature band
Other error sources	0.087 lbf	Rectangular, $\div \sqrt{3}$	Alignment and side load; can be multiple lines



Device resolution	0.0029 lbf	Rectangular, $\div \sqrt{12}$	Same device typically used 10 % to 100 %
Reference standard resolution	0 lbf	Rectangular, $\div \sqrt{12}$	Deadweight reference has no readout, so 0 here

4.1 Why the GUM retired “random” and “systematic”

Why the GUM sorts by method, not by “random” vs “systematic”



A “random” component in one calibration becomes “systematic” in the next. Type A (from repeated data) and Type B (from everything else) do not flip.

Figure 8. The GUM sorts by how you got the number, not by how the effect behaves. The label that flips is the one it dropped.

If you learned uncertainty before about 1993, you learned to sort errors into two bins: random and systematic. The GUM reserved those words for errors and removed them from uncertainty. People still trip on this, so here is what I’ve been told, as even though I remember Paul Harvey, I was in high school in 1993 and certainly not reading the GUM before 1999. And get this, I slept fine back then. Now I read the GUM and have issues staying asleep.

Random and systematic describe how an effect behaves. A random effect scatters your readings; average more, and it shrinks. A systematic effect pushes them all one way; averaging does nothing. Those are useful words for effects and for errors (GUM 3.2.2). They are bad words for uncertainty components, for one reason: the label moves. Calibrate a reference yesterday, and its scatter looked random. Use that reference today to calibrate a customer’s cell, and yesterday’s scatter is now a fixed



offset baked into every reading, a systematic effect (GUM 3.3.3, note). Same number, different label, depending on how it is used. A classification that flips like that is no classification at all.

So the GUM sorts by a question that never flips: how did you get the number? If you got it from repeated data, it is Type A. If you got it from anything else, a certificate, a spec, a handbook, or judgment, it is Type B (GUM 3.3.3). That is the whole distinction.

Think of it as the artist formerly known as Prince. The music did not change when the name became a symbol nobody could pronounce. The catalog was the same; only the marquee changed, and people fumbled the new label for years. The GUM did the same thing to doubt. A reference-stability term off a certificate is the same number it always was. What changed is the question you ask to assign it.

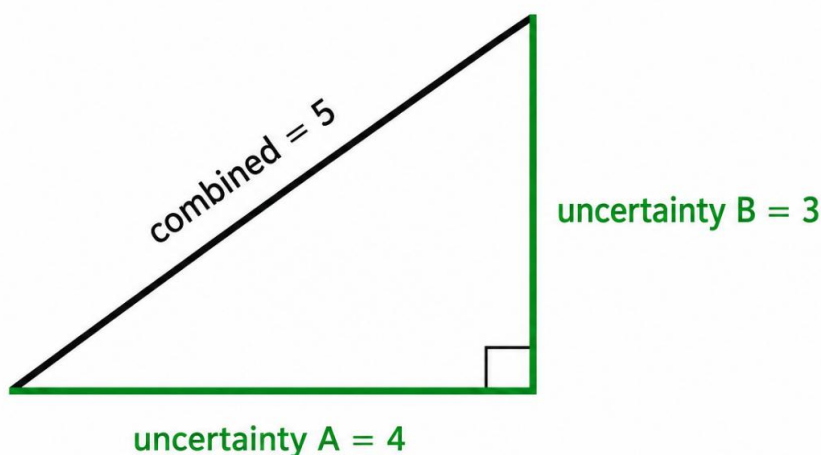
This matters in practice for one reason above all. Both types are real uncertainty, both are standard deviations, and both go into the combination the same way (GUM 3.3.4). There is no penalty box for Type B and no gold star for Type A. The repeatability from four runs and the reference stability off a certificate sit side by side in the budget and add in exactly the same way. Which brings us to how they add.



5. Combining uncertainty components

Parts 3 and 4 sized all nine inputs in our 5000 lbf budget: four Type A pieces earned from data, five Type B pieces built from what we already knew. Now you need one number for the result. You do not get it by adding the pieces straight up.

Combining uncertainty: the law of propagation (root-sum-of-squares)



Independent doubts add in quadrature: $\sqrt{4^2 + 3^2} = 5$ (not 7).

Figure 9. Doubt adds like the sides of a right triangle. Independent uncertainties of 3 and 4 combine to 5, not 7.

Doubt does not add like money. It adds up like the sides of a right triangle. Two uncertainties of 3 and 4 combine not to 7 but to 5, the square root of 3 squared plus 4 squared. This is the root-sum-of-squares, and it is the heart of what the GUM calls the law of propagation of uncertainty (GUM 5.1.2). The reason is that independent doubts partly cancel: when one input happens to read high, another is as likely to read low, so the total grows more slowly than a straight sum.

Before you square and add, though, every piece has to be in the same units, the units of the result. A temperature uncertainty in degrees Celsius cannot be added to a force uncertainty in pounds until you convert it. The converter is the sensitivity coefficient: how much the result moves when one input moves by one unit (GUM 5.1.3).



You already used one. In Section 4 we turned a temperature bound into 0.043 lbf of force by multiplying degrees by the cell's temperature coefficient. That coefficient is the sensitivity coefficient. It answers the recipe question: if this one ingredient is off, how far off is the finished dish? A pinch of salt has a small coefficient. A cup of flour has a large one. Get the big-coefficient inputs right and the small ones barely matter.

For our instrument every input is already expressed in pounds of force at the 5000 lbf point, so the sensitivity coefficients are all 1, and the combination is a clean root-sum-of-squares: square each contribution, add them, take the square root. The nine contributions combine to a combined standard uncertainty of $u_c = 0.666$ lbf.

One number, one sigma, the honest spread of the result. Notice what dominates. Repeatability alone, at 0.58 lbf, accounts for about three-quarters of the total variance. Reference standard stability, at 0.29 lbf, takes most of what is left. Everything else, including the reference standard's own calibration uncertainty at 0.05 lbf, barely registers once it is squared. That is the root-sum-of-squares telling you exactly where to spend your next effort: hold the force point steadier and shorten the calibration interval on the reference. The small terms are not worth a second look.

5.1 Correlation between inputs

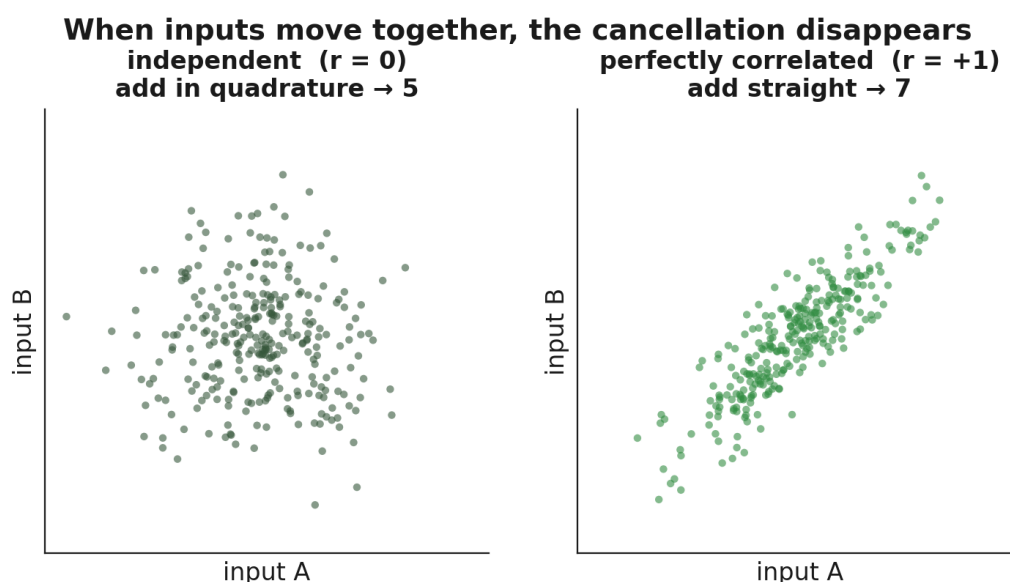


Figure 10. Independent inputs on the left ($r = 0$) add in quadrature. Perfectly correlated inputs on the right ($r = +1$) add straight up.

The root-sum-of-squares has a catch, and it is the one most budgets quietly get wrong. It works only when the inputs are independent. When two inputs move together, the partial cancellation that makes doubt add slowly stops happening, and the real uncertainty is larger than the square-root sum.

Two readings are correlated when something they share makes them drift the same way. Measure two forces with the same reference whose value is slightly off, and both readings are off in the same direction by the same fraction. Read two channels through the same thermometer that is reading 0.5 °C high, and both temperature corrections lean the same way. The shared cause links them (GUM 5.2.2).

The GUM measures the link with a correlation coefficient, a number from -1 to +1 (GUM 5.2.2). Zero means independent, and you are back to plain root-sum-of-squares. Plus one means the two move in perfect lockstep. The extreme case is the one to remember, because it sets the worst-case penalty. When two components are perfectly correlated, they no longer add in quadrature. They add straight up, in a simple sum (GUM 5.2.2). That is the difference between 3 and 4 combining to 5 when they are independent and combining to 7 when they are locked together. Correlation can cost you the whole cancellation.

In our single-point budget the nine inputs are effectively independent, so plain root-sum-of-squares holds and u_c stays 0.666 lbf. But the moment you reuse one reference across several force points, or one thermometer across several channels, the correlation is real and you owe it an honest accounting. The practical fix is usually the cleanest one: remove the shared cause, and the correlation goes with it.

Since 2008: the root-sum-of-squares method taught here now carries a formal name. The JCGM calls it the GUM uncertainty framework, or GUF, to set it apart from the Monte Carlo method of Supplement 1 (JCGM 101:2008), which propagates the whole distributions rather than combining standard uncertainties. Both are current. Part 6 shows the day the GUF stops being enough.



6. Expanded uncertainty and the coverage factor

Part 5 combined our nine inputs into a combined standard uncertainty of $u_c = 0.666$ lbf. That is one sigma, and by itself, it is too thin to act on.

From Standard Uncertainty to Expanded Uncertainty: The Coverage Factor k

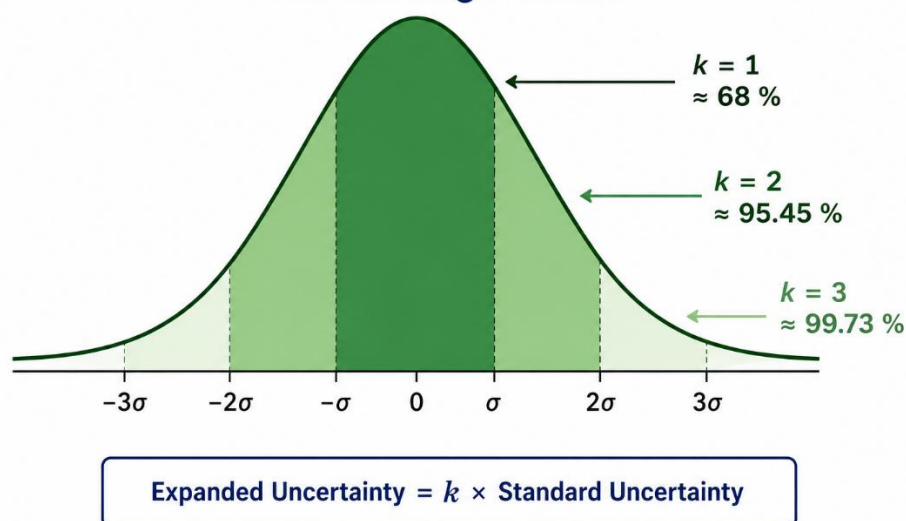


Figure 11. One sigma catches about 68 %. Widen to $k = 2$ for about 95 %, to $k = 3$ for about 99 %.

A one-sigma interval catches only about 68 % of where the value could be, which means it misses almost a third of the time. Nobody ships a tolerance on a coin flip plus a bit.

So you widen it. Multiply the combined standard uncertainty by a coverage factor, called k , and you get the expanded uncertainty U (GUM 6.2.1). The interval $y \pm U$ is meant to capture a large stated fraction of where the value could reasonably lie. Stretch to $k = 2$, and you catch about 95 %. Stretch to $k = 3$, and you catch about 99 % (GUM 6.3.3).

Two things about that multiply are easy to miss, and both matter. First, multiplying by k adds no new information. It only presents the doubt you already had in a more usable form (GUM 6.2.3). The honesty of your U is never better than the u_c underneath it. A confident-looking $k = 2$ on a sloppy budget is still a sloppy budget. Second, the bare number means nothing without its label. $U = 1.71$ lbf is not a result. $U = 1.71$ lbf, $k = 2.57$, approximately 95 % confidence is the result. Without k and the coverage probability, the next person cannot recover your one-sigma number to combine with theirs. State k every time (GUM 7.2.3).



For our instrument, the comfortable move is $U = 2 \times 0.666 \text{ lbf} = 1.33 \text{ lbf}$, $k = 2$, approximately 95 % confidence. Comfortable, and, as the rest of this part shows, not quite honest. The $k = 2$ shortcut comes with a condition almost nobody checks.

This is the Jaws moment. You head out with $k = 2$ because it is the boat tied to every dock. Then you look over the side at how little data is holding your largest term up, and the line writes itself: you are going to need a bigger boat. For our instrument, the bigger boat is $k = 2.57$. Here is why $k = 2$ was never big enough.

6.1 Effective degrees of freedom

Why $k = 2$ sometimes lies: the t -value at low degrees of freedom

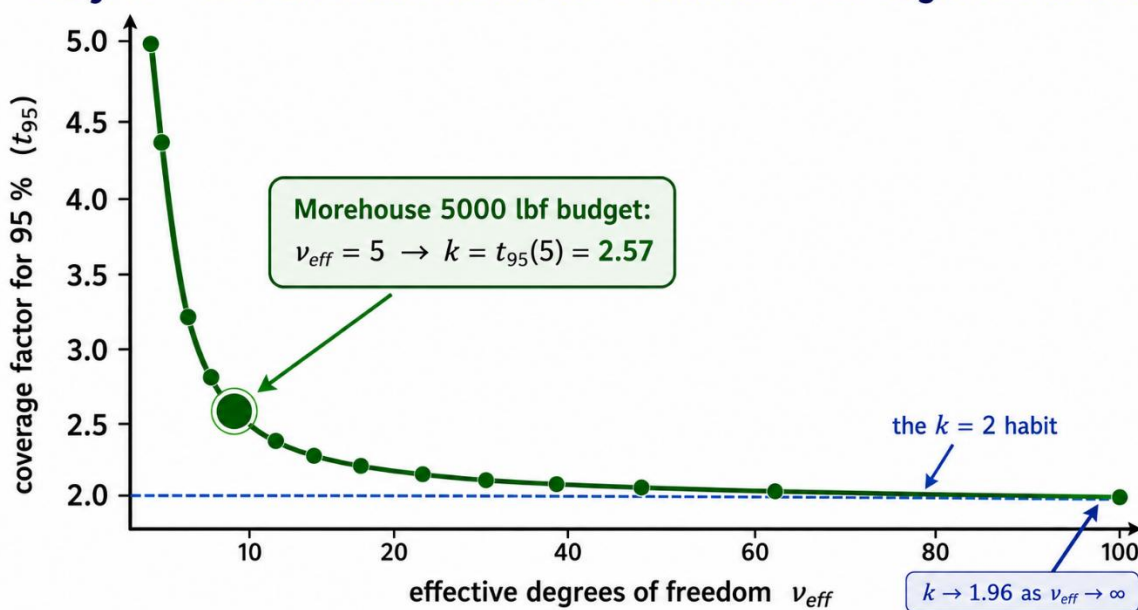


Figure 12. The 95 % coverage factor against effective degrees of freedom. Our budget lands at $\nu_{eff} = 5$, where $k = 2.57$, not 2.

Ask a baseball fan whether a hitter batting 0.400 is better than one batting 0.250, and the answer comes fast. Ask whether a player who went 2 for 5 last night is better than one who went 60 for 240 on the season, and the fan slows down. Two hits in five tries tells you almost nothing. Sixty in two hundred forty tells you a lot. Your uncertainty budget has the same problem, and most people never see it.

Every standard uncertainty carries a second number you rarely write down, its degrees of freedom. Think of it as how many independent looks you got before you trusted the spread. Weigh a part twice, and you barely know its scatter. Weigh it fifty times, and you know it well. Degrees of freedom is the bookkeeping for how much you should trust the uncertainty you just calculated.

For a Type A component the rule is simple: if you took n readings, the degrees of freedom is $n - 1$. Our four repeatability runs give 3. For a Type B component, you assign a large value because you are not going to second-guess a certificate or a resolution limit run to run. The Morehouse budget uses 200 for the Type B pieces, which is large enough that they barely impact the result.

So why does any of this matter? It matters because $k = 2$ is a promise with a condition. When your combined standard uncertainty is built from components you know well, and the result is roughly normal, then $k = 2$ buys you about 95 % (GUM 6.3.3). That is the comfortable case. But when a small-sample Type A component dominates the budget, you do not know u_c as well as $k = 2$ assumes. You calculated a spread from a handful of numbers. The honest interval has to be wider, and the multiplier has to be bigger than 2.

How much bigger? That is the one question the Welch-Satterthwaite formula answers, and it is the part NIST Technical Note 1297 leaves you to look up. In plain language: take each component's contribution to the combined uncertainty, raise it to the fourth power, divide by that component's degrees of freedom, and add those up. Divide the fourth power of the total by that sum. The result is the effective degrees of freedom of the whole budget (GUM G.4.1).

The fourth power is the part to feel. It is brutal. It means the largest, least-trusted component sets the answer almost by itself. A big component you measured four times will swamp a dozen small components you know cold.

Watch it land on our running budget. Repeatability is both the largest contribution, at 0.58 lbf, and the least-trusted, resting on only four runs and 3 degrees of freedom. Everything else all but disappears under the fourth power. Run the numbers, and the effective degrees of freedom comes to about 5.2. Truncate to the next lower integer, and you have 5. Now look up the t -value for 95 % confidence at 5 degrees of freedom. It is 2.57, not 2.00.²

That gap is the whole point. Report $U = 2 \times 0.666 \text{ lbf} = 1.33 \text{ lbf}$, and you have quietly claimed more confidence than your data supports. The honest number is $U = 1.71 \text{ lbf}$, $k = 2.57$, approximately 95 % confidence, nearly 30 % more interval than the lazy version. You only see that cost if you carry the degrees of freedom, and you only shrink it by taking more than four runs at the points that matter.

This is where the ASTM E74 question from Part 3 comes home. Run a full E74 calibration and repeatability does not rest on four runs at one point. It rests on the fit across the whole range, roughly

² A common shortcut skips Welch-Satterthwaite and carries the repeatability term's own degrees of freedom, $n - 1 = 3$, as if it belonged to the whole budget. The two numbers sit close only because repeatability dominates; they are not the same number, and the gap is not cosmetic. Welch-Satterthwaite never returns fewer degrees of freedom than the smallest contributor carries, and the other terms, small as they are, lift the combined uncertainty above the repeatability alone. For this budget $v_{\text{eff}} \approx 3 \times (0.666 / 0.58)^4 \approx 5.2$. At 3 degrees of freedom the 95 % t value is 3.18, not 2.57, so the shortcut would report $U = 2.12 \text{ lbf}$ and overstate the interval by about 24 %. Carry every term's degrees of freedom and do the division.

35 points and 32 degrees of freedom. Rely on the E74 fit for your repeatability, and the dominant term would carry 32 degrees of freedom rather than 3; v_{eff} would climb and k would settle back near 2. We did not do that here. We sized repeatability from a single-point sample on purpose, because that is the doubt in one determination at 5000 lbf, and the price of that honesty is $v_{eff} = 5$ and $k = 2.57$. Same instrument, same machine, two defensible numbers. The only difference is how much data you let stand behind the term that dominates, and you should always know which one you are reporting.

Simple, right? So carry them. Put a degrees-of-freedom column in every budget you build. When the effective degrees of freedom comes back large, use $k = 2$ and move on. When it comes back small, reach for the t table and tell the truth.

There is a second way $k = 2$ can lie, and it runs in the opposite direction. Welch-Satterthwaite handles a small-sample Type A term that dominates. But sometimes the term that dominates is Type B and rectangular, most often a coarse indicator resolution whose half-step bound swamps everything else. When that happens, the combined result is no longer normal, and $k = 2$ over-claims instead of under-claiming (M3003 Appendix C). M3003 gives a quick test: divide the standard uncertainty of all the remaining terms by the standard uncertainty of the dominant rectangular term. If that ratio falls below 0.70, the rectangular term sets the shape, and you read the coverage factor from a convolution table rather than the t table. For a budget that is almost purely rectangular, the 95 % factor falls to about 1.65, not 2. Our running budget is safe here, because the largest term is repeatability, a Type A piece that behaves normally. On an instrument with coarse resolution, run the test before you trust $k = 2$.

Since 2008: the escape hatch at the end of this part is now a full method. When the effective degrees of freedom machinery, the GUF, breaks down, JCGM 101:2008 propagates the input distributions by Monte Carlo and reads the coverage interval straight off the result, with no Welch-Satterthwaite and no t table. You will rarely need it for a force budget. It is the JCGM's own answer when k cannot be trusted at all. And as of 2026 the advice sits in the GUM itself: Amendment 1 (JCGM 100:2008/Amd.1:2026) adds a note at 4.1.4 that when nonlinearity is significant, you either carry the higher-order Taylor terms or go straight to the Monte Carlo of JCGM 101.

7. Reporting the result

Six parts in, we have an expanded uncertainty we can defend: $U = 1.71$ lbf, $k = 2.57$, approximately 95 % confidence. This last part covers what most certificates get wrong, reporting it, and then puts the whole budget on one page.

Report it so someone can rebuild it



Report all three. Drop one and the next person cannot use your number.

Figure 13. A complete statement carries three things: the expanded uncertainty, the coverage factor, and the coverage probability.

A measurement result is a recipe. Leave out the amounts, and no one can reproduce the dish.

The GUM's reporting rules come down to one test: could a competent stranger, given only your report, rebuild your uncertainty and check it? (GUM 7.1.1) If yes, you reported enough. If no, you did not, however tidy the final number looks.

At minimum, every uncertainty statement carries three things together: the expanded uncertainty U , the coverage factor k , and the coverage probability. For our instrument: $U = 1.71$ lbf, $k = 2.57$, approximately 95 % confidence. Drop any one of the three, and you have handed the reader a number they cannot use. Round the expanded uncertainty to two significant figures, and round the measured value to the same decimal place (GUM 7.2.6). Our $U = 1.71$ lbf is two figures already; printing 1.7124 lbf would claim a resolution the budget cannot support.

Above that minimum, and especially as you move up the calibration chain, you also list the components, say how each was evaluated, Type A or Type B, and give the degrees of freedom where it matters (GUM 7.2.7). That is what makes a budget auditable instead of merely asserted.

There is a field check worth reporting alongside the budget. Compare two reference standards that were both calibrated by primary standards, and compute the normalized error, E_n , across the range. If any E_n exceeds one, the budget is not telling the truth, and you do not fix that by inflating the CMC until the number falls in line. You review the whole budget. A clean E_n is the closest thing to a proficiency test you can run on your own floor.

One more report, the one most likely to get you written up: the conformity statement. The instant you say a result passes or fails a tolerance, you have made a decision, and a decision needs a stated rule. A pass or fail with no decision rule, and no word on whether you accounted for uncertainty, is not a conformity statement (ISO/IEC 17025:2017, Clause 7.8.6). State the rule. “Accept if the measured value falls within the tolerance” is one rule. “Accept only if the measured value plus its expanded uncertainty falls within the tolerance” is a stricter one, a guard band, and it shifts who carries the risk. Either can be correct. Silent is not.

Since 2008: the conformity statement this part warns about now has a dedicated rulebook. JCGM 106:2012, The role of measurement uncertainty in conformity assessment, works out decision rules and guard bands in full, and ISO/IEC 17025:2017 (Clause 7.8.6) makes stating the rule a requirement, not a courtesy.

7.1 Worked example: the 5000 lbf budget



Where the doubt comes from: Morehouse 5000 lbf uncertainty budget

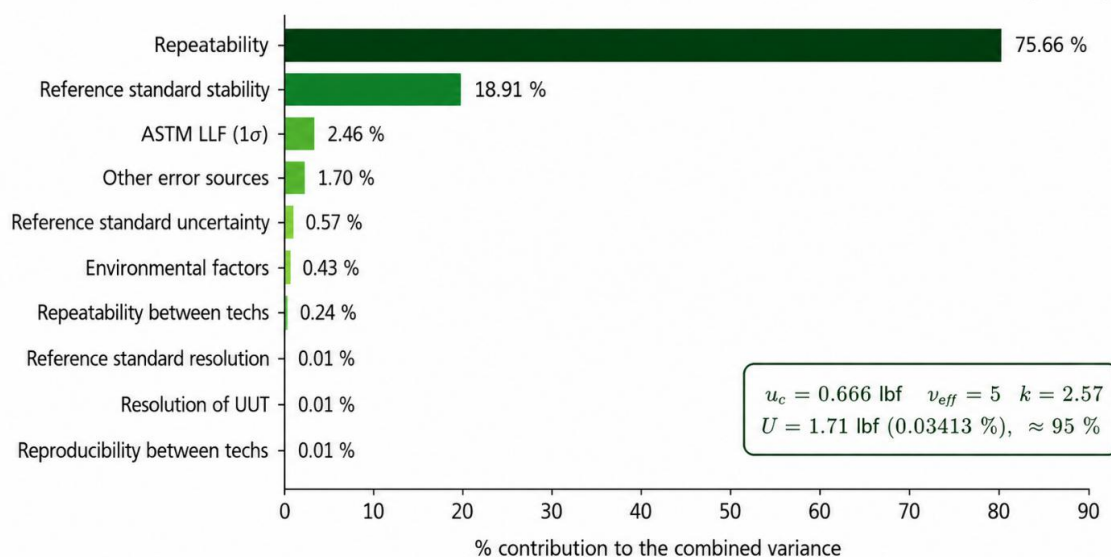


Figure 14. The complete 5000 lbf budget. Repeatability sets three-quarters of the variance; the reference standard sets most of the rest.

Here is everything from the last six parts on one page: a force-measuring instrument calibrated on a Morehouse machine at 5000 lbf (22.241 kN), following ASTM E74, built into a complete uncertainty budget the way you would actually run it.

Source	Type	How it was sized	Std. uncert. (lbf)	df	% of variance
Repeatability	A	Std. dev. of 4 runs at 5000 lbf	0.58	3	75.66
Reference standard stability	B	0.5 lbf bound, rectangular $\div \sqrt{3}$	0.29	200	18.91
ASTM LLF (1σ)	A	Reported LLF 0.25 lbf $\div 2.4$	0.10	32	2.46
Other error sources	B	Side load 0.05 % $\times 0.0625 \text{ in}$, $\div \sqrt{3}$	0.087	200	1.70
Reference standard uncertainty	B	Deadweight CMC 0.1 lbf at $k = 2$, $\div 2$	0.05	200	0.57
Environmental factors	B	0.075 lbf bound, rectangular $\div \sqrt{3}$	0.043	200	0.43
Repeatability between techs	A	Pooled std. dev. across techs	0.032	10	0.24
Resolution of UUT	B	0.025 lbf step, $\div \sqrt{12}$	0.007	200	0.01
Reference standard resolution	B	0.025 lbf step, $\div \sqrt{12}$	0.007	200	0.01
Reproducibility between techs	A	Std. dev. of technician means	0.006	1	0.01

Combine by root-sum-of-squares: $u_c = 0.666 \text{ lbf}$. Effective degrees of freedom by Welch-Satterthwaite: repeatability, with 3 degrees of freedom and the largest contribution, dominates the fourth-power sum, so v_{eff} comes to about 5.2, truncated to 5. Coverage factor from the t -table at 95 % and 5 degrees of

freedom: $k = 2.57$. Expanded uncertainty: $U = 1.71$ lbf (0.0342 %), $k = 2.57$, approximately 95 % confidence.

Read that result against the lazy version. The $k = 2$ habit would have reported $U = 1.33$ lbf and felt fine about it. The honest budget reports 1.71 lbf, nearly 30 % wider, because the one contributor that swamps the budget rests on only four runs. Neither number is a typo. One tells the truth about how much you actually know, and the other does not.

Now zoom out one level, because the single biggest lever is not in this table. It is the reference standard itself. Calibrate this same instrument against deadweight primary standards known to 0.0016 %, and the expanded uncertainty at 100 000 lbf is about 9.78 lbf. Calibrate it instead against a secondary machine known to 0.025 %, and the same point jumps to about 26.32 lbf. Same instrument, same care, more than double the uncertainty, purely from the standard above you. No transfer method beats deadweight. The reference you choose sets the floor for everything below it.

One caveat about this budget: it sizes a single ascending point. A real certificate covers every calibration point, because repeatability, resolution, and the reference's own uncertainty all change with applied force. ASTM E74 handles that by fitting an equation across the range, and a result that lands between calibrated points inherits the added doubt of that interpolation (M3003 E.4). Morehouse publishes a CMC that builds the budget for every force point at once, so you do not work them one at a time. And if the cell will be used to calibrate descending points, hysteresis enters the picture, and you develop a separate budget for the descending direction.



Coverage Factor k for 95 % Confidence

Based on Student's t distribution (two-sided)

 Readings (n) number of observations	 Degrees of freedom (ν) $\nu = n - 1$	 k for 95 % two-sided coverage
2	1	12.71
4	3	3.18
6	5	2.57
11	10	2.23
21	20	2.09
∞ (many)	∞	1.96



As the number of readings increases, the degrees of freedom increase and k approaches 1.96 (the normal value). With few readings, k must be larger to give 95 % coverage.



Note: Use this table to select the appropriate coverage factor k for a 95 % two-sided interval based on your effective degrees of freedom (ν).

Figure 15. The 95 % coverage factor k against the number of readings. As observations accumulate the effective degrees of freedom rise and k falls toward 1.96; with only a few readings, k must be larger to hold 95 % coverage.

That is the GUM, end to end. Write the model. Size each input, by data (Type A) or by what you already know (Type B). Convert each to the units of the result with its sensitivity coefficient. Combine by root-sum-of-squares, watching for inputs that move together. Carry the degrees of freedom, pick the coverage factor honestly, and report all three numbers so the next person can check you.

Seven parts of plain English for one disciplined page. The official GUM, JCGM 100:2008, has the full treatment, the proofs, and the harder examples, and you should keep it close. But the next time a certificate crosses your desk reading $U = 1.71$ lbf, $k = 2.57$, you will know exactly where the 2.57 came from, what it promised, and what would have made it bigger.

You do not have to grind through any of this by hand. Morehouse built the whole calculation, effective degrees of freedom and coverage factor included, into a free Excel CMC workbook at mhforce.com.

If you build force budgets, build the next one this way: every input written down, every contribution in the same units, a degrees-of-freedom column you actually fill in, and a coverage factor you can defend. **Write down every input, fill in every degree of freedom, and the coverage factor walks out of the math on its own, the way the ballplayers walked out of the corn when he built it. Build your budget correctly, instill confidence, and customers will come.**



Appendix A. Summary of formulas

Symbols are italic for quantities (U , k , u , s , n) and roman for units (N, kN, lbf). The combined standard uncertainty is u_c , and degrees of freedom is ν .

Standard uncertainty from repeated readings (Type A). For n repeated readings with experimental standard deviation s , the dispersion of a single reading is s itself; the standard uncertainty of their mean is $u = s / \sqrt{n}$. In force calibration, the repeatability carried is usually s , the doubt in a single determination, not $s / \sqrt{n}$.

Type B from a stated coverage factor. $u = U / k$.

Type B from a bound of half-width a . Rectangular: $u = a / \sqrt{3}$. Triangular: $u = a / \sqrt{6}$. Normal 99.73 % bound: $u = a / 3$. Normal 95.45 % bound: $u = a / 2$. U -shaped: $u = a / \sqrt{2}$.

Combined standard uncertainty. For independent inputs, root-sum-of-squares: $u_c(y) = \sqrt{[\sum (c_i \times u(x_i))^2]}$. With correlation, add a covariance term under the root; for perfectly correlated inputs the result reduces to a straight sum.

Expanded uncertainty. $U = k \times u_c(y)$.

Effective degrees of freedom (Welch-Satterthwaite). $\nu_{eff} = u_c(y)^4 / \sum [u_i(y)^4 / \nu_i]$. Truncate to the next lower integer, then take k as the t value at your confidence level for that ν_{eff} . Note that ν_{eff} can never be smaller than the smallest ν_i in the sum, so a shortcut that carries only the dominant term's degrees of freedom understates ν_{eff} and overstates k .

ν	1	2	3	4	5	6	8	10	20	∞
t	12.71	4.30	3.18	2.78	2.57	2.45	2.31	2.23	2.09	1.96

Selected t values at 95 % confidence (NIST Technical Note 1297, Table B.1).

Worked combination for the running example. $u_c = \sqrt{(0.58^2 + 0.29^2 + 0.10^2 + 0.087^2 + 0.05^2 + 0.043^2 + 0.032^2 + 0.0029^2 + 0.006^2)}$ lbf = 0.666 lbf. The fourth-power sum is dominated by repeatability, $0.58^4 / 3$, so $\nu_{eff} \approx 5.2$, truncated to 5. $U = 2.57 \times 0.666$ lbf = 1.71 lbf (0.0342 %).

References

JCGM 100:2008, Evaluation of measurement data – Guide to the expression of uncertainty in measurement (the GUM), including Amendment 1:2026. Joint Committee for Guides in Metrology.

JCGM 101:2008, Evaluation of measurement data – Supplement 1 to the GUM: Propagation of distributions using a Monte Carlo method. Joint Committee for Guides in Metrology.

JCGM GUM-5:2026, Examples of uncertainty evaluation. Joint Committee for Guides in Metrology.

JCGM 200:2012, International vocabulary of metrology – Basic and general concepts and associated terms (VIM). Joint Committee for Guides in Metrology.

Taylor, B. N., and Kuyatt, C. E., Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results. NIST Technical Note 1297, 1994.

ISO/IEC 17025:2017, General requirements for the competence of testing and calibration laboratories, Clauses 7.6 and 7.8.6.

ASTM E74, Standard Practices for Calibration and Verification for Force-Measuring Instruments. ASTM International.

UKAS M3003, The Expression of Uncertainty and Confidence in Measurement, Edition 6, 2024. United Kingdom Accreditation Service.

Thompson, A., and Taylor, B. N., Guide for the Use of the International System of Units (SI). NIST Special Publication 811, 2008.