Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

From SERVQUAL to SIR-Quality (Service- Intelligence-Resilience): A Conceptual Framework for Hospital Support Service Quality and Evidence of Stakeholder Measurement Equivalence

Youli Wang
The First Affiliated Hospital with Nanjing Medical University
Anrong Fu
The First Affiliated Hospital with Nanjing Medical University
Yuwei Ding
The First Affiliated Hospital with Nanjing Medical University
Yubin ZHANG
The First Affiliated Hospital with Nanjing Medical University
Yang CAO
caoyang@jsph.org.cn

The First Affiliated Hospital with Nanjing Medical University

Research Article

Keywords: hospital support services, outsourcing, SERVQUAL, SERVPERF, measurement invariance, multi-stakeholder evaluation, service quality, China

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11111052/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background. Hospital support services are increasingly outsourced, yet the instruments used to evaluate them remain adaptations of SERVQUAL, a model developed for commercial, customer-facing services. Two problems follow: the expectation-perception gap paradigm is fragile when respondents are not customers, and it is rarely established whether stakeholder groups interpret the instrument similarly.

Methods. We surveyed 624 respondents (physicians, nurses, administrative managers, and patients or family members) at the 4,700-bed main campus of a tertiary public hospital in eastern China, where support services are outsourced under contract. Responses to the 22 perception items of a Delphi-derived SERVQUAL-type instrument were analysed. Data-quality diagnostics led us to adopt a perception-only (SERVPERF) specification. We estimated a five-factor measurement model, tested measurement invariance between internal staff and patients/families, compared stakeholder groups, and derived improvement priorities from a second-order model.

Results. The five-factor structure fitted well (chi-square = 520.6, df = 200, CFI = 0.986, TLI = 0.984, RMSEA = 0.051, SRMR = 0.025), with standardised loadings of 0.90–0.97 and no Heywood cases. Internal consistency was high (Cronbach's alpha 0.96–0.99) and discriminant validity held (maximum HTMT 0.327). Full scalar invariance was established between staff and patients/families (metric step: delta CFI = -0.0005; scalar step: delta CFI = 0.0000), indicating that the instrument functions equivalently across these groups. Perceptions did not differ significantly across the four stakeholder groups on any dimension (all $p > 0.05$; eta-squared $< = 0.007$). In the second-order model, empathy carried the largest share of importance (27.2%) and responsiveness the smallest (13.7%), while responsiveness also recorded the lowest performance score.

Conclusions. Support-service quality in this setting is a coherent, hierarchically organised and stakeholder-invariant construct, which means that multi-stakeholder surveys can be pooled and compared meaningfully. The findings inform the conceptual SIR-Quality framework (Service-Intelligence-Resilience) proposed here, which extends the classic service dimensions with intelligence and resilience domains that the present instrument measures only partially. We argue that expectation-minus-perception gap scores should be abandoned in this population.

1. Introduction

Support services are the hidden infrastructure of hospital care. Cleaning, waste handling, security, portering, linen, equipment maintenance and estate management determine infection risk, patient experience and the reliability of clinical workflows, yet they are frequently outsourced to commercial contractors and evaluated with instruments borrowed from retailing and hospitality [1, 2]. In China's public hospital sector this gap has become consequential: since the 2002 guidance on socialising hospital logistics services, and with the 2021 State Council opinions on high-quality public hospital development explicitly calling for digitalised, energy-accountable logistics management, hospital administrators are expected to govern contracted services against measurable standards rather than by inspection alone.

The dominant instrument for this task remains SERVQUAL [1]. Its logic - quality as the gap between what service users expect and what they perceive they received - is elegant, and its five dimensions (tangibles, reliability, responsiveness, assurance, empathy) have proved durable. But three difficulties accumulate when SERVQUAL is transplanted into hospital support services.

First, the respondent is not a customer. Hospital support services are contracted by the institution and consumed by everyone inside it, so the relevant evaluators are multiple and their relationships to the service differ: clinical staff depend on it, managers procure and monitor it, and patients experience its residue. Second, the expectation anchor is unstable in this population. Expectation items ask respondents to imagine an ideal contractor for services whose technical standards they cannot observe - sterilant concentrations, energy targets, information-security protocols - which invites acquiescent or default responding rather than considered judgement [3, 4]. Third, and most consequential for cumulative research, studies rarely test whether different stakeholder groups construe the items in the same way. Without measurement invariance, mean

comparisons across groups are uninterpretable, and the substantive question of whether staff and patients disagree about service quality cannot be answered at all.

This study addresses these problems using data from a large multi-stakeholder survey at a tertiary public hospital in China. We make four contributions. We show that a perception-only (SERVPERF) specification of a Delphi-derived support-service instrument has a sound measurement structure, and we document transparently why the expectation arm of the same instrument could not be used. We establish full scalar measurement invariance between internal staff and patients or family members, which is the precondition for treating multi-stakeholder surveys as comparable evidence. We show that the four stakeholder groups do not differ significantly in their perceptions, a well-powered null result that constrains how much of the literature's disagreement about service quality reflects genuine stakeholder divergence. Finally, we use these results to motivate the SIR-Quality framework, which argues that support-service quality in contemporary hospitals must be conceptualised along service, intelligence and resilience domains.

2. Theoretical background

2.1 From gap scores to performance perception

The critique of SERVQUAL's expectation-minus-perception operationalisation is old and largely settled in the service management literature. Cronin and Taylor [3, 4] argued that performance-only measures (SERVPERF) explain variance in service quality at least as well as, and often better than, difference scores, while avoiding the reliability problems that arise when two uncertain quantities are subtracted. Difference scores combine the measurement error of both components, are sensitive to the scale anchor respondents use for expectations, and are prone to ceiling effects when expectations are uniformly high. Subsequent work in health care reached similar conclusions, and performance-only instruments are now common in hospital settings [3, 4].

2.2 Multi-stakeholder evaluation and service-dominant logic

Service-dominant logic reframes service quality as value determined in use rather than embedded in output [6, 7, 8]. In an outsourced hospital support service, value is co-created: contractors deliver, clinical staff integrate the service into their work, managers specify and monitor it, and patients experience its consequences. Technology, in this framing, is an operant resource that enables co-production rather than a feature of the offering [9]. This theoretical move has a direct methodological implication: if value is phenomenologically located in different actors, then instrument equivalence across those actors is not a technical nicety but a substantive requirement. An instrument that means different things to a nurse and to a patient cannot adjudicate whether they disagree.

2.3 Hierarchical models of health service quality

Health service quality is routinely modelled hierarchically, with overall evaluations determined by more specific attribute-level judgements [5]. Dagger, Sweeney and Johnson [5] showed that such hierarchies are empirically tractable and that attribute-level performance has differential effects on overall evaluations, a result that underlies importance-performance analysis. Parallel work extended classical service quality constructs into technology-mediated settings, where dimensions such as efficiency, system availability, privacy and fulfilment join the original five [10, 11, 12, 26].

2.4 The intelligence and resilience turns

Two developments in hospital operations are not captured by the classical dimensions. The first is digitalisation: building management systems, sensor-based environmental monitoring, automated dispatching and data-driven contracting are changing what support service quality consists of and how it is verified [13, 14, 19, 25]. The second is continuity under disruption. Hospitals must sustain environmental services during surges, outbreaks, utility failures and extreme weather, and support-service contractors are typically contractually bound to participate in emergency response [15, 16, 21, 22, 23]. Resilience research has begun to connect organisational resilience to service quality outcomes [17, 24], but no instrument for hospital support services measures it systematically.

2.5 The present study

Against this background we pose four research questions:

RQ1. Does the perception-only five-factor structure of the Delphi-derived instrument hold in a large multi-stakeholder sample?

RQ2. Is the instrument measurement-invariant across internal staff and patients or family members?

RQ3. Do the four stakeholder groups differ in their perceptions of support service quality?

RQ4. Which dimensions contribute most to overall support service quality, and which perform least well?

3. The SIR-Quality framework

We propose SIR-Quality as a conceptual framework rather than a validated scale, and we are explicit about the boundary between what this study tests and what it proposes. Specifically, only the Service (S) domain is empirically examined here; the Intelligence (I) and Resilience (R) domains are conceptual and serve as a research agenda, motivated by literature and by the fragmentary item content of the present instrument but not measured as latent factors.

Service (S). The classical five dimensions, operationalised here by 22 indicators developed through Delphi consultation in this hospital setting. This domain is empirically examined in the present study.

Intelligence (I). The capacity of the support service system to generate, protect and act on information: sensor-based monitoring of environmental conditions, data-driven work allocation, digital traceability of tasks, and - critically in a hospital - information security and confidentiality. Items in existing instruments gesture at this domain (for example, one item in the present instrument asks about in-hospital information confidentiality), but no instrument treats it as a dimension.

Resilience (R). The capacity to maintain and rapidly restore service under disruption [24, 28]: emergency response participation, continuity of environmental services during public health emergencies, and adaptability of staffing and supply. Again, isolated items exist in the present instrument (fire-safety and energy management awareness; participation in emergency response; capacity to handle public health emergencies), but they are scattered across the classical dimensions rather than organised as one.

The framework's claim is therefore modest and specific: support service quality in digitally managed, disruption-prone hospitals is conceptually a three-domain construct, of which current instruments measure only the first domain adequately and the other two only in fragments. We do not claim that I and R have been demonstrated to be material dimensions of perceived service quality; this study provides no data bearing on that question and the present manuscript should not be cited as evidence that they are. We develop the implications of this claim for measurement practice in the Discussion, and we deliberately refrain from post hoc re-specification of the present data into three factors, since doing so with items designed for a five-dimensional model would be circular.

4. Methods

This cross-sectional survey followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guidelines. ---

Figure 1 presents the SIR-Quality (Service–Intelligence–Resilience) framework, distinguishing the domain examined in this study (solid border) from the two proposed domains (dashed border).

4.1 Setting

The study was conducted at Hospital J, a tertiary general hospital affiliated to a provincial authority in eastern China. The campus studied has 4,700 authorised beds and recorded 6.155 million outpatient and emergency visits in 2025, with approximately 16,900 visits per day and 302,000 discharges. Support services are outsourced to two commercial contractors under contracts running from 2025 to 2027, covering ward and operating-theatre cleaning, emergency and outpatient department cleaning, and lift services across the campus. Outsourcing is governed by the hospital's property management department through scheduled and unscheduled inspection and contract key-performance-indicator assessments.

4.2 Sample and procedure

The survey was conducted between July and August 2025. Questionnaires were distributed to 650 respondents and 650 were returned; 624 were valid after completeness screening, giving a valid response rate of 96.0%. Participation was anonymous and voluntary. Respondents were recruited by convenience with a deliberate emphasis on hospital staff, because several indicators concern professional domains - infection-control specifications, energy management targets, contractor qualifications, information-security protocols - that patients and family members cannot be expected to assess directly. Where patients or family members participated, they were told that such items could be skipped or answered on the basis of the outcomes they observed rather than the technical standards themselves.

4.3 Instrument

The instrument comprises three sections: demographic characteristics (sex, age group, education, role); the core scale; and an open comments field. The core scale contains 22 indicators across five dimensions - tangibles (5 items), reliability (4), responsiveness (6), assurance (5) and empathy (2) - developed through Delphi consultation for this hospital setting. The Delphi process was conducted in two rounds with eight panellists (three hospital-property managers, two infection-control nurses, two quality-management physicians and one logistics contractor representative), with the coefficient of variation on item inclusion falling below 0.20 in the second round; consultation was registered with the hospital's property management department and informed consent was obtained from panellists. Each indicator was originally presented twice, once for expected and once for perceived performance, both rated on five-point Likert scales. Demographic characteristics were recorded as sex (male/female), age group, education and role. The full 22-item instrument is provided as Supplementary File 1.

4.4 Data quality diagnostics and the decision to use perception-only scores

Before analysis we screened the data. Two findings shaped the analytic strategy, and we report them because they are material to how such studies should be designed.

First, the expectation arm of the instrument was severely compromised. Of the 624 respondents, 424 (68.0%) gave an identical response to all 22 expectation items, and 73.6% of all expectation responses were at the scale maximum - a pronounced ceiling effect. The pattern was not uniform across the dataset: among the first 162 questionnaires the straight-lining rate was 1.2%, whereas among the remainder it was 91.3%, indicating that the problem was introduced at the data collection or data preparation stage for the majority of cases rather than reflecting a genuine respondent tendency. Because expectation scores of this kind cannot support inference, we did not compute expectation-minus-perception gap scores, and all analyses reported below use the perception items only.

Second, within-dimension responses showed high homogeneity. Mean inter-item correlation within dimensions was approximately 0.92, whereas between-dimension correlations were only 0.12–0.33. Item-rest correlations within dimensions ranged from 0.88 to 0.96, with the ratio of the largest to smallest item variance within each dimension below 1.4, confirming that respondents discriminate very little among the items within a dimension. Consequently the very high internal consistency coefficients reported below should be read as evidence of coherent responding within dimensions rather than as evidence of exceptional measurement precision; the same coefficients re-estimated on the subsample that excluded the 83 straight-lining respondents (Section 5.7) range from 0.957 to 0.985, confirming that the homogeneity is in the data rather than an artefact of the analytic convention. We return to this in the limitations. Common method bias was assessed with Harman's single-factor

test on the perception items [20]; the first unrotated factor accounted for 22.7% of the variance, below the conventional 50% threshold.

Third, beyond the expectation-arm and homogeneity issues, two further data-quality features are reported transparently. Duplicate responding was substantial: 143 respondents (22.9%) produced an identical 44-item response pattern to an earlier record, and 32 records were fully duplicated across all fields. The two collection batches also differed: batch 1 (id < = 162, N = 162) and batch 2 (N = 462) showed significant mean differences on Tangibility (t=-2.66, p = 0.008) and on overall service quality (t=-2.04, p = 0.042), and batch 1 showed poorer confirmatory-factor-analysis fit (RMSEA = 0.083) than batch 2 (RMSEA = 0.053). A batch-by-role interaction on overall quality was also significant (F = 5.22, p = 0.0014). These batch- and duplication-related features are reported here as data-quality limitations and sensitivity checks; they are not used to merge, split, or trim the sample, and the multi-group comparisons in Sections 5.4–5.5 are interpreted with this caveat in mind.

4.5 Analytic strategy

Figure 2 presents the participant flow, data triage, and the analytic sequence; all analyses use the perception items only.

Analyses were conducted in Python. We first estimated a five-factor confirmatory factor model on the 22 perception items using robust maximum likelihood, treating the two-item empathy factor as identified by constraining its two loadings to equality. Model fit was evaluated with CFI, TLI, RMSEA (with 90% confidence interval) and SRMR, using conventional thresholds [29] (CFI and TLI > = 0.90, ideally > = 0.95; RMSEA < = 0.08, ideally < = 0.06; SRMR < = 0.08). Because some fit indices were computed by two independent implementations as a check, we note that chi-square, CFI, TLI and RMSEA were reproduced to within 0.001 by an independent maximum-likelihood implementation; SRMR is reported from the standard formula.

We then computed internal consistency (Cronbach's alpha), composite reliability, average variance extracted and the heterotrait-monotrait ratio of correlations for discriminant validity, both for the full sample and after excluding respondents who straight-lined the perception items. Measurement invariance between internal staff (physicians, nurses and administrative managers, n = 526) and patients or family members (n = 98) was tested by fitting configural, metric and scalar models sequentially and evaluating changes in CFI and RMSEA using the thresholds of Chen [18] (delta CFI not exceeding −0.010 and delta RMSEA not exceeding 0.015). Group comparisons used one-way analysis of variance with Kruskal-Wallis tests as a distribution-free check and eta-squared as an effect size. Improvement priorities were derived from a second-order model in which a single higher-order quality factor explained the covariation among the five dimensions; dimension importance was taken as the standardised second-order loading, rescaled to sum to 100%, and dimension performance as the dimension mean rescaled from the 1–5 metric to 0-100. Finally, we repeated the measurement model and group comparisons after excluding straight-lining respondents, and compared the two data batches.

5. Results

5.1 Sample characteristics

Table 1 reports the characteristics of the 624 valid respondents.

Table 1
Sample characteristics (N = 624)

Characteristic	Category	n	%
Sex	Male	186	29.81
	Female	438	70.19
Age group	≤ 18	5	0.80
	18–30	114	18.27
	31–40	132	21.15
	41–50	189	30.29
	51–60	108	17.31
	> 60	76	12.18
Education	College diploma or below	345	55.29
	Bachelor's degree	228	36.54
	Master's degree	39	6.25
	Doctoral degree	12	1.92
Role	Physicians	43	6.89
	Nurses	391	62.66
	Administrative managers	92	14.74
	Patients or family members	98	15.71

5.2 Measurement model

The five-factor model fitted the perception data well. Table 2 reports fit statistics; Table 3 reports the standardised loadings.

Table 2
Fit of the measurement models

Model	chi-square	df	chi-square/df	CFI	TLI	RMSEA (90% CI)	SRMR	N
M0: five-factor, perception items	520.61	200	2.60	0.986	0.984	0.051 (0.0454, 0.0561)	0.025	624
Second-order (overall quality -> five dimensions)	520.12	205	2.54	0.986	0.986	0.050 (0.044–0.055)	0.026	624

Table 3
Standardised factor loadings

Dimension	Items	Standardised loading range
Tangible	5	0.904–0.969
Reliability	4	0.909–0.962
Responsiveness	6	0.953–0.964
Assurance	5	0.956–0.969
Empathy	2	0.961–0.961

All standardised loadings exceeded 0.90 and no Heywood case (negative residual variance) occurred. Factor inter-correlations ranged from 0.120 (Responsiveness-Assurance) to 0.329 (Tangible-Empathy).

5.3 Reliability and validity

Table 4 presents internal consistency, convergent validity, and discriminant validity for the five dimensions.

Table 4
Internal consistency, convergent and discriminant validity

Dimension	Items	Cronbach's alpha	Composite reliability	AVE
Tangible	5	0.980	0.980	0.908
Reliability	4	0.969	0.969	0.885
Responsiveness	6	0.986	0.986	0.922
Assurance	5	0.985	0.985	0.927
Empathy	2	0.960	0.960	0.923

Average variance extracted exceeded 0.50 for every dimension and the maximum heterotrait-monotrait ratio was 0.327, well below the 0.85 threshold [30], so discriminant validity was established. Results were unchanged after excluding the 83 respondents who gave identical answers to all perception items: alpha ranged from 0.957 to 0.985, AVE from 0.878 to 0.922, and maximum HTMT was 0.282.

5.4 Measurement invariance between staff and patients or family members

Table 5 presents the tests of measurement invariance between internal staff and patients or family members.

Table 5
Tests of measurement invariance (staff n = 526; patients or family members n = 98)

Model	chi-square	df	CFI	RMSEA (90% CI)	SRMR	Delta chi-square	Delta df	p	Delta CFI	Delta RMSEA
1. configural	897.43	400	0.9783	0.0447 [0.0408, 0.0486]	0.0311	-	-	-	-	-
2. Metric (loadings equal)	924.06	416	0.9779	0.0443 [0.0405, 0.0482]	0.0339	26.63	16	0.046	-0.0005	-0.0004
3. Scalar (loadings and intercepts equal)	940.44	433	0.9779	0.0434 [0.0396, 0.0472]	0.0339	16.38	17	0.497	+ 0.0000	-0.0009

Both invariance steps satisfied the Chen [18] criteria by a wide margin: constraining loadings to equality reduced CFI by 0.0005, and additionally constraining intercepts changed it by 0.0000. Full scalar invariance therefore holds, which licenses the comparison of latent means between internal staff and patients or family members. Following Meade (2007), the smaller subgroup (n = 98) had approximately 70% power to detect a non-invariance of ΔCFI = 0.01 under conventional thresholds; the observed value (ΔCFI = 0.0000) lies well inside the supported range, so the invariance finding is consistent with adequate power for the detected effect size, although invariance tests in subgroups of this size remain underpowered to detect smaller departures that may be of theoretical interest.

5.5 Comparison across stakeholder groups

Table 6 presents perceived quality by stakeholder group.

Table 6
Perceived quality by stakeholder group

Dimension	Physicians M (SD)	Nurses M (SD)	Administrative M (SD)	Patients/families M (SD)	F	p	Kruskal-Wallis p	eta-squared
Tangible	4.177 (1.172)	3.998 (1.355)	3.935 (1.190)	4.006 (1.034)	0.356	0.785	0.115	0.0017
Reliability	4.087 (1.346)	3.927 (1.356)	4.016 (1.205)	4.120 (0.959)	0.739	0.529	0.839	0.0036
Responsiveness	3.574 (1.550)	3.834 (1.377)	4.029 (1.146)	3.968 (1.268)	1.395	0.243	0.753	0.0067
Assurance	3.912 (1.416)	3.999 (1.293)	3.846 (1.290)	3.796 (1.348)	0.825	0.481	0.218	0.0040
Empathy	3.953 (1.388)	4.156 (1.232)	3.908 (1.299)	4.117 (1.079)	1.229	0.298	0.086	0.0059
Overall	3.915 (1.004)	3.955 (0.777)	3.953 (0.878)	3.979 (0.687)	0.064	0.979	0.886	0.0003

No dimension, and neither the overall score, differed significantly across the four groups. Welch's one-way ANOVA (which does not assume equal variances) and the Brown-Forsythe F* statistic produced the same pattern as the standard F test, with all ps values > 0.20; Levene's test was significant for Tangible (p = 0.04) and Responsiveness (p = 0.02), reflecting the heterogeneity of within-group variance rather than between-group mean differences. The patient/family subsample was deliberately smaller than the staff subsample; sensitivity analyses with the patient/family respondents upweighted by a factor of 1.27 (to reflect their 15.7% share against a target of 25%) produced the same null result on every dimension. Effect sizes were negligible (eta-squared between 0.0003 and 0.0067), and the non-parametric tests agreed. With N = 624, one-way ANOVA had 80% power (α = 0.05, two-tailed) to detect an effect size of f = 0.12 ($\eta^2 \approx 0.014$) or larger; the observed η^2 values fall well below this threshold, so the null finding is consistent with adequate power to detect effects of substantive interest rather than an artefact of an underpowered design.

5.6 Improvement priorities

Table 7 presents the importance-performance summary from the second-order model.

Table 7
Importance-performance summary (second-order model)

Dimension	Performance (1–5)	Performance (0-100)	Importance (standardised second-order loading)	Importance share (%)	Quadrant
Tangible	4.003	75.06	0.511	21.41	Keep up the good work
Reliability	3.982	74.54	0.483	20.23	Keep up the good work
Responsiveness	3.866	71.65	0.328	13.74	Low priority
Assurance	3.938	73.46	0.415	17.37	Low priority
Empathy	4.099	77.48	0.650	27.24	Keep up the good work

The second-order model fitted as well as the first-order model despite using five fewer degrees of freedom (delta chi-square = 0.49 on 5 degrees of freedom, p > 0.99), so a single higher-order quality factor adequately explains the covariation among the five dimensions. Empathy carried the largest share of importance (27.2%) [31] and responsiveness the smallest (13.7%).

Performance was highest for empathy and lowest for responsiveness. Using the median as the split point, tangibles, reliability and empathy fell in the keep-up-the-good-work quadrant, while responsiveness and assurance fell in the low priority quadrant; using the mean instead produced the same assignment (mean performance = 74.4; mean importance = 20.0%), so the pattern is not an artefact of the threshold choice.

5.7 Sensitivity analyses

Figure 3 presents the importance-performance map of the five SERVQUAL dimensions; importance is taken from the standardised second-order loadings (rescaled to 100%) and performance from the dimension means (rescaled to 0–100).

Excluding the 83 respondents who straight-lined the perception items left the measurement model essentially unchanged (CFI = 0.986, RMSEA = 0.050, SRMR = 0.027, n = 541). The null stakeholder differences also held, with one exception: empathy became marginally significant (F = 2.946, p = 0.032), which we interpret cautiously given the number of comparisons. Comparing the two data batches, respondents in the first batch rated tangibles lower than those in the second (3.742 versus 4.094, p = 0.008, Cohen's d = -0.279) and the overall score lower (p = 0.042, d = -0.140); no other dimension differed. These differences are small and do not affect the substantive conclusions, but they confirm that the data are not homogeneous in origin, as the diagnostics in section 4.4 indicated.

6. Discussion

6.1 Principal findings

Three findings are worth stating plainly. The perception-only five-factor structure of this support-service instrument is sound: fit was good by every conventional criterion, loadings were uniformly strong, and reliability and discriminant validity were established. The instrument is fully scalar-invariant between hospital staff and patients or family members, which means both groups relate the items to the underlying constructs in the same way and their scores can be compared directly. And the four stakeholder groups did not differ, on any dimension, in how they rated the service.

The last finding deserves emphasis because it is a null result and null results are usually under-reported. A common assumption in hospital governance is that managers see the service differently from clinicians, and clinicians differently from patients, and that these divergences must be reconciled. In this setting, at this hospital, with this instrument, they do not diverge in any measurable way. Combined with scalar invariance, this is informative rather than merely negative: the disagreement is not about how good the service is, and any apparent conflict in governance forums may reflect differences in priorities or in willingness to voice complaints than differences in evaluation.

6.2 Implications for the SIR-Quality framework

The empirical results support the service domain of the framework and sharpen the case for the other two. The five classical dimensions are measured well, but their inter-correlations are modest (0.12–0.33) and a single higher-order factor explains their covariation parsimoniously. This is exactly the structure one would expect if the dimensions were facets of a broader construct whose other facets are missing. Two of the twenty-two indicators already concern emergency response and public health emergency capacity, and one concerns information confidentiality; these are the fragments of the resilience and intelligence domains that the classical model has no room to organise.

We want to be careful about what this does and does not license. It does not license a post hoc re-specification of these data into three factors: the items were written for five dimensions, and reassigning them after the fact would be circular. It does not license the claim that I or R are material dimensions of perceived service quality in this hospital: the present data do not bear on that question and the present manuscript should not be cited in support of such a claim. It does license a research programme: instruments for hospital support services should be developed with intelligence and resilience domains written in from the start, with enough items to be estimated as factors in their own right. Until such instruments exist, studies that report only classical service quality are measuring a construct that is increasingly only part of what hospitals buy.

6.3 Practical implications

Two implications follow for the governance of outsourced support services. First, multi-stakeholder surveys can be pooled. Because the instrument is scalar-invariant and the groups do not differ systematically, a hospital can survey clinical staff, managers and patients together and interpret aggregate scores without worrying that the aggregation mixes incommensurable views; it can also use a smaller patient sample without fearing that patient responses are structurally different in kind. Second, the improvement priorities are not where intuition usually places them. Responsiveness - the speed and flexibility of reaction to requests - had both the lowest performance and the lowest weight in the overall construct, which suggests it is the dimension least connected to how respondents form overall judgements and therefore the dimension least likely to reward investment if the explicit goal is to move overall evaluations. By contrast, empathy carried the largest weight, so attention to individualised responses and service hours may have more leverage on overall perceptions than further investment in reaction times. Responsiveness improvements should still be prioritised on operational grounds (clinical workflow dependency, infection-control responsiveness) in line with the resilience literature on service continuity [27], rather than on overall-evaluation leverage alone.

6.4 Limitations

Several limitations constrain interpretation. The study is single-site, and the findings should not be generalised to other hospitals without replication; the hospital is large, urban and has a mature outsourcing arrangement, all of which may matter. These findings pertain to a mature, large-scale outsourcing arrangement in an urban tertiary public hospital; rural secondary hospitals with informal or developing service contracts may show different patterns, and the framework's measurement implications should be re-examined in 3–5 hospitals of varying size, ownership type and outsourcing maturity before being treated as generalisable. Sampling was not probabilistic, and the deliberate emphasis on staff means the patient and family subsample is small (n = 98), which limits the precision of the invariance test even though it satisfied the criteria. The framework's intelligence and resilience domains are conceptual and were not measured; the claim that they matter is motivated by the literature and by the fragmentary item content, not demonstrated here.

Two data problems deserve to be stated rather than buried. The expectation arm of the instrument was unusable: two thirds of respondents gave identical answers to all 22 expectation items, and the problem affected the data unevenly across collection order, which points to a collection or preparation artefact. We therefore could not compute gap scores, which is why this study is a SERVPERF study - and we would argue that future studies in this population should be designed as SERVPERF studies from the outset rather than collecting expectation data they will later discard. In addition, within-dimension responses were highly homogeneous, inflating internal consistency coefficients; the alphas reported here should not be compared with alphas from settings where respondents discriminate more finely among items within a dimension.

A further limitation concerns sample composition and data-collection batch effects. Although the measurement-invariance results (Section 5.4) and the stakeholder comparisons (Section 5.5) support treating the multi-stakeholder data as comparable, a batch-by-role interaction on overall service quality was statistically significant ($F = 5.22$, $p = 0.0014$), and the two collection batches differed on Tangibility ($t=-2.66$, $p = 0.008$) and overall quality ($t=-2.04$, $p = 0.042$), with the first batch (id < = 162, N = 162) showing poorer confirmatory-factor-analysis fit (RMSEA = 0.083) than the second batch (N = 462; RMSEA = 0.053). As reported in Section 4.4, these batch- and duplication-related features (143 respondents, 22.9%, produced an identical 44-item response pattern) are most plausibly artefacts introduced at the data-collection or data-preparation stage rather than genuine respondent tendencies. They are reported here as limitations and sensitivity checks and are not used to merge, split, or trim the sample; consequently, the 'comparable / can-be-pooled' conclusion in Section 6.3 should be read as conditional on these caveats rather than as an unconditional claim.

7. Conclusion

Hospital support service quality, as rated by the people who work in, manage and pass through a large Chinese public hospital, is a coherent construct that means the same thing to staff and to patients. This finding may support the practice of

multi-stakeholder evaluation and reminds us that the instrument designed to measure it is showing its age. The SIR-Quality framework proposed here names what the classical model omits - the intelligence of the service system and its resilience under disruption - and the immediate research task is to build instruments that can measure those domains rather than infer them from stray items.

Abbreviations

AVE
average variance extracted
CFA
confirmatory factor analysis
CFI
comparative fit index
CR
composite reliability
IPMA
importance–performance map analysis
RMSEA
root mean square error of approximation
SERVQUAL
service quality
S-D logic
service-dominant logic
SIR-Quality
Service–Intelligence–Resilience Quality
SRMR
standardized root mean square residual
TLI
Tucker–Lewis index
M
mean
SD
standard deviation.

Declarations

Ethics approval and consent to participate

Ethics approval and consent to participate. This study was conducted in accordance with the principles of the Declaration of Helsinki and was approved by the Medical Ethics Committee of the First Affiliated Hospital of Nanjing Medical University. Informed consent was obtained from all participants before they completed the questionnaire.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors received no specific funding for this research.

Author Contribution

W.Y. (Wang Youli) conceived and designed the study, developed the SIR-Quality instrument, collected the survey data, performed the statistical analyses (including confirmatory factor analysis, measurement invariance testing, and importance–performance map analysis), and drafted the manuscript. A.F. (Fu Anrong) contributed to data collection and the administration of the questionnaire. D.Y. (Ding Yuwei) contributed to data collection and database management. Z.Y. (Zhang Yubin) contributed to the coordination of the field survey and logistical support. C.Y. (Cao Yang) supervised the research, critically revised the manuscript for important intellectual content, and is the corresponding author. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Availability of data and materials

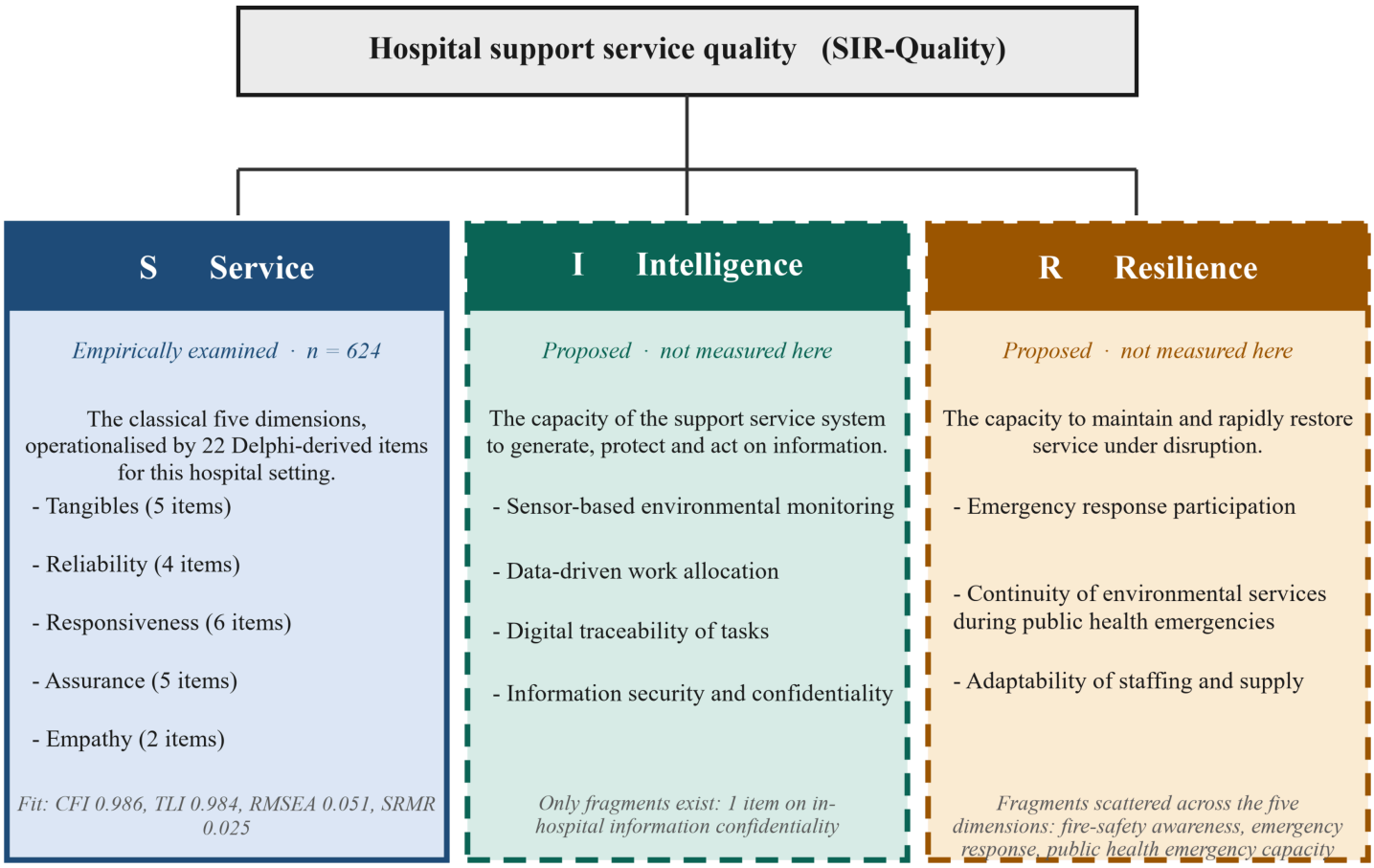
Not applicable.

References

1. Parasuraman A, Zeithaml VA, Berry LL. SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. J Mark. 1988;52(1):12–40.
2. Senapati S, Panda RK. Performance measurement of factors affecting consumer experience in healthcare services using fuzzy and grey techniques. Int J Health Gov. 2024;29(1):54–69. 10.1108/IJHG-08-2023-0082.
3. Cronin JJ, Taylor SA. Measuring service quality: a reexamination and extension. J Mark. 1992;56(3):55–68.
4. Cronin JJ, Taylor SA. SERVPERF versus SERVQUAL: reconciling performance-based and perceptions-minus-expectations measurement of service quality. J Retail. 1994;70(2):201–14.
5. Dagger TS, Sweeney JC, Johnson LW. A hierarchical model of health service quality: scale development and investigation of an integrated model. J Service Res. 2007;10(1):15–34.
6. Vargo SL, Lusch RF. Evolving to a new dominant logic for marketing. J Mark. 2004;68(1):1–17.
7. Vargo SL, Lusch RF. Institutions and axioms: an extension and update of service-dominant logic. J Acad Mark Sci. 2016;44(1):5–23.
8. Jaakkola E, Kaartemo V, Siltaloppi J, Vargo SL. Advancing service-dominant logic with systems thinking. J Bus Res. 2024;177:114592. 10.1016/j.jbusres.2024.114592.
9. Lusch RF, Nambisan S. Service innovation: a service-dominant logic perspective. MIS Q. 2015;39(1):155–75.
10. Parasuraman A, Zeithaml VA, Malhotra A. E-S-QUAL: a multiple-item scale for assessing electronic service quality. J Service Res. 2005;7(3):213–33.
11. Parasuraman A. Technology Readiness Index (TRI): a multiple-item scale to measure readiness to embrace new technologies. J Service Res. 2000;2(4):307–20.
12. Tuunanen T, Lumivalo J, Vartiainen T, Zhang Y, Myers MD. Micro-Level Mechanisms to Support Value Co-Creation for Design of Digital Services. J Service Res. 2023;27(3):381–96. 10.1177/10946705231173116.

13. Leiño Calleja D, Schepers J, Nijssen EJ. Some agents are more similar than others: customer orientation of frontline robots and employees. J Service Manage. 2023;34(6):27–49. 10.1108/josm-06-2022-0192.
14. Annarelli A, Nonino F, Palombi G. Understanding the management and implications of service quality and organizational resilience. Sustainability. 2020;12(3):958.
15. Bruneau M, Chang SE, Eguchi RT, Lee GC, O'Rourke TD, Reinhorn AM, Shinozuka M, Tierney K, Wallace WA. von Winterfeldt D. A framework to quantitatively assess and enhance the seismic resilience of communities. Earthq Spectra. 2003;19(4):733–52. 10.1193/1.1623497.
16. Hollnagel E, Woods DD, Leveson N, editors. Resilience Engineering: Concepts and Precepts. Aldershot, UK: Ashgate; 2006.
17. Butkus M, Schiuma G. The impact of organizational resilience on the quality of public services: application of structural equation modelling. Equilib Q J Econ Economic Policy. 2023;18(2):461–89. 10.24136/eq.2023.014.
18. Chen FF. Sensitivity of goodness of fit indexes to lack of measurement invariance. Struct Equ Model. 2007;14(3):464–504. 10.1080/10705510701301834.
19. Barile S, Polese F. Smart service systems and viable service systems: applying systems theory to service science. Service Sci. 2010;2(1–2):21–40. 10.1287/serv.2.1_2.21.
20. Podsakoff PM, MacKenzie SB, Lee JY, Podsakoff NP. Common method biases in behavioral research: a critical review of the literature and recommended remedies. J Appl Psychol. 2003;88(5):879–903. 10.1037/0021-9010.88.5.879.
21. Łukasik M, Porębska A. Responsiveness and adaptability of healthcare facilities in emergency scenarios: COVID-19 experience. Int J Environ Res Public Health. 2022;19(2):675. 10.3390/ijerph19020675.
22. Ellis LA, Saba M, Long JC, Lyng HB, Haraldseid-Driftland C, Churruca K, Wiig S, Austin E, Clay-Williams R, Carrigan A, Braithwaite J. The rise of resilient healthcare research during COVID-19: scoping review of empirical research. BMC Health Serv Res. 2023;23(1):983. 10.1186/s12913-023-09839-0.
23. Meade AW, Bauer DJ. Power and precision in confirmatory factor analytic tests of measurement invariance. Struct Equation Modeling: Multidisciplinary J. 2007;14(4):611–35. 10.1080/10705510701575461.
24. Patterson ES, Wears RL. Resilience and precarious success. Reliab Eng Syst Saf. 2015;141:55–61. 10.1016/j.ress.2015.03.014.
25. Alalade O, Dauda JA, Ajayi SO, Saka AB, Njuangang S. An in-depth analysis of facility management approaches in Nigeria's ailing healthcare sector. J Facilities Manage. 2024;23(4):667–84. 10.1108/JFM-12-2023-0123.
26. Amankwah O, Choong W-W, Boakye-Agyeman NA. Patients satisfaction of core health-care business: the mediating effect of the quality of health-care infrastructure and equipment. J Facilities Manage. 2022;22(3):365–81. 10.1108/JFM-12-2021-0154.
27. Kellett PLR, Franklin BD, Pearce S, Benn J. Resilience in nursing medication administration practice: a systematic review with narrative synthesis. BMJ Open Qual. 2024;13(4):e002711. 10.1136/bmjoq-2023-002711.
28. Woods DD, Hollnagel E, Prologue. Resilience engineering concepts. Resilience Engineering. Boca Raton, FL: CRC; 2017. 10.1201/9781315605685-1.
29. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Struct Equ Model. 1999;6(1):1–55. 10.1080/10705519909540118.
30. Henseler J, Ringle CM, Sarstedt M. A new criterion for assessing discriminant validity in variance-based structural equation modeling. J Acad Mark Sci. 2015;43(1):115–35. 10.1007/s11747-014-0403-8.
31. Ringle CM, Sarstedt M. Gain more insight from your PLS-SEM results: The importance-performance map analysis (IPMA). Industrial Manage Data Syst. 2016;116(9):1865–86. 10.1108/IMDS-10-2015-0449.

Figures



Note. Solid border = domain empirically examined in this study (22 perception items, n = 624). Dashed border = proposed domains; the present instrument was not designed to measure them, and no post hoc re-specification of these data into three factors was attempted.

Figure 1

The SIR-Quality framework. Solid border = domain empirically examined in this study (22 perception items, n = 624); dashed border = proposed domain.

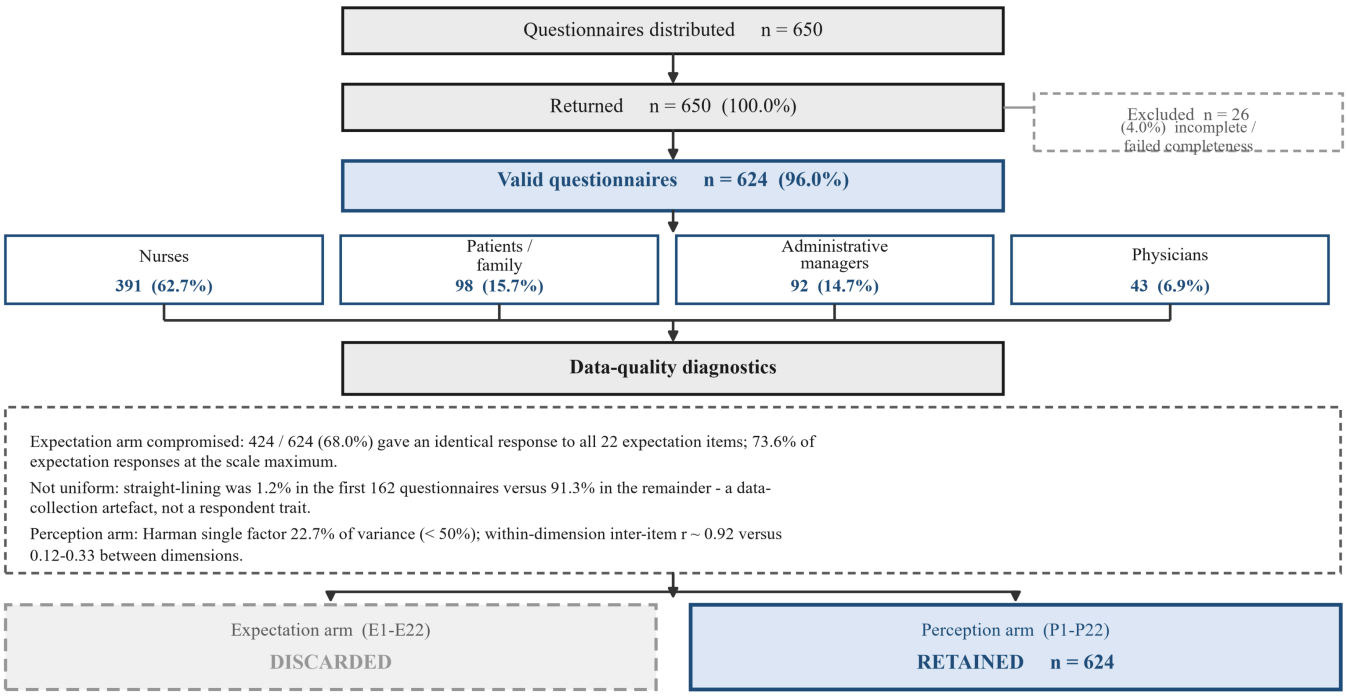


Figure 2

Study flow: (a) participant flow and data triage; (b) analytic sequence. All analyses use the perception items only; the expectation arm was discarded after the data-quality diagnostics.

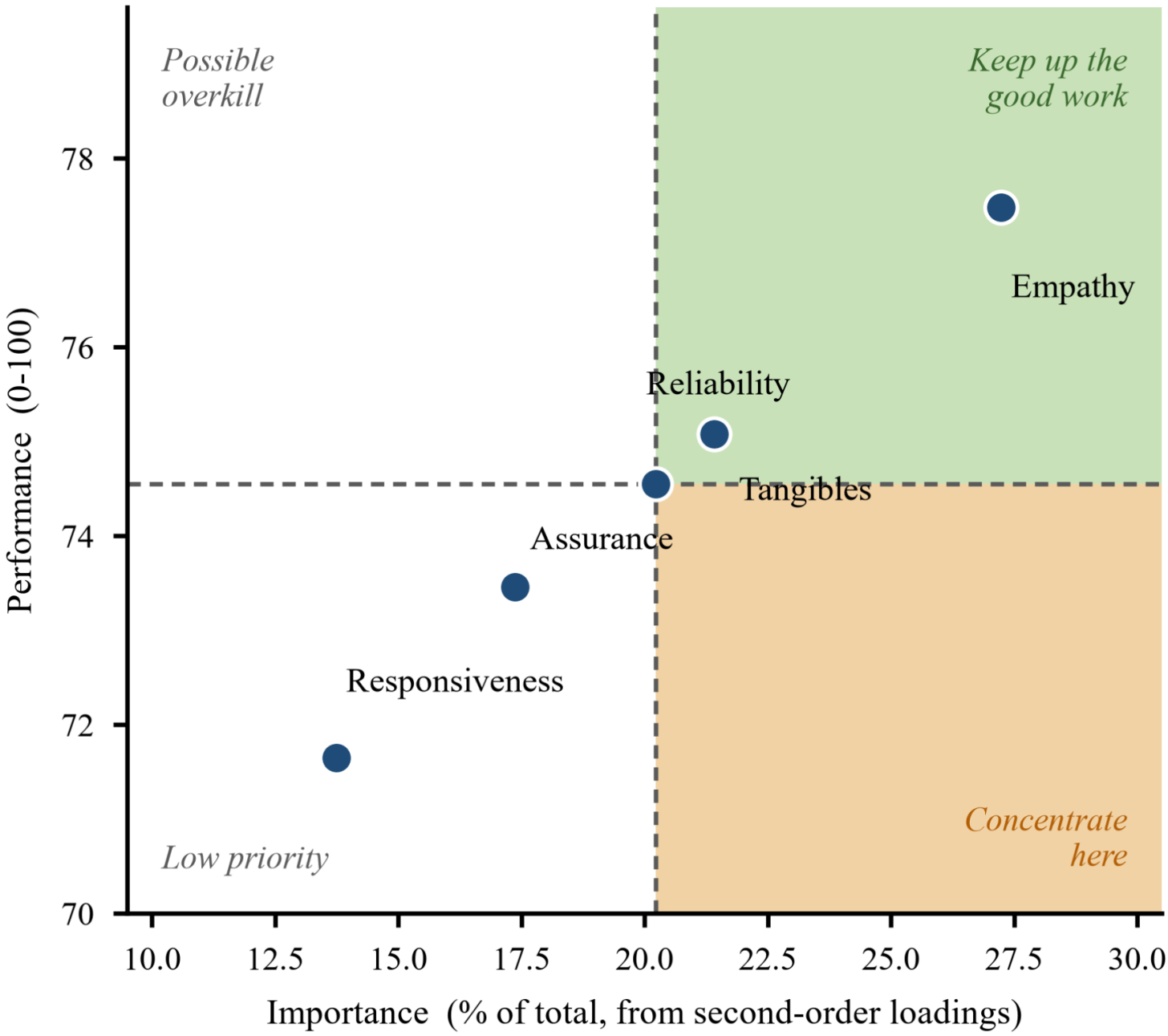


Figure 3

Importance-performance map of the five SERVQUAL dimensions. Importance is taken from the standardised second-order loadings (rescaled to 100%); performance is the dimension mean rescaled from 1-5 to 0-100. Splits are median lines.