When the Model Writes the Questionnaire: An Item-Level Defect Taxonomy for LLM-Generated 360-Degree Feedback Instruments

Evidence from two production pilots. Preprint v1 (field study with a defect taxonomy)

Alessio Biancheri*

12 July 2026

Status: Work in progress. Open research agenda in Section 9. Collaboration invited.

Abstract

Large language models are now used to generate the items of psychological and organisational assessment instruments. The emerging literature on automatic item generation evaluates such items psychometrically, under laboratory conditions, on constructs chosen by the researcher. It does not tell us what happens when an LLM writes the questionnaire that a real company will actually send to its employees.

We report a field audit of exactly that. Across two production pilots of an AI-native 360-degree feedback platform, we audited the items the system generated and the behaviour of the surrounding pipeline. At one site, the system generated four instruments (manager, peer, direct-report and self perspectives) from the organisation's own competency framework. We audited every generated item against established item-writing criteria. **37 of 48 items (77%) carried at least one item-writing defect.**

We organise the defects into eight classes: relevance failure (the item presumes a role or a customer-facing context the respondent does not have), double-barrelled items, content-validity imbalance, social-desirability exposure, leading items, format mismatch (an open-ended stem attached to a rating response), redundancy, and binary constructs forced onto a Likert scale. Relevance failure (15 items) and double-barrelling (12) were the most common.

Beyond item writing, the other site surfaced a distinct and, we argue, more serious class of failure. A real-time assistant intended to help assessors write better feedback systematically steered them to rewrite legitimate negative feedback in positive terms. A multi-source feedback instrument whose assistant suppresses negative valence does not merely produce noisy data. It produces data that is wrong in a specific and undetectable direction, while appearing to work.

We report additional pipeline-level failures (locale propagation across a shared session, a single model serving every pipeline stage, and central-tendency bias induced by an odd-numbered scale) and we set out mitigation directions at the level of design principle. We do not report our implementation.

*ignostiq (Alesserg Technology OÜ), Tallinn, Estonia.
a.biancheri@ignostiq.com
ORCID 0009-0009-9653-7091

We are explicit about what this study is not. It is a retrospective single-rater audit of two pilots of one product. There is no inter-rater reliability statistic, no blinding, no control condition, and no comparison against human-written items from the same organisations. It establishes that these defects occur and are common. It does not establish their rate in general, nor that LLM-written items are worse than human-written ones. We set out the study that would.

Keywords: automatic item generation; large language models; 360-degree feedback; multi-source feedback; questionnaire design; item-writing; automation bias; human-AI interaction; HR technology

---

1. Introduction

Multi-source, or 360-degree, feedback rests on a fragile assumption: that the instrument asks questions whose answers mean something. Decades of work on feedback interventions show that the effect on performance is variable and can be negative [Kluger & DeNisi, 1996; Smither, London & Reilly, 2005], and that instrument quality is one of the conditions separating the cases where it helps from the cases where it does not [Bracken & Rose, 2011]. The rules for writing such items are old and well established: do not ask two things at once, do not lead the respondent, match the response format to the stem, do not expose the respondent to social-desirability pressure, and ensure the item is answerable by the person being asked [Haladyna, Downing & Rodriguez, 2002; Bradburn, Sudman & Wansink, 2004; Podsakoff et al., 2003].

A recent and rapidly growing literature applies large language models to *automatic item generation* (AIG), producing candidate items for personality inventories, situational judgement tests and educational assessments, then screening them psychometrically [Li et al., 2024; Russell-Lasalandra, Christensen & Golino]. This work is careful and it is largely optimistic, with the important caveat that LLM item quality varies markedly by construct and that items often align poorly with the intended construct.

But that literature evaluates items in the laboratory, on constructs the researcher selected, screened by the researcher before anyone answers them. It does not describe the situation that is now common in commercial HR software: **a company uploads its own competency framework, the model writes the questionnaire, and the questionnaire goes out to employees.** Nobody screens the items. The generation step and the deployment step are the same step.

We had the opportunity to observe this directly. This paper reports what we found.

1.1 Contributions

1. **An item-level defect taxonomy** for LLM-generated multi-source feedback items, grounded in classical item-writing criteria and derived from a real deployment (Section 5).
2. **A defect rate from a production instrument set**: 37 of 48 generated items (77%) carried at least one defect, with the distribution across eight classes (Section 6).
3. **A pipeline-level failure catalogue** covering locale propagation, single-model-for-all-stages, scale-induced central tendency, and the assistant valence-steering failure (Section 7).
4. **The valence-steering finding**, which we consider the most consequential result here and which we believe is novel: an assistive LLM that nudges assessors toward positively phrased feedback corrupts the instrument in a direction that its own outputs cannot reveal (Section 8).

5. **An honest account of what this study cannot support**, and the controlled study that would (Sections 10 and 11).

---

2. Related work

Item-writing rules. The criteria we audit against are not novel and are not ours. Haladyna, Downing and Rodriguez [2002] consolidated the item-writing guideline literature into a validated taxonomy. Classic survey-methodology texts codify the same rules for attitude and behaviour items [Bradburn, Sudman & Wansink, 2004]. Podsakoff and colleagues [2003] catalogue the method biases that item wording introduces, including social desirability and common-method variance. Krosnick [1991] explains why respondents satisfice when items are cognitively demanding, which is the mechanism connecting bad items to bad data.

Rating scales and the midpoint. Whether an odd-numbered scale with an explicit midpoint invites central-tendency responding is a long-running question [Garland, 1991; Preston & Colman, 2000; Weng, 2004]. The literature does not fully settle it, but the concern is well founded and standard practice in high-stakes instruments is to consider it deliberately rather than by default.

Multi-source feedback. Kluger and DeNisi [1996] showed feedback interventions have highly variable effects, including negative ones. Smither, London and Reilly [2005] found only modest average improvement following multi-source feedback and identified moderators. Bracken and Rose [2011] argue explicitly that instrument and process design determine whether 360 feedback changes behaviour.

Automatic item generation with LLMs. Recent work generates items for personality and situational judgement tests with LLMs and then screens them psychometrically. Li and colleagues [2024] generate situational judgement items and report that prompt design and sampling temperature materially affect content validity. Russell-Lasalandra, Christensen and Golino integrate generation with network psychometrics (AI-GENIE) to select high-quality, non-redundant items, and validate the resulting scales against expert-developed measures. This line of work is careful and broadly encouraging, with the consistent caveat that item quality varies substantially by construct and that many generated items align poorly with the intended construct.

Note what these methods share: **a screening stage between generation and the respondent.** That stage is the contribution. Our study is complementary and less comfortable. We look at what reaches respondents when that stage is absent, because in commercial deployment it usually is.

Automation bias and assistive AI. The literature on automation bias and complacency [Parasuraman & Manzey, 2010] and on cognitive forcing in AI-assisted decisions [Buçinca, Malaya & Gajos, 2021] provides the frame for our valence-steering finding. An assistant that reduces the cost of agreeing with it, and raises the cost of disagreeing, will shift behaviour even when the human retains formal control.

---

3. Setting

The system under study is a commercial AI-native 360-degree feedback platform. The relevant pipeline stages are:

1. **Ingest** the customer's competency or values framework.
2. **Generate** items for each assessor perspective (self, manager, peer, direct report), with a response format per item.
3. **Calibrate** the evaluation scale.
4. **Distribute** the campaign, including localisation into the respondent's language.
5. **Assist** the assessor in real time while they write free-text feedback.
6. **Analyse** the responses and generate an insight report.

At the time of both deployments, a single general-purpose LLM served every generative stage.

Site 1 ran the platform for a leadership feedback cycle. Findings were captured in a structured pilot analysis produced jointly with the organisation's HR lead. The record documents system behaviour across the pipeline rather than an item-by-item audit.

Site 2 supplied a competency framework and the system generated four instruments from it, one per assessor perspective. A structured item-by-item review of all four instruments was produced.

3.1 Ethics and confidentiality

Both organisations reviewed this paper and consented to publication **on the express condition that they are not identifiable**. We have therefore removed the pilot dates, the sectors, the countries, the organisation sizes, and the structure of the competency frameworks, and we refer to the sites only as Site 1 and Site 2 in no meaningful order. Where an item is used to illustrate a defect class, it is an **illustrative reconstruction**, written by us to be typical of the class. It is not a customer's item and does not reproduce any customer's proprietary competency wording.

No respondent-level data is reported anywhere in this paper. No individual's feedback, ratings, or free text is reported, quoted, or characterised. The unit of analysis throughout is **the item**, not the person.

We accept that this restricts what the reader can verify, and we regard that as the correct trade. A field study of failures is only possible if the organisations that permit it are protected.

---

4. Method

For Site 2 we treat the four generated instruments as the corpus. Each item was reviewed against a coding frame derived from the item-writing literature (Section 2), and coded for the presence of one or more defects. An item could carry multiple defects; the classes are not mutually exclusive.

For Site 1 we treat the structured pilot analysis as a field-note corpus and extract failures at the pipeline level rather than the item level, because that record documents system behaviour rather than an item-by-item audit.

The coding was performed by a single rater, the author, retrospectively, and not blind to the fact that the items were machine-generated. This is a serious methodological limitation and we return to it in Section 10. We report it here rather than in a footnote because it conditions everything that follows.

The coding frame is reproduced in full in Appendix A so that a second rater can replicate or contest the coding.

5. The defect taxonomy

We define eight classes. Every example below is an **illustrative reconstruction**: written by us to be typical of the class, not quoted from a customer's instrument. See Section 3.1.

D1. Relevance failure (item is not universally answerable). The item presumes a context, a role or a relationship that the respondent does not have. Generated items repeatedly asked *all* respondents to rate a colleague's handling of customers or external stakeholders. A large share of the respondent population held technical and specialist roles with no customer contact of any kind. Items also presumed managerial authority, asking whether the person ensures their team has the resources it needs, when the item was addressed to individual contributors who have no team. The respondent must then guess, satisfice, or decline, and each of those choices damages the data differently.

The mechanism is worth naming precisely, because it is not a wording problem and no amount of item-level polish will fix it. The generator was given the competency framework, which describes *what* to measure. It was not given the respondent population, which determines *who can answer*. Lacking the second, it wrote for a plausible default: a customer-facing manager. The organisation was full of people who were neither. **The most common defect in our audit is therefore a grounding failure, not a writing failure**, and it is the direct consequence of an architecture that treats item generation as a function of the construct alone.

D2. Double-barrelled item. The item names two distinct constructs and admits one answer. The generator repeatedly conjoined pairs that a psychometrician would separate: accountability for oneself with accountability for others; fairness with inclusion; honesty with reliability; a positive climate with an inclusive one; recognising others' feelings with acting on them; setting priorities with allocating resources. The response is uninterpretable, because the analyst cannot know which half of the item the respondent answered.

D3. Content-validity imbalance. The generated instrument distributed items across the source framework's areas unevenly. In one instrument, a single area received four items while three other areas received one item each. In another, an entire area of the customer's framework received **no items at all**. The instrument therefore does not measure the framework it was generated from. This defect is invisible at the item level. It appears only when the instrument is viewed as a whole, which is exactly the view an item-by-item generator never takes.

D4. Social-desirability exposure. The item invites the respondent to report something nobody will admit. Observed cases include asking respondents to rate whether their own decisions are free of bias, asking self-assessors to describe how well they handle difficult conversations, and asking direct reports to rate their own manager's integrity. The last is compound: the construct is socially loaded *and* the power relationship makes an honest low rating personally costly.

D5. Leading item. The stem presupposes the behaviour it claims to measure. Generated items asked respondents to describe a time the person *effectively* resolved a conflict, *successfully* seized an opportunity, or *actively* listened. An item that presupposes the positive outcome cannot detect its absence. The respondent who has never seen the behaviour has no way to say so.

D6. Format mismatch (stem and response type disagree). Items phrased as open questions ("How does this person…", "How do you…") were attached to a numeric rating scale. In at least

one case an item explicitly requesting an example ("Give an example of…") was assigned a rating scale. The respondent is asked a question they cannot answer in the form they are given. This is the most mechanical defect in the list and, notably, the easiest to detect automatically without any model at all.

D7. Redundancy. Two items in the same instrument measured substantially the same construct in different words: two separate items on setting priorities and managing workload, and two separate items on customer-centred decision making. Redundancy inflates internal-consistency statistics while adding no information, which makes the instrument look *more* reliable than it is. This is the one defect in the taxonomy that flatters the vendor.

D8. Binary construct on a graded scale. Some constructs do not admit degrees. The generator produced graded rating items asking how well a person maintains confidentiality. Confidentiality is respected or it is breached. A five-point scale on a binary construct manufactures variance that does not exist.

---

6. Results

Site 2, item audit. The system generated 48 items across four instruments (manager, peer, direct-report and self perspectives) from the organisation's competency framework. **37 items (77%) carried at least one defect.** Because items can carry more than one defect, the class counts below sum to more than 37.

Class	Defect	Items affected
D1	Relevance failure	15
D2	Double-barrelled	12
D3	Content-validity imbalance	11
D4	Social-desirability exposure	10
D5	Leading item	7
D6	Format mismatch	6
D7	Redundancy	3
D8	Binary construct on a graded scale	2

Flagged items were distributed evenly across the four perspectives (manager 9, peer 9, direct report 10, self 9), which is worth noting: the defects are not an artefact of one hard perspective. They are a property of the generation step.

Two observations sharpen the picture.

First, **the most common defect is not a wording error.** Relevance failure (D1) is a *grounding* failure. The generator had access to the competency framework but not to a usable model of who would answer, so it wrote items for a generic corporate respondent who interacts with customers and manages a team. Most respondents were neither.

Second, **the defect the generator is structurally incapable of noticing is D3.** Content-validity imbalance is a property of the instrument, not of any item. A pipeline that generates items one at a time, or one competency area at a time, cannot see it. This is an architectural point, not a prompt-quality point.

7. Pipeline-level failures

The item audit is only part of the picture. Site 1 surfaced failures elsewhere in the pipeline.

F1. Locale propagation across a shared session. Campaign emails were dispatched in the wrong language despite the campaign being configured otherwise, and translated output appeared corrupted in the insight report. The output language was coupled to the administrator's session locale rather than to the campaign or the recipient. There was no per-respondent language attribute at all. A multilingual feedback cycle in which some assessors receive the questionnaire in a language they do not read is not a cosmetic bug: those assessors drop out, and they do not drop out at random.

F2. Machine-translated items lose their psychometric properties. Items translated literally by a general-purpose translation service produced stilted and sometimes meaning-shifted questions. An item that has been carefully worded to avoid double-barrelling in English can acquire a conjunction in translation. **Item quality is not translation-invariant**, and to our knowledge no AIG work addresses this.

F3. One model for every stage. A single general-purpose LLM served item generation, real-time assistance, translation and analysis. These are different tasks with different failure profiles and different quality-latency-cost trade-offs. A single model choice is a single point of compromise across all of them.

F4. Scale-induced central tendency. The default evaluation scale was 1 to 5. Assessors clustered on the midpoint. This is the oldest and best documented problem in the list, and the platform reproduced it by default.

F5. Assistant valence steering. Treated separately below, because we think it is the finding that matters.

---

8. The valence-steering failure

The platform included a real-time assistant that reviewed an assessor's free-text feedback as they wrote it and suggested improvements. Its intent was benign: help people give feedback that is specific, behavioural and constructive rather than vague or abusive.

In the pilot, it systematically **pushed assessors to rewrite legitimate negative feedback in positive terms.** It did not distinguish between feedback that was *inappropriate* (abusive, personal, unactionable) and feedback that was merely *negative* (accurate, specific, unwelcome). It treated negative valence itself as the defect to be corrected.

We want to state plainly why this is worse than the item defects.

An item defect adds noise. A reviewer, an analyst or a factor structure can eventually detect noise. **Valence steering adds bias, in a known direction, and erases its own evidence.** The assessor's original, accurate, negative observation is not recorded anywhere. What is recorded is the softened version they accepted. The resulting dataset is internally consistent, complete, well formatted, and systematically wrong. Every downstream artefact, the scores, the themes, the sentiment analysis, the insight report, the development plan, inherits the bias and none of them

can reveal it. The instrument reports that the organisation's feedback culture is more positive than it is, and it reports this *because of a feature that was added to improve feedback quality.*

This is automation bias [Parasuraman & Manzey, 2010] operating on the input side of an assessment instrument. The assessor retains formal control, and can reject the suggestion. But the suggestion is free to accept and effortful to refuse, and the assistant's framing casts the softer version as the more *professional* one. That asymmetry is the mechanism.

The implication generalises well beyond our product. Any assistive LLM placed between a human and an assessment instrument is a potential source of directional bias in the resulting measurement, and the bias will not appear in the instrument's own outputs. We do not think this is currently a recognised threat to validity in the multi-source feedback literature, and we think it should be.

---

9. Mitigation directions

We state these as design principles. We deliberately do not describe our implementation, our prompts, or our system architecture, and nothing below requires them. Every principle here is either standard practice in psychometrics or a straightforward engineering consequence of the failures above.

1. **Screen generated items before they are deployed, not after.** The AIG literature already assumes a screening step. Commercial pipelines frequently omit it. D6 (format mismatch), D2 (double-barrelling) and D8 (binary constructs) are substantially detectable by automated checks that do not require a model.
2. **Give the generator the respondent, not just the construct.** D1 is a grounding failure. An item generator that does not know the respondent's role, level and functional context will write for a generic manager.
3. **Validate the instrument, not only the item.** D3 is only visible at instrument scope. Coverage against the source framework should be an explicit, checked constraint.
4. **Choose the response scale deliberately.** An odd-numbered default with an explicit midpoint is a choice, and it should be made rather than inherited.
5. **Decouple output language from session state, and treat the item as the unit of localisation.** An item's psychometric properties must survive translation, which means translation must be quality-checked at the item level.
6. **Match the model to the pipeline stage.** Item generation, real-time assistance, translation and analysis are different tasks. Treating them as one task guarantees that at least one of them is served badly.
7. **An assistant that touches assessment input must distinguish inappropriate from unwelcome.** This is the strongest recommendation in the paper. If the distinction cannot be made reliably, the assistant should not rewrite; it should prompt the assessor to be more specific and leave the valence alone. Where the assistant does alter text, the original should be retained so that the direction and magnitude of the shift can be audited.

A subsequent version of the platform addressed the defect classes reported here. We report no results from it, because a before-and-after comparison across two different customers, two different frameworks and two different time points would not be a controlled comparison, and we decline to present it as one.

10. Limitations

We list these plainly, because most of them are severe.

1. **Single rater, no reliability statistic.** The item coding was performed by one person, the author, who is also the vendor. There is no second rater and therefore no inter-rater reliability. This is the single biggest weakness of the study. The coding frame is published in Appendix A precisely so that this can be corrected by someone else.
2. **Not blind.** The rater knew the items were machine-generated and knew which product produced them.
3. **No human-written control.** We did not audit human-written items from the same firms against the same frame. We therefore **cannot claim that LLM-generated items are worse than human-written ones.** Human-written 360 items are also frequently double-barrelled and leading; that is why the item-writing literature exists. What we can claim is that the LLM did not avoid these defects, and that no screening layer caught them.
4. **Two sites, one product, one model.** The defect rate of 77% is a property of these instruments, this model, this prompt configuration and this moment. It is not an estimate of a population parameter and should not be cited as one.
5. **Retrospective.** The audit was conducted on records produced for commercial purposes, not to a research protocol specified in advance.
6. **The instrument pool was reconstructed from the audit record**, not from a generation log. The denominator of 48 items should be treated as accurate but not independently verified.
7. **The Site 1 record is a field-note corpus, not an item audit.** The pipeline failures in Section 7 are reported as observed and are not quantified.
8. **The valence-steering finding is qualitative.** We observed the behaviour and the organisation reported it. We did not measure the magnitude of the induced shift, because the original text was not retained. This is the most important thing in the paper and it is the least well measured, which is an uncomfortable but accurate thing to say.
9. **Vendor authorship.** The author has a commercial interest in this product. We have tried to counteract this by reporting only our own failures and by declining to report the comparison that would flatter us. The reader should nevertheless discount accordingly.

---

11. Open research agenda

Each of these is a study we would like to run with an academic partner, and several of them we cannot run alone.

1. **The controlled item-quality study.** Same competency framework, three conditions: items written by a trained I/O psychologist, items generated by an LLM without screening, items generated by an LLM with an automated screening layer. Blind, multi-rater coding against a pre-registered frame. This settles the question our study can only raise.
2. **Measuring valence steering.** Randomise assessors to an assistant that rewrites, an assistant that prompts without rewriting, and no assistant. Retain the original text in all conditions. Measure the shift in sentiment, specificity and behavioural anchoring between first draft and submitted text. **We believe this is the highest-value experiment in this space and we know of no one who has run it.**

3. **Does instrument defect rate actually degrade the measurement?** Our audit establishes defects against a normative standard. It does not establish that the resulting scores are worse. Correlating defect density with the psychometric properties of the returned data (internal consistency, factor structure, rater agreement, missingness) would close that gap.
4. **Item quality under translation.** Do the item-writing properties of a generated item survive machine translation, and if not, which defects are introduced by translation rather than by generation?
5. **Automated item screening.** How much of D2, D6, D7 and D8 can be caught by deterministic checks, and how much needs a model? Cheap wins here would improve every commercial AIG pipeline immediately.
6. **The instrument-scope problem.** D3 suggests item-by-item generation is the wrong unit. What does a generation procedure that optimises the instrument as a whole look like?

---

12. Conclusion

Large language models are already writing the questionnaires that organisations use to evaluate their people, and in commercial deployment they are frequently doing so without a screening step between generation and the respondent. In two production pilots, we found that most generated items (37 of 48) violated at least one long-established item-writing rule, most often by presuming a role or context the respondent did not have, and by asking two things at once.

We consider the item defects to be the lesser finding. The more serious one is that an assistive model placed between the assessor and the instrument steered assessors away from negative feedback, producing a dataset that is complete, coherent and biased in a direction that nothing downstream can detect. The rules for writing good items are old and available. The failure mode introduced by putting a helpful model next to the respondent is new, and it is not yet on anyone's list of threats to validity.

We publish these findings about our own product, with our own methodological weaknesses stated, because a field that only publishes what worked will not find these problems.

References

- Bracken, D. W. & Rose, D. S. (2011). When does 360-degree feedback create behavior change? And how would we know it when it does? *Journal of Business and Psychology*, 26(2), 183-192.
- Bradburn, N. M., Sudman, S. & Wansink, B. (2004). *Asking Questions: The Definitive Guide to Questionnaire Design.* Jossey-Bass.
- Buçinca, Z., Malaya, M. B. & Gajos, K. Z. (2021). To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1).
- Garland, R. (1991). The mid-point on a rating scale: is it desirable? *Marketing Bulletin*, 2, 66-70.
- Haladyna, T. M., Downing, S. M. & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309-333.

- Kluger, A. N. & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254-284.
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3), 213-236.
- Parasuraman, R. & Manzey, D. H. (2010). Complacency and bias in human use of automation: an attentional integration. *Human Factors*, 52(3), 381-410.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y. & Podsakoff, N. P. (2003). Common method biases in behavioral research. *Journal of Applied Psychology*, 88(5), 879-903.
- Preston, C. C. & Colman, A. M. (2000). Optimal number of response categories in rating scales. *Acta Psychologica*, 104(1), 1-15.
- Smither, J. W., London, M. & Reilly, R. R. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology*, 58(1), 33-66.
- Li, C.-J., Zhang, J., Tang, Y. & Li, J. (2024). Automatic item generation for personality situational judgment tests with large language models. *arXiv preprint* arXiv:2412.12144.
- Russell-Lasalandra, L. L., Christensen, A. P. & Golino, H. Generative psychometrics via AI-GENIE: automatic item generation and validation with network-integrated evaluation. *PsyArXiv preprint*. https://doi.org/10.31234/osf.io/fgbj4
- Weng, L. J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, 64(6), 956-972.

Appendix A. Coding frame

An item is coded into a class if it meets the criterion. Classes are not mutually exclusive.

Class	Criterion for coding
D1 Relevance failure	The item presumes a role, relationship or context (customer contact, managerial authority, team ownership) that some intended respondents do not have.
D2 Double-barrelled	The item names two distinct constructs joined by "and" or by implication, and admits a single response.
D3 Content-validity imbalance	At instrument scope: a source-framework area receives disproportionately many or few items, or none.
D4 Social-desirability exposure	An honest unfavourable response carries social or professional cost to the respondent, or the item asks the respondent to disavow a common failing.
D5 Leading item	The stem contains a term ("effectively", "successfully", "actively") presupposing the behaviour or outcome being measured.

Class	Criterion for coding
D6 Format mismatch	The stem is phrased as an open question or requests an example, and the assigned response type is a rating scale.
D7 Redundancy	Two items within the same instrument measure substantially the same construct.
D8 Binary construct on a graded scale	The construct does not admit degrees (confidentiality, disclosure) and is assigned a graded response scale.

Appendix B. Reproducibility and data availability

The a