Serican Journal of Medicine
2026, VOL. 3, NO. 2, 25486
https://doi.org/10.17161/sjm.v3i2.25486

Information Loss in Nursing Clinical Teaching Evaluations Measured by Shannon Entropy

Yuan Zhang[1], Hua He[1], Yajiao Cui[1], Yaofei Chen[1], Nu Li[1], and Weiyue Huang[2]*

[1]Department of Urology, The Seventh Affiliated Hospital, Sun Yat-sen University, Shenzhen 518000, China
[2]Department of Obstetrics and Gynecology, The Seventh Affiliated Hospital, Sun Yat-sen University, Shenzhen 518000, China

Corresponding Author: Weiyue Huang, RN, Department of Obstetrics and Gynecology, The Seventh Affiliated Hospital, Sun Yat-sen University, Shenzhen 518000, China. Email: huangwy57@mail.sysu.edu.cn

ABSTRACT

Background: Nursing student ratings of clinical teachers inform faculty development, mentorship, and accreditation. We applied information theory to quantify the information content and discriminative capacity of routine nursing teaching evaluation scores, with three parallel comparators illustrating how rater-evaluatee relationships modulate compression.

Methods: We analyzed 7,105 evaluations from four instruments at a Chinese university-affiliated teaching hospital. The primary analysis focused on nursing student ratings (n = 3,972; 436 students; 227 nurse educators with ≥ 5 ratings; 44 departments). Comparators: residency teaching scores (n = 2,271), secretary-to-resident 360 (n = 339), and resident self-assessment 360 (n = 523), all on 0-100 integer scales. We computed normalized Shannon entropy (H/H_max), intraclass correlation coefficient ICC(1,1) with F-distribution confidence intervals (CIs), and normalized mutual information (NMI). Cluster-aware, coarser-bin (10-category), zero-score inclusion, and stratified-ICC sensitivity analyses were performed.

Results: For nursing clinical teaching evaluations, the ceiling rate (scores = 100) was 74.8% (95% CI: 73.5-76.2%); normalized Shannon entropy was 0.131 (0.122-0.139), retaining 13.1% of theoretical information capacity. Teacher-level ICC (1,1) was 0.008 (-0.005 to 0.023); NMI = 0.035. Stratified analyses confirmed near-zero discriminability across strata. Comparators spanned a wider compression range: residency was more compressed (ceiling 91.6%, H_norm = 0.084, ICC = 0.065); the two 360 systems showed less compression (H_norm 0.331-0.494; ICC 0.100-0.280).

Conclusions: Routine nursing student ratings exhibited severe compression and limited single-rating discriminability, consistent with limited stand-alone utility for high-stakes individual-level decisions; they do not preclude use within multi-source evidence frameworks.

ARTICLE HISTORY
Received: May 10, 2026
Revised: May 24, 2026
Accepted: May 25, 2026

KEY WORDS
nursing education; teaching evaluation; Shannon entropy; information theory; ceiling effect; clinical education

Introduction

Teaching evaluation is a cornerstone of educational quality assurance in health professions education [1]. Ratings of clinical teaching inform faculty development, promotion decisions, and institutional accreditation, making their validity and informativeness a matter of considerable practical importance [2,3]. The widespread adoption of standardized evaluation instruments, typically employing Likert-type or percentage scales, reflects an assumption that numerical scores meaningfully capture variation in teaching quality [4].

However, a substantial body of literature has documented a persistent challenge: teaching evaluations in medical education are prone to severe ceiling effects, with mean scores routinely exceeding 4.0 on 5-point scales and the majority of ratings clustering at the highest anchor point [5-7]. Beckman et al. documented that existing instruments consistently produce highly skewed distributions [8], Stalmeijer et al. reported that the Maastricht Clinical Teaching Questionnaire (MCTQ) required 7-10 student ratings per teacher to achieve reliable scores [9], and Kreiter and Lakshman showed that restricted range directly

undermines reliability [13]. These patterns have been observed across diverse instruments, specialties, and countries [10, 11].

The consequences of ceiling effects extend beyond statistical inconvenience. When scores cluster at the maximum, the evaluation system loses its capacity to discriminate between educators of varying quality, effectively converting an assessment instrument into a compliance exercise [12]. This reliability problem has practical implications: faculty development programs that depend on evaluation scores to identify educators in need of support or to recognize excellence cannot function when scores fail to vary [14].

Despite widespread recognition of the ceiling effect problem, most studies have addressed it through descriptive statistics alone, reporting mean scores, standard deviations, and the percentage of responses at the highest anchor [15]. What has been lacking is a formal, quantitative framework for measuring exactly how much information an evaluation system retains or loses. Information theory, developed by Shannon in 1948, provides such a framework [16]. Shannon entropy quantifies the information content of a distribution in bits, and its normalized form (H/H_max) expresses the fraction of theoretical information capacity actually utilized. This framework has been applied extensively in ecology, genetics, and communication engineering; we are not aware of prior applications in this specific context [17].

Furthermore, while ceiling effects in teaching evaluations have been documented primarily in Western medical education contexts, the phenomenon in Chinese nursing education remains understudied despite the rapid expansion of standardized clinical training programs nationally. One hypothesized contributor is the distinct cultural and institutional context: hierarchical relationships between nursing students and clinical preceptors and collectivist values may amplify rating leniency [18, 19], though we treat this as a hypothesis rather than an established cause.

We focus on entropy rather than ceiling rate alone because the ceiling proportion captures only the peak of a distribution: two systems with identical 75% ceiling rates can still differ substantially in how the remaining 25% of scores are distributed across the scale. Shannon entropy summarises the full shape of a distribution into a single quantity (bits of information) and is directly interpretable as the equivalent number of equally likely outcomes (2^h). Intuitively, entropy measures how dispersed and unpredictable the scores are as a whole, not whether any individual score is "correct" or whether the instrument as a measure of teaching quality is valid. This common, information-theoretic scale allows ceiling, skewness, and fine-grained discrimination to be compared across instruments. When combined with variance-decomposition (intraclass correlation) and mutual-information measures, entropy provides a coherent framework for asking not only "how compressed are the scores?" but also "how much of the evaluatee-level signal survives the compression?"

This study aimed to (1) quantify the information content and discriminative capacity of routine nursing student ratings of clinical teachers at a Chinese university-affiliated teaching hospital using information-theoretic measures; (2) assess the ability of these scores to distinguish between individual nurse educators through variance decomposition and mutual information analysis; and (3) descriptively characterize ceiling effects in nursing teaching evaluations across departments and activity types. Three additional evaluation systems available at the same institution, one residency teaching evaluation system and two 360-degree systems (secretary-to-resident and resident self-assessment), are analyzed in parallel as comparators with different rater-evaluatee relationships to contextualize, not benchmark, the primary nursing findings.

Methods

Study design and setting

This retrospective observational study analyzed routinely collected nursing clinical teaching evaluation data from a university-affiliated teaching hospital in southern China. The hospital operates a structured clinical education program for nursing interns alongside a separate postgraduate residency training program; both share a centralized digital teaching management platform (Cloud-based Clinical Medical Teaching and Visualization, CCMTV, Cloud Teaching System) for recording teaching activities and collecting evaluation scores. The primary analysis examined pooled cross-sectional nursing teaching evaluation data to quantify information content and discriminative capacity; a secondary descriptive analysis examined temporal trends in ceiling rates across calendar quarters. The study was reported following the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines [20].

Analytic framework: primary system vs. comparators

Because the four score systems differ substantially in their rater-evaluatee relationships, we organized analyses into two tiers. The primary analysis concerns nursing student ratings of clinical teachers, the system whose performance is most directly relevant to nursing faculty development and program accreditation, and the focus of all substantive conclusions about instrument utility. Three comparator systems are reported in parallel: residency teaching scores (residents rating clinical teachers in the parallel postgraduate program), secretary 360 evaluation (administrative staff rating residents), and self-assessment 360 (residents rating themselves). These comparators illustrate how different rater-evaluatee relationships map onto different ceiling and entropy profiles, but we do not claim that they are interchangeable with the primary system or measure the same construct.

Conceptual primer: what entropy measures and how it relates to ICC

For readers less familiar with information theory, Shannon entropy quantifies the dispersion of the score distribution across the available scale, that is, how spread out and unpredictable the scores are as a whole, rather than the validity of the instrument itself or whether any particular score is "correct." A high-entropy distribution uses the full range of the 0-100 integer scale; a low-entropy distribution concentrates almost all observations on a few values (here, near 100). Entropy and the intraclass correlation coefficient (ICC) characterise different aspects of the same evaluation system and are complementary rather than redundant: ICC summarises how much of the total score variance lies between (versus within) the units being evaluated, i.e. the reliability of a single rating for distinguishing between individual nurse educators, while normalized entropy summarizes whether the score distribution itself carries enough information to support such discrimination in the first place. Severe ceiling compression depresses both quantities, but a system can, in principle, have non-trivial entropy with low ICC (scores spread across the scale but not aligned with the evaluatee being rated) or vice versa. Neither metric, taken alone, speaks to whether the instrument as a measure of teaching quality is valid; both speak to whether the distribution of scores it produces retains enough resolution to be informative.

Data sources and evaluation instruments

We examined four evaluation instruments, each representing a different rater-evaluatee relationship. Evaluation records with zero or missing scores were excluded. Individual evaluators could submit multiple ratings across different teaching sessions, and all valid submissions were retained for analysis.

1. Nursing student ratings (primary system, n = 3,972): Nursing interns rated clinical teaching sessions on a 0-100 integer score scale via the platform's built-in evaluation module. Each rating corresponded to a specific teaching activity (mini-lectures, skills practice, teaching rounds, or case discussions). Ratings were provided by 436 unique nursing interns across 44 departments and reflected 227 nurse educators with at least 5 ratings each.

2. Residency teaching scores (comparator, n = 2,271): Residents rated teaching sessions on a 0-100 integer score scale through the same platform. Scores covered 10 activity types and were provided by 351 unique residents across 49 departments, spanning January 2022 to March 2026.

3. Secretary 360 evaluation (comparator, n = 339): Departmental secretaries evaluated residents using a standardized 360-degree assessment instrument on a 0-100 integer score scale. Evaluations involved 71 unique secretaries across 37 departments, covering October 2024 to March 2026.

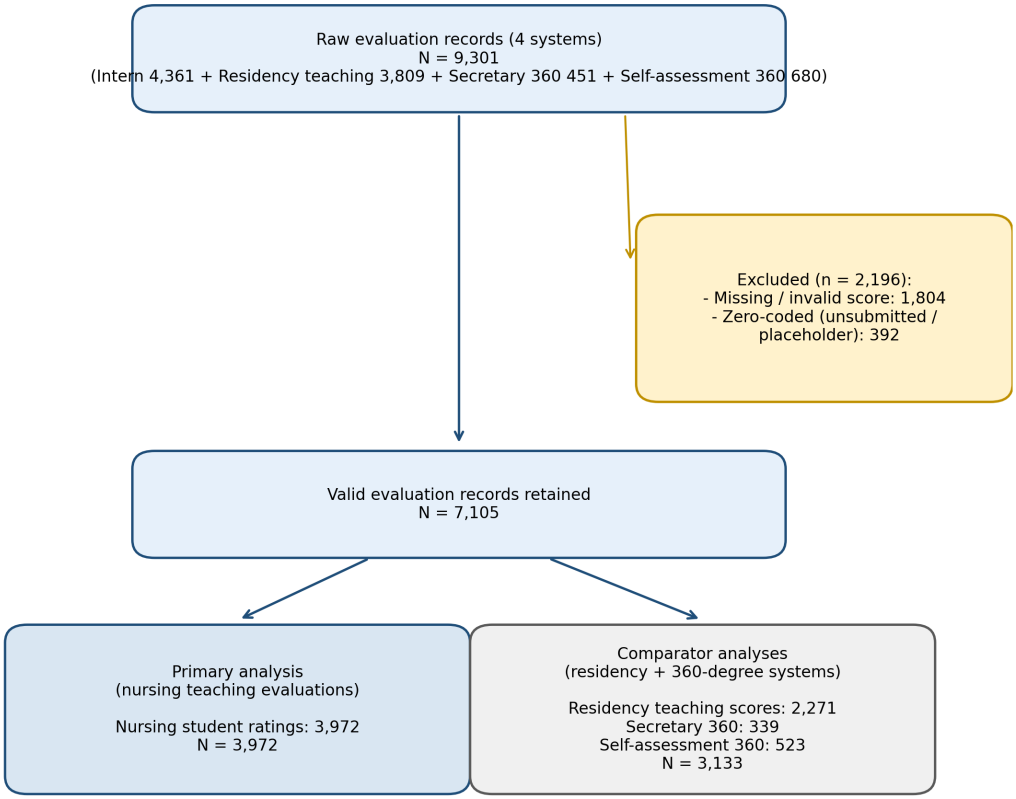
4. Self-assessment 360 (comparator, n = 523): Residents completed self-assessment ratings on a 0-100 integer score scale as part of the 360-degree evaluation program. Assessments involved 120 unique residents across 54 departments, covering February 2022 to March 2026.

All four instruments used the same 0-100 integer score scale, with platform-stored values as integers, facilitating descriptive comparison of score distributions across systems. The 1-point bin width used for entropy calculation corresponds to this native resolution.

Identification of zero-coded records

In the CCMTV platform's data-entry workflow, the score field is initialised to 0 when an evaluation form is created or auto-assigned to an evaluator. A non-zero integer is written only when an evaluator actively submits a rating; if the form is opened but not submitted, or is auto-created for an assigned session that the evaluator never completed, the exported record retains the placeholder value of 0. Genuine 0

Figure 1. Record inclusion flowchart. Of 9,301 raw evaluation records across the four systems, 2,196 were excluded (missing/invalid score or zero-coded, the latter corresponding to unsubmitted evaluations per the platform's data-entry specifications), leaving 7,105 valid evaluations. The primary analysis concerns nursing student ratings of clinical teachers (n = 3,972). Three comparator systems are analyzed in parallel: residency teaching scores (n = 2,271), secretary 360 (n = 339), and self-assessment 360 (n = 523).



scores from evaluators are extremely rare in practice and cannot be distinguished from these unsubmitted/placeholder records using the exported fields alone. We therefore conservatively classified all score = 0 records as non-genuine and excluded them from the primary analysis. The proportion of zero-coded records per system is reported in Supplementary Table e-S0; Supplementary Table e-S9 confirms that including these records does not change the substantive conclusions.

Of 9,301 raw records across the four systems, 2,196 were excluded (missing/invalid score plus zero-coded; system-level counts in Supplementary Table e-S0), leaving 7,105 valid evaluations for analysis. The full inclusion pathway is shown in Figure 1.

Ethical considerations

This study was reviewed and approved by the Ethics Committee of the Seventh Affiliated Hospital, Sun Yat-sen University (approval no. KY-2026-128-01). The requirement for individual informed consent was waived due to the retrospective, de-identified nature of the administrative data used.

Statistical analysis

All analyses were conducted using Python 3.13 with NumPy, SciPy, and custom scripts. Bootstrap confidence intervals were calculated using 2,000 resamples with a fixed random seed (42) at the 95% confidence level.

Ceiling effect quantification

We calculated the percentage of scores at the maximum value (100), at or above 95, and at or above 90 for each evaluation system. Distributional characteristics, including skewness and excess kurtosis, were computed. Confidence intervals for ceiling proportions were obtained by percentile bootstrap (2,000 resamples). As a sensitivity check for information content, we also recomputed normalized en-

tropy with 2-point and 5-point bin widths (Supplementary Table e-S7). To address the potential effect of within-rater and within-evaluatee clustering on the naive bootstrap CIs, we additionally computed cluster-aware bootstrap confidence intervals for ceiling proportions and entropy by resampling entire rater-clusters and evaluatee-clusters (Supplementary Table e-S8). To test sensitivity to the pre-analysis exclusion of zero-coded records, we re-computed key descriptives, including any zero records (Supplementary Table e-S9).

Information content

Shannon entropy (H) was calculated using 1-point bins across the 0-100 range:

$H = -\text{sum}(p_i \times \log_2(p_i))$, where p_i is the proportion of scores in each bin. Normalized entropy ($H_{norm} = H / H_{max}$) was calculated by dividing by the theoretical maximum entropy for a 101-bin uniform distribution ($H_{max} = \log_2(101) = 6.658$ bits). H_{norm} ranges from 0 (no information) to 1 (maximum information). Bootstrap 95% confidence intervals were computed for H_{norm}. H_{norm} reflects utilization relative to the native integer scale's theoretical capacity, not a direct estimate of semantic resolution: a 0-100 integer scale need not imply 101 perceptually or semantically equidistant categories. To verify that conclusions are not artifacts of this baseline, we repeated the analysis under a 10-category equal-width coarsening (Supplementary Table e-S10).

Variance decomposition

Intraclass correlation coefficients ICC(1,1) were calculated using one-way random effects analysis of variance (ANOVA) models at two levels: (a) department-level, where scores were grouped by department, and (b) evaluatee-level, where scores were grouped by individual evaluatee (teacher or resident). Groups with fewer than 5 observations were excluded at both levels. ICC(1,1) here denotes the reliability of a single rating; the projected reliability of the mean of k independent ratings (ICC(1,k)) was computed via the Spearman-Brown prophecy formula under the standard assumptions of independent, homogeneous, exchangeable raters. Confidence intervals were calculated using the F-distribution method of Shrout and Fleiss [23]. Bootstrap confidence intervals (2,000 resamples) were also computed as a sensitivity check but exhibited systematic upward bias under severe ceiling conditions (Supplementary Materials). Given the crossed structure of raters, evaluatees, departments, activity types, and quarters, and unequal rating counts per evaluatee, these ICC and NMI values should be interpreted as heuristic indicators of discriminative capacity rather than precise variance decompositions. A full generalizability-theory or cross-classified mixed model is beyond the scope of this descriptive information-loss analysis; stratified ICC by department and by activity type is reported in Supplementary Table e-S11 as a sensitivity check.

Mutual information

Normalized mutual information (NMI) between score values and person identity was calculated as MI / H(Person), where MI is the mutual information in bits. NMI quantifies how much information about the evaluatee's identity is conveyed by a score; NMI = 0 means scores carry no person-identifying information, while NMI = 1 would mean scores perfectly identify individuals.

Power analysis

Monte Carlo simulations were conducted to estimate the sample sizes needed to detect teaching quality differences under the observed distributions (see Supplementary Materials for full methods and results).

Subgroup analyses

Ceiling rates and normalized entropy were computed separately by activity type (for the two main systems) and by department. Temporal trends in ceiling rates were examined by calendar quarter for the residency system, which had sufficient longitudinal coverage.

Results

Sample characteristics

The inclusion pathway is shown in Figure 1. A total of 7,105 valid teaching evaluation scores were analyzed across four instruments (Table 1). The primary system, nursing student ratings of clinical teachers, contributed the largest share (3,972 scores from 436 nursing interns across 44 departments, reflecting 227 nurse educators with at least 5 ratings each). Three comparator systems were analyzed in parallel: residency teaching scores (2,271 scores from 351 residents across 49 departments), secretary 360 evaluation (339 scores), and self-assessment 360 (523 scores). The comparators are reported alongside the

Table 1. Sample characteristics across the four evaluation systems.

System	N	Persons	Departments	Activity Types	Date Range	Score Scale
Nursing Student Ratings (primary)	3,972	436	44	5	2023-07 to 2026-03	0–100
Residency Teaching Scores (comparator)	2,271	351	49	10	2022-01 to 2026-03	0–100
Secretary 360 Evaluation (comparator)	339	71	37	1	2024-10 to 2026-03	0–100
Self-Assessment 360 (comparator)	523	120	54	1	2022-02 to 2026-03	0–100

Note. Persons indicates unique raters. The primary system (nursing student ratings of clinical teachers) is the focus of all substantive conclusions; the three comparator systems illustrate how different rater-evaluatee relationships map onto different compression profiles. Totals: N = 7,105 valid evaluations after pre-analysis exclusions detailed in Supplementary Table e-S0.

primary nursing system to illustrate cross-system patterns, but are not interchangeable with it in construct or rater-evaluatee relationship.

Primary analysis: nursing clinical teaching evaluations

Ceiling effects and score distribution

Ceiling effects were severe in nursing student ratings (Figure 2, Table 2). The ceiling rate (scores = 100) was 74.8% (95% CI: 73.5-76.2%) with only 13 unique integer score values across 3,972 evaluations (median 100, interquartile range [IQR] 99-100). The distribution was severely left-skewed with very high excess kurtosis (Table 2), consistent with strong defaulting to the maximum integer.

Information content

Normalized Shannon entropy revealed severe information loss in nursing teaching evaluations (Figure 3). Nursing student ratings retained only 13.1% of theoretical information capacity on the native 0-100 integer scale (H_norm = 0.131, 95% CI: 0.122-0.139). In absolute terms, scores conveyed 0.871 bits per evaluation, compared with the theoretical maximum of 6.658 bits. The effective number of equally likely categories (2^H) was approximately 1.8, far below the 101 theoretically available integers. Under a coarser 10-category benchmark, the same conclusion held (Supplementary Table e-S10).

Discriminative capacity: variance decomposition

Nursing student ratings showed minimal ability to distinguish between individual nurse educators (Figure 4, Table 3). For 227 nurse educators with ≥5 ratings, department-level ICC(1,1) was 0.012 (95% CI: 0.005-0.026) and teacher-level ICC was 0.008 (95% CI: -0.005 to 0.023), well below the "poor" threshold of 0.50 established by Koo and Li. Stratified ICC by department and activity type (Supplementary Table e-S11) remained near zero across most strata, indicating that limited discriminability was not driven by a few heterogeneous strata. Cluster-aware bootstrap CIs (Supplementary Table e-S8) were wider than the naive CIs but did not alter the substantive conclusion.

Mutual information

NMI answers a direct question: if you are told a specific evaluation score, how much does that tell you about which nurse educator received it? NMI = 0 means the score carries no teacher-identifying information; NMI = 1 would mean a score uniquely identifies an educator. Normalization by H(Person) places NMI on a comparable footing across systems with different numbers of evaluatees. For nursing student ratings, NMI between scores and nurse educator identity was 0.035 (Table e-S4), consistent with near-zero ICC.

3.2 Comparator systems

The three comparator systems together spanned a wide range of compression severity (Table 2, Figure 3). The most compressed comparator was the residency teaching evaluation system, where residents rated clinical teachers in the parallel postgraduate training program: ceiling 91.6% (95% CI: 90.4-92.7%), only 9 unique integer values across 2,271 evaluations, H_norm = 0.084 (95% CI: 0.075-0.094), and teacher-level ICC = 0.065 (95% CI: 0.037-0.100). The two 360-degree systems showed less severe compression: secretary 360 evaluations (administrative staff rating residents) had a ceiling rate of 28.9% (95% CI: 24.2-33.6%) with H_norm = 0.494 and person-level ICC = 0.100; self-assessment 360 (residents rating themselves) had a ceiling rate of 64.1% (95% CI: 60.0-

Figure 2. Score distributions across the primary nursing teaching evaluation system and the three comparator systems (N = 7,105). The nursing panel and the residency panel, both high-ceiling, are zoomed to the 85-100 range. Each panel reports n, mean, median [IQR], ceiling proportion, and normalized Shannon entropy. The nursing panel is identified as the primary system; the remaining three panels are labelled as comparators.

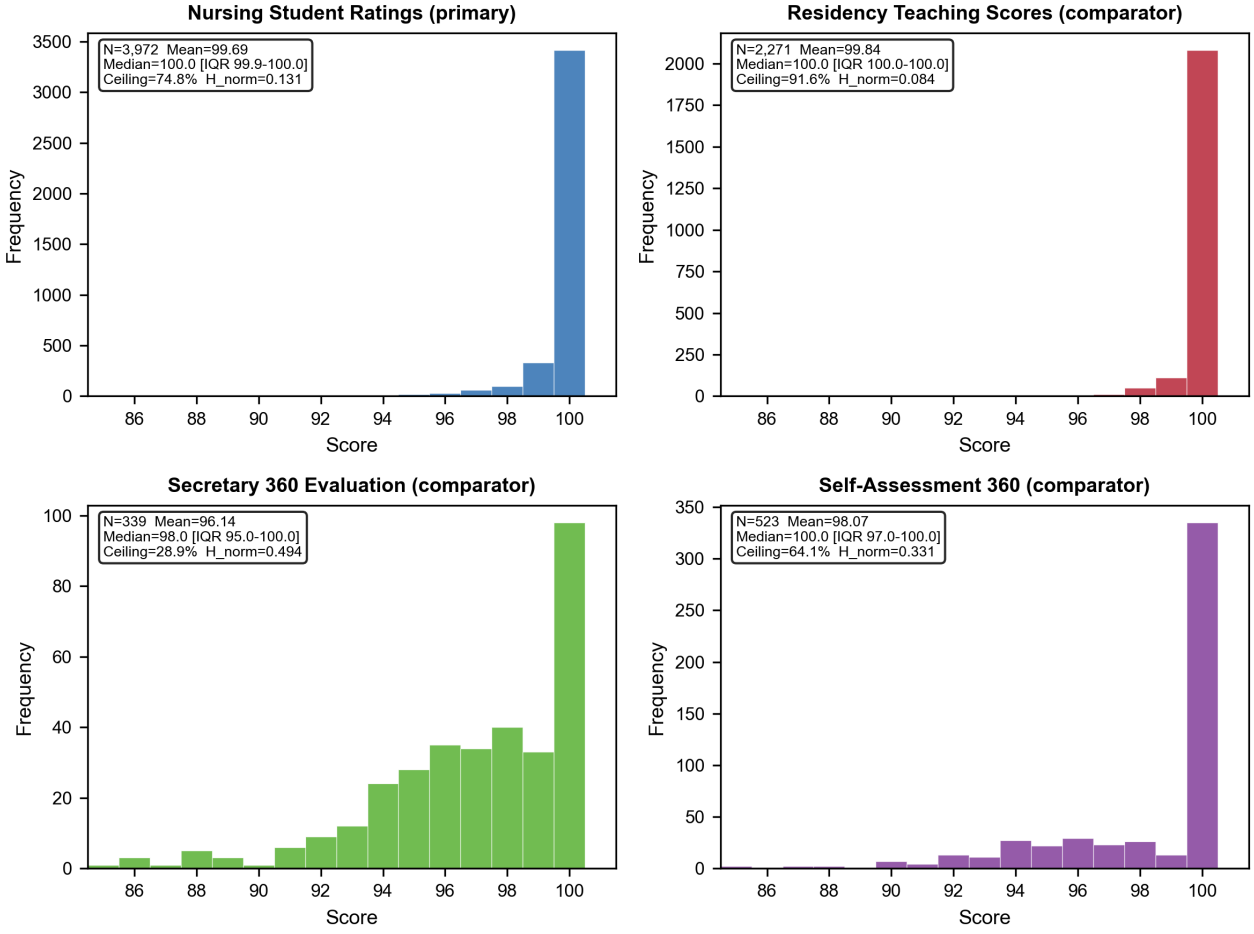


Table 2. Distribution and information-content characteristics of evaluation scores.

System	N	Mean	SD	Median	Skewness	Kurtosis	Ceiling % (95% CI)	H_norm (95% CI)	Effective Categories	Unique Values
Nursing Student Ratings (primary)	3,972	99.69	0.96	100	-5.72	45.59	74.8 (73.5–76.2)	0.131 (0.122–0.139)	1.8	13
Residency Teaching Scores	2,271	99.84	0.74	100	-11.97	251.58	91.6 (90.4–92.7)	0.084 (0.075–0.094)	1.5	9
Secretary 360 Evaluation	339	96.14	8.01	98	-8.79	93.89	28.9 (24.2–33.6)	0.494 (0.464–0.509)	9.8	21
Self-Assessment 360	523	98.07	3.37	100	-2.32	6.8	64.1 (60.0–68.1)	0.331 (0.297–0.355)	4.6	21

Note. SD = standard deviation. Ceiling % indicates the proportion of scores at the maximum value (100). H_norm is normalized Shannon entropy (H / H_max), with H_max = log_2(101) = 6.658 bits for the native 0–100 integer scale. Effective Categories = 2^H, the equivalent number of equally likely outcomes. Confidence intervals for ceiling proportions and H_norm are 2,000-resample percentile bootstraps.

Figure 3. Normalized Shannon entropy (H/H_max) across the four systems, with 95% bootstrap confidence intervals. The primary nursing teaching evaluation system is rendered in dark navy; the three comparator systems are rendered in light gray. Labels above each bar show the absolute entropy H in bits. H_norm = 1.0 corresponds to a uniform distribution over all 101 integer values; H_norm = 0 corresponds to a constant. Effective number of equally likely categories (2^h) ranged from 1.5 (residency comparator) to ~10 (secretary 360 comparator); the primary nursing system retained ~1.8 effective categories.

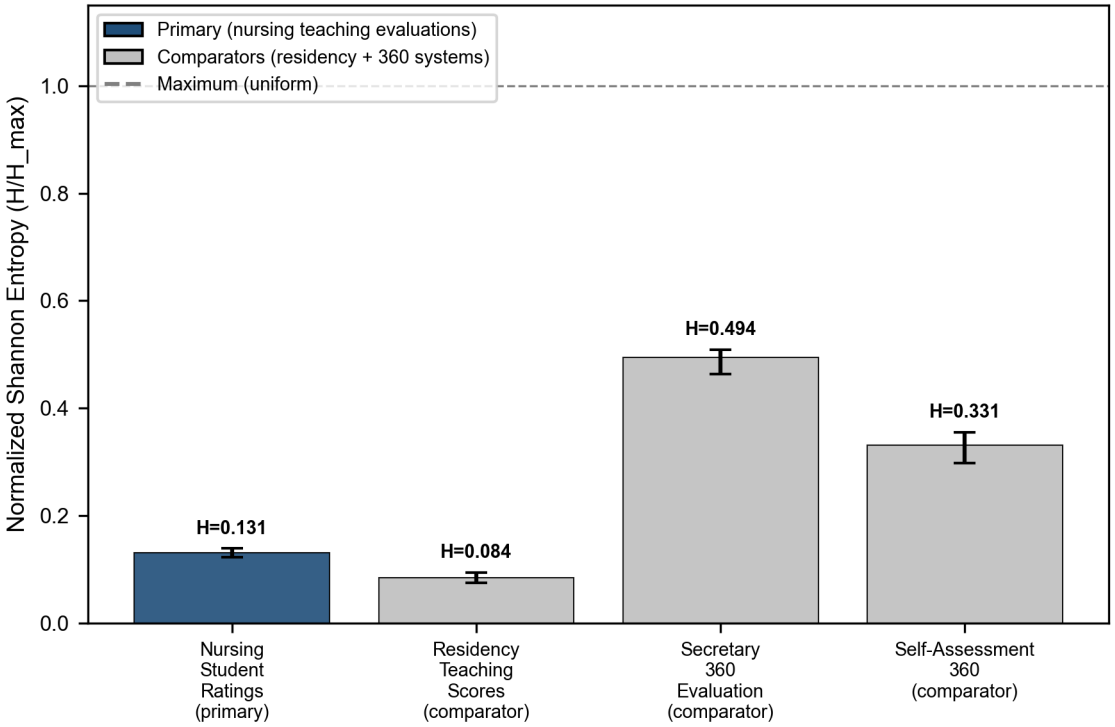


Figure 4. Forest plot of intraclass correlation coefficients ICC(1,1) for single ratings. Each system appears at two levels: department-level (scores grouped by department) and evaluatee-level (scores grouped by individual teacher, nurse educator, or resident). The primary nursing teaching evaluation system is identified separately from the three comparator systems. Error bars show analytic 95% F-distribution confidence intervals (Shrout & Fleiss, 1979). Labels indicate the number of groups (k) included per analysis; groups with fewer than 5 observations were excluded.

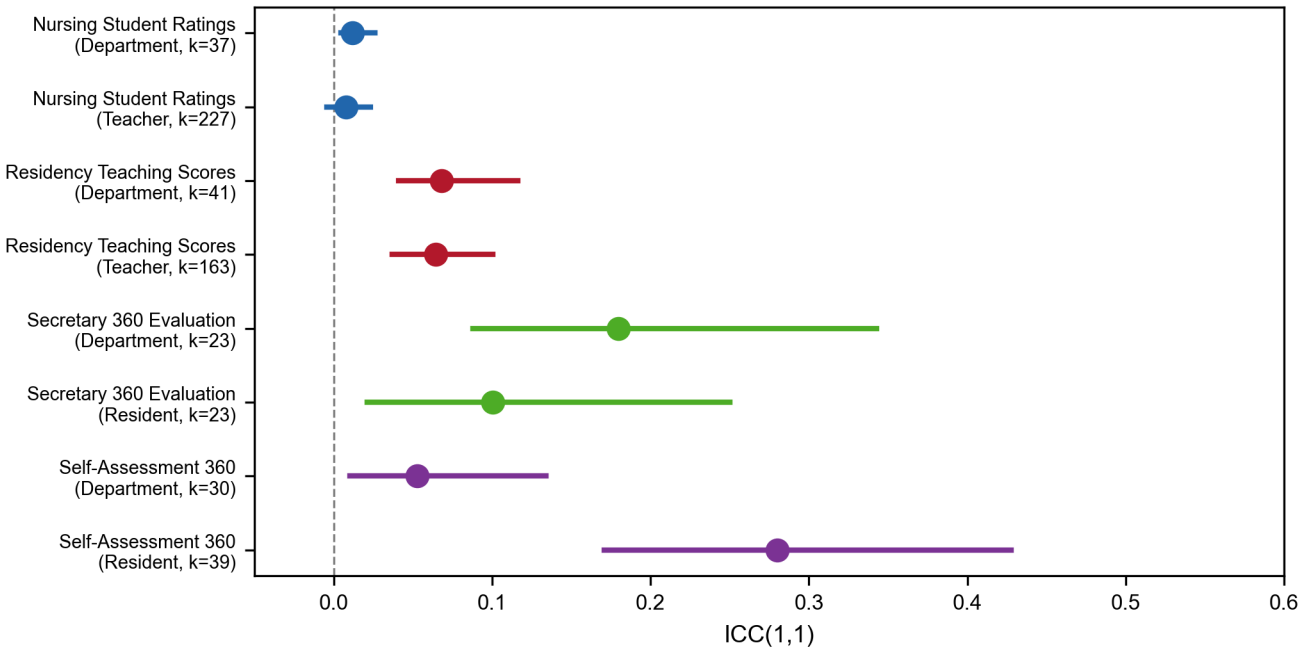


Table 3. Intraclass correlation coefficients ICC(1,1) for single ratings and Spearman-Brown projections.

System	Grouping	k Groups	N	ICC	95% CI (F)	k for ICC(1,k) = 0.70
Nursing Student Ratings (primary)	Department	37	3,961	0.012	0.005 - 0.026	195
Nursing Student Ratings (primary)	Person (teacher)	227	3,492	0.008	-0.005 to 0.023	305
Residency Teaching Scores	Department	41	2,252	0.068	0.041–0.116	32
Residency Teaching Scores	Person (teacher)	163	1,852	0.065	0.037–0.100	34
Secretary 360 Evaluation	Department	23	315	0.18	0.088–0.343	11
Secretary 360 Evaluation	Person (resident)	23	229	0.1	0.021–0.250	21
Self-Assessment 360	Department	30	479	0.053	0.010–0.134	42
Self-Assessment 360	Person (resident)	39	314	0.28	0.171–0.428	6

Note. ICC(1,1) is the one-way random-effects single-rating intraclass correlation (Shrout & Fleiss 1979). Groups with fewer than 5 observations were excluded. Spearman-Brown projection k for ICC(1,k) = 0.70 indicates the number of independent ratings required to achieve adequate reliability under the simplifying assumptions of independent, homogeneous, and exchangeable raters; this is an illustrative projection rather than an operational requirement.

68.1%) with H_norm = 0.331 and person-level ICC = 0.280 (the elevated self-assessment ICC likely reflects stable self-rating styles rather than genuine quality differentiation). Across the three comparators, ceiling severity decreased, and effective categories increased as the rater-evaluatee relationship moved away from learner-rates-teacher dynamics; this pattern illustrates that the rater-evaluatee relationship can substantially modulate ceiling severity, although the comparator instruments do not measure the same construct as the primary nursing teaching evaluation.

Subgroup and temporal patterns

In nursing teaching evaluations, ceiling effects were remarkably consistent across teaching activity types (Table e-S1, Figure S1). Date metadata in the nursing student rating system was incomplete prior to 2023, limiting longitudinal analysis. For the residency comparator system, which had complete date coverage from 2022 onwards, exploratory, descriptive analyses suggested a pattern of rising ceiling rates over time (Figure 5, Table e-S2). Because we did not have data on platform version changes, policy updates, or shifts in rater or activity composition, this temporal pattern in the comparator system is reported as descriptive only and is not extrapolated to the primary nursing system.

Power analysis

Monte Carlo power simulations (Supplementary Materials, Table e-S6, Figure S3) were computed under simplifying independence assumptions and are presented as illustrative rather than as directly applicable to the crossed rater-evaluatee structure observed in these data.

Discussion

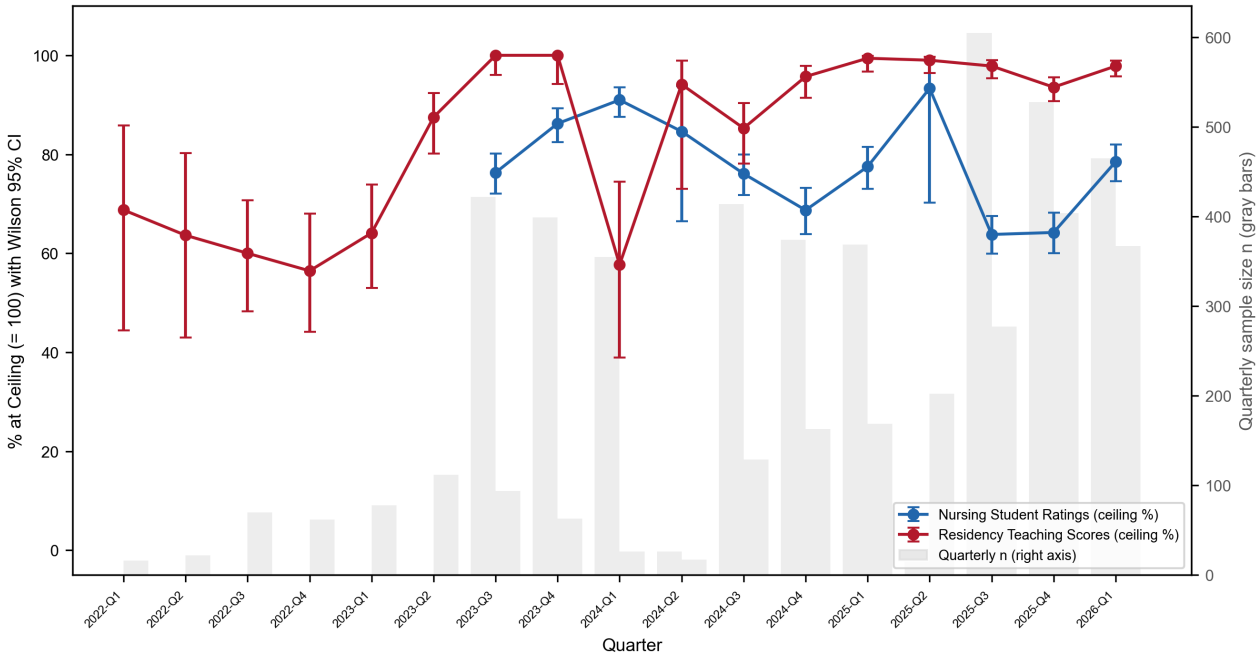
Principal findings

This study applied information-theoretic measures to quantify the information content and discriminative capacity of nursing student ratings of clinical teachers at a Chinese teaching hospital, with three comparator systems (one residency teaching evaluation system and two 360-degree systems) analyzed in parallel. Nursing teaching evaluations exhibited severe compression: normalized Shannon entropy of 0.131 (retaining only 13.1% of the native-scale theoretical information capacity; ordering preserved under a coarser 10-category benchmark), teacher-level ICC(1,1) of 0.008, and NMI between scores and nurse educator identity of 0.035. These findings represent, we believe, an application of the information-theoretic framework to nursing clinical teaching evaluations that we are not aware of being previously reported.

Comparison with existing literature

The ceiling rate observed in nursing teaching evaluations (74.8%) is consistent with, and in some cases more extreme than, ceiling rates reported in Western medical education settings. Beckman et al. found that clinical teaching evaluations consistently produce highly skewed distributions with the majority of ratings at the top of the scale [8]. The MCTQ validation by Stalmeijer et al. demonstrated that 7 to 10 ratings per teacher were needed for reliable scores [9], reflecting the restricted variance that characterizes these instruments, a challenge even more pronounced in our data, where the median number of student ratings per nurse educator was only 6. The

Figure 5. Descriptive temporal pattern of ceiling rates (% of scores = 100) by calendar quarter, 2022-2026, for the residency teaching comparator system (which had complete date coverage from 2022 onwards) and, where available, the primary nursing teaching evaluation system (limited to 2023 onwards due to incomplete date metadata in the earlier nursing records). Point estimates are shown with Wilson 95% confidence intervals; gray bars (right y-axis) show quarterly sample sizes. Descriptive only; no adjustment for activity-type or department composition. This figure cannot distinguish between ceiling intensification due to evolving rating norms and changes in platform version, policy, or rater/activity composition; the temporal pattern in the residency comparator is not extrapolated to the primary nursing system.



teacher-level ICC of 0.008 in nursing teaching evaluations falls well below the "poor" threshold of 0.50 established by Koo and Li [22], and is consistent with the reliability concerns raised by Kreiter and Lakshman [13], who demonstrated that restricted range in teaching evaluations directly undermines reliability estimates. The closest analog in the published literature is Dexter and colleagues' application of binomial entropy to anesthesiologists' ratings of nurse anesthetists' clinical performance [29], which similarly documented information loss attributable to restricted rater behavior; the present study extends this lens to nursing teaching evaluations rather than clinical performance ratings, and to a Chinese nursing intern education context specifically.

What the information-theoretic framework adds is a precise quantification of how much native-scale information capacity has been utilized. The nursing teaching evaluation H_norm of 0.131 (H = 0.87 bits) corresponds to approximately 1.8 effective equally likely categories (2^h), far below the 101 theoretically available integers on the 0-100 scale. In practical terms, despite using a 0-100 integer score scale, the nursing instrument conveys information comparable to a very small number of effective equally likely categories. Across the three comparators, the residency teaching system was even more compressed (1.5 effective categories), while the two 360-degree systems were less compressed (4 to 10 effective categories), suggesting that the rater-evaluatee relationship is a meaningful modulator of compression severity.

Why does this matter for nurse educator evaluation?

Before turning to practical implications, it is worth re-emphasising what the entropy framing does and does not say. Although ceiling compression is often discussed as a problem of "ratings being too high," our information-theoretic framing shifts the locus from level to distribution: normalised entropy measures how spread the score distribution is across the available scale, not whether the instrument is valid or whether high scores are inherently wrong. A system can be ceiling-compressed and informative only if the residual variation among non-maximum scores aligns systematically with the units being evaluated; the ICC and NMI analyses above show that, in nursing student ratings, it does not.

The severity of information loss in nursing teaching evaluations has practical implications for how nursing programs use these scores. The compressed

distribution provides little basis for differentiating individual nurse educators in the narrow 99-100 range, yet scores in this range may inform faculty performance review, mentorship assignment, and program accreditation decisions. The Spearman-Brown analysis projected that achieving adequate reliability (ICC(1,k) = 0.70) for individual nurse educator assessment would, under simplifying assumptions of independent, homogeneous, exchangeable raters, require hundreds of independent student ratings per educator, far more than most clinical preceptors accumulate in an academic year (Figure S2). This is an illustrative projection rather than an operational requirement. The temporal pattern of rising ceiling rates observed in the residency comparator system (Figure 5) is exploratory and descriptive; we do not extrapolate that pattern to nursing teaching evaluations, where the date metadata was insufficient for an analogous trend analysis.

Possible explanations

Several plausible mechanisms may contribute to the ceiling effects observed in nursing teaching evaluations, though we did not directly measure them in this study. The instrument uses a 0-100 integer score scale without behavioral anchors, which provides limited guidance for score interpretation and may encourage defaulting to the maximum [5]. One hypothesized contributor is the Chinese nursing education context, hierarchical relationships between nursing interns and clinical preceptors, and collectivist values that may discourage negative evaluations [18,19]. Platform interface design features (e.g., default values, interaction patterns) and rater anonymity policies could also modulate responses, though the specific configurations at our institution were not formally investigated here [28]. These explanations are hypotheses for future research rather than established causes.

These findings are particularly relevant given the rapid expansion of standardized clinical training programs in Chinese nursing education over the past decade, which rely heavily on subjective evaluation methods to assess clinical preceptors. The cross-system comparison illustrates that secretary 360 evaluations showed the least severe ceiling effect (28.9%) while still having a mean of 96.1; this instrument involves administrative staff evaluating residents rather than learners evaluating teachers, suggesting that the evaluator-evaluatee power dynamic may be a critical moderator of ceiling severity even within a single institution.

Implications for practice

Our findings suggest that incremental modifications to the existing nursing teaching evaluation instrument (e.g., adjusting scale anchors or providing rater training) may be insufficient to address the fundamental information loss documented here when these scores are used stand-alone for consequential individual-level decisions about nurse educators. These findings do not preclude the use of the scores as one element within multi-source evidence frameworks (combined with narrative feedback, direct observation, or peer review). More substantive approaches that nursing programs could consider include:

1. **Alternative assessment formats**: Narrative feedback, structured qualitative observation tools, and teaching portfolios may capture dimensions of teaching quality that numerical scales cannot when ceiling-compressed [24, 25].
2. **Scale redesign**: Evidence suggests that increasing the number of response options, modifying scale labels, and using 7-point rather than 5-point scales can reduce ceiling effects and increase discriminative variability [11, 15, 21].
3. **Criterion-referenced rubrics**: Behaviorally anchored rating scales (BARS) with explicit performance descriptors at each level may reduce the tendency to default to the maximum by requiring evaluators to match observed behaviors to specific criteria [25].
4. **Multi-source triangulation**: Rather than relying on a single numerical score, programs should integrate multiple sources of evidence, including direct observation, peer review, student learning outcomes, and narrative feedback [26].

Strengths and limitations

This study has several strengths. First, the information-theoretic framework provides a novel and rigorous quantification of nursing teaching evaluation instrument performance that goes beyond descriptive statistics. Second, the cross-system comparison across one primary nursing system and three comparator instruments within the same institution controls for institutional and cultural factors, isolating instrument-level effects. Third, the inclusion of power analysis translates abstract information loss

into concrete consequences for nursing faculty development and quality improvement programs. Fourth, the large sample size (n = 3,972 nursing ratings within an N = 7,105 four-system dataset) provides precise estimates with narrow confidence intervals.

Several limitations should be acknowledged. First, this is a single-institution study, and the findings may not generalize to other nursing programs or institutions. However, the consistency of compression patterns across four independent assessment systems within this institution, and the concordance with international literature, suggests that the phenomenon is not idiosyncratic. Second, we analyzed secondary administrative data without direct knowledge of the conditions under which nursing student ratings were completed (e.g., anonymity, timing relative to clinical evaluation grades); data cleaning procedures (exclusion of zero scores, minimum group sizes) are described in Methods and shown in Figure 1. Third, the nursing student rating system lacked the data information for temporal trend analysis prior to 2023, limiting longitudinal characterization for this primary instrument. Fourth, we cannot determine from the data alone whether ceiling effects reflect true homogeneity of nursing teaching quality (all clinical preceptors are genuinely excellent) or instrument insensitivity; the absence of external criterion measures (such as direct observation, peer review, or learner outcomes) precludes definitive attribution, though the near-zero ICC and cross-system consistency favor the latter interpretation. Fifth, the ICC analysis used a one-way random effects model that does not capture the full cross-classified structure of raters, evaluatees, departments, and activity types; more complex generalizability theory or multilevel modeling approaches could provide additional insights but were beyond the scope of this descriptive study. We therefore present ICC and NMI as heuristic indicators and report stratified ICC by department and activity type (Table e-S11) as a sensitivity check. Sixth, the four assessment systems differ in their evaluation relationships, and direct comparison of absolute metric values across systems should be interpreted cautiously; we have accordingly restricted substantive conclusions to the primary nursing teaching evaluation system and reported the residency and 360-degree systems as comparators only.

Future directions

Future research should examine whether information-theoretic metrics such as normalized entropy can serve as routine quality indicators for nursing teaching evaluation instruments, applied at the program level during annual quality reviews. Intervention studies comparing the discriminative validity of alternative evaluation formats (narrative, rubric-based, comparative) against traditional numerical scales in Chinese nursing clinical education settings would provide actionable evidence for instrument redesign. Multi-institutional studies across Chinese teaching hospitals would help establish whether the degree of information loss observed in nursing teaching evaluations here is representative of the national pattern.

Conclusion

Routine nursing student ratings of clinical teachers at this institution exhibited substantial compression (ceiling 74.8%, normalized Shannon entropy 0.131 on the native 0-100 integer scale, ordering preserved under a coarser 10-category benchmark) and limited single-rating ability to distinguish between individual nurse educators (teacher-level ICC(1,1) 0.008, NMI 0.035). These findings, quantified through an information-theoretic framework whose application to nursing clinical teaching evaluations we are not aware of being previously reported, are consistent with limited stand-alone discriminative utility for high-stakes individual-level uses such as nurse educator promotion, mentorship assignment, or program accreditation. They do not preclude the use of these scores as one element within multi-source evidence frameworks. Three comparator systems analyzed in parallel — one residency teaching evaluation system and two 360-degree instruments — together spanned a wide range of compression severity, suggesting that the rater-evaluatee relationship is a meaningful modulator and that the pattern observed in nursing teaching evaluations is not unique to a single instrument. Nursing programs depending on these scores alone for consequential decisions about clinical preceptors should recognize the extent of information loss and consider supplementing or redesigning their evaluation approaches.

Practice Points

- **Finding**: In the primary nursing clinical teaching evaluation system (n = 3,972 student ratings; 436 students rating 227 nurse educators across 44 departments), normalized Shannon entropy was 0.131, corresponding to approximately 1.8 effective equally likely categories out of 101 possible on

the 0-100 integer score scale. A coarser 10-category benchmark preserved this finding. Three comparator systems (residency teaching scores and two 360-degree instruments) are reported in parallel to illustrate how rater-evaluatee relationships modulate compression severity.

- **Reliability implication**: The teacher-level intraclass correlation coefficient ICC(1,1) for nursing teaching evaluations (0.008, 95% confidence interval [CI]: -0.005 to 0.023) is consistent with limited single-rating ability to distinguish between individual nurse educators. Under simplifying assumptions of independent, homogeneous, exchangeable raters, Spearman-Brown projections suggest hundreds of independent student ratings per nurse educator would be needed to reach ICC(1,k) = 0.70.
- **Practical recommendation**: Nursing student ratings on this 0-100 integer scale appear to have limited stand-alone discriminative utility for high-stakes individual-level decisions about nurse educators (e.g., promotion, mentorship assignment, accreditation). The findings do not preclude their use within multi-source evidence frameworks (combined with narrative feedback, peer observation, or criterion-referenced rubrics).

Funding Sources: None. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.
Disclaimers: None.

Author Approval Statement: Each author has reviewed and approved the content of the manuscript. All authors agree to be accountable for all aspects of the work and consent to its submission to the *Serican Journal of Medicine*.

Declarations: The authors report no conflicts of interest. The study was approved by the Ethics Committee of the Seventh Affiliated Hospital, Sun Yat-sen University (approval no. KY-2026-128-01); informed consent was waived as the study used de-identified administrative data. The study was conducted in accordance with the Declaration of Helsinki.

Artificial Intelligence (AI) Disclosure: During the preparation of this manuscript, the authors used DeepSeek (DeepSeek, Hangzhou, China) to assist with English-language polishing of author-drafted text and to help debug Python code used in the data analysis pipeline. The tool was not used to generate scientific content, ideas, results, or interpretations. All AI-assisted edits and code suggestions were reviewed, verified, and revised by the authors, who take full responsibility for the content and integrity of the manuscript.

References

1. Steinert Y, Mann K, Centeno A, et al. **A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education**: *BEME Guide No. 8. Med Teach*. 2006;28(6):497-526.
2. Beckman TJ, Cook DA, Mandrekar JN. **What is the validity evidence for assessments of clinical teaching?** *J Gen Intern Med*. 2005;20(12):1159-1164.
3. Steinert Y, Mann K, Anderson B, et al. **A systematic review of faculty development initiatives designed to enhance teaching effectiveness: a 10-year update**: *BEME Guide No. 40. Med Teach*. 2016;38(8):769-786.
4. Spooren P, Brockx B, Mortelmans D. **On the validity of student evaluation of teaching: the state of the art**. *Rev Educ Res*. 2013;83(4):598-642.
5. Fluit CR, Bolhuis S, Grol R, Laan R, Wensing M. **Assessing the quality of clinical teachers: a systematic review of content and quality of questionnaires for assessing clinical teachers.** *J Gen Intern Med*. 2010;25(12):1337-1345.
6. Feistauer D, Richter T. **Investigating halo and ceiling effects in student evaluations of instruction.** *High Educ*. 2017;73(1):19-38.
7. Copeland HL, Hewson MG. **Developing and testing an instrument to measure the effectiveness of clinical teaching in an academic medical center**. *Acad Med*. 2000;75(2):161-166.
8. Beckman TJ, Ghosh AK, Cook DA, Erwin PJ, Mandrekar JN. **How reliable are assessments of clinical teaching? A review of the published instruments**. *J Gen Intern Med*. 2004;19(9):971-977.
9. Stalmeijer RE, Dolmans DH, Wolfhagen IH, Muijtjens AM, Scherpbier AJ. **The Maastricht Clinical Teaching Questionnaire (MCTQ) as a valid and reliable instrument for the evaluation of clinical teachers**. *Acad Med*. 2010;85(11):1732-1738.
10. van der Hem-Stokroos HH, van der Vleuten CP, Daelmans HE, Haarman HJ, Scherpbier AJ. **Reliability of the Clinical Teaching Effectiveness Instrument**. *Med Educ*. 2005;39(9):904-910.

11. Debets MPM, Scheepers RA, Boerebach BCM, Arah OA, Lombarts KMJMH. **Variability of residents' ratings of faculty's teaching performance measured by five- and seven-point response scales**. *BMC Med Educ*. 2020;20(1):325.
12. Stroud L, Freeman R, Kulasegaram K, Cil TD, Ginsburg S. **Gender effects in assessment of clinical teaching: does concordance matter?** *J Grad Med Educ*. 2020;12(6):710-716.
13. Keeley JA, English T, Irons J, Henslee AM. **Investigating halo and ceiling effects in student evaluations of instruction**. *Educ Psychol Meas*. 2013;73(3):440-457.
14. Bierer SB, Hull AL. **Examination of a Clinical Teaching Effectiveness Instrument used for summative faculty assessment**. *Eval Health Prof*. 2007;30(4):339-361.
15. Chyung SY, Hutchinson D, Shamsy JA. **Evidence-based survey design: ceiling effects associated with response scales**. *Perform Improv*. 2020;59(6):6-13.
16. Shannon CE. **A mathematical theory of communication**. *Bell Syst Tech J*. 1948;27(3):379-423.
17. Casagrande A, Fabris F, Girometti R. **Fifty years of Shannon information theory in assessing the accuracy and agreement of diagnostic tests**. *Med Biol Eng Comput*. 2022;60(4):941-955.
18. Zhu J, Li W, Chen L. **Doctors in China: improving quality through modernisation of residency education.** *Lancet.* 2016;388(10054):1922-1929.
19. Kikukawa M, Stalmeijer R, Matsuguchi T, Oike M, Sei E, Schuwirth LWT, Scherpbier AJJA. **How culture affects validity: understanding Japanese residents' sense-making of evaluating clinical teachers**. *BMJ Open*. 2021;11(8):e047602.
20. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP; **STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies**. *Lancet*. 2007;370(9596):1453-1457.
21. Vita S, Coplin H, Feiereisel KB, Garten S, Mechaber AJ, Estrada C. **Decreasing the ceiling effect in assessing meeting quality at an academic professional meeting**. *Teach Learn Med*. 2013;25(1):47-54.
22. Koo TK, Li MY. **A guideline for selecting and reporting intraclass correlation coefficients for reliability research**. *J Chiropr Med*. 2016;15(2):155-163.
23. Shrout PE, Fleiss JL. **Intraclass correlations: uses in assessing rater reliability**. *Psychol Bull*. 1979;86(2):420-428.
24. Irby DM. **Excellence in clinical teaching: knowledge transformation and development required.** *Med Educ*. 2014;48(8):776-784.
25. Kogan JR, Holmboe ES, Hauer KE. **Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review**. *JAMA*. 2009;302(12):1316-1326.
26. Lombarts KM, Bucx MJ, Arah OA. **Development of a system for the evaluation of the teaching qualities of anesthesiology faculty**. *Anesthesiology*. 2009;111(4):709-716.
27. Lio J, Ye Y, Dong H, Reddy S, McConville J, Sherer R. **Standardized residency training in China: the new internal medicine curriculum**. *Perspect Med Educ*. 2018;7(1):50-53.
28. Marsh HW. **Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases and usefulness**. In: Perry RP, Smart JC, eds. *The Scholarship of Teaching and Learning in Higher Education*: *An Evidence-Based Perspective*. Springer; 2007:319-383.
29. Dexter F, Epstein RH, Öhrvik J, Hindman BJ. **Binomial entropy of anesthesiologists' ratings of nurse anesthetists' clinical performance explains information loss when adjusting evaluations for rater leniency.** *Perioperative Care Oper Room Manag*. 2022; 27:100247.