EPiC Series in Built Environment

Volume XXX, 2026, Pages 1–10

Proceedings of Associated Schools of Construction 62nd Annual International Conference

Testing AI Accuracy in Quantity Takeoff: A Methodological Case Study in Commercial Construction

Hooman Sadeh [1], Dominick Geloso [1], and Dimitar Todorov[1]
[1]Utica University

This study empirically evaluated the accuracy of an AI-based quantity takeoff (QTO) platform, Togal AI, by comparing its automated measurements against contractor-produced takeoffs for a live commercial construction project. The analysis included 29 line items spanning exterior, floor, and ceiling finishes, as well as windows and doors, quantified in square feet (SF), each (EA), and linear feet (FT). Using non-parametric statistical methods—including the Wilcoxon signed-rank and Kruskal–Wallis tests—the study examined whether systematic deviations existed between AI- and contractor-generated quantities. It was found Togal AI's overall deviations were small but statistically significant, showing a consistent underestimation for area-based quantities while achieving near-perfect alignment for count-based items. Further analysis indicated that measurement accuracy varied significantly by finish type: ceiling finishes exhibited the highest consistency with contractor data, floor finishes demonstrated moderate agreement, and exterior finishes showed the greatest deviation. The results prove that AI performs best when quantifying repetitive, clearly bounded, and orthogonal elements but becomes less reliable when interpreting irregular geometries and complex façades. The study provides the first quantitative validation of Togal AI in professional practice and concludes that while AI can accelerate takeoff workflows, estimator oversight remains essential for accuracy assurance in complex building elements.

Keywords: AI-based quantity takeoff, computer vision, construction estimating, Togal AI

Introduction

Quantity takeoff (QTO) is a fundamental activity in the preconstruction phase, as errors in material quantities directly influence project budgets, procurement strategies, and downstream cost control. Despite increasing digitalization across the construction industry, many firms continue to rely on manual measurement and user interpretation, making QTO both time-consuming and susceptible to human error (Sands et al., 2025). Digital tools such as Bluebeam Revu have improved workflow efficiency by enabling on-screen measurement and markups; however, these systems still depend on estimators to interpret drawings, define boundaries, and perform repetitive tracing—particularly for complex finish systems (Zhao et al., 2025). Consequently, estimators often spend substantial time on measurement rather than higher-value tasks such as pricing analysis, constructability review, and risk assessment.

Recent developments in artificial intelligence (AI) have introduced new opportunities to automate estimating tasks. Studies have shown that machine learning models, neural networks, and hybrid fuzzy-based cost prediction systems can reduce estimator bias and enhance accuracy during early-stage project estimation (Elmousalami, 2020; García de Soto & Adey, 2016). A comprehensive review by Pan and Zhang (2020) identified automated quantification and measurement as an emerging subfield within AI applications for construction management. Among these innovations, platforms such as Togal AI utilize computer vision to automatically detect building components, classify elements, and extract measurements directly from digital drawings—minimizing the need for manual tracing (Zhao et al., 2025).

At the same time, newer studies are beginning to explore the role of generative AI in construction decision-making. Ghasemi and Dai (2024) found that generative AI models, such as GPT-4, can perform reasoning-related estimating tasks and produce reasonable cost outputs, but they caution that reliability and output consistency remain significant concerns. Similarly, broader digital transformation research indicates that adoption of advanced construction technologies is hindered by fragmented project data environments, digital maturity issues, and resistance to new processes (Abioye et al., 2021; Regona et al., 2022). Together, these studies demonstrate both the promise of AI and the ongoing skepticism among practitioners: while automation can accelerate estimating processes, construction firms remain reluctant to trust AI-generated outputs without objective performance validation.

Within academic settings, recent empirical work has evaluated the performance of Togal AI in construction education. Zhao et al. (2025) reported that Togal AI reduced takeoff time by 51.3%, improved measurement accuracy by 20.4%, and enhanced collaboration and student confidence in an undergraduate estimating course. Similarly, Sands et al. (2025) found that students using AI-based tools demonstrated higher digital proficiency and perceived efficiency compared to those using manual methods. However, these studies primarily focused on user experience and learning outcomes rather than verifying the accuracy of AI-generated quantities against professional contractor takeoffs. Moreover, Zhao et al. (2025) cautioned that automation may lead to "over-reliance," with students potentially bypassing critical evaluation when results are produced automatically.

Outside of construction education, AI research shows that the performance of generative or recognition-based automation depends heavily on the complexity and structure of building information. Ploennigs and Berger (2024) examined the use of diffusion-based generative AI for producing construction floor plans and found that the models frequently produced invalid or incomplete plan geometries unless they were trained using semantic encodings. Their results is an indication that generative AI tools struggle when architectural elements require interpretation of spatial relationships or boundary conditions, and they often lack semantic understanding of building logic. After introducing a semantic encoding strategy—explicitly labeling spatial elements in the input, the validity of generated plans improved dramatically (from approximately 6% to 90%), which shows the need for structured representations when AI interacts with architectural geometry. Although these findings relate to computer vision segmentation, they offer a theoretical foundation suggesting that Togal AI's measurement accuracy may vary depending on whether the element being quantified is horizontal, semi-repetitive, or visually complex.

Despite these advancements, a clear research gap remains. Prior studies have examined predictive cost modeling (Elmousalami, 2020; García de Soto & Adey, 2016), digital adoption barriers (Abioye et al., 2021; Regona et al., 2022), and educational applications of AI-based takeoff tools (Zhao et al., 2025; Sands et al., 2025). However, no published research has empirically validated whether an AI-based takeoff platform—such as Togal AI—produces quantity values that are statistically comparable to

those generated by professional estimators during live preconstruction. Furthermore, no studies have tested whether measurement deviations differ by finish type, even though computer vision theory suggests that geometric complexity directly affects extraction accuracy (Ploennigs & Berger, 2024).

Research Problem and Objective

Although AI-based takeoff platforms demonstrate strong potential for improving efficiency, their real-world accuracy and reliability remain unverified. Construction firms currently lack empirical evidence demonstrating whether AI-generated takeoffs can be trusted for bidding and procurement. Existing research focuses on either predictive cost modeling or user perceptions but not on quantitative fidelity relative to contractor measurements. Therefore, this study aims to statistically compare Togal AI's quantity takeoff results with contractor-produced takeoffs on a commercial building project and to determine whether systematic deviations exist across different measurement types and finish categories.

Research Methods

Case Study Project and Data Collection: This case study involves Thurston Hall at Utica University, located in Utica, New York, a commercial academic building constructed to support the university's Construction Management program. The project consists of a 16,000-square-foot, two-story facility including classrooms, instructional laboratories, and a multi-use auditorium, delivered on an accelerated schedule of approximately seven months and incorporating a modular design intended to accommodate future expansion. The general contractor provided detailed take off spreadsheets for exterior finishes, windows, floor finishes, base finishes, door schedules and ceiling finishes. Quantities were reported in units of square feet (SF), each (EA) and linear feet (FT). For example, exterior finishes included metal siding (A and B), glass/aluminum curtain walls, stone and brick; the window schedule listed nine window types; floor finishes consisted of epoxy concrete (PCON), carpet (CPT 1 & 2), vinyl composition tile (VCT 1), concrete floor and walk off mats; and ceiling finishes included acoustic ceiling tiles (ACT 1–4) and painted exposed ceilings. A total of 29 items were examined, consisting of 15 based on area, 13 based on count, and one measured linearly.

The authors used Togal AI to perform a QTO on the same architectural drawings. Togal AI's workflow involves uploading PDF plans, selecting relevant drawings and letting the software automatically detect and measure building elements based on computer vision algorithms. Users then review and correct AI generated mark ups before exporting quantities. For each item, we calculated the percentage deviation between Togal AI and contractor quantities as:

$$Deviation\ \% = \frac{(Togal - Contractor)}{Contractor} \times 100$$

Statistical Analysis: Because the sample sizes were small and we could not assume normality, non-parametric tests were used. The Wilcoxon signed rank test is a non-parametric alternative to the paired t test; it assesses whether the median of paired differences differs from zero and accounts for both the sign and magnitude of differences. We first performed a one sample Wilcoxon signed rank test on all 29 percent deviations to determine whether Togal AI systematically over or underestimates quantities. We then ran paired Wilcoxon signed rank tests separately for area based (SF) and count based (EA) items. Linear foot data (FT) were excluded due to the single observation.

To investigate whether accuracy varied by finish type, we performed a Kruskal–Wallis H test. The Kruskal–Wallis test is the non-parametric analogue of one way ANOVA and compares the medians of

three or more independent groups using ranked data. Here, percent deviations for exterior finishes, floor finishes and ceiling finishes were treated as independent samples. Post hoc pairwise comparisons were conducted using Bonferroni adjusted Mann–Whitney U tests to identify specific group differences. A significance level of $\alpha = 0.05$ was used throughout. Statistical analyses were performed in SPSS 30.

Results And Discussion

Table 1 below shows the actual quantities from the contractor versus those from Togal AI.

Table 1. Quantities Contractor vs. Togal AI

Category	Item	Unit	Contractor Qty	TogalAI Qty	Deviation %
Exterior Finishes	Metal Siding A	SF	4056.8	3988.12	-1.69
Exterior Finishes	Metal Siding B	SF	438	431.28	-1.53
Exterior Finishes	Glass/Alum	SF	1943.3	1963.93	1.06
Exterior Finishes	Stone	SF	519.5	589.1	13.40
Exterior Finishes	Brick	SF	3008.4	3095.82	2.91
Window Schedule	Type A Window	EA	10	10	0.00
Window Schedule	Type B Window	EA	5	5	0.00
Window Schedule	Type C Window	EA	4	4	0.00
Window Schedule	Type D Window	EA	6	6	0.00
Window Schedule	Type E Window	EA	3	3	0.00
Window Schedule	Type F Window	EA	2	2	0.00
Window Schedule	Type G Window	EA	1	1	0.00
Window Schedule	Type H Window	EA	4	4	0.00
Window Schedule	Type I Window	EA	1	1	0.00
Window Schedule	Type SF 0 Window	EA	10	10	0.00
Floor Finishes	PCON	SF	8087.6	7976.95	-1.37
Floor Finishes	CPT 1&2	SF	5399	5172.76	-4.19
Floor Finishes	VCT 1	SF	752.2	725.97	-3.49
Floor Finishes	CONC. FLR.	SF	293.7	269.97	-8.08
Floor Finishes	Walk Off Mat	SF	84.1	83.07	-1.22
Base Finish	RB 1	FT	12978	2699.32	-79.20
Door Schedule	FG	EA	12	12	0.00
Door Schedule	Type F	EA	22	20	-9.09
Door Schedule	Type G	EA	6	10	66.67
Ceiling Finish	ACT 1	SF	4764.9	4603.87	-3.38
Ceiling Finish	ACT 2	SF	7020.9	6635.38	-5.49
Ceiling Finish	ACT 3	SF	931.8	826.51	-11.30
Ceiling Finish	ACT 4	SF	621.1	561.5	-9.60
Ceiling Finish	Painted Exposed	SF	1602.2	1517.15	-5.31

One-Sample Wilcoxon Signed-Rank Test

A one-sample Wilcoxon signed-rank test was conducted to assess whether the median percentage deviation between Togal AI and contractor quantities differed from zero.
Results indicated a statistically significant difference ($Z = -1.98$, $p = .048$), suggesting that Togal.AI quantities tended to be slightly lower than contractor estimates overall. Although the median deviation was 0.0%, the significant test result reflects a systematic negative bias in rank distribution. This is shown in Table 2.

Paired Wilcoxon Signed-Rank Tests by Measurement Unit

As shown in Table 2, paired Wilcoxon signed-rank tests were performed by measurement unit to identify potential bias patterns. For square-footage (SF) quantities, results revealed a significant underestimation by Togal AI compared to contractor values ($Z = -2.22$, $p = .027$), with 12 of 15 items lower in Togal AI's output. For count-based (EA) quantities, no significant difference was found ($Z = -0.45$, $p = .655$), as 11 of 13 items were identical between both sources. Linear-foot (FT) quantities were excluded due to insufficient sample size ($n = 1$).

Jointly, both the one-sample and paired Wilcoxon tests indicate that while Togal AI's overall deviations are small, they are statistically biased toward underestimation, particularly for area-based (SF) quantities. Count-based (EA) items, by contrast, show near-perfect alignment with contractor quantities.

Kruskal Wallis Test

A Kruskal–Wallis H test was performed to examine whether the percentage deviation between Togal AI and contractor quantity takeoffs varied across three finish categories—Exterior Finishes, Floor Finishes, and Ceiling Finishes. The results showed a statistically significant difference among the groups, $H(2) = 8.420$, $p = .015$, which suggests that Togal AI's measurement deviations differ depending on the type of finish.

As shown in Table 2, Exterior Finishes had the highest mean rank (12.20), followed by Floor Finishes (7.80), while Ceiling Finishes had the lowest (4.00). Post-hoc pairwise comparisons using Bonferroni-adjusted Mann–Whitney tests revealed that the difference between Exterior and Ceiling Finishes was statistically significant ($p = .022$). The difference between Floor and Exterior Finishes was marginally significant ($p = .214$), whereas Ceiling and Floor Finishes showed no significant difference ($p = 1.000$).

Table 2. Statistical Analysis Contractor v. Togal AI

Test	Variable	Z/H	p-value	Direction/Mean Rank	Interpretation
One-Sample Wilcoxon	All Items	−1.982	0.048	Negative	Overall underestimation
Paired Wilcoxon	SF	−2.215	0.027	Togal.AI < Contractor	Significant
Paired Wilcoxon	EA	−0.447	0.655	None	Not significant
Kruskal–Wallis	Finish Type	8.42	0.015	Exterior > Ceiling	Significant difference by finish type

Post-hoc pairwise comparisons (using Bonferroni-adjusted Mann–Whitney tests) showed that the difference between Exterior Finishes and Ceiling Finishes was statistically significant ($p < .05$).
Exterior Finishes exhibited the highest percent deviations (mean rank = 12.20), while Ceiling Finishes had the lowest (mean rank = 4.00).
No significant difference was observed between Floor and Ceiling Finishes ($p > .05$).
These results suggest that Togal AI's takeoff accuracy is more consistent for horizontal elements (ceilings, floors) but diverges for vertical façade components, possibly due to geometric or visual complexity in wall delineation. Table 3 demonstrates the post-hoc analysis shown below.

Table 3. Post-Hoc Analysis

Finish Type	Mean Rank	Comparison	Test Statistic	Adjusted Sig.	Interpretation
Exterior Finishes	**12.20**	Exterior – Ceiling	15.200	**.022**	Significant
Floor Finishes	**7.80**	Floor – Exterior	11.000	.214	Marginal
Ceiling Finishes	**4.00**	Ceiling – Floor	4.200	1.000	Not significant

These results show that Togal AI generally overestimates quantities for exterior finishes, likely because of the greater geometric complexity and segmentation challenges associated with façades. In contrast, the tool demonstrates more consistent accuracy when measuring ceiling finishes. Floor finishes showed intermediate variability, suggesting a moderate level of agreement with contractor estimates. Overall, it can be interpreted that Togal AI's quantity extraction accuracy differs significantly across finish types—performing best on ceiling finishes, moderately on floors, and least accurately on exterior surfaces. This pattern points to the role of geometric complexity and visual obstructions in influencing the precision of AI-based takeoff measurements.

Togal AI versus Contractors' Takeoffs

Figures 1 and 2 show the exterior elevations quantity takeoffs by the contractor and Togal AI, respectively, Figures 3 and 4 show the reflected ceiling plans takeoffs by contractor and Togal AI.

Figure 1. Exterior Elevations Contractor's Takeoff in PlanSwift

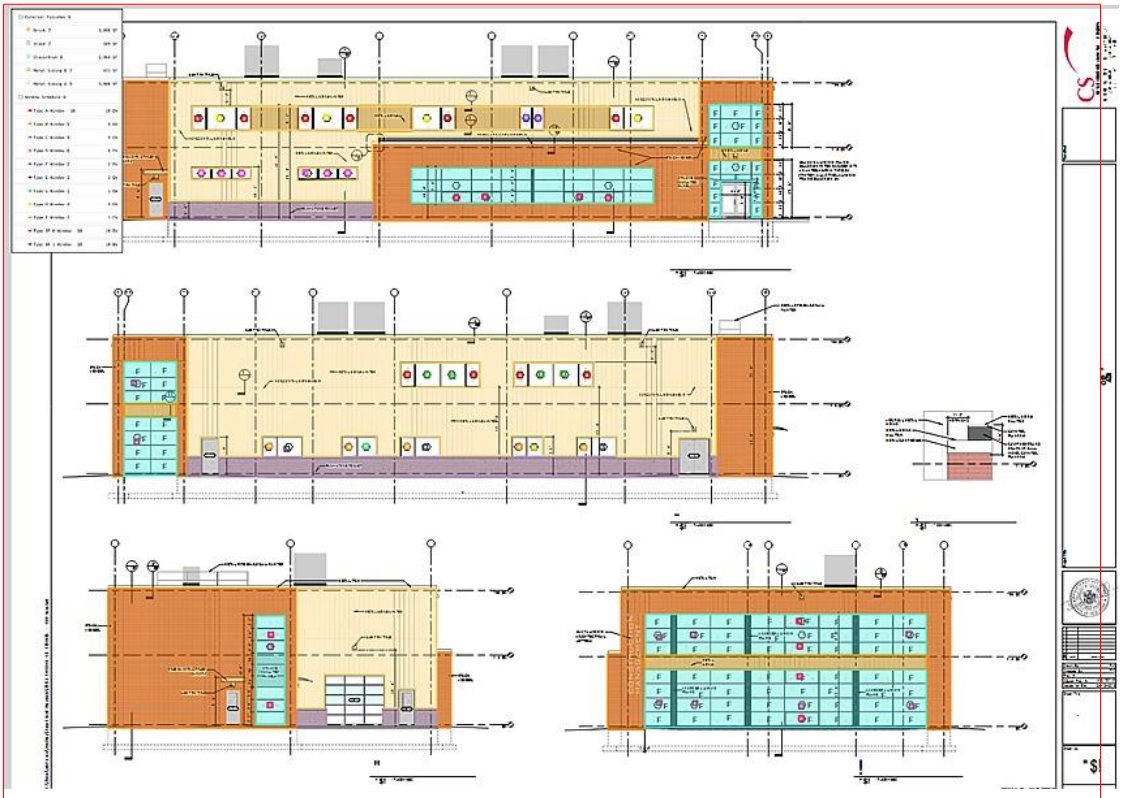

Figure 2. Exterior Elevations Togal AI Takeoff

Figure 3. Reflected Ceiling Plan Takeoff Contractor

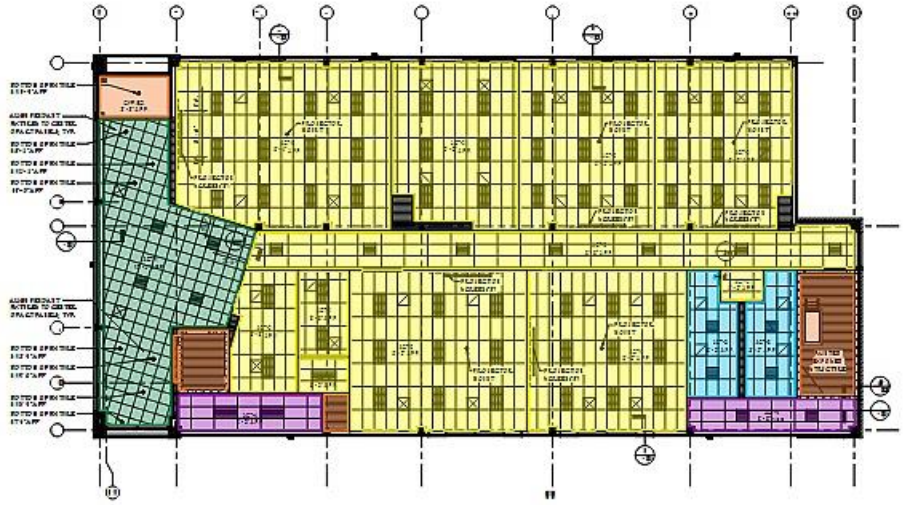

Figure 4. Reflected Ceiling Plan Takeoff Togal AI

Togal AI's performance differences between count-based and area-based items can be explained by how the software detects quantities. For discrete objects such as windows and doors, Togal AI uses symbol recognition and text detection, allowing it to match contractor counts exactly. In contrast, area-based measurements rely on computer vision segmentation, where the AI must determine polygon boundaries around complex shapes. When façade geometry includes offsets, material transitions, or irregular edges, boundary interpretation becomes more difficult and increases the likelihood of over- or underestimation. This explains the near-perfect performance on count-based quantities and the systematic underestimation observed in area-based finishes.

This study provides the first empirical validation of Togal AI's performance on a live commercial project by comparing AI-generated quantities to the contractor's actual bid takeoff. The findings show that Togal AI produces measurements that are very close to contractor quantities overall; however, deviations are not random. Instead, the AI demonstrates a statistically significant tendency to underestimate area-based quantities, while count-based quantities align almost perfectly.

This distinction points out how Togal AI interprets different types of geometry. Count-based items are discrete objects that computer vision can easily identify, leading to perfect or near-perfect accuracy—consistent with Zhao et al. (2025) and Sands et al. (2025), who reported that Togal AI improves efficiency and user confidence in classroom settings. However, unlike those studies, our results reveal limitations when applied to real construction documents. The Kruskal–Wallis test confirms that finish type significantly influences measurement accuracy. Togal AI was most accurate extracting ceiling finishes and least accurate for exterior finishes. This aligns with Ploennigs and Berger (2024), who showed that AI struggles when architectural components involve irregular geometry or fragmented boundaries. Exterior façades often include offsets, window penetrations, material transitions, and complex segmentation, requiring semantic interpretation—something current vision-based AI struggles to fully understand. Ceilings, however, are typically orthogonal planes with clear extents, making automatic detection easier.

Thus, this study contributes to the literature by demonstrating that the accuracy of AI-based takeoffs depends on the geometric complexity of the building element being quantified. Furthermore, the statistical underestimation detected in area-based quantities reinforces the ongoing concern that estimators remain reluctant to trust AI without validation (Abioye et al., 2021; Regona et al., 2022). Our results suggest a balanced perspective: (1) Togal AI is reliable enough to reduce labor for object-based takeoffs, (2) Human review is still required for area-based quantities, particularly on complex façades.These findings align with Ghasemi and Dai (2024), who caution that generative AI tools can produce confident—but not always correct—outputs. The present study confirms that AI can accelerate quantity takeoff but should be used to *augment*, not replace, estimator judgment.

Conclusion

This study assessed the performance of an AI-based quantity takeoff tool on a live commercial construction project and compared its output against the contractor's actual bid quantities. The results exhibit that the AI platform can generate takeoffs that are very close to contractor values, providing strong evidence that automated measurement can reduce the amount of manual tracing traditionally required during preconstruction. However, the statistical analysis conveyed that deviations are systematic rather than random. The AI consistently underestimated area-based quantities, whereas count-based items matched contractor quantities almost exactly. When takeoff accuracy was analyzed by finish type, ceiling finishes showed the highest alignment with contractor measurements, floor finishes showed moderate consistency, and exterior finishes produced the largest deviations. These patterns show that the AI performs best when extracting quantities from repetitive, clearly bounded, and orthogonal elements, and becomes less reliable when interpreting irregular geometries or visually complex façade conditions.

The results magnify the value and importance of AI in automating quantity takeoff but also reveal current boundaries of reliability. The tool is well suited for accelerating estimating workflows, especially for discrete, easy-to-identify components; however, estimator oversight remains essential for area-based measurements and façade-related finishes where segmentation challenges are more likely. In practice, the most effective use of AI in quantity takeoff may be a hybrid approach: allowing the AI to perform initial extraction, while the estimator reviews, validates, and adjusts measurements as necessary. By reducing time spent on measurement, estimators can shift focus to more strategic tasks such as pricing, bid leveling, and risk management.

The study has several limitations as follows. The analysis was based on a single commercial project and a limited number of quantity categories, which restricts the generalizability of the findings. Only one AI platform was evaluated, so performance differences across other AI tools or workflow configurations were not explored. Additionally, although the AI generated initial quantities automatically, users were still required to review and correct markups, meaning that the tool does not yet fully eliminate human involvement. The accuracy analysis was also limited to comparing quantities, not the time or cost savings achieved. Future research should examine multiple projects, additional finish categories, and performance across different AI takeoff platforms to determine consistency and generalizability across building types and design complexities.

References

Abioye, S. O., Oyedele, L. O., Akanbi, L., Ajayi, A., Delgado, J. M. D., Bilal, M., Akinade, O. O., & Ahmed, A. (2021). Artificial intelligence in the construction industry: A review of present

status, opportunities and future challenges. *Journal of Building Engineering*, *44*, 103299. https://doi.org/10.1016/j.jobe.2021.103299.

Elmousalami, H. H. (2019). Artificial Intelligence and Parametric Construction Cost Estimate Modeling: State-of-the-Art Review. *Journal of Construction Engineering and Management*, *146*(1). https://doi.org/10.1061/(asce)co.1943-7862.0001678.

García De Soto and B. T. Adey (2016). Preliminary Resource-based Estimates Combining Artificial Intelligence Approaches and Traditional Techniques. Procedia Engineering, vol. 164, pp. 261-268.

Ghasemi, A., & Dai, F. (2024). Can ChatGPT assist in cost analysis and bid pricing in construction estimating? A pilot study using a bridge rehabilitation project. *Smart Construction*. https://doi.org/10.55092/sc20240009.

Pan, Y., & Zhang, L. (2020). Roles of artificial intelligence in construction engineering and management: A critical review and future trends. *Automation in Construction*, *122*, 103517. https://doi.org/10.1016/j.autcon.2020.103517.

Ploennigs, J., & Berger, M. (2024). Automating computational design with generative AI. *Civil Engineering Design*, *6*(2), 41–52. https://doi.org/10.1002/cend.202400006.

Regona, M., Yigitcanlar, T., Xia, B., & Li, R. Y. M. (2022). Opportunities and adoption Challenges of AI in the construction industry: A PRISMA review. *Journal of Open Innovation Technology Market and Complexity*, *8*(1), 45. https://doi.org/10.3390/joitmc8010045.

Sands, K., Wang, X., & Zhao, T. (2025). Use of AI in a Graduate Construction Estimating Course. In *2025 ASEE Annual Conference & Exposition* (Construction Engineering Division: Best of Construction). https://doi.org/10.18260/1-2--57755

Zhao, T., Lin, X., & Na, R. (2025). Integrating AI in Construction Estimation Education: A Comparative Study of Togal AI and Bluebeam ReVu 20. *European Journal of Education*, *60*(4). https://doi.org/10.1111/ejed.70287