Technical Note: Internal Service Quality, Pilot and Final Form

Instrument Development, stage 4, with the two-group invariance add-on. Sample deliverable on a synthetic pilot, prepared by Dr Milos Kankaras, Center for Psychology, September 2026.

This is a demonstration on synthetic data. The pilot export was simulated from the blueprint in the companion sample with a known answer built in (two dimensions, four facets that do not separate, four weak items and one item that behaves differently at one site), so that the reader can see the analysis find it. Every number below is the actual output of the analysis on that export; nothing is illustrative. A client receives this note on their own pilot.

1. Verdict

The pilot supports a 24-item instrument scored on two dimensions, Delivery and Relationship, with twelve items each. Both dimensions are reliable (alpha and omega 0.92 and 0.92), the two-dimension model fits the data closely (CFI 0.999, RMSEA 0.009), and a single-score model does not. The four facets are content areas, not separate scores: modelled separately they correlate at or above .99 within their dimension and the model is not admissible. Of the forty items fielded, four were dropped by the selection rule for the reasons the blind review had anticipated, and one more, CM7, was dropped after the two-site invariance test showed it answered differently at Site B than its dimension warranted; the next-ranked reserve item took its place. On the final form, full strict invariance holds across the two sites under both estimators, so the sites can be compared directly. They do not differ: Site B is within sampling error of Site A on both dimensions.

2. The pilot sample

384 respondents completed the forty-item form at two sites (Site A 204, Site B 180). The export arrived unreversed under survey-tool column names with a duration column, 2.2 percent of item cells missing, three out-of-range values and no attention check. Data preparation was rule-based and recorded: reverse-keyed items flipped once, out-of-range cells set missing, 6 respondents flagged as straight-liners (constant response as given) and 5 as speeders (under two seconds per item). Both groups were excluded before analysis, a rule set before the data were opened, leaving 373 respondents (Site A 199, Site B 174). 2 respondents carried a multivariate-outlier flag and were retained. Missing cells were handled by full-information estimation in the continuous models and pairwise in the ordinal ones.

3. Item analysis

Table 1 gives, for each fielded item, the mean, standard deviation, share of responses in the top and bottom categories, corrected item-total correlation within its intended dimension, standardised loading in the forty-item two-dimension confirmatory model, and its loading on the other dimension in the exploratory solution. The flag column applies the selection rule of section 5. Four items fail it:

RS16 and OW14 on loading and item-total, RL13 on cross-loading, CM4 on a ceiling (53 percent in the top category). All four were the items the blind review had marked as worth piloting rather than assuming.

Table 1. Item analysis of the forty fielded items (N = 373).

Item	Key	M	SD	low %	top %	r(it)	Loading	Other	Flag
RS1	+	3.48	0.99	3	15	0.69	0.76	0.09	
RS2	+	3.36	1.00	3	12	0.70	0.75	0.00	
RS3	-	3.26	1.00	3	11	0.65	0.70	0.00	
RS4	+	3.27	0.98	3	11	0.66	0.71	0.00	
RS5	-	3.15	0.95	4	6	0.68	0.73	-0.01	
RS6	+	3.38	1.01	3	14	0.58	0.61	-0.06	
RS8	-	3.19	0.99	4	8	0.59	0.66	0.09	
RS12	+	3.50	0.94	2	14	0.67	0.70	-0.10	
RS15	+	3.39	1.00	4	14	0.59	0.65	0.07	
RS16	-	3.00	0.95	6	5	0.31	0.34	0.03	loading 0.34; item-total 0.31
RL1	+	3.25	0.98	3	10	0.73	0.78	-0.01	
RL2	+	3.16	1.00	5	8	0.66	0.72	0.08	
RL3	-	3.24	1.00	4	10	0.67	0.72	0.02	
RL4	+	3.36	1.03	4	14	0.77	0.82	0.00	
RL5	-	3.14	0.97	4	8	0.58	0.62	-0.03	
RL6	+	3.54	0.89	1	15	0.64	0.72	0.07	
RL7	-	3.38	0.99	3	14	0.55	0.60	0.09	
RL8	+	3.31	0.96	3	11	0.60	0.66	0.06	
RL15	+	3.34	0.97	3	10	0.69	0.71	-0.07	
RL13	-	3.07	1.00	5	8	0.56	0.73	0.39	cross-loading 0.39
CM1	+	3.25	0.99	3	11	0.72	0.78	0.02	
CM2	+	3.07	1.00	5	7	0.61	0.67	0.07	
CM3	-	2.95	1.00	6	6	0.64	0.69	-0.01	
CM4	+	4.39	0.73	0	53	0.47	0.54	-0.07	ceiling 53%
CM5	-	3.51	0.99	2	17	0.56	0.62	0.05	
CM6	+	3.27	1.02	3	12	0.70	0.75	-0.01	
CM7	+	3.30	1.02	4	13	0.68	0.74	0.04	
CM8	-	3.30	1.05	5	12	0.77	0.84	0.04	
CM12	+	3.36	1.02	3	13	0.66	0.70	-0.12	

Item	Key	M	SD	low %	top %	r(it)	Loading	Other	Flag
CM13	+	2.88	1.00	8	5	0.67	0.70	-0.05	
OW1	+	3.13	1.00	4	9	0.71	0.77	0.07	
OW2	-	3.04	0.98	4	8	0.71	0.77	-0.01	
OW3	+	3.24	0.98	4	9	0.64	0.69	0.02	
OW4	-	3.14	0.94	4	7	0.64	0.71	0.03	
OW5	+	3.21	1.03	6	10	0.65	0.72	0.11	
OW6	+	3.07	0.95	4	6	0.63	0.69	0.01	
OW7	-	3.37	0.96	2	12	0.56	0.62	0.05	
OW9	+	3.29	1.01	4	12	0.63	0.68	-0.01	
OW12	+	3.28	0.98	3	12	0.71	0.77	-0.03	
OW14	+	2.58	0.99	15	2	0.31	0.36	0.08	loading 0.36; item-total 0.31

Note. Loading: standardised loading on the intended dimension, WLSMV, forty-item model (CFI 0.990, RMSEA 0.022, dimension correlation 0.64). Other: pattern loading on the other dimension, two-factor exploratory solution with oblimin rotation on polychoric correlations.

4. Dimensionality

Parallel analysis on the polychoric correlations of the forty items retains 2 factors: the first five eigenvalues are 15.61, 3.58, 0.31, 0.28, 0.27, and only the first two exceed their random counterparts. The two-factor exploratory solution explains 48 percent of the variance, the factors correlate at 0.59, and every Delivery item loads on one factor and every Relationship item on the other, with the single exception of RL13, whose loadings split. The blueprint's two dimensions are what the data return.

5. Selection of the final form

The rule, stated before the analysis: an item is eligible if its standardised loading is at least .50, its corrected item-total at least .40, its loading on the other dimension below .30 and neither extreme category holds more than half the responses. From the eligible items the six highest-loading per facet are taken, with at least two reverse-keyed per facet. 36 of forty items were eligible; the rule selected a 24-item form and left 12 items in reserve (RS6, RS8, RS15, RL7, RL8, RL15, CM2, CM5, CM13, OW6, OW7, OW9).

Table 2. Items dropped at selection

Item	Facet	Reason	Stem
RS16	Responsiveness	loading 0.34; item-total 0.31	Simple requests take as long as complicated ones.
RL13	Reliability	cross-loading 0.39	The outcome of a request depends on who happens to handle it.
CM4	Communication	ceiling 53%	The support team explains things in language I understand.

Item	Facet	Reason	Stem
OW14	Ownership	loading 0.36; item-total 0.31	The support team follows up after a fix to check that it worked.

6. Structure of the final form

Table 3 compares three models on the 24 retained items. The one-dimension model is rejected outright. The two-dimension model fits closely. The four-facet model fits no better and is not admissible: the facet correlations within each dimension are 0.99 (Responsiveness with Reliability) and 1.02 (Communication with Ownership), so the facets carry no variance of their own. Scores are therefore reported for the two dimensions, and the facets serve as the content map that keeps each dimension's item set balanced.

Table 3. Fit of alternative models for the 24-item final form (WLSMV, N = 373).

Model	chi2	df	CFI	TLI	RMSEA [90% CI]	SRMR	Admissible
M1 One dimension	1678.74	252	0.838	0.823	0.123 [0.118, 0.129]	0.112	yes
M2 Two dimensions (retained)	258.66	251	0.999	0.999	0.009 [0.000, 0.023]	0.034	yes
M3 Four facets	252.92	246	0.999	0.999	0.009 [0.000, 0.023]	0.033	no

7. Reliability

Table 4. Reliability and descriptives of the two dimension scores, final form.

Dimension	Items	alpha	omega	M	SD
Delivery	12	0.920	0.920	3.31	0.71
Relationship	12	0.922	0.924	3.17	0.74

Note. Scores are item means on the 1 to 5 scale after reverse-keying. Omega from the single-dimension continuous model.

8. Two-site invariance and the replacement step

The cascade was run with the robust continuous estimator and full-information missing-data handling, because with about 174 respondents per site the bottom category of several items is empty at one site, which the ordinal estimator cannot carry across groups; five ordered categories under robust maximum likelihood is the documented fallback (Rhemtulla et al., 2012). An ordinal sensitivity run with the two lowest categories merged is reported at the end of the section. Decision rule: a more constrained model is retained when CFI falls by no more than .010 and RMSEA rises by no more than .015 (Chen, 2007).

On the form as first selected (Table 5), loadings are invariant and scalar invariance fails; the score test locates the misfit in the intercept of CM7, and freeing it restores fit (partial scalar). The item reads "I can see the status of my request at any time.". Its intercept is 3.05 at Site A and 3.60 at Site B, against near-identical means on every other Communication item: Site B respondents endorse it

more than their Relationship score warrants, which is what a site with a status-tracking portal produces. A reviewer's note on the blind sheet had raised exactly this possibility. Rather than carry a freed intercept into an operational instrument, CM7 was dropped and the next-ranked reserve item in its facet, CM13, took its place.

Table 5. Invariance across sites, form as first selected (MLR, N = 373).

Model	chi2	df	CFI	RMSEA [90% CI]	dCFI	dRMSEA	Decision
M1 Configural	521.52	502	0.995	0.015 [0.000, 0.030]	-	-	reference
M2 Metric	541.09	524	0.996	0.013 [0.000, 0.029]	+0.001	-0.002	supported
M3 Scalar	626.57	546	0.982	0.029 [0.015, 0.039]	-0.014	+0.015	rejected
M4 Partial scalar (CM7 intercept freed)	556.71	545	0.997	0.011 [0.000, 0.027]	+0.001	-0.002	partial
M5 Strict	662.05	570	0.980	0.029 [0.017, 0.039]	-0.018	+0.018	rejected

Note. dCFI and dRMSEA relative to the last retained model. Strict invariance was not reached on this form within five releases; it is not required for mean comparison and the form was not carried forward.

Table 6. Invariance across sites, final form (MLR, N = 373).

Model	chi2	df	CFI	RMSEA [90% CI]	dCFI	dRMSEA	Decision
M1 Configural	506.83	502	0.999	0.008 [0.000, 0.026]	-	-	reference
M2 Metric	526.50	524	0.999	0.006 [0.000, 0.025]	+0.001	-0.003	supported
M3 Scalar	542.69	546	1.000	0.000 [0.000, 0.024]	+0.001	-0.006	supported
M4 Strict	577.48	570	0.999	0.008 [0.000, 0.025]	-0.001	+0.008	supported

On the final form every level holds, strict included. The ordinal sensitivity run (WLSMV, categories 1 and 2 merged) gives the same verdict at every step (dCFI +0.003, +0.000 and +0.000 for metric, scalar and strict), so the conclusion does not depend on the estimator.

9. Site comparison

Under the scalar-invariant model with Site A as reference, Site B's latent mean is +0.03 (SE 0.08, p = 0.69, d = +0.05) on Delivery and -0.09 (SE 0.09, p = 0.34, d = -0.10) on Relationship. Neither differs from zero. Observed dimension means tell the same story (Delivery 3.29 and 3.33, Relationship 3.20 and 3.13 at Sites A and B). Had CM7 been kept, Site B's Relationship total would have read higher than its construct level by the item's own shift.

10. Final form and scoring key

Twenty-four items, six per facet, twelve per dimension, eight reverse-keyed. Administer in the order below or randomised within the whole form; do not block by facet. Reverse-keyed items (R) are scored 6 minus the response. Each dimension score is the mean of its twelve items, on the 1 to 5 scale, computed when at least ten of the twelve are answered. An overall Internal Service Quality index, where one number is needed, is the mean of the two dimension scores; report the two

dimensions wherever the reader can take two numbers. No norms are claimed: scores are for comparison across units, sites and waves within the organisation, not against an external reference.

Table 7. Final form: items, keying and standardised loadings (two-dimension model).

No.	Id	Dimension	Facet	R	Loading	Stem
1	RS1	Delivery	Responsiveness		0.77	When I raise a request, the support team acknowledges it the same day.
2	RS2	Delivery	Responsiveness		0.75	It is easy to reach someone on the support team when I need to.
3	RS5	Delivery	Responsiveness	R	0.74	I have to chase the support team to get any response.
4	RS4	Delivery	Responsiveness		0.72	The support team starts work on my request soon after I send it.
5	RS3	Delivery	Responsiveness	R	0.71	My requests sit unanswered for days.
6	RS12	Delivery	Responsiveness		0.71	The support team replies to my messages promptly.
7	RL4	Delivery	Reliability		0.83	I can rely on the support team to follow through on what it agrees.
8	RL1	Delivery	Reliability		0.79	When the support team says something will be done by a date, it is done by that date.
9	RL2	Delivery	Reliability		0.73	The support team gets my request right the first time.
10	RL3	Delivery	Reliability	R	0.72	I often have to go back to the support team because the fix did not work.
11	RL6	Delivery	Reliability		0.73	The support team's work is accurate.
12	RL5	Delivery	Reliability	R	0.63	Promised deadlines slip without warning.
13	CM8	Relationship	Communication	R	0.84	I am left guessing about where my request stands.
14	CM1	Relationship	Communication		0.78	The support team tells me what will happen next with my request.
15	CM6	Relationship	Communication		0.75	When my request is closed, I am told what was done.
16	CM13	Relationship	Communication		0.69	Updates from the support team arrive without my having to ask for them.
17	CM12	Relationship	Communication		0.70	If my request cannot be done, the support team tells me why.
18	CM3	Relationship	Communication	R	0.69	I only find out about delays when I ask.
19	OW1	Relationship	Ownership		0.77	One person takes charge of my request until it is resolved.
20	OW2	Relationship	Ownership	R	0.77	My request gets passed from person to person.
21	OW12	Relationship	Ownership		0.77	The support team sees my request through to the end.
22	OW5	Relationship	Ownership		0.72	When something goes wrong, the support team takes responsibility.
23	OW4	Relationship	Ownership	R	0.71	I have to re-explain my request each time someone new picks it up.
24	OW3	Relationship	Ownership		0.70	The person handling my request keeps it moving even when others are involved.

11. Methods

Data preparation followed a fixed, recorded procedure (reverse-keying, range check, straight-line and speeder flags, missingness), with the resolved specification kept. Analyses ran in R 4.5.3 with lavaan 0.6.21, semTools 0.5.8 and psych 2.6.1. Polychoric correlations, parallel analysis (minimum residual, fifty random datasets) and the exploratory solution (oblimin) used psych. Confirmatory models used WLSMV with theta parameterisation and pairwise missing handling; fit is reported as scaled CFI, TLI and RMSEA with its 90 percent interval, and SRMR. The invariance cascades used MLR with full-information maximum likelihood (primary) and WLSMV with the Wu and Estabrook (2016) identification on merged categories (sensitivity); when a level failed, the score test identified the worst constraint and it was released one at a time, to a maximum of five, until the criterion was met. Reliability is coefficient alpha and McDonald's omega. The selection rule and the replace-and-rerun step are stated in sections 5 and 8. The export, the data-preparation specification, the analysis code and every model setting are supplied with this note so any number can be re-run.

12. Limits

One pilot, one occasion, one organisation: the structure and reliability reported here are evidence from a single sample and were estimated on the same respondents used to select the items, which flatters fit. The first operational wave should re-fit the two-dimension model on new data before the scores are used for decisions. Nothing here speaks to whether the scores predict anything outside the questionnaire, and no cut score or norm is implied. The site comparison is valid for these two sites; a third site needs its own invariance test.

References

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. Structural Equation Modeling, 14(3), 464-504. https://doi.org/10.1080/10705510701301834

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. Psychological Methods, 17(3), 354-373. https://doi.org/10.1037/a0029315

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. Psychometrika, 81(4), 1014-1045. https://doi.org/10.1007/s11336-016-9506-0

Psychometrician's verdict

The 24-item form measures two things, how dependably the support function delivers and how it treats the person asking, and measures them reliably and in the same way at both sites. Its scores can be used now to compare units, sites and waves within the organisation. They should not be read against any external benchmark, and their structure should be confirmed on the first operational wave before they inform decisions about individuals or teams.

Dr Milos Kankaras, Founder and Director, Center for Psychology

Specimen: a client report carries the reviewer's dated signature here.

Dr Milos Kankaras, Founder and Director, Center for Psychology | milos@centerforpsychology.me | miloskankaras.com | centerforpsychology.me | ORCID | Google Scholar | LinkedIn