Academic Journal of Management and Social Sciences
ISSN: 2958-4396 | Vol. 18, No. 1, 2026

AI Adoption and Audit Quality in Public Survey and Firm Inspection Data

Yanjun Liu [1, a], Xiaohang Li [2, *]

[1] College of International Education, Hong Kong Baptist University, Hong Kong SAR, China
[2] School of Economics and Management, Beihua University, Jilin, Jilin, 132013, China

*** Corresponding author:** Xiaohang Li (Email: 15584040772@163.com), [a] 25644831@life.hkbu.edu.hk

Abstract: This study evaluates the association between AI-related questionnaire scores and audit-quality perceptions and examines what that association can establish about audit firms. A secondary analysis uses 98 respondent records and 25 items released by Handoko (2025). Equal-weight block scores are analysed using ordinary least squares, HC3 uncertainty estimates, rank correlations, alternative scoring, item-omission checks and row resampling. The AI–AQ correlation is 0.819. The AI coefficient is 0.735 in the unadjusted model and 0.731 after adjustment for four additional questionnaire blocks, with a 95% HC3 interval of 0.523–0.939. However, the AI–AQ heterotrait–monotrait ratio is 0.903, and several other block pairs exceed 1.00. Strong internal consistency therefore coexists with weak empirical separation. A separate descriptive analysis of UK Big Four inspection reports records 190 satisfactory or limited-improvement outcomes across 223 inspection instances in three cycles. These outcomes are not linked to the questionnaire respondents or to matched AI expenditure. The contribution is a reproducible assessment of the gap between observable association and an investment-effect claim. The evidence supports a positive association in the released questionnaire matrix, while neither its construct validity nor a causal return to firm-level AI expenditure is established.

Keywords: Artificial Intelligence; Audit Quality; Secondary Data; Measurement Validity; Audit Inspection; Reproducibility.

1. Introduction

AI can change the tasks auditors perform, the information they examine and the way they document judgements. Whether those changes improve audit quality is an empirical question. Expenditure on technology, use of a particular tool, confidence in its usefulness and the quality of a completed audit are related concepts, but they are observed at different stages of the audit process. Treating them as interchangeable can produce a precise numerical answer to an incorrectly specified question.

This study starts with data that can be inspected and reanalysed. Handoko (2025) publicly released a questionnaire workbook containing 98 rows and 25 coded items. According to its title, the deposit concerns AI adoption and sustainable audit quality in non-Big Four accounting firms. With the rows available, the association can be estimated directly; no spending figures have to be invented, and nothing has to be rebuilt from a published regression table. The limits are just as plain. The workbook has no firm identifiers and no investment variable.

Our main result is a strong positive relationship between the AI and AQ item blocks ($r = 0.819$), and it stays positive across the alternative calculations in Section 4.3. A second result complicates it. The blocks are not cleanly separated, and the background blocks that could serve as controls are highly correlated with one another. So the slope speaks first to this released score matrix. What it says about distinct organisational capabilities, or about audit performance that others could verify, is far less clear.

At the firm level we look separately at the public inspection reports of the Financial Reporting Council (FRC) for Deloitte, EY, KPMG and PwC. These record actual inspection outcomes. They cannot identify the survey respondents, however, and they contain no AI expenditure series common to the four firms. Because the two sources answer different questions, we compare them as measurement systems and never merge them into one sample.

What the paper adds is a reproducible assessment in three steps. We measure the AI–AQ association, check whether it holds when the calculation changes, and ask what evidence would be needed to read it as a firm-level investment effect. The last question is answered by documenting a measurement gap, with no treatment variable assumed, so the study does not compete with causal work built on matched organisational data. All source observations, calculations and numerical outputs accompany the manuscript. The analysis is exploratory and was not preregistered.

The research question depends on keeping adoption and investment apart. Investment commits resources to software, data infrastructure, training or new ways of organising work; adoption is about having access to a tool and using it. Audit quality is the outcome of professional work, and different observers may assess it differently. A survey can show how adoption is perceived without measuring the money behind it or the quality of finished engagements. That is why the title names only what can be observed.

The two sources fall short in opposite ways. The questionnaire has many respondent-level scores but thin documentation of sampling and measurement. The inspection reports have externally assessed firm outcomes but no matched AI exposure. No defensible link joins them, so neither can fill the other's gap. The contrast is still useful: it shows what is gained, and what stays unidentified, when the outcome shifts from perceived quality to a regulator's judgement.

2. Related Research and Interpretation

Archival auditing research separates the quality of assurance from the proxies used to observe it. DeFond and

Zhang (2014) review archival measures, and Rajgopal et al. (2021) discuss why audit quality is hard to measure at all. For this study the lesson is narrow: a measure must be read through the process that produced it. An inspection rating is a regulator's judgement on a selected audit; a questionnaire score is a coded answer. Neither covers every dimension of audit quality.

Fedyk et al. (2022) and Law and Shen (2025) examine AI within audit firms, each with data suited to their own setting. Their results cannot simply be carried over to a different dataset. We do not reproduce their firm samples, labour measures or identification strategies, so our conclusions stay within the questionnaire records and inspection aggregates used here.

Technology-acceptance research offers a narrower link. Venkatesh et al. (2003) organise acceptance around performance expectancy, effort expectancy, social influence and facilitating conditions, and the workbook uses matching prefixes (PE, EE, SI and FC) next to AI and AQ. We use these labels only as a provisional guide. They reveal nothing about item wording, coding direction or local adaptation of the scale, so all tables keep the original codes.

Why might AI and AQ scores move together? Several accounts are possible, and Section 2.3 sets them out. One point matters from the start: a single cross-section cannot tell them apart. Adjusting for other questionnaire scores collected at the same moment gives a conditional association, but it creates no time order and no exogenous change in adoption.

Measurement separation is particularly relevant when the explanatory and outcome variables come from one questionnaire. Henseler et al. (2015) propose the heterotrait–monotrait ratio, or HTMT, as a diagnostic of discriminant validity. This study uses that diagnostic alongside internal consistency and the observed correlations. The objective is not to certify a latent-variable model. It is to test whether a simple association interpretation remains credible when the empirical overlap between item blocks is made visible.

2.1. Technology Can Change Different Parts of the Audit Process

Four stages of the audit process can be distinguished. Resources make a tool available, and implementation decides whether staff can use it properly within the audit methodology. Task performance covers what information is processed and how findings are judged; assurance outcomes concern the support behind the final audit conclusion. We use this sequence as a conceptual device and estimate no mediation model. Its point is that a measure taken at one stage is not automatically evidence about the later ones.

AI tools might help retrieve documents, recognise patterns or pick out transactions that need a closer look. None of that resolves an accounting judgement or decides whether evidence is sufficient. An alert improves audit quality only when the engagement team follows it up and ties the result to the relevant assertion. Time saved is not quality gained, either, if it is absorbed by other tasks or takes attention away from the difficult judgements. The present questionnaire cannot identify which functions respondents had in mind.

The expected relationship is therefore conditional. Benefits may depend on the relevance of the underlying data, the reliability of tool outputs, staff understanding and the way results enter engagement review. Problems can arise when users treat an output as more authoritative than its evidence warrants, or when a tool identifies many anomalies that cannot be investigated adequately. These are analytical possibilities, not observed deficiencies in the surveyed firms. They explain why a general adoption score should be interpreted alongside information about use, oversight and the particular audit tasks affected.

Fedyk et al. (2022) and Law and Shen (2025) are relevant because they place AI within the organisation of audit work. Their research also highlights the importance of choosing data that correspond to the organisational claim being made. This secondary analysis does not equate a coded questionnaire mean with their measures or present agreement in coefficient direction as replication. A meaningful comparison concerns the unit of observation, the definition of technology exposure and the outcome process, before it concerns whether estimates have the same sign.

2.2. Perceptions, Professional Performance and Observable Audit Quality

A respondent's assessment can contain useful information. Auditors may observe changes in their own workflow that are not visible in published accounts or regulatory statistics. Perceptions are therefore not intrinsically uninformative. However, an assessment of usefulness can differ from independently demonstrated quality, and respondents may have unequal access to engagement outcomes. The released AQ codes do not resolve this distinction because their item wording is unavailable. For this reason we call it an AQ-related score and read no more professional meaning into it than the documentation allows.

External outcomes fix some measurement problems and create others. Because an inspection rating comes from an independent review, the person judging quality is not the person reporting adoption. The rating still covers only the audits chosen for inspection, within that review's scope, and says nothing about every audit or every dimension of assurance. Following DeFond and Zhang (2014) and Rajgopal et al. (2021), we read each audit-quality proxy through its own measurement setting.

Aggregation makes this concrete. An engagement-level finding, a respondent's perception and a firm's inspection proportion have different denominators, and adding up one does not produce another. A firm should not count for more in an investment analysis just because its staff returned more questionnaires, and a larger number of inspected engagements partly reflects how the regulator works. Without firm identifiers, even the first aggregation is impossible here. The unit of analysis remains the released row.

2.3. Competing Explanations for the Positive Association

Three accounts could produce the positive association. Under technology complementarity, AI use comes with training, suitable data and organisational backing, and respondents who have all of these score higher on both focal blocks. A common-evaluation account traces the covariance to a broad favourable view carried across related questions. Then there is selection: auditors in better-resourced settings may report more adoption and more confidence in quality even if the tool is not what makes the difference.

Podsakoff et al. (2003) show how a shared measurement method can shape relationships in behavioural research. Their review is a reason for caution about scores collected at one time with one questionnaire. It cannot tell us how large the method influence is in this release, or even which way it

pushes. The present overlap diagnostics are consistent with several explanations. They cannot isolate an unobserved response tendency from substantively related constructs. Accordingly, no claim is made that common-method bias has been proved, quantified or removed.

These explanations need not be mutually exclusive. A real improvement in work could coexist with favourable response tendencies, and capable organisations could be more likely to adopt useful technology. The empirical task is therefore to establish what remains observable without choosing a preferred causal story. So we report the raw association, then what survives adjustment for the other scores, and finally how far the blocks can be told apart empirically. These steps are not a chain of confirmed hypotheses about investment, implementation and audit performance.

3. Data and Methods

3.1. Public Data and Provenance

The primary dataset is version 1 of Handoko's (2025) Zenodo release (DOI 10.5281/zenodo.15034566), published on 16 March 2025 under CC BY 4.0. We use the original file, Auditor Using AI.xlsx, exactly as deposited; its MD5 checksum (5469242e0e21f4c6cf6986da6c41b01f) matches the repository listing. Liu Yanjun did not collect these responses. Credit for the original collection stays with the dataset creator.

The workbook has a single sheet with 98 data rows and 25 numeric columns. Each observed cell holds an integer between 1 and 5. We found no missing cells and no exact duplicate rows, and no record was removed, imputed, winsorised or replicated. Sequential row numbers were added purely to trace calculations. They are not respondent identifiers from the original collection.

The deposit describes the file as questionnaire results from respondents. It includes no item-level codebook, respondent demographics, firm identifiers, collection dates or sampling frame. A matching checksum and clean numeric entries help reproducibility, and a public repository shows where a file came from. None of this verifies individual responses, recruitment or representativeness. We do not infer the country or the number of firms from the author's affiliation, nor do we assume the survey was run where the conference was held.

Table 1 shows how the coded blocks enter the calculations. No reverse-coding key was deposited, so the primary analysis averages items in the direction in which they were released. This is our operational assumption. It does not mean the original instrument has been validated, and every coefficient below depends on it. The supplementary package keeps the original workbook, so a later rescoring based on verified item documentation would need no change to the source data.

Table 1. Item blocks and operational definitions

Code	Deposited columns	Operational role
AI	AI.1–AI.4	Mean score; AI-related predictor, not monetary investment
AQ	AQ.1–AQ.6	Mean score; questionnaire outcome, not inspection quality
PE	PE.1–PE.4	Additional coded block; provisionally performance expectancy
EE	EE.1–EE.4	Additional coded block; provisionally effort expectancy
SI	SI.1–SI.3	Additional coded block; provisionally social influence
FC	FC.1–FC.4	Additional coded block; provisionally facilitating conditions

Source: Handoko (2025), original column names. Construct expansions are provisional; item wording and response anchors are not supplied in the deposit. All items are averaged as coded, with equal weight within each block.

Block means keep the common 1–5 response range and weight each item in a block equally. They are easy to reproduce and avoid data-dependent weights that could maximise fit in such a small release. Equal weighting cannot make up for unknown item content, though: an item that needed reverse scoring would still contribute the wrong information. The alternative scores test numerical sensitivity under the available coding. They cannot show that an undocumented instrument is correct.

No cases are excluded for unusual score patterns. Without response times, attention checks or the original instructions, deleting rows after the fact could remove genuine answers that merely look unexpected. All rows stay in, and the influence analyses show how much single records move the focal result.

3.2. Estimation and Diagnostics

Four nested ordinary least squares (OLS) models take the AQ mean as the outcome. Model 1 has an intercept and AI only. Model 2 adds PE, Model 3 adds PE and EE, and Model 4 includes PE, EE, SI and FC. Nesting the models in this way shows how the AI slope moves as overlapping score information comes in. Model 4 is the fullest model, but we do not call it causal: no covariate is known to precede adoption, and there is neither an instrumental variable nor a treatment date.

Coefficients are reported unstandardised. A value of 0.70, for example, means that a one-point difference in the AI mean goes with a 0.70-point difference in the AQ mean, given the other scores in the model. HC3 covariance estimates adjust the squared residual contribution by the squared complement of each observation's leverage. Two-sided p-values and 95% intervals use the residual degrees of freedom, with 92 degrees of freedom in Model 4. These are approximate inferential summaries, following the finite-sample motivation of MacKinnon and White (1985).

The uncertainty estimates assume independent rows. That assumption cannot be verified because the release omits firm identifiers. If respondents are clustered within firms, row-based HC3 intervals and row-resampling intervals may understate uncertainty. The absence of identifiers also rules out firm fixed effects, clustered covariance estimates and a panel analysis. These limitations are not repaired by the sample containing many questionnaire items.

Internal consistency is calculated with Cronbach's alpha using the item covariance matrix. HTMT is calculated as the mean absolute cross-block item correlation divided by the geometric mean of the two within-block mean absolute correlations. The diagonal is omitted from within-block means. No HTMT entry is capped at 1.00, because the statistic can exceed one. The values are diagnostic rather than an automatic pass–fail certification.

Sensitivity analysis covers Spearman correlation, median instead of mean block scores, all 98 leave-one-row-out fits of Model 4, and ten item-omission fits that exclude one AI or AQ item at a time. All 16 subsets of the four additional blocks are also estimated, always retaining AI and the intercept. Finally, 2,000 paired row resamples with replacement produce a percentile interval for the Model 4 AI slope. We fixed the random seed at 20260928. Every one of these checks measures stability inside the released matrix; unobserved confounding and undocumented coding errors are beyond their reach.

We also ask how much AI adds beyond the other four block means. A reduced regression contains PE, EE, SI and FC; the full regression adds AI. The gain in R-squared is the difference between their in-sample fits, and partial R-squared divides that gain by the variation the reduced model left unexplained. Neither figure assigns a share of real-world audit quality to technology.

Residualisation provides a second check. AI and AQ are each regressed on the four other blocks plus an intercept, and the slope between the two sets of residuals equals the Model 4 AI coefficient, an arithmetic confirmation of the fitted result. It does not make AI exogenous. Only the linear variation shared with the included scores is removed; omitted influences and measurement overlap remain.

We did not fit a structural equation model or a causal mediation chain, and we did not hunt for a statistically favourable subset to use as the main specification. Each would need assumptions about constructs and causal order that the release cannot support. Instead, all 16 control-subset estimates are disclosed, so no favourable combination can be presented as a preselected test of theory.

4. Questionnaire Results

4.1. Distribution and Score Association

Table 2 gives descriptive statistics for the full sample. The AI mean is 3.217 (SD 1.187) and the AQ mean is 3.228 (SD 1.066). Since the original response anchors were not deposited, a mean above three is described only as a number. We do not read it as agreement with any particular statement.

Table 2. Descriptive statistics and internal consistency

Block	Items	Mean	SD	Minimum	Maximum	Alpha
PE	4	3.130	1.218	1.250	4.750	0.918
EE	4	3.222	1.175	1.250	5.000	0.894
SI	3	3.238	1.188	1.333	5.000	0.874
FC	4	3.117	1.213	1.250	4.750	0.915
AI	4	3.217	1.187	1.250	5.000	0.913
AQ	6	3.228	1.066	1.333	4.667	0.901

N = 98 for every block. SD uses the sample denominator N − 1. Alpha is calculated from released item covariances. High alpha does not establish content validity, discriminant validity or objective audit quality

Correlations are strong across several blocks (Table 3). For AI and AQ the Pearson correlation is 0.819 and the Spearman correlation is 0.777. Figure 1 plots the underlying pairs as a distribution of derived scores, not 98 independently verified firms. Each average comes from a few integer responses, so coordinates repeat; marker area grows with frequency, and no jitter is added because it would suggest false precision. The line is the fitted Model 1 relationship, and the spread around it shows that the two blocks are not numerically identical.

Table 3. Pearson correlations between block means

Block	PE	EE	SI	FC	AI	AQ
PE	1.000	0.905	0.813	0.823	0.838	0.697
EE	0.905	1.000	0.910	0.920	0.814	0.752
SI	0.813	0.910	1.000	0.938	0.673	0.722
FC	0.823	0.920	0.938	1.000	0.707	0.752
AI	0.838	0.814	0.673	0.707	1.000	0.819
AQ	0.697	0.752	0.722	0.752	0.819	1.000

N = 98. These are correlations between derived block means, not latent-factor correlations. Source: author calculations from Handoko (2025).

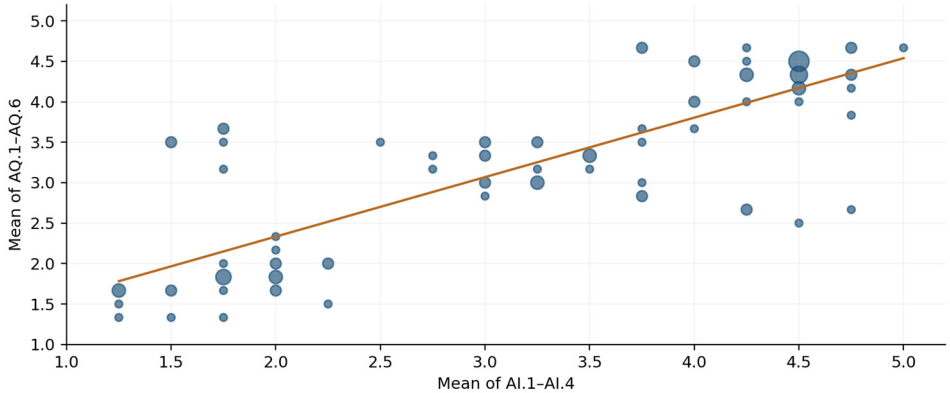

Figure 1. Observed AI and AQ block means

Each marker represents an observed score combination; area is proportional to its frequency. The line is Model 1. Both axes retain the original 1–5 item-score scale.

The strong correlations are not confined to AI and AQ. SI and FC correlate at 0.938, PE and EE at 0.905, and EE and FC at 0.920. Knowing where a respondent sits on one block often tells you a good deal about another. That pattern fits overlapping constructs, a broad shared disposition, or some mix of substantive and response-method influences. Which of these dominates, the released data cannot say.

All six alpha coefficients are above 0.87, so the items within each block covary strongly under the chosen scoring rule. Sijtsma (2009) warns against treating alpha as a general certificate of reliability or unidimensionality, and high alpha does not cancel out the high correlations between blocks. A scale can be internally consistent and still share much of its variation with another, which matters once several blocks enter the same regression. Whether each block measures one distinct construct, or matches independently assessed audit quality, is a question alpha cannot answer.

The rank correlation of 0.777 tells a similar story: respondents with higher AI means generally sit higher in the AQ distribution. The gap from the Pearson value does not by itself signal nonlinearity or an outlier, since ranking changes the scale and creates ties. Agreement on direction helps, up to a point. Both statistics come from the same respondents and share the same documentation limits.

4.2. Regression Estimates

Table 4. Ordinary least squares estimates for the AQ mean

Predictor	Model 1	Model 2	Model 3	Model 4
Intercept	0.863 (0.211)	0.851 (0.219)	0.699 (0.211)	0.556 (0.206)
AI	0.735 (0.059)	0.707 (0.085)	0.631 (0.106)	0.731 (0.105)
PE	—	0.033 (0.086)	-0.268 (0.163)	-0.287 (0.110)
EE	—	—	0.415 (0.165)	-0.218 (0.145)
SI	—	—	—	0.245 (0.109)
FC	—	—	—	0.362 (0.096)
N	98	98	98	98
R^2	0.670	0.671	0.706	0.764
Adjusted R^2	0.667	0.664	0.697	0.751
AI p-value	<0.001	<0.001	<0.001	<0.001

Unstandardised coefficients; HC3 standard errors in parentheses. All models include an intercept and retain the same 98 rows. No significance stars are used. Model 4 AI 95% interval: 0.523–0.939. See Section 3.2 for independence and scoring assumptions.

Table 4 reports the estimates. Without adjustment, the AI slope is 0.735, with an HC3 standard error of 0.059 and an interval of 0.619–0.852, and Model 1 accounts for 67.0% of the sample variance in the AQ mean. We treat that fit as descriptive. Part of it may simply reflect the fact that both scores summarise related answers given by the same people on the same instrument.

Adding PE moves the AI estimate to 0.707. With PE and EE both included it drops to 0.631, and adding SI and FC brings it back up to 0.731, so the adjustment path is not monotonic. In Model 4 the interval stays positive (0.523–0.939) and the adjusted R-squared is 0.751. On the fitted scale, a one-point difference in AI goes with roughly three-quarters of a point in AQ when the other block means are held constant. That is a difference in scores, not a percentage gain in actual audit quality.

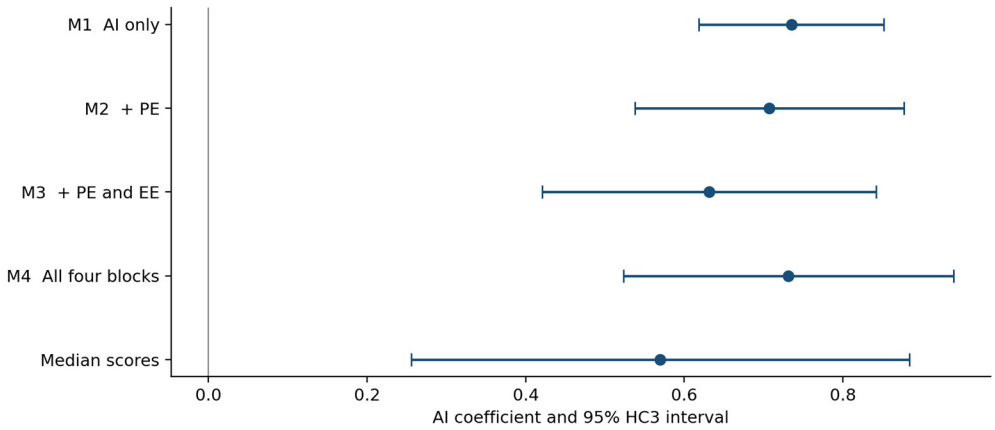

Figure 2. AI coefficient across scoring and adjustment choices

The background coefficients need particular care. PE is slightly positive in Model 2 but negative in Models 3 and 4, while EE is positive in Model 3 and negative in Model 4. Sign changes like these are what suppression and overlapping

predictor information tend to produce. They are no evidence that performance expectations or ease of use harm audit quality. The Model 4 variance inflation factors make the point: 3.86 for AI, 6.70 for PE, 14.37 for EE, 9.95 for SI and 10.62 for FC. Each background coefficient rests on small differences between strongly correlated scores.

Figure 2 shows the focal coefficient's uncertainty across the nested specifications and under median scoring. The intervals are not adjusted for specification search, because the figure discloses sensitivity and is not used to pick an estimate. The direction of association holds throughout, within the limits of the measurement and sampling documentation.

Points and intervals use actual fits. The first four models use mean scores; the last uses medians for every block. All intervals use HC3 covariance estimates and the relevant residual degrees of freedom. Coefficients across scoring rules are not identical estimands.

4.3. Incremental Information and the Meaning of the Adjusted Coefficient

With PE, EE, SI and FC alone, the reduced model has an R-squared of 0.592. Once AI is added the value reaches 0.764, a gain of 0.172. Expressed relative to the variation left unexplained by the reduced model, the partial R-squared is 0.422. Thus AI adds substantial linear information about AQ within this matrix, even after the other four block means are included. This result is more specific than reporting the full model's fit, because it distinguishes the focal block's incremental contribution from the variation already represented by the background scores.

The residualised AI–AQ correlation is 0.649. It remains positive after removing the linear relationships with the four other blocks. This finding limits one simple explanation of the original correlation: it is not entirely reproduced by linear adjustment for those available scores. Omitted characteristics, nonlinear relationships, shared wording and common response tendencies all remain possible explanations. In particular, the blocks included as controls are not independently documented pre-adoption characteristics. Including them is no proof that the relevant confounding is gone.

Standardising AI and AQ gives a Model 4 coefficient of about 0.815, so one standard deviation more on AI goes with roughly 0.82 standard deviations more on AQ, holding the included scores constant. This describes size within the observed distributions only. The standard deviation here is the spread of the AQ mean in this file, not a unit of inspection performance, financial-statement reliability or monetary return.

The residual standard deviation of AI after adjustment is approximately 0.604, compared with 1.187 before adjustment. Much of its original dispersion is shared with the other predictors. This makes the interpretation of a conditional one-point comparison more demanding than the original score range suggests: the model compares respondents who differ in AI while having the same values on several highly related dimensions. Such comparisons are mathematically available, but they need not correspond to a practical intervention that changes adoption while leaving every related perception unchanged.

The numerical persistence of the AI coefficient should therefore be separated from the instability of the background coefficients. The latter reveal how overlapping information is redistributed across predictors as the model changes. Their sign reversals do not invalidate the arithmetic of the focal estimate, but they caution against interpreting each coefficient as an independent behavioural mechanism. The full model is a description of conditional score relationships. It is not evidence that the six coded blocks form a well-separated causal system.

4.4. Sensitivity and Measurement Separation

Table 5. Sensitivity calculations

Calculation	Result	Interpretation
Spearman AI–AQ	0.777	Positive rank association
Median-scored Model 4	0.570 [0.256, 0.884]	HC3 interval; different scoring rule
Paired row bootstrap	[0.388, 0.870]	95% percentile interval; 2,000 resamples
Leave one row out	0.678–0.758	Range of AI slopes over 98 fits
Omit one AI or AQ item	0.586–0.747	Range over ten Model 4 fits
All control subsets	0.513–0.735	Range over all 16 adjustment sets

Ranges describe specification or observation sensitivity, not confidence intervals. Row deletion is diagnostic only; the reported main estimates retain all records. Full fit results and the random seed are in the reproducibility package.

Table 5 summarises the sensitivity calculations. The median-score estimate is 0.570, and the paired bootstrap interval is 0.388–0.870. Deleting any single row leaves the AI slope between 0.678 and 0.758. These results reduce concern that the association depends on one observation or one averaging convention. They do not test whether missing firm clusters, response styles or item wording account for the association.

The measurement diagnostic points to a more consequential limitation. As Figure 3 and Table 6 show, AI–AQ HTMT is 0.903. PE–EE is almost 1.00, while EE–SI, EE–FC and SI–FC exceed 1.00. Values this high raise concerns about the empirical distinction between the item blocks. The study therefore reports a score association and does not interpret Model 4 as a validated structural model among six distinct latent constructs.

Table 6. Selected measurement and collinearity diagnostics

Diagnostic	Value	Consequence
AI–AQ HTMT	0.903	Focal predictor and outcome overlap
PE–EE HTMT	0.999	Very weak separation of these blocks
EE–SI HTMT	1.025	Cross-block correlation exceeds within-block benchmark
EE–FC HTMT	1.014	Separation concern
SI–FC HTMT	1.048	Separation concern
Largest predictor VIF	14.368	Model 4 EE coefficient is sensitive to overlap

HTMT is calculated from absolute item correlations and can exceed one. VIF is calculated from auxiliary regressions among Model 4 predictors. These diagnostics do not establish fabrication or identify a specific response bias.

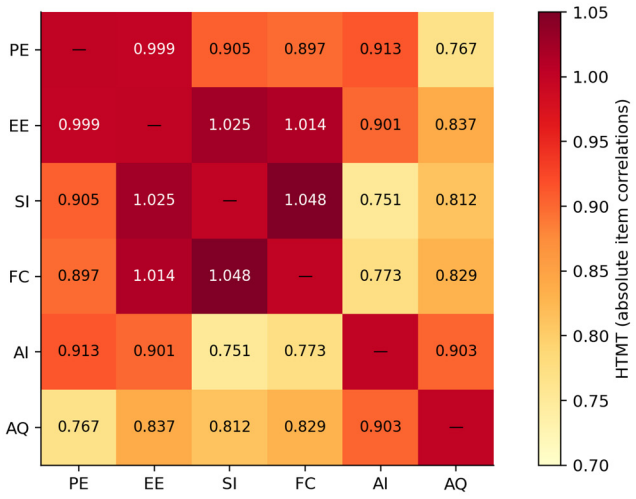

Figure 3. Full matrix of item block separation diagnostics

HTMT values are calculated without truncation. Diagonal entries are not interpreted. High internal consistency and high cross-block overlap can coexist; the original questionnaire wording is required for a complete construct assessment.

The sensitivity exercises address different forms of fragility. Leave-one-row-out estimation tests whether one record drives the AI slope, and item omission tests whether any single focal item is essential. Median scoring summarises the responses differently, while the control subsets vary which blocks are held constant. The checks agree. That gives the positive relationship some numerical stability, but they are all checks on one dataset.

The bootstrap interval is wider than the HC3 interval and less symmetric around the estimate, and we report both. Resampling changes the joint score distribution in each fit, while HC3 adjusts uncertainty around the observed design matrix. Neither handles undocumented firm clustering. Both ranges are positive, supporting the direction of the association under their assumptions, but no single interval captures all of the uncertainty.

More regressions cannot settle the problem that HTMT raises. The AI–AQ value of 0.903 is close to a commonly discussed diagnostic boundary, and several other pairs are higher still. A threshold is no replacement for examining the items, and without a codebook we cannot tell whether the overlap comes from similar wording, broad construct definitions or genuine conceptual closeness. Henseler et al. (2015) motivate the diagnostic; they do not license treating an unexplained matrix as a validated measurement model.

No post hoc correction is used to make the blocks look distinct. Dropping correlated items until HTMT falls, without knowing their content, could quietly change what a score means, and merging all blocks into one index would abandon the research question. The deposited structure stays and the interpretation is narrowed to an association between two coded response summaries, with incremental predictive content but unresolved construct separation.

5. Firm-level Inspection Evidence

For the firm-level evidence we transcribed a separate dataset from the FRC's 2025 inspection and supervision reports on Deloitte LLP, Ernst & Young LLP, KPMG LLP and PricewaterhouseCoopers LLP (FRC, 2025a–d). It holds the published, rounded percentages for five cycles and exact counts of inspected audits for the three most recent. An inspection instance is an audit examined during a reporting cycle; repeated inspection of an entity across years is not ruled out. For that reason we do not describe the counts as unique clients. Figure 4 plots the published five-cycle rate series.

Table 7 gives the count-based results. In 2022/23, 58 of 73 inspection instances were rated good or as requiring limited improvements. The figures were 59 of 70 in 2023/24 and 73 of 80 in 2024/25, giving pooled rates of 79.45%, 84.29% and 91.25%. Over the three cycles this comes to 190 out of 223. Pooling weights each firm by its number of inspected audits; an average of the four rounded published percentages would answer a different question.

Table 7. Actual firm inspection counts

Firm	2022/23	2023/24	2024/25
Deloitte	14/17 (82.35%)	16/17 (94.12%)	19/20 (95.00%)
EY	16/20 (80.00%)	13/17 (76.47%)	18/20 (90.00%)
KPMG	14/19 (73.68%)	17/19 (89.47%)	18/20 (90.00%)
PwC	14/17 (82.35%)	13/17 (76.47%)	18/20 (90.00%)
Pooled	58/73 (79.45%)	59/70 (84.29%)	73/80 (91.25%)

Numerator: good or limited improvements; denominator: inspected audits. Source: FRC (2025a–d), report summaries and underlying inspection-count disclosures. This table is separate from the 98-row questionnaire dataset.

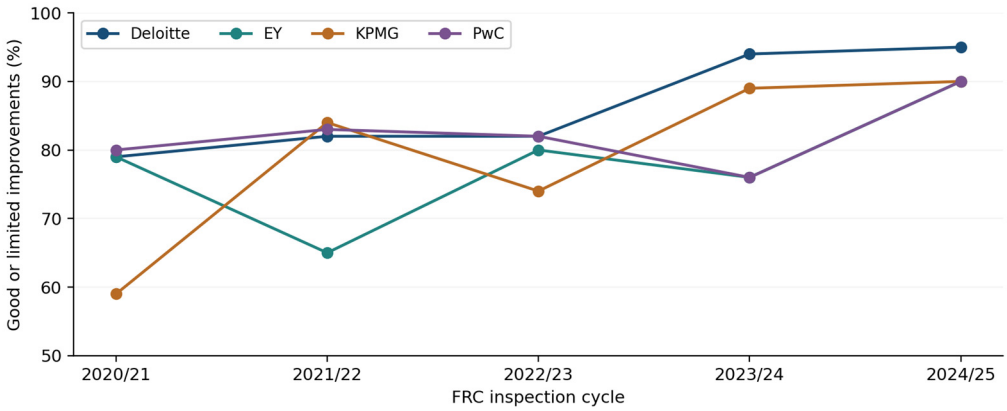

Figure 4. Published firm inspection rates across five cycles

Source: FRC (2025a–d), page 4 of each report. These are published rounded percentages. Exact pooled values in Table 7 are calculated from counts and are not obtained by averaging this figure.

Cycle labels need care as well. They do not necessarily match the year in which a firm invested in AI or used a tool. The 2024/25 review mainly concerns audits completed earlier, so a 2025 announcement cannot be assigned as exposure for those engagements. Selection is also risk-based, which means the rate describes the inspected set and is not an unbiased estimate for all of a firm's audits.

The rising inspection rates are an observed pattern in outcomes. They are not an estimated AI effect, since this dataset has no matched expenditure variable. Nor can the questionnaire supply one: its respondents are not identified as employees of the inspected UK entities, and its scores do not measure monetary spending. Joining the two sources by year, broad firm category or an arbitrary average would invent a link between exposure and outcome.

Investment announcements cannot fill this gap either. A global multi-year technology commitment differs from annual UK audit expenditure, and a product deployment count differs from money spent. Even an accurately quoted announcement may be an invalid regressor if its organisational perimeter, timing or technology scope differs from the outcome. The present manuscript therefore makes no firm-level investment regression and does not code missing expenditure as zero.

5.1. What Changes in Inspection Proportions Establish

The count-based rates show a rise of 11.80 percentage points between the first and last of the three recent cycles. This is a descriptive difference between inspected sets. It is not the improvement of a fixed cohort of engagements followed under a constant intervention. The denominator changes from 73 to 80, and the reports do not make the questionnaire or an AI investment series available for those cases. A higher proportion can reflect changes in underlying audit work, the engagements selected, the severity of issues encountered, or several influences together.

In the final cycle, each of the four firms has 20 inspected audits in the extracted counts. Deloitte has 19 outcomes rated satisfactory or limited improvements and the other three firms have 18 each, giving the pooled 73 of 80. Earlier cycles have unequal denominators, so pooling engagement counts and averaging firm percentages diverge more there. When selection probabilities are unknown, neither replaces a population-weighted estimate.

The historical percentages in Figure 4 add temporal context, but they are rounded published values. We do not expand them into invented counts, and a one-point change in that series is not read as a fixed number of extra satisfactory audits. Exact count-based statements are limited to the three cycles with extracted numerators and denominators.

5.2. Why the Sources Remain Separate

The two sources differ in too many ways to be combined. One holds coded assessments from respondents described as working in non-Big Four firms; the other holds regulatory evaluations of selected engagements at four Big Four UK entities, aggregated by firm and review cycle. With no shared identifier, jurisdiction or exposure measure, a combined regression could not be defended.

The comparison still shows what the stronger claim would require: spending, or another clearly defined exposure, assigned to the same organisational unit and period as the outcome. Public announcements tend to bundle technologies, service lines or years, and inspected audits may predate them. So the inspection trend is reported without attribution to AI, and the questionnaire is used only for the association it contains.

6. Discussion

The questionnaire association holds up through several changes in calculation, and that is a substantive finding about the released data. Its stability rules out some narrow explanations, such as reliance on one record or one focal item. It cannot tell us which behavioural process produced the covariance, and more than one could have.

Measurement is the most direct implication. Internal consistency above 0.90 for AI and AQ might look reassuring on its own; the HTMT result shows why it is not enough. If two blocks partly capture the same favourable view of technology, their association could stay stable across many regressions without reflecting any effect of one distinct construct on another. Extra controls do not necessarily help when they overlap strongly as well, as the sign reversals in Table 4 show.

For research with public data, the paper offers a chain from source observations to interpretation that others can audit. A published dataset can support new secondary analysis if its

creator is credited and its actual variables are respected. Its openness does not justify supplying missing demographics, inferring a questionnaire wording, or relabelling respondents as firms.

6.1. Interpreting a Strong Result without Overstating its Reach

Two findings must be read together. The AI-related score is strongly associated with the AQ-related score, and the association carries information beyond the four other block means. Yet the item blocks overlap heavily and their content is not fully documented. The second finding qualifies the first without erasing it, which leaves open whether the scores capture distinct aspects of adoption and audit performance.

Reproducibility is not replication. With the accompanying code, another analyst can recover our means, coefficients, intervals and diagnostics from the same workbook. Replicating the relationship itself would need new evidence from a documented measurement process, ideally with comparable constructs and a separately observed outcome. Reproducing the coefficient shows that the arithmetic can be checked. It does not validate the fieldwork or the causal reading.

Our contribution to the literature is to make an accessible evidence gap measurable. In the score matrix, a large association sits alongside weak separation; in the regulatory data, an external outcome appears with no treatment measure to match it. Fixing one obstacle would leave the other in place, which is why adding more public information does not always strengthen identification.

6.2. Implications for Audit-firm Evaluation and Governance

For audit-firm management, the results argue for keeping resource commitments, staff assessments of implementation and completed-work outcomes apart. A favourable staff score may reflect acceptance, confidence or perceived usefulness, and it can inform organisational learning. Our coefficients do not convert it into an expected inspection improvement. The incremental fit of AI in Model 4 is especially easy to misread: 17.2 additional percentage points of explained questionnaire variation are not a 17.2-percentage-point rise in the probability of a satisfactory audit.

A meaningful evaluation would connect technology use to the decisions and evidence within an engagement. Were outputs reviewed? Were important exceptions investigated, and were conclusions supported? These questions follow from the gap between task assistance and assurance quality; they are not findings about any firm in the data. We rank no vendors, spending levels or implementation programmes. The practical message is that evaluation units must correspond to the outcome being claimed.

The inspection data raise a parallel problem. Independent assessment is valuable because it comes from a different reporting process, but a better outcome measure does not create treatment variation. A firm's published rate may rise for many reasons, including staffing, engagement selection, remediation and quality management, all of which can change alongside technology. Without linked exposure information, crediting the rise to AI would swap an unknown relationship for a preferred explanation. The evidence supports tracking these indicators separately, with no return-on-investment claim.

6.3. Transparency in Secondary Empirical Work

The public release made this analysis possible without proprietary firm data, and it obliges us to keep deposited observations apart from derived variables. The block means, regressions and diagnostics are newly derived; the respondent records remain credited to their creator. Where context is missing we say so, without guessing at the creator's location or the likely structure of a familiar questionnaire.

Explanation should also be kept apart from prediction. The R-squared values come from the same sample used to estimate the coefficients, so they measure fitted association, not performance on new respondents, firms or inspection outcomes. A model that predicts one questionnaire block from another could fit well simply because the instrument captures overlapping evaluations, and still say little about independently observed audit deficiencies. We make no deployment claim for the regression.

Dropping an unsupported causal label makes the main conclusion more precise, not weaker. The public responses contain a strong, stable association that the other coded blocks cannot fully reproduce. They do not show a monetary investment effect or better independently assessed engagements. The inspection trend supplies real outcomes and shows why an exposure measure matched to them is still needed.

7. Limitations and Conclusion

Several limitations bound the findings. The deposit does not document the sampling frame, fieldwork dates or firm clusters behind the questionnaire, and item wording and any reverse-scoring instructions are unavailable, so the scoring rule stays conditional. Simultaneous self-reports, overlap among blocks and unobserved respondent characteristics all stand in the way of a causal interpretation. The inspection sample covers four large UK firms selected through a regulatory process. It does not represent all auditors or jurisdictions.

The uncertainty estimates assume independent rows and the stated scoring rule; they are not population guarantees. Resampling, leave-one-out and item-omission checks show numerical stability, but they cannot add missing firms, remove common-method influences or establish that the original collection is authentic. We attribute the deposit as respondent data as its creator describes it and claim no independent verification of the fieldwork.

Within these bounds the conclusion is clear. In the 98-row release, AI-related and AQ-related questionnaire scores are strongly and positively associated, with a fully adjusted slope of 0.731. The relationship stays positive across multiple sensitivity calculations, although the separation diagnostics argue against a structural interpretation. The independently collected inspection evidence shows improving outcomes among selected Big Four audits but cannot connect them to matched AI investment. The study establishes association and measurement limits, not a causal monetary return to AI expenditure.

8. Data Availability and Attribution

Handoko (2025) makes the original questionnaire workbook available at https://doi.org/ 10.5281/ zenodo. 15034566 under CC BY 4.0. The accompanying package

includes an unchanged copy, repository metadata, item-level CSV, derived scores, complete model outputs and executable analysis code. Derived calculations and manuscript interpretation are new secondary analyses; the original dataset is not presented as data collected by Liu Yanjun. FRC report URLs and page locators are retained in the source register. The two datasets are never merged. The incremental-fit and residualisation checks are reproduced by expanded/analyze_depth.py and recorded in expanded/depth_results.json.

No new human participants were recruited for this secondary analysis. The public deposit does not document original consent or institutional review arrangements, and this manuscript makes no claim about them. Funding, competing-interest and authorship-contribution declarations require the named author's confirmation before submission. AI assistance was used in retrieval, calculation, drafting and exhibit preparation; the author should verify the source material and analysis before submitting.

Acknowledgments

This research was funded by Enerstay Sustainability Academy (Singapore) Grant Call (Call 7/2025) _SDG (Project ID CUMT-001), Singapore.

References

[1] DeFond, M., & Zhang, J. (2014). A review of archival auditing research. Journal of Accounting and Economics, 58(2–3), 275–326. https://doi.org/10.1016/j.jacceco.2014.09.002.

[2] Fedyk, A., Hodson, J., Khimich, N., & Fedyk, T. (2022). Is artificial intelligence improving the audit process? Review of Accounting Studies, 27, 938–985. https://doi.org/10.1007/s11142-022-09697-x.

[3] Financial Reporting Council. (2025a). Deloitte LLP audit quality inspection and supervision. https://media.frc.org.uk/documents/Deloitte_LLP_Audit_Quality_Inspection_and_Supervision_2025.pdf.

[4] Financial Reporting Council. (2025b). Ernst & Young LLP audit quality inspection and supervision. https://media.frc.org.uk/documents/Ernst__Young_LLP_Audit_Quality_Inspection_and_Supervision_2025.pdf.

[5] Financial Reporting Council. (2025c). KPMG LLP audit quality inspection and supervision. https://media.frc.org.uk/documents/KPMG_LLP_Audit_Quality_Inspection_and_Supervision_2025.pdf.

[6] Financial Reporting Council. (2025d). PricewaterhouseCoopers LLP audit quality inspection and supervision. https://media.frc.org.uk/documents/PricewaterhouseCoopers_LLP_Audit_Quality_Inspection_and_Supervision_2025.pdf.

[7] Handoko, B. L. (2025). Artificial intelligence adoption for sustainable audit quality in non-big four accounting firm (Version 1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.15034566.

[8] Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. Journal of the Academy of Marketing Science, 43, 115–135. https://doi.org/10.1007/s11747-014-0403-8.

[9] Law, K. K. F., & Shen, M. (2025). How does artificial intelligence shape audit firms? Management Science, 71(5), 3641–3666. https://doi.org/10.1287/mnsc.2022.04040.

[10] MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. Journal of Econometrics, 29(3), 305–325. https://doi.org/10.1016/0304-4076(85)90158-7.

[11] Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. Journal of Applied Psychology, 88(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879.

[12] Rajgopal, S., Srinivasan, S., & Zheng, X. (2021). Measuring audit quality. Review of Accounting Studies, 26, 559–619. https://doi.org/10.1007/s11142-020-09570-9.

[13] Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika, 74, 107–120. https://doi.org/10.1007/s11336-008-9101-0.

[14] Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. MIS Quarterly, 27(3), 425–478. https://doi.org/10.2307/30036540.