From demonstration to evidence

An evaluation protocol for PARA, an AI-augmented method for culinary development

James Won
French Borneo · Nari Concepts Sdn. Bhd.
Adjunct Professor, Taylor's Culinary Institute, Kuala Lumpur
Correspondence: the author

ABSTRACT

A companion case study established, by demonstration, that PARA — a structured, AI-augmented method for fine-dining menu development in the French Borneo idiom — can produce a coherent, canon-compliant, service-ready degustation. A demonstration is not an evaluation. It shows that an artefact can be used; it does not show that the artefact's outputs are good, that the method helps practitioners who did not build it, or that the conservation claim made for the idiom holds in the judgement of those who hold the knowledge. This paper specifies the study that would test those claims. It is a protocol, not a set of results: no findings are reported, and the document is written to be executed, and ideally pre-registered, by hands other than the designer's. The protocol is structured by the Framework for Evaluation in Design Science Research, and proceeds through four evaluation episodes that move from artificial, controlled settings toward naturalistic, real-use ones. It separates the evaluation of the method as an artefact — can independent chefs operate it, and does it reduce error against unaided development — from the evaluation of the menus the method produces, assessed by trained sensory panels, by guests, by expert raters of culinary creativity, and by the knowledge-holders whose techniques the menus engage. Instruments are specified for each strand: a canon-compliance audit, descriptive and hedonic sensory testing, the Consensual Assessment Technique for creativity, and a conservation-fidelity rubric governed by free, prior, and informed consent. The protocol states plainly what a completed evaluation would, and would not, license one to claim, and holds the method's roadmap capability — that it will learn across menus — outside its scope.

Keywords: design science evaluation; FEDS; sensory evaluation; consensual assessment technique; culinary creativity; menu development; research protocol; indigenous food knowledge; informed consent

1 Purpose and scope

A companion case study examined PARA, an AI-augmented method for developing fine-dining menus, through a single worked example: a five-course chef's table built on durian read across two cultivars. That study was explicit about the limit of its own evidence. It was a demonstration in the design-science sense — a showing that the artefact can be used to address the problem it was built for — and not a summative evaluation, which would compare the method's stated objectives against measured outcomes of use (Peffers et al., 2007). The demonstration established that PARA can carry a menu from brief to costed, plated, service-ready output under a governing canon. It established nothing about quality, about transferability to other hands, or about the conservation claim that gives the idiom its purpose.

This paper specifies the evaluation that would supply that evidence. Three commitments govern it.

First, it is a protocol and not a study. No results are reported, because none have been gathered. The document is written to be executed, and its analysis plan is written to be pre-registered before any data are collected, so that the claims it can support are fixed before the evidence is in hand.

Second, the designer is removed from the evidence. The principal threat to the demonstration's validity was that one person designed the artefact, operated it, and analysed the result. The evaluation answers that threat structurally: the method is operated by chefs who did not build it, the outputs are rated by judges blind to the conditions, and the designer takes no part in rating.

Third, the claims to be tested are the four design objectives the method sets itself — externalisation, conservation through technique, coherence under a governing idea, and trustworthiness — restated here as testable questions. The evaluation does not invent new criteria; it measures the artefact against the criteria by which it was built.

The scope has one deliberate exclusion. PARA is described, in its author's roadmap, as a method that will learn and compound across menus. That capability is not a present property of the artefact, was not claimed in the demonstration, and is not evaluated here. Testing it would require a longitudinal study of a different design, and Section 8 marks it as out of scope.

2 Evaluation framework

The protocol is structured by the Framework for Evaluation in Design Science Research, hereafter FEDS (Venable, Pries-Heje and Baskerville, 2016). FEDS was developed to remedy the thinness of guidance on how design-science artefacts should be evaluated, and it characterises any particular evaluation along two dimensions: its functional purpose, which is either formative — improving the artefact during development — or summative — assessing the finished artefact's performance; and its paradigm, which is either artificial, in a controlled setting, or naturalistic, in real use. FEDS sets out a four-step process for designing an evaluation: explicate its goals, choose a strategy, determine the properties to be evaluated, and design the individual evaluation episodes (Venable, Pries-Heje and Baskerville, 2016).

This protocol is principally summative, and it adopts the strategy FEDS describes as moving the evaluation from artificial toward naturalistic settings as confidence accumulates — beginning with controlled inspection and ending in real kitchens with real guests and the communities whose knowledge is at stake. The reasoning is that PARA's value rests on human use: a chef must be able to operate it, a guest must find the result worth eating, and a knowledge-holder must judge the technique faithfully carried. An evaluation confined to controlled settings could not reach those judgements, so the trajectory ends in naturalistic evaluation even though it begins away from it.

A distinction runs through the whole design, and it is worth stating before the questions are posed. Evaluating PARA means evaluating two different things. One is the method as an artefact: whether it can be operated, by whom, with what reliability, and at what cost in error and effort. The other is the menus the method produces: whether they are good, as food, as creative work, and as carriers of indigenous technique. A method could be usable and produce poor menus, or produce fine menus but only in the hands of the chef who built it. Both must be evaluated, and the protocol keeps them apart.

3 Evaluation questions

The four design objectives of the artefact, restated as questions, with a fifth on efficiency.

EQ1 — Operability and externalisation. Can chefs other than the designer operate PARA to produce menus that are coherent, canon-compliant, and judged feasible for service? This tests the externalisation objective: a method whose reasoning is genuinely inspectable should be operable by a second practitioner. *(Artefact.)*

EQ2 — Sensory quality of outputs. Are menus developed in PARA judged, by trained descriptive panels and by guests, to reach fine-dining sensory quality? *(Output.)*

EQ3 — Creative and conceptual quality of outputs. Are PARA-developed menus judged creative and conceptually coherent by domain experts? *(Output.)*

EQ4 — Conservation fidelity. Do the knowledge-holders and indigenous-foodways scholars whose techniques a menu engages judge that the technique is faithfully and respectfully carried, and the attribution properly made? *(Output and ethics.)*

EQ5 — Efficiency and trustworthiness. Relative to unaided development, does PARA reduce errors of canon and feasibility, and does its verification gate prevent unverified or fabricated content from reaching the output? *(Artefact.)*

4 Design: four evaluation episodes

The evaluation is built as four episodes, in the FEDS sense of particular evaluations, sequenced to move from artificial toward naturalistic settings (Venable, Pries-Heje and Baskerville, 2016). Each episode addresses specified questions and feeds the next. Sample sizes are given as design parameters drawn from the norms of the methods cited, not as commitments; they would be fixed at pre-registration against a power analysis where inferential testing is intended.

4.1 Episode A — Artificial, formative: expert inspection of outputs

A panel of independent chefs and culinary educators inspects a set of completed PARA dossiers — including the durian dossier and others developed to varied briefs — against a structured rubric for coherence, canon compliance, and feasibility. The purpose is formative: to surface faults in the method's outputs before operators are asked to use it, and to refine the instruments that the later episodes will apply. This episode does not test operability; it tests whether the artefact's outputs, as artefacts, withstand expert reading.

4.2 Episode B — Artificial, summative: the independent-operator study

The core artefact evaluation, addressing EQ1 and EQ5. A sample of chefs and senior students who did not build PARA — a design parameter of the order of twelve to twenty operators — each develops a menu to a common brief. Operators are assigned to one of three conditions: developing with PARA; developing to the same staged scaffold and canon by hand, without the system; or developing by their ordinary practice without either. The three conditions are what allow the method's effects to be attributed. The contrast between the structured-manual condition and PARA isolates the system's own contribution — the question the case study could not answer from a single development; the contrast between the structured-manual and the unaided conditions isolates the contribution of the staging. Outputs are assessed, by judges blind to condition, against the canon-compliance audit and the coherence and feasibility rubrics of Section 5. Operability is measured by a standard usability instrument and by the operator's own account; effort is measured by time to a service-ready output; trustworthiness is measured by an audit of whether any unverified or fabricated content reached the output, and by the completeness of the verification-gate record. The three-arm comparison is what allows EQ5 to be answered, and what prevents an effect of the staging from being mistaken for an effect of the model.

4.3 Episode C — Naturalistic, summative: kitchen, panel, and guest

The principal output evaluation, addressing EQ2 and EQ3. A set of PARA-developed menus is cooked and served under real service conditions. Three measurements are taken. A trained descriptive panel profiles the dishes by quantitative descriptive analysis, characterising what is on the plate independently of whether it is liked (Lawless and Heymann, 2010; Meilgaard, Civille and Carr, 2016). A panel of guests rates overall liking on a nine-point hedonic scale and rates the balance of specific attributes — sweetness, acidity, salt, the intensity of the durian — on just-about-right scales, which together indicate not only whether a dish is liked but in which direction it might be adjusted (Lawless and Heymann, 2010). And a panel of domain experts rates the creative and conceptual quality of the menus by the method of Section 5.3. Where a comparison is sought, PARA-developed menus are set against menus developed without the method, with raters and panellists blind to which is which.

4.4 Episode D — Naturalistic: the conservation-fidelity panel

The evaluation specific to the idiom's purpose, addressing EQ4, and the episode that carries the heaviest ethical weight. The indigenous techniques a menu engages are assessed for fidelity and for the propriety of their attribution by those best placed to judge: the knowledge-holders of the communities concerned, and scholars of the relevant foodways. The assessment is both qualitative, through structured conversation, and structured, through the fidelity rubric of Section 5.4. This episode cannot be run as a detached measurement. The episode is governed by the communities, not merely consulted: under the authority-to-control and collective-benefit principles of the CARE Principles for Indigenous Data Governance (Carroll et al., 2020), the prior informed consent and mutually agreed terms required by the Nagoya Protocol (Secretariat of the CBD, 2011), and the rights affirmed in Article 31 of UNDRIP (United Nations, 2007), the communities hold the standing to shape the rubric, to decline, and to determine how the resulting knowledge is held and used. A distinction is drawn that the episode must not blur. Fidelity is the faithful carrying of a technique in one menu, which a panel can judge at a sitting. Conservation is the continued living transmission of that technique across time, which no single episode can certify and which remains a community-held, longitudinal outcome. Episode D can speak to the first; it can only point toward the second.

5 Instruments and measures

5.1 Canon-compliance audit (artefact)

A checklist derived directly from the method's governing constructs, scored by a blind assessor against each output. It records whether the protein-class base rule holds in every course — meat preparations on a chicken base, non-meat on a dashi base, the bases never crossed; whether the standing exclusions are observed; whether the cultivar canon is respected, each course naming and holding to its assigned variety; whether each indigenous technique is attributed to its community; and whether the verification gate was completed, with every assumption and unconfirmed figure flagged. A failure in any item is recorded as a canon breach, and the breach rate is the audit's principal measure.

5.2 Coherence and feasibility rubrics (artefact)

Coherence is rated by expert judges against a rubric covering the integrity of the arc, the fit of each course to the governing idea, and the soundness of the pairing logic where a flight is present. Feasibility is rated by an executing chef against a rubric covering whether the recipes are specified completely enough to cook, whether the timings and holds are sound, and whether the menu could be produced under the stated brigade and conditions.

5.3 Creativity and conceptual quality (output)

Creativity is assessed by the Consensual Assessment Technique (Amabile, 1982; Amabile, 1996). A panel of domain experts — chefs and critics familiar with the fine-dining register — independently rates each menu for creativity, where the object rated is the realised, plated, and tasted dishes at a sitting and not the development dossier, using their own implicit conceptions of what is creative rather than a definition imposed by the researcher, and without consulting one another, so that the average rating across experts becomes the measure and the idiosyncrasy of any single rater is diluted. The technique was validated on artefacts that judges apprehend directly; applying it to the experience of a served menu is an assumption the protocol makes openly, and the tasting format is chosen to honour it, so that the experts judge the eating rather than the paper that planned it. To guard against the known tendency of such ratings to be pulled by technical polish, raters separately rate technical quality and conceptual coherence, and are instructed to hold creativity apart from technical execution; rating distinct dimensions separately is part of the technique's own guard against one quality colouring another (Amabile, 1996). Inter-rater reliability is reported. Raters are blind to whether a menu was developed with PARA.

5.4 Conservation-fidelity rubric (output and ethics)

A rubric, developed with rather than for the communities concerned, by which knowledge-holders and foodways scholars judge whether the technique engaged in a menu is faithfully carried — whether, for instance, a fermentation reproduces the action the traditional method achieves — and whether the attribution names the source correctly and with respect. The rubric yields a structured judgement, but the qualitative account that accompanies it is treated as the primary evidence, because fidelity of this kind is not fully reducible to a scale.

5.5 Operability and effort (artefact)

Operability is measured by a standard, validated usability instrument completed by each operator, supplemented by a short structured interview. Effort is measured by time from brief to service-ready output. Both are compared across the three conditions of Episode B.

6 Analysis plan

The analysis is specified in advance, and the hypotheses and the analysis are pre-registered before any data are collected. Artefact measures — canon-breach rates, coherence and feasibility scores, usability, time, and the verification audit — are compared across the three conditions of Episode B (PARA, structured-manual, and unaided) by analysis of variance, or its mixed-effects equivalent where operators contribute repeated outputs, with the structured-manual arm isolating the system's contribution from that of the staging. Sensory measures are analysed by the standard procedures of descriptive and consumer testing, with mixed-effects models accommodating panellist and product effects (Lawless and Heymann, 2010). Creativity ratings are analysed for inter-rater reliability by the intraclass correlation coefficient before any aggregate is reported, and the aggregate is the mean across experts as the technique requires (Amabile, 1982). Conservation-fidelity evidence is analysed as mixed data, the qualitative account leading. The integration of the strands is interpretive: no single number is treated as a verdict on the method, and the protocol is explicit that a method may pass on one strand and fail on another.

7 Ethics, consent, and benefit return

The conservation strand makes a claim on the knowledge of specific communities, and the case study was clear that a method can name a source but cannot, on its own, constitute consent. The evaluation therefore treats consent and benefit return as conditions of the research, not as courtesies appended to it.

No technique is engaged in an evaluated menu without the free, prior, and informed consent of the community from which it comes, secured before development begins and not after, and on mutually agreed terms, in the manner the Nagoya Protocol requires for access to the traditional knowledge associated with genetic resources (Secretariat of the CBD, 2011). The terms of attribution and of any return of benefit are agreed with the community, in keeping with the collective-benefit and authority-to-control principles of the CARE framework (Carroll et al., 2020) and the rights affirmed in Article 31 of UNDRIP (United Nations, 2007); the design of Episode D is settled with the community's participation. The standing of safeguarding here follows the processual understanding of heritage, in which what is kept alive is a living practice held by its bearers rather than an object to be extracted (UNESCO, 2003); an evaluation that treated those bearers as a data source rather than as partners would contradict the very claim it set out to test.

The human-tasting episodes carry the ordinary obligations of research on human participants: institutional ethics approval, informed consent, the declaration and management of allergens, and food-safety governance for every item served, including the indigenous proteins whose handling the production documents specify. Participant data are held under the applicable protection law.

8 Threats to validity and their mitigation

The designer-as-evaluator threat, fatal to a demonstration run by one person, is removed by structure: independent operators in Episode B, judges and panellists blind to condition throughout, and the designer excluded from all rating. Blinding is applied where it can be — the expert raters of creativity, the trained sensory panel, and the canon and coherence judges are blind to which condition produced an output — and it is not pretended where it cannot. An operator necessarily knows whether they used the system; a guest may recognise the provenance of an avant-garde durian menu. These residual unblinded points are recorded as limitations rather than presented as solved.

Construct validity — whether the instruments measure the claims — is addressed by deriving the canon audit directly from the governing constructs, by separating creativity from technical quality in the creativity panel to prevent a halo effect (Amabile, 1996), and by treating conservation fidelity as primarily qualitative rather than forcing it onto a scale that would not bear it.

Several threats are acknowledged rather than dissolved. Fine-dining contexts yield small samples, which limits the inferential weight any single episode can carry; the protocol's answer is the convergence of independent strands rather than the power of one. Guest novelty bias — the tendency of an unfamiliar ingredient to be rated for its novelty rather than its quality — is mitigated by the blind comparison against menus developed without the method, but not eliminated. A single-site evaluation cannot speak to whether the method travels; multi-site replication is the proper remedy and is named here as the study beyond this one. And the roadmap capability, that the method learns across menus, remains outside scope; a longitudinal evaluation of a different design would be required to test it, and nothing in the present protocol should be read as testing it.

9 Phasing and feasibility

The four episodes are sequenced rather than simultaneous, each informing the next. Episode A is the lightest and comes first, refining the instruments. Episode B requires a cohort of operators and a controlled comparison, and sits naturally within a culinary institute, where independent operators and senior students are at hand and the method's pedagogical purpose makes the study double as teaching. Episode C requires a working kitchen, a trained sensory panel, and a guest cohort, and is the most demanding in resource. Episode D moves at the pace of the relationships it depends on, and cannot be hurried to a timetable set by the others. The protocol is feasible within a culinary-institute research setting with sensory-panel access and standing community relationships, which is the setting in which PARA is already used and taught.

10 Conclusion

A demonstration showed that PARA can be used. This protocol specifies what it would take to show that PARA is worth using — that its outputs are good as food and as creative work, that the method serves chefs who did not build it, that it reduces error rather than merely producing acceptable results, and that the indigenous technique on which the idiom rests is, in the judgement of those who hold it, faithfully carried. A completed evaluation built on this protocol would license those claims to the extent its evidence supported them, and no further. It would not license a claim that the method learns, which it does not yet do, and it would not, from a single site, license a claim that the method travels. Those are the studies beyond this one. What this protocol offers is the turn from a worked example to gathered evidence, designed so that the evidence is produced by hands other than the designer's, and judged by those with the standing to judge it.

A note on this document's status

This is a research protocol. It reports no results, and none have been gathered. Sample sizes and design parameters are drawn from the norms of the cited methods and would be fixed at pre-registration. This is the second revision, strengthened after a steelman assessment; it awaits the author's final editorial pass.

References

Amabile, T. M. (1982) 'Social psychology of creativity: A consensual assessment technique', *Journal of Personality and Social Psychology*, 43(5), pp. 997–1013.

Amabile, T. M. (1996) *Creativity in Context: Update to the Social Psychology of Creativity*. Boulder, CO: Westview Press.

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., et al. (2020) 'The CARE Principles for Indigenous Data Governance', *Data Science Journal*, 19(1), 43, pp. 1–12.

Flyvbjerg, B. (2006) 'Five misunderstandings about case-study research', *Qualitative Inquiry*, 12(2), pp. 219–245.

Gregor, S. and Hevner, A. R. (2013) 'Positioning and presenting design science research for maximum impact', *MIS Quarterly*, 37(2), pp. 337–355.

Hevner, A. R., March, S. T., Park, J. and Ram, S. (2004) 'Design science in information systems research', *MIS Quarterly*, 28(1), pp. 75–105.

Lawless, H. T. and Heymann, H. (2010) *Sensory Evaluation of Food: Principles and Practices*. 2nd edn. New York: Springer.

Meilgaard, M. C., Civille, G. V. and Carr, B. T. (2016) *Sensory Evaluation Techniques*. 5th edn. Boca Raton, FL: CRC Press.

Peffers, K., Tuunanen, T., Rothenberger, M. A. and Chatterjee, S. (2007) 'A design science research methodology for information systems research', *Journal of Management Information Systems*, 24(3), pp. 45–77.

Secretariat of the Convention on Biological Diversity (2011) *Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity*. Montreal: Secretariat of the CBD.

UNESCO (2003) *Convention for the Safeguarding of the Intangible Cultural Heritage*. Paris: UNESCO.

United Nations (2007) *United Nations Declaration on the Rights of Indigenous Peoples*. New York: United Nations.

Venable, J., Pries-Heje, J. and Baskerville, R. (2016) 'FEDS: a framework for evaluation in design science research', *European Journal of Information Systems*, 25(1), pp. 77–89.

Yin, R. K. (2018) *Case Study Research and Applications: Design and Methods*. 6th edn. Thousand Oaks, CA: Sage.

Appendix — Instrument outlines

The instruments are sketched here at the level a research office would need to commission their full development. Each would be finalised and, where appropriate, piloted in Episode A before summative use.

Canon-compliance audit. A per-course checklist: protein class and base correct and uncrossed; standing exclusions observed; cultivar named and held; indigenous techniques attributed; verification gate completed with assumptions and unconfirmed figures flagged. Output: a breach count and rate per menu.

Coherence rubric. Arc integrity; fit of each course to the governing idea; soundness of pairing logic where present. Expert-rated on a fixed scale with anchored descriptors.

Feasibility rubric. Completeness of recipe specification; soundness of timings and holds; producibility under the stated brigade and conditions. Rated by an executing chef.

Sensory battery. Trained descriptive panel by quantitative descriptive analysis for the profile; guest panel on a nine-point hedonic scale for overall liking and just-about-right scales for the balance of key attributes.

Creativity panel (Consensual Assessment Technique). Independent domain experts rate creativity by their own implicit conceptions, separately rate technical quality and conceptual coherence, and hold creativity apart from technical execution; inter-rater reliability reported; raters blind to condition.

Conservation-fidelity rubric. Developed with the communities concerned: fidelity of the technique to the action of the traditional method; correctness and respect of attribution. Structured judgement supported by a qualitative account treated as primary.