

TAKEOFF AGREEMENT BENCHMARK

A switchgear takeoff to level gear quotes against: how often Makeoff's gear list matches a working estimator's, on 32 real projects

SEP 28, 2026 · WILL ZACHER

99%	90%	30 / 32	13
of the estimator's gear listed, after review	of Makeoff's gear also on the estimate	projects agree on every gear unit priced	contractors, eight states, nothing curated

Executive summary

Every switchgear quote a contractor receives has to be leveled: checked, unit by unit, against what the drawings call for, so that a low number is a real number and not a missing transformer. That check needs an independent gear list to level against, and producing one is the estimator's first pass on every bid. Makeoff now produces that list from the drawings. This paper measures whether it is good enough to level against, by asking the only question that matters for the job: when Makeoff reads the same drawings a working estimator was given, does it list the same gear?

On 32 real projects from 13 contractors, it listed 97% of the gear the estimator listed, and 88% of what it listed the estimator had listed too, measured on the gear slice (panels, switchboards, transformers, transfer switches, generators, UPS, surge devices and CT cabinets) at the type level, averaged per project. After the eleven disagreements were reviewed against the drawings, agreement is 99% on the estimator's side and 90% on Makeoff's, and 30 of the 32 projects agree on every gear unit the estimator priced. Two of the projects used a switchgear vendor's quote as the reference rather than an estimator's takeoff, which is the leveling case itself; on both, the list Makeoff produced from the drawings found every unit the quote priced and surfaced the units the quote left to the contractor.

Nothing was curated. The projects are what contractors uploaded over the past year with their own takeoffs attached, in the state they uploaded them, across eight states and twelve reference layouts, from Excel parts lists and Bluebeam reports to vendor quotes and handwritten pages. The paper does not claim the software is right and the estimator is wrong. It reports how often two readers of the same drawings agree, prints every case where they do not with its cause, states which pairs could not be

scored and why, and publishes the rules, the code and the reviewer decisions so the number can be regenerated by anyone with the files. Breakers and disconnects are reported separately and sit lower; those gaps are named and are product work.

Why a takeoff to level against

The switchgear package is the largest single purchase on most commercial electrical bids, and it is bought on quotes. A contractor sends the drawings to two or three distributors, each comes back with a list of gear and a price, and the estimator has to level them: confirm that each quote carries every panel, switchboard, transformer, transfer switch and generator the drawings call for, at the rating the drawings state, with nothing left to the contractor that the contractor did not plan to buy. A quote that omits a 112.5 kVA transformer is not the low quote; it is a change order in the field. Leveling is how that gets caught before bid day.

Leveling needs a reference list that came from the drawings and not from any vendor: the estimator's own gear takeoff. Producing it is the first pass on every bid and it is skilled work, reading the one-line, the riser and every panel schedule, and it is the pass that gets squeezed when six bids are due in a week. Makeoff has released an automatic takeoff for that package. The claim this paper tests is narrow and useful: that the list Makeoff produces from a drawing set is close enough to what a working estimator would produce that a contractor can level quotes against it.

The estimator is the reference. There is no answer key for a set of construction drawings; the closest thing to one is the takeoff a working estimator produced under bid conditions and submitted. So each project here pairs Makeoff's output with that estimator's own document, and agreement is measured item by item in both directions. Recall is the share of the estimator's gear that Makeoff also found: for leveling, the units the reference list would have to carry. Precision is the share of Makeoff's gear that the estimator also had: for leveling, whether the reference list is padded with units a vendor would rightly refuse to quote. Recall is the headline because a unit missing from the reference is a unit no quote gets checked for; both are printed on every project.

An estimator can miss a unit or price one the drawings do not show, and when that happens the score records a disagreement, not an error on either side; the disagreement list is published so the reader can judge each case. The test is deliberately unforgiving in three other ways. The projects were not selected for quality; they are whatever contractors uploaded, including partial sets and handwritten takeoffs. The rules that pair items were settled on a separate calibration set and then checked, never tuned to a project in the reported set. And every pair that could not be scored is counted and explained rather than dropped.

What was tested

Thirty-four projects were attempted and 32 were scored. They came from 13 electrical contractors who uploaded bid documents to Makeoff between March and August 2026 and attached their own takeoff to the project record. Every such pair on the platform that passed a document gate went into the set; none was chosen for how it would score. The two that fell out are described under Outside the score.

The spread is the point. The scored projects include a 1.1 million square foot distribution warehouse in Pennsylvania with two 4000A service switchboards, a pharmaceutical manufacturing fit-out in South San Francisco, a Publix in Key West, a Microtel, a six-story residential building over parking on Long Island, an athletic center addition and a kitchen addition at two public schools, an elementary school modernization in California, a community solar farm in Maryland, a bank branch, a collision repair facility, a pool house, a single-family residence, and a run of quick-service restaurants, medical and dental tenant improvements and retail fit-outs. Eight states are represented. Gear counts run from one panel to thirty units; the median project priced two gear units, and nine projects priced six or more.

The estimator references were used exactly as submitted, and they came in twelve layouts. Handwritten pages were transcribed to CSV with the scan filed beside the transcription; nothing else was reformatted.

REFERENCE LAYOUT	PROJECTS	EXAMPLE
Excel or CSV parts list	17	one gear line per row with quantity and description; includes three handwritten pages transcribed to CSV
PDF material summary	3	two-column summaries
Vendor gear quote	2	Siemens and ABB quotes standing in for the takeoff (reported as their own class)
Priced item list	2	quotes with quantity, unit and price
PDF parts list	1	assemblies with cost and hours, Aldi Dillsburg
Estimate detail	1	Fifth Third Bank
Bluebeam measurement report	1	Baldwin High School
QuickBooks estimate	1	Meadow Elementary
Count and description sheet	1	Family Initiative
Material order	1	16 Handles
Tag and quantity worksheet	1	gear quote request, Exton Carstar
Narrative scope	1	Half Price Books

Two projects use a switchgear vendor's quote as the reference rather than an estimator's document. A quote is what was bought, not what was taken off, so those two are marked as a vendor-quote class throughout and their scores are shown with that label. On the drawings side, every project is Makeoff's

standard export for the uploaded set: the extracted schedules, one-line and riser text, plan artifacts and scope-of-work items, with no manual correction.

How agreement is measured

Both sides of a pair are reduced to the same shape before anything is compared: a list of items, each with a type (panelboard, switchboard, transformer, transfer switch, generator, UPS, surge device, CT cabinet, disconnect, breaker), a designation where the document gives one, the ratings the document states (amps, kVA, kW, voltage, poles, main type), a quantity, and a scope mark. The estimator's document is parsed by a layout-specific reader and then normalized with one set of rules; Makeoff's export goes through a reader for its schedules, diagrams and scope text and then the same rules. From that point on neither side is a document anymore, and the comparison cannot favor one format over another.

Items are paired in stages. Panels and boards pair on designation first, one to one, so an estimator's "Panel MDP-2" and a schedule titled "Distribution Panel MDP-2" are one thing. Unnamed panels pair by family and rating with the closest unpaired candidate. Unit gear pairs on type and rating, and restated units (the same transformer on the one-line and again in scope text) fold to one before pairing. A pair is scored at three strictness levels: type, which asks whether both sides list the unit; type plus rating, which adds the amps or kVA; and type plus rating plus quantity. The headline is the type level, and rating agreement is reported on its own, on the units both sides found, rather than blended in.

Recall on a project is the number of estimator units that found a partner divided by the number of estimator units. Precision is the number of Makeoff units that found a partner divided by the number of Makeoff units. Both are computed per project and averaged, with the median and the middle half shown, and pooled across all units as a second figure, so a thirty-unit grocery store and a one-panel residence each count once in the mean and by weight in the pool.

Two slices are always reported. The gear slice is the boxes on the one-line: panels, boards, load centers, meter centers, transformers, transfer switches, generators, UPS, surge devices, CT cabinets and generator docking stations. The all slice adds breakers and disconnects, which are numerous and would otherwise dominate a unit count. Every number in this paper says which slice it comes from.

Scope marks keep the score honest about what a contractor buys. Gear the drawings mark as existing, demolished, relocated or furnished by others is carried through the comparison but never counted against Makeoff when the estimator left it out, and the estimator's own demolition and by-others lines go to an audit list rather than the score. Feeder lengths, wire, conduit, fixtures and devices are outside the gear package and are not scored at all.

A gate decides whether a pair can be scored. The estimator's document must name at least one panel, board, transformer, transfer switch or generator; a breaker-only count sheet is a device count, not a gear takeoff. The export must contain distribution equipment. And the two documents must describe the same drawings, checked by whether they share any panel designation when both name three or more. A pair that fails the gate is recorded with its reason and counted in the completion rate.

The pairing and normalization rules were settled on a separate calibration set of seven projects with a distribution package and regions, run before any reported project. As reported pairs came in and exposed a new document layout or an unhandled phrasing, a rule was added only if it was general (a way estimators write, not a fact about one job), and every change was re-run against the calibration set and kept only when the rows it moved there were checked against the drawings. The calibration set is scored on the final rules and reported beside the result. It never appears in the headline number.

After the harness scored the set, the disagreements were reviewed against the drawings by an estimator. Each decision is recorded in a file with the reviewer, the date and the drawing evidence quoted, and the harness applies that file on top of the raw match, writing the decision into every row it touches. Both numbers are published: the harness alone, and the harness with the review applied.

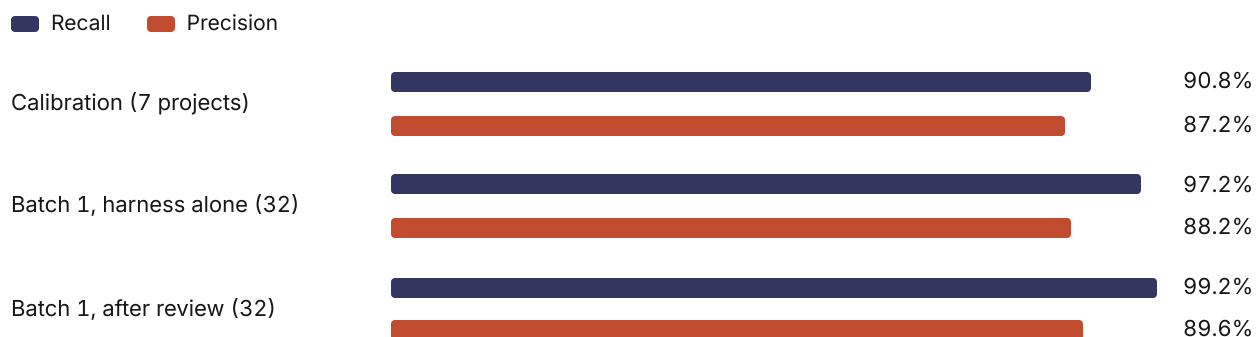
Results

On the gear slice at the type level, the harness alone scores 97.2% recall and 88.2% precision as means per project, with a median of 100% on both and the middle half of projects at 100% recall. Pooled across units, 199 of the 210 gear units the estimators priced were found (94.8%), and 199 of the 238 units Makeoff listed had been priced (83.6%). Twenty-seven of 32 projects reached 100% recall and twenty agreed exactly in both directions.

After the estimator's review of the eleven disagreements, recall is 99.2% (pooled 203 of 208, 97.6%) and precision 89.6% (pooled 203 of 239, 84.9%). Thirty of 32 projects reach 100% recall; the two that do not are the Publix, where three surge devices stand as misses, and the Grantville Microtel, where two panels named on the material summary have no counterpart on the drawings. The review is described in full in the next section and in Appendix B, and every figure in this paper is given both ways.

FIGURE 1

Mean agreement per project, by set

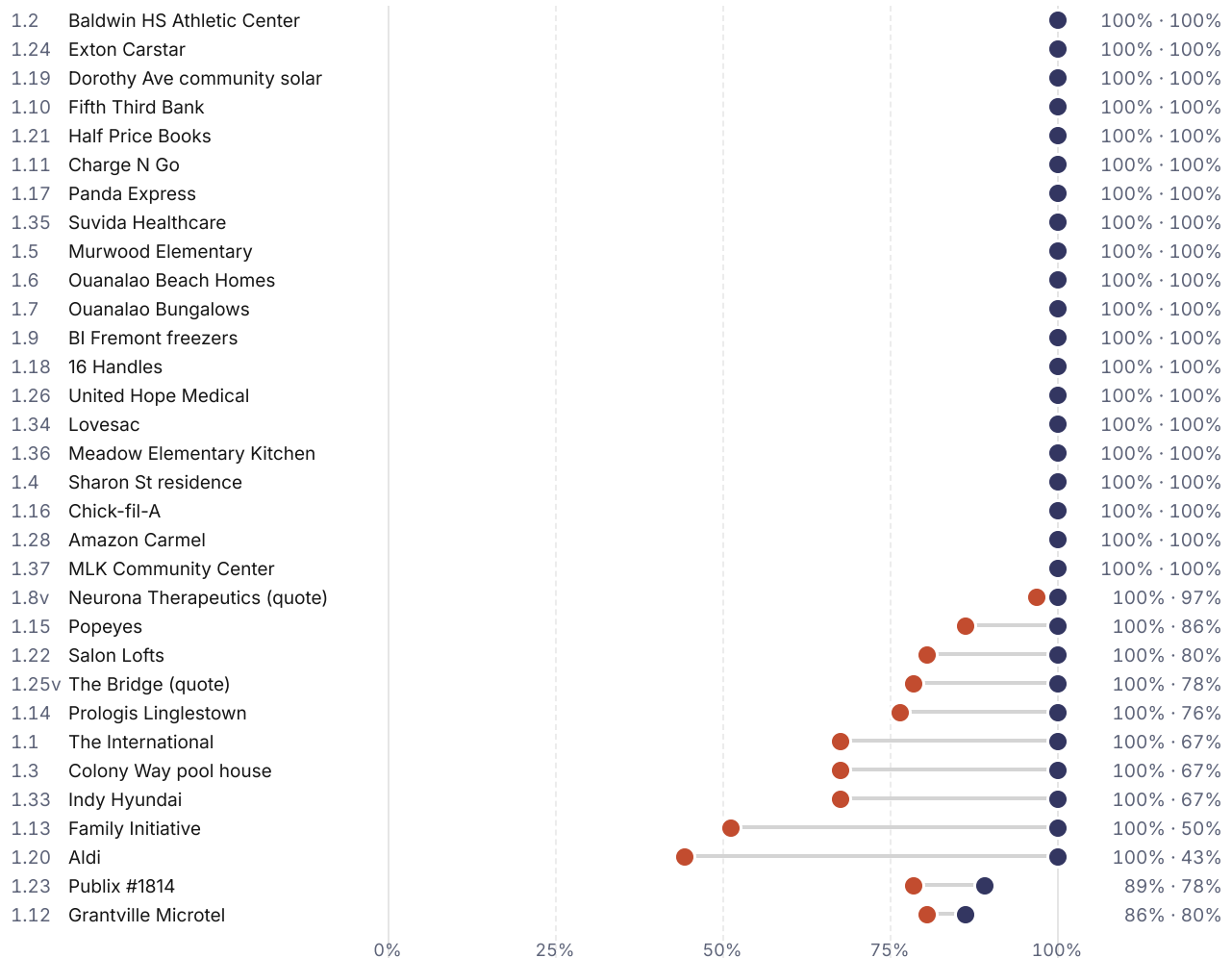


Same harness (makeoff-bench 0.3.1), same rules hash. The calibration set was used to settle the rules and is kept out of the headline. makeoff-bench 0.3.1 · results.json, raw and adjudicated runs · 32 scoreable of 34 attempted

FIGURE 2

Gear agreement by project, after review

● Recall ● Precision



Recall = share of the estimator's gear Makeoff found. Precision = share of Makeoff's gear the estimator priced. Type level, gear slice. Sorted by recall, then precision.

Ratings agree on the units both sides found. At the type plus rating level, recall is 97.4% raw and 99.1% after review; on found units the stated amperage or kVA agreed 93.5% of the time (mean per project). Most of the rating disagreements are one estimator habit: on the Publix, the material summary carries every panel at 250A while the schedules rate them from 100A to 400A. Quantity agreed on 84.4% of found units, with the differences concentrated on the two vendor-quote projects, where a quote's count and the drawings' count are not the same thing.

The calibration set of seven projects, scored on the same rules, sits at 90.8% recall and 87.2% precision. It is lower than the reported set because it carries the hardest layouts (a sectioned PDF estimate, a hotel with 19 gear units on a partial set) and because it was never allowed to benefit from the review; it is

shown so a reader can see that the reported number is not an artifact of tuning.

Where the two readers disagreed

Every unit the estimator priced that the harness did not find, and every unit Makeoff listed that the estimator did not price, is printed here with its cause. This is the part of the paper an estimator should read most closely, because it is where the number comes from.

FIGURE 3

Every disagreement, one block per unit

Priced by the estimator, not found		11 of 210	Listed by Makeoff, not priced		39 of 238
Credited on review		5	Drawn, not priced		16
Left the denominator		1	Out of the reference's scope		8
Stands		5	Named in text only		5
			Riser OCR variant		4
			Label variant		3
			Restatement		3

Left: the eleven units the harness did not find, by outcome after review. Right: the 39 units Makeoff listed that the estimator did not price, by what the drawings say about each one. Both are itemized in the tables below.

Units the estimator priced that were not found

Eleven of 210. They fall on six projects, and seven of the eleven sit on the two vendor-quote references, where the quote and the drawings disagree about how many units there are.

PROJECT	MISSED	THE REFERENCE PRICED	THE DRAWINGS SHOW	AFTER REVIEW
1.8v Neurona (gear quote)	3	four 112.5 kVA transformers, T-1 to T-4	all four named on the one-line; the harness emitted T-1 only	Credited: a harness rule gap, not an extraction miss
1.23 Publix Key West	3	three SPD units on the quote	none reached the export; surge devices are outside Makeoff's distribution definition today	Stands
1.23 Publix Key West	1	two generators	one 125 kVA emergency unit; the 500 kW standby is marked by others	Left the denominator: not this contractor's to buy
1.12 Grantville Microtel	2	panels 4C and AB	the drawings schedule 4A and 4B and never name AB	Stands, not yet reviewed
1.13 Family Initiative	1	a second 200A panel enclosure	in the riser text as the 200A UNIT 6 device at the existing gear	Credited: in the extracted text, dropped by the harness
1.25v The Bridge (quote)	1	panel K21	the switchboard schedule lists retail panel K21 at 800A; the riser labels the kitchen panel K1	Credited: same panel, label transposed

The pattern matters more than the count. Six of the eleven were in Makeoff's own extracted text and were lost between the export and the score by a comparison rule that read a transformer clause, a switchboard-schedule position or a meter-center device too narrowly. Those are logged as harness changes for the next version and the rules were left alone for this paper so the raw number stands as run. One is a scope call the drawings make explicitly. The four that stand are two real product gaps: surge devices, which the extraction does not yet define, and two panels a contractor listed that the drawings do not show, which may be the estimator's error or a naming the reviewer has not yet traced.

Units Makeoff listed that the estimator did not price

Thirty-nine of 238, on 13 projects. They are classified by what the drawings say about each one.

CLASS	UNITS	WHAT IT MEANS
Drawn, not priced	16	The plans show it and the estimate leaves it out: apartment load centers on The Bridge, panel-mounted SPDs and a scheduled 60A panel on Prologis, two service switchboard sections on the Microtel that the material summary never lists, a 22 kW generator at the pool house.
Out of the reference's scope by its own terms	8	Owner-furnished panels (Aldi), gear a quote leaves to the contractor (The Bridge), by-others units.
Named in text only	5	A designation on a riser or in scope text with no schedule and no rating behind it.
Riser OCR variant	4	One label on the Publix riser read four ways.
Label variant	3	The same panel under two names: LPN1A on the Hyundai riser is most likely the scheduled Panel C; a 60A sub panel on United Hope is most likely the new 60A breaker in MDP-7.
Restatement	3	The same unit read twice, a schedule and a riser label at different ratings. Marked as duplicates in the review; this is where precision moves from 88.2% to 89.6%.

The first class is the one that carries product value. Sixteen units on the drawings that an estimate left out is, for a contractor, a list of questions to settle before bid day: is the load center furnished by the owner, is the SPD in the panel spec, did we see that second switchboard section. The paper does not say those estimates were wrong; several are scope decisions a reader of the specification would make the same way. It says the software saw them and the estimate did not, and lists them.

Leveling in practice: the two vendor-quote projects

Two of the 32 references are not estimator takeoffs but the switchgear vendor's quote itself, attached to the project record in place of a takeoff. Scoring Makeoff's list against a quote is the leveling job done backwards: instead of checking a quote against the estimator's list, the paper checks Makeoff's list against a quote and reports where the two disagree. What comes out is what a contractor leveling that quote would want to know.

FIGURE 4

Leveling a vendor quote, unit by unit

The Bridge, Mineola NY

Siemens quote



28 / 28

QUOTED UNITS FOUND

8

BEYOND THE QUOTE

Neurona Therapeutics

ABB quote



30 / 30

FOUND, AFTER REVIEW

1

BEYOND THE QUOTE

■ Quoted, found ■ Quoted, credited on review ■ On the drawings, not on the quote

On The Bridge, a six-story residential building over parking in Mineola, the Siemens quote carries four switchboards and 24 panels. Makeoff's list from the drawings found every one of the 28 units the quote priced. It also listed eight units the quote did not carry: three apartment load-center types the quote's notes leave to the contractor, a 600A kitchen panel the riser labels K1 and the switchboard schedule labels K21 at 800A (the quote carries the 800A unit; the rating conflict on the drawings is itself a leveling question), and three panels named on the riser with no schedule behind them. A contractor leveling that quote against Makeoff's list would have the load-center question and the K1/K21 rating question on the table before the quote was accepted.

On the Neurona Therapeutics manufacturing fit-out in South San Francisco, the ABB quote carries two switchboards, 24 panels and four transformers. Makeoff's list found all 30 units after review, and listed one unit beyond the quote. Here the disagreement ran the other way in the raw run: the harness had emitted only one of the four transformers named on the one-line, a rule gap in the comparison, not the extraction, and the review credited the other three from the export's own text. A leveling check that read the export directly would have had all four.

The same pattern appears on the estimator-takeoff projects wherever the estimate itself was a gear list. On the Publix in Key West, the contractor's gear export prices two generators; the drawings mark the 500 kW standby unit as by others, and Makeoff's list carries the scope mark. On Prologis Linglestown, Makeoff listed five panel-mounted surge devices and a scheduled 60A panel the material summary never carries. On the Grantville Microtel, it listed two service switchboard sections the material summary omits. Sixteen such units across the batch, drawn on the plans and absent from the estimate, are catalogued under Where the two readers disagreed. Some are scope calls a reader of the specification would make the same way; every one is a question a leveled quote should have answered.

Below the gear line

Breakers and disconnects are scored the same way and reported separately, because they are numerous, because they reach the extraction by a different path than the one-line does, and because a unit count that blended them with the gear would say nothing about either.

On the all slice (gear plus breakers and disconnects) the mean recall per project is 55.0% and the median 54.1%, with precision at 70.5%. Pooled by category, Makeoff found 206 of the 921 breakers the estimators priced (22%) and 20 of the 162 disconnects (12%). Precision on disconnects is 100%: every disconnect the export produced had been priced. That asymmetry is the finding. The export does not invent disconnects; it sees one only where a schedule or the one-line names it, and most disconnects live on the power plans as symbols beside the equipment they serve. On at least seven projects (Popeyes, Panda Express, Salon Lofts, United Hope, Amazon Carmel, Lovesac, MLK Community Center) the estimator counted plan-level disconnects the export never carried.

Breakers fall short for a different reason. Panel schedules reach the export, but flattened. Some arrive as prose with the trip column dropped (Salon Lofts, Hyundai Panel C), some carry a “new breaker” legend without the marks that say which positions are new (Hyundai P2 to P4), and some state a pole count only where the drawing spelled it out (Meadow Elementary). Where a schedule arrived intact, as at Prologis or the Baldwin athletic center, breaker recall on that project is high; where it arrived as a paragraph, the harness recovers what it can and the rest is lost.

These are product gaps, and they are stated here as such. A structured circuit table per panel, with position, trip, poles and lifecycle mark as fields, would close most of the breaker gap; reading disconnect symbols off the power plans is a separate extraction problem and the larger one. Neither is a finding about estimators, and neither changes the gear result above, which is the package the switchgear vendor prices.

Two smaller extraction issues recur and are listed for completeness: the same sheet extracted three times (Rise York, E-601), and a riser label read four ways on the Publix. Four of the 34 exports were empty on the first processing pass (Grantville Microtel, Prologis, Suvida, Meadow Elementary) and complete on a re-pull, so that failure sits in the pipeline run rather than in the drawings; all four are scored on the re-pulled export.

Outside the score

Thirty-four projects were attempted and 32 scored, a completion rate of 94%. The two outside the score are reference problems, not extraction problems, and the distinction matters for how the number should be read.

PROJECT	WHY IT IS NOT SCORED
1.8 Neurona Therapeutics (estimate)	The takeoff attached to the record names panels A, B and MDP; the drawings name DB1 to DB6. No designation in common: the takeoff on file belongs to another job. The ABB gear quote attached to the same record is scored instead, as 1.8v.
1.38 Rise Dispensary, York	The estimator document is breaker counts and labour on three existing panels, with no gear unit to compare. It is a device count, and the gate sets it aside.

Where an export existed and the reference was a gear takeoff, 32 of 32 scored. Four exports were empty on the first processing pass and came back complete when re-run on 23 September; all four are in the scored set on the re-run export, and one of them, Prologis Linglestown, is the largest estimator-takeoff project in the batch at 19 gear units and 100% recall.

Seven roster entries were never filed and are not in any number here: five handwritten pages that carry no job name and have not yet been matched to their projects (1.27, 1.29 to 1.32), and two projects whose takeoff sits on the Makeoff record and was not pulled in time (1.39 Streetsboro Commons, 1.40 Seagoville Elementary). They can be added as an addendum without changing the method.

Twenty further projects on the platform were screened and dropped before scoring because their attached document was not a gear takeoff: lighting-only lists, conduit measurement reports, device counts, a scope letter, a single-item ATS quote, and two New Zealand projects. Nineteen more had the drawings themselves uploaded in the takeoff slot and no estimator document at all. Those 39 are listed in the project roster that accompanies the harness, so the reader can see what was left out and why.

Limitations

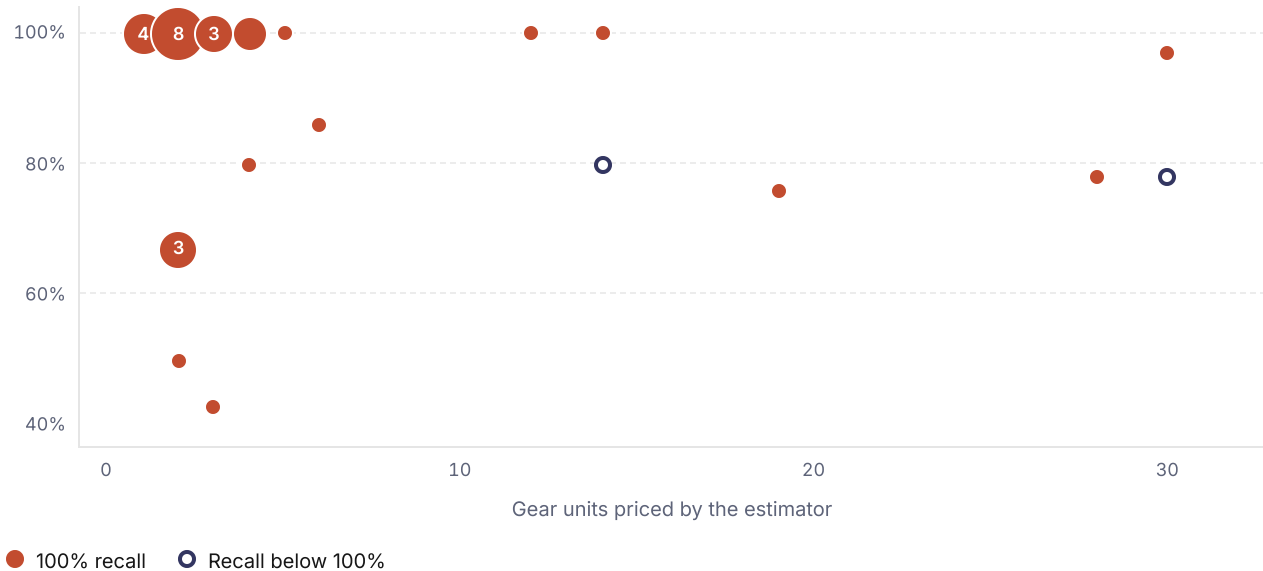
The reference is one estimator per project. Their takeoff was produced under bid conditions, with the specification and any addenda in front of them, and it is the best available record of what a competent reader took off those drawings. It is not an answer key. Estimators leave units off and add units the drawings do not show, and the disagreement list above contains examples of both. Where the paper credits Makeoff after review, the review is one estimator's reading of the drawings, recorded with its evidence; a second reviewer could disagree on a row, and the raw number is printed so that a reader who trusts no reviewer still has a figure.

The drawings are whatever was uploaded. Several sets were partial, one was a progress bid set, and the schedules on some sheets reached the extraction as prose. That is the condition under which the product is used and the paper does not correct for it, but it means the result describes Makeoff on the documents contractors actually have, not on clean sets.

The sample is 32 projects and 210 gear units, and the median project priced two units. The mean per project is therefore driven by many small jobs that agree exactly, and the pooled figure is driven by four large ones (Publix, The Bridge, Neurona, Prologis) that carry more than half the units. Both are printed for that reason. A reader who works on large jobs should weight the pooled figure and the per-project table; a reader who works on tenant improvements should weight the mean.

FIGURE 5

Project size against precision, after review



Each bubble is one or more projects; the number inside is the count where more than two share a point. The four largest projects (Publix, The Bridge, Neurona, Prologis) sit at the right.

Thirteen contractors contributed, but not evenly: six projects come from one Maryland contractor and four from one Florida contractor. Estimator habit shows up in the results (every Publix panel priced at 250A, a vendor quote's unit counts).

The rules that pair items were written while reading Makeoff's exports as well as the estimators' documents. They were checked against a held-out calibration set at every change and no rule refers to a project, but a harness built by the vendor is not an independent instrument, and the paper's answer to that is to publish it: the code, the rule tables, their hash, the filed pairs and the decisions file, so anyone can re-run the comparison or change a rule and see what moves.

Finally, the paper measures the reference list, not the leveling outcome. It shows that the list Makeoff produces agrees with an estimator's closely enough to level against; it does not measure how many quote omissions contractors caught by using it, which would need a different study with vendor quotes and purchase records on both sides. Agreement on the gear list is also not the whole takeoff. Breakers and disconnects are reported separately and sit far lower, and feeders, wire, conduit, fixtures and devices are not measured here at all.

Reproducing the result

Everything behind the figures in this paper is a file. The harness is makeoff-bench 0.3.1, a Python package with the document readers, the normalization and pairing rules, the gate, the scorer and the report writer. Its rule tables are hashed, and the hash (3e98800f311c) is written into every results file; it has not changed since version 0.2.33, when the last rule was added, and it is the same for the raw run, the reviewed run and the calibration set. Forty-one regression tests pin the calibration projects, so a change to any rule that moves a calibration row fails the suite.

The filed pairs are the 34 Makeoff exports and the 34 estimator documents, named by project and unaltered except for the three handwritten transcriptions, which are filed beside their scans. The reviewer decisions are one CSV with the project, the row, the decision, the reviewer, the date and the drawing evidence. The changelog records every rule added during the batch, the pair that motivated it, and the calibration rows it moved.

To regenerate the paper's numbers: run the harness on the filed pairs for the raw figures, run it again with the decisions file for the reviewed figures, and run it on the calibration fixtures for the comparison. Each run writes a graded workbook with one sheet per project showing every pairing, its verdict and its note, and a results file with the per-project and aggregate figures. The per-project table in Appendix A is a direct reading of that workbook.

A reader who wants to test a different rule can change it, run the calibration set, and see which rows move before running the reported set. That is the same discipline the batch was run under, and it is the paper's answer to the question of whether a vendor-built harness can be trusted: the harness is not asked to be trusted, it is handed over.

What comes next

The second batch asks a different question of the same drawings: how much of the gear list a general-purpose chat model already produces. ChatGPT, Claude and Gemini will each be given the full drawing set and one fixed prompt in a single pass, as a person would use them today, and their replies filed and scored beside Makeoff's export on fresh pairs with the rules frozen at the hash above. Every system faces the same drawings, so a system that returns nothing usable for a project scores zero for it rather than being set aside. The protocol, including how each app reads drawing PDFs, is published with the harness before the first pair runs.

On the product side, the extraction gaps this batch exposed are the work: a structured circuit table per panel, disconnect symbols read off the power plans, surge devices added to the distribution definition, and the empty-first-pass failure traced in the pipeline. Each will be measured the same way when it lands, on the same pairs, against the same rules hash or a new one with its changelog entry.

Appendix A: the 32 scored pairs

Gear slice, type level. Gear units priced is the estimator's count after scope marks (the denominator of recall before review). Vendor-quote references are marked.

PAIR	PROJECT	REFERENCE	GEAR UNITS PRICED	HARNESS ALONE (R / P)	AFTER REVIEW
1.1	The International, Indianapolis IN	parts list	2	100% / 67%	100% / 67%
1.2	Baldwin HS Athletic Center, NY	Bluebeam report	14	100% / 100%	100% / 100%
1.3	Colony Way pool house, Jupiter FL	parts list	2	100% / 40%	100% / 67%
1.4	Sharon St residence, Hobe Sound FL	parts list	1	100% / 100%	100% / 100%
1.5	Murwood Elementary, Walnut Creek CA	parts list	2	100% / 100%	100% / 100%
1.6	Ouanalao Beach Homes, FL	priced item list	2	100% / 100%	100% / 100%
1.7	Ouanalao Bungalows, FL	priced item list	2	100% / 100%	100% / 100%
1.8v	Neurona Therapeutics, S. San Francisco CA (vendor quote)	vendor quote	30	90% / 96%	100% / 97%
1.9	BI Fremont freezers, CA	parts list	2	100% / 100%	100% / 100%
1.10	Fifth Third Bank, Neptune Beach FL	estimate detail	4	100% / 100%	100% / 100%
1.11	Charge N Go, Tucson AZ	parts list	3	100% / 100%	100% / 100%
1.12	Grantville Microtel, PA	material summary	14	86% / 80%	86% / 80%
1.13	Family Initiative, Punta Gorda FL	count sheet	2	50% / 33%	100% / 50%
1.14	Prologis Linglestown, PA	parts list	19	100% / 76%	100% / 76%
1.15	Popeyes, Ellicott City MD	parts list	6	100% / 86%	100% / 86%
1.16	Chick-fil-A, Catonsville MD	parts list	1	100% / 100%	100% / 100%
1.17	Panda Express, Linthicum MD	parts list	3	100% / 100%	100% / 100%
1.18	16 Handles	material order	2	100% / 100%	100% / 100%
1.19	Dorothy Ave community solar, MD	parts list	5	100% / 100%	100% / 100%
1.20	Aldi, Dillsburg PA	PDF parts list	3	100% / 43%	100% / 43%
1.21	Half Price Books, Austin TX	narrative	4	100% / 100%	100% / 100%
1.22	Salon Lofts, Gambrills MD	parts list	4	100% / 80%	100% / 80%
1.23	Publix #1814, Key West FL	parts list	30	87% / 79%	89% / 78%
1.24	Exton Carstar, PA	tag worksheet	12	100% / 100%	100% / 100%
1.25v	The Bridge, Mineola NY (vendor quote)	vendor quote	28	96% / 75%	100% / 78%
1.26	United Hope Medical, Indianapolis IN	parts list	2	100% / 100%	100% / 100%
1.28	Amazon Carmel, IN	parts list	1	100% / 100%	100% / 100%
1.33	Indy Hyundai, IN	parts list	2	100% / 67%	100% / 67%
1.34	Lovesac, Webster TX	material summary	2	100% / 100%	100% / 100%
1.35	Suvida Healthcare, Houston TX	material summary	3	100% / 100%	100% / 100%
1.36	Meadow Elementary Kitchen, Baldwin NY	QuickBooks estimate	2	100% / 100%	100% / 100%
1.37	MLK Community Center, Hagerstown MD	parts list	1	100% / 100%	100% / 100%

Appendix B: reviewer decisions

Reviewed by Will Zacher, 28 September 2026, against the filed exports. Each row is applied by the harness on top of the raw match and written into the graded workbook's note column. "Credited" counts the unit as found on both sides; "excluded" removes it from the denominator; "duplicate" removes a second read of an already-matched unit from the generated side; "stands" records the row as reviewed with no change.

PAIR	ROW	DECISION	EVIDENCE
1.13	second 200A panel enclosure	Credited	Riser text in the export: "NEW PANEL, 120/208 3PH., 200 AMP., 42 CIR." and the existing gear's UNIT 6 enclosure marked 200/200; both priced 200A units have counterparts in the extracted text.
1.23	three SPD2 units	Stands	No SPD on the riser or panel schedules in the export; surge devices are outside Makeoff's distribution definition.
1.23	two 200A transfer switches	Stands, no change needed	Both units (ATS-E, ATS-EM) already matched at the type level.
1.23	second generator	Excluded (one unit)	The 500 kW optional-standby generator is marked by others on the drawings; the quote's second unit is not this contractor's to buy.
1.23	500 kW standby generator (generated side)	Excluded	Same unit; leaves both sides of the score.
1.25v	panel K21	Credited	MSWBD schedule, section M2: "RETAIL PANEL K21, frame 800A, trip 800A, 3 poles"; the riser labels the 600A kitchen panel K1. Same panel, label transposed.
1.8v	transformers T-1 to T-4	Credited	One-line text names T-1 (112.5 kVA, feeds L1P1), T-2 (feeds DB2), T-3 (feeds DB3) and T-4 (feeds a Level 2 225A panel); the quote prices four. The harness emitted T-1 only; rule gap logged.
1.3	"90 AMP POOL PANEL" and "Proposed Electrical Panel (100 amp)"	Duplicates	The riser label and the schedule title are the same pool-house panel at two ratings; the estimate prices one, matched.
1.25v	"RESIDENTIAL SWITCHBOARD SWBD-RESI"	Duplicate	The SB2 switchboard the quote names SWBD-RESI--SB2 and the harness matched under that name, read a second time under the short label.

Not yet reviewed: panels 4C and AB on 1.12 Grantville Microtel (stand as misses); LPN1A on 1.33 Hyundai and the 60A sub panel on 1.26 United Hope (stand as generated-only units).

Appendix C: terms

Takeoff	The list of material and equipment an estimator pulls from the drawings. Not an estimate, which adds pricing and labour.
----------------	--

Switchgear package, gear	The distribution equipment as a group, from the service switchboard down to the branch panels, and in this paper the transformers, transfer switches, generators, UPS, surge devices and CT cabinets around them. Contractors say gear. The breakers inside each panel are part of what a vendor prices but are reported below the gear line here.
One-line, riser	The diagram that shows power flowing from the utility through the service equipment to each panel. The source of most of the gear list.
Panel schedule	The table for each panelboard listing its breakers by position, trip and poles. The source of the breaker count.
Recall, precision, agreement	Recall: the share of the estimator's units that Makeoff also listed. Precision: the share of Makeoff's units that the estimator also listed. Agreement: either, in headline copy; body copy names which.
Type level, type plus rating, type plus rating plus quantity	The three strictness levels at which a pair is scored. The headline is the type level; rating and quantity agreement are reported on their own.
Gear slice, all slice	The gear package alone, or the gear package plus breakers and disconnects. Every figure names its slice.
Scope mark	New, existing, demolished or by others, as the drawings state it. Units that are not new are carried but never counted against Makeoff when the estimator left them out.
Scoreable, completion rate	Whether a pair could be compared at all, and the share of attempted pairs that could.
Vendor-quote class	A project where the reference is the switchgear vendor's quote rather than an estimator's document. Two of the 32, marked throughout.
Calibration set	Seven projects used to settle the pairing rules, scored on the final rules and reported beside the result, never in it.
Harness, rules hash	Harness: makeoff-bench, the code that reads both documents, pairs items and scores them. Rules hash: a fingerprint of its rule tables, written into every results file.
Review, adjudication	An estimator's decisions on the disagreements, recorded with evidence and applied on top of the raw score. Both numbers are published.