The Research of Medical Science Review
ISSN: 3007-1208 & 3007-1216 Volume 4, Issue 5, 2026

FROM SCORES TO SENSE-MAKING: HOW SUPERVISORS' FEEDBACK PRACTICES MEDIATE THE RELATIONSHIP BETWEEN WORKPLACE-BASED ASSESSMENT AND TRAINEE LEARNING IN INTERNAL MEDICINE RESIDENCY TRAINING IN PUBLIC SECTOR TEACHING HOSPITALS OF KPK, PAKISTAN

Dr Qismat Ullah[*1], Dr. Ayesha Junaid[2], Dr.Junaid Sarfraz Khan[3], Dr Saad Hamid[4]

[*1]*Resident physician in department of Medicine of Khalifa Gul Nawaz MTI Bannu Ex Students pf MHPE in Health Services academy, Islamabad*
[2]*Assistant Professor of Linguistics, Department of English language and Literature, college of humanities. Prince Sattam bin Abdul Aziz university , Al Kharj, KSA*
[3]*Dean School of Health Professionals' Education, Research and Entrepreneurship Rector/ Director Academics,Health Services Academy, Islamabad*
[4]*MBBS student in Bannu medical College, Bannu.*

[*1]uqismat966@gmail.com,[2]a.sarfrazkhan@psau.edu.sa,[3]junaidsarfraz@hsa.edu.pk,
[4]saadhamid8888@gmail.com

DOI: https://doi.org/10.5281/zenodo.20527119

Keywords
Competency based medical education; feedback; Pakistan; postgraduate education, Medical, Workplace-based assessment.

Article History
Received: 11 March 2026
Accepted: 03 May 2026
Published: 30 May 2026

Copyright @Author
Corresponding Author: *
Dr Qismat ullah

Abstract

Purpose: *To examine how supervisor feedback practices mediate the relationship between workplace-based assessment (WBA) data - numerical scores and narrative comments- and trainee learning in the fellowship of College of Physicians and Surgeons (FCPS) Internal medicine training programme and to determine the association between written feedback quality and the observed dynamics of trainee-supervisor feedback dialogue in peripheral public-sector hospitals in Khyber Pakhtunkhwa (KPK), Pakistan.*

Method: *A sequential explanatory mixed-method design was employed. In the Quantitative phase, 120 WBA forms completed by 30 trainee and 15 supervisors were evaluated for Likert-scale, narrative comments quality, specificity, actionability and word count; Spearmen rank correlation and clustered regression analysis examined associations. In the qualitative phase, 10 purposively sampled trainee-supervisor dyads were directly observed during post-WBA feedback conversation, with structured rating of conversation duration, dialogue exchange, trainee-initiated questions, action planning and sense-making. Inter-rater reliability was established using Cohen's kappa prior to independent coding.*

Result: *Mean WBA score was 3.50 (SD, 0.87), while mean narrative feedback quality was notably lower at 2.75 (SD, 0.94); specifically and actionability sub-scores fell further below the midpoint at 2.49 (SD, 1.00) and 2.52 (SD, 0.94) respectively. Median comment word count was 14*

words (IQR, 5-25). Feedback quality correlated with specificity (ρ=0.56, p<0.01) and word count (ρ=0.48, p<0.01). At dyad level, higher written feedback was associated with longer conversations (ρ=0.72), more dialogic turns (ρ=0.71), greater trainee question frequency (ρ=0.86), and stronger observed sense-making (ρ=0.59)

Conclusion: *written feedback on WBA forms meaningfully mediate from assessment data to trainee learning in the FCPS training context. Where supervisors wrote specific, behaviorally grounded comments with a clear next step, trainees engaged more- they asked more questions, stayed longer in conversation and understood more of what was expected of them.*

INTRODUCTION

Something becomes apparent when you spend time in peripheral training hospitals in Khyber Pakhtunkhwa: WBA forms get completed, but trainees leaves the feedback encounter looking no clearer about what to do differently than when they walked in. The form was ticked, the score assigned, a few words written. Formally, the assessment was done. Educationally, very little had happened. This study grew out of that observation. WBA instrument- the Mini-CEX, DOPS and case-based discussion- were in traduced into the FCPS Internal medicine training programme to give clinical supervisors a structured method for observing and documenting trainee performance. Their adoption has been formally mandated and widely implemented [1]. What has been carefully examined is whether the written records they generate – the narrative comments that accompany each numerical score- are actually doing the developmental work they were intended to do.

The feedback literature is unambiguous on this point: a number alone teaches nothing. The educational valve of any assessment instrument is carried primarily by the quality of the feedback that follows it- specifically by whether that feedback is grounded in observed behavior, identifies what needs to change and points towards how [2]. What the evidence also shows, consistently across settings, is that this standard is rarely met. Narrative comments WBA forms tend to be brief, generic and evaluative rather than developmental- phrases that confirms a judgment but offer no map for improvement [3].

This matter not just as an abstract quality concern but as a practical one, particularly in setting like KPK peripheral hospitals where time is scarce and hierarchical culture is strong. Contemporary feedback theory argues that the moment of educational valve is not when the supervisor writes the comment but when the trainee actively makes sense of it- when they can interrogate it, connect it to what they experienced and form a plan [4]. That sense making process depends on conversation and conversation, as this study shows, depends heavily on what was written down first.

In FCPS training sites across KPK, supervisors simultaneously carry full ward responsibilities and are expected to function as assessors and educators. Protected teaching times is minimal. The cultural authority gradient between senor consultants and junior trainees is steep and largely unchallenged. In these conditions, feedback is often rushed or perfunctory and power differential makes it difficult for trainee to ask questions even when they want to [5, 6].

Against that backdrop, this study examined specifically how supervisor feedback practice- both written and spoken – Mediate the pathway between WBA data and trainee learning and what features of written narrative comments are most closely associated with meaningful feedback dialogue.

Methods

Study design:

A sequential explanatory mixed-methods design was employed [7]. Quantitative analysis of WBA forms established the range and pattern of feedback quality features and their associations. A

subsequent qualitative phase – direct observation of feedback conversations- provided explanatory depth, allowing examination of how difference in written feedback quality corresponded with observable difference in what actually happened between supervisors and trainees. The study is reported in accordance with the good reporting of mixed methods study (GRAMMS) guidelines. Ethical approval was obtained from institutional Review Board prior to data collection (IRB No. 11101-2026894-9), and written informed consent was secured from all the participants.

Setting and participants:
The study was conducted across three public – sector tertiary care hospitals in KPK, Pakistan all accredited as FCPS Internal Medicine training sites. These hospitals have bed occupancy consistently above 90%, minimal protected teaching time and heavy clinical service demands on both supervisors and trainees. For the quantitative phase, 120 completed WBA forms from 30 trainees and 15 supervisors were included; all forms had at least one written narrative comments. For the qualitative phase, purposive sampling selected 10 trainees – supervisor dyads designed to represent a board range of written feedback quality, from specific and behaviorally anchored comments to brief evaluated statements.

Quantitative data collection and measures:
Each of the 120 WBA forms was independently evaluated by two trained assessors using a standardized coding protocol. Dimensions included: overall WBA performance score (1-5); narrative comment quality (1-5 global rating); specificity; actioability; behavioural anchoring; forward guidance; and comment word count. Disagreements were resolved by consensus. Inter-rater was calculated using Cohen's kappa prior to independent coding (Table 1).

Qualitative data collection and measure:
Each of the 10 dyads was absorbed during a naturally occurring post-WBA feedback conversation. Observation covered: conversation duration (minutes); dialogic turns (reciprocal exchanges); trainee-initiated questions frequency; action planning (yes/ partial/ No); and observed sense- making rated on a 1-5 scale anchored to demonstrated trainee comprehension, ability to articulate next steps, and active cognitive engagement. Conversations were audio-recorded with consent and transcribed verbatim.

Statistical analysis:
Data was analyzed using SPSS version 25 (IBM Corp., Armonk, NY, USA). Descriptive statistics were computed for all variables. Spearman's correlation coefficient (ρ) was used for bivariate associations; $\rho < 0.05$ was considered statistically significant. Clustered regression analysis accounted for non-independence from multiple trainee rated by the same supervisor. Cross-phase spearman correlations linked written feedback quality scores with qualitative observation variables at the dyad level.

Qualitative analysis:
Transcripts were analyzed using inductive thematic coding. An initial coding framework derived from the research questions was refined through constant comparative analysis across dyads. Qualitative ratings were then examined against written feedback quality scores to explore convergence across strands.

Results

Inter-rater reliability:
Cohen's kappa valves ranged from k=0.71 to k=0.78 across all narrative coding dimensions indicating substantial agreement (Table 1)

Table 1. Inter-rater reliability for narrative feedback coding dimensions

Narrative dimension	Cohen's κ	Agreement level
Overall narrative quality	0.78	Substantial

Narrative dimension	Cohen's κ	Agreement level
Specificity	0.74	Substantial
Actionability	0.76	Substantial
Behavioural anchoring	0.71	Substantial
Forward guidance	0.73	Substantial

Agreement categories follow Landis and Koch (1977): κ 0.61–0.80 = substantial.

WBA form characteristics and feedback quality:
Mean overall WBA performance score was 3.50 (SD, 0.87) – notably higher than mean narrative feedback quality at 2.75 (SD, 0.93). Specificity (mean, 2.49; SD, 1.000) and actionability (mean 2.52; SD 0.94) both fell down the scale midpoint. Median comment word count was 14 words (IQR, 5- 25). Fewer than 20% of the comments contained behavioural anchor; forward guidance was present in less than one quarter (Table 2)

Table 2. WBA form characteristics and narrative feedback quality (N=120)

Measure	Statistic	Value
Trainees	N	30
Supervisors	N	15
Overall WBA score (1–5)	Mean (SD)	3.50 (0.87)
Narrative comment quality (1–5)	Mean (SD)	2.75 (0.93)
Specificity (1–5)	Mean (SD)	2.49 (1.00)
Actionability (1–5)	Mean (SD)	2.52 (0.94)
Comment word count	Median (IQR)	14 (5–25)

IEER
Institute for Excellence in Education & Research

WBA, workplace-based assessment; SD, standard deviation; IQR, interquartile range. All ratings on a 1–5 Likert scale unless otherwise stated.

Associations between feedback quality and dialogue indicators:
Spearman correlations are presented in Table 3. At the form level, feedback quality correlated with specificity (ρ =0.56, p<0.01) and word count (ρ=0.48, p<0.01). The correlation between overall WBA score and narrative quality was weak (ρ=0.32, p<0.05), and between WBA score and actionability weaker still (ρ=0.23, p<0.05) – a consistent pattern of decoupling between how trainee was rated and how richly that rating was explained. At the dyad level, written quality correlated with conversation duration (ρ=0.72), p<0.01), dialogic turns (ρ=0.71, p<0.01), trainee-initiated questions (ρ=0.86, p<0.01) and observed sense-making (ρ =0.59, p<0.05).

Table 3. Spearman correlations between WBA feedback features and dialogue indicators

Sample	Variable pairing	ρ	Significance
WBA forms (N=120)	Comment quality vs. Specificity	0.56	p<0.01
WBA forms (N=120)	Comment quality vs. Word count	0.48	p<0.01
WBA forms (N=120)	WBA score vs. Comment quality	0.32	p<0.05

Sample	Variable pairing	ρ	Significance
WBA forms (N=120)	WBA score vs. Actionability	0.23	p<0.05
Dyads (n=10)	Written quality vs. Conversation duration	0.72	p<0.01
Dyads (n=10)	Written quality vs. Dialogic turns	0.71	p<0.01
Dyads (n=10)	Written quality vs. Trainee questions	0.86	p<0.01
Dyads (n=10)	Written quality vs. Observed sense-making	0.59	p<0.05

ρ, Spearman rank correlation coefficient; WBA, workplace-based assessment. *p<0.05; **p<0.01.

Dyad-level profiles:

Table 4 presents Individual data for all 10 observed dyads ordered by descending written feedback quality. A clear gradient is visible: as written quality declined from 4.2 to 1.5, conversation duration fell from 12 to 2 minutes, trainee questions dropped from 7 to 0 and sense-making ratings declined from 4 to 1. No dyad with written quality below 2.5 produced a sense-making rating above 2.

Table 4. Individual dyad profiles: written feedback quality and observed conversation indicators (n=10)

Dyad	Written quality (1–5)	Duration (min)	Dialogic turns	Trainee Qs	Sense-making (1–5)
1	4.2	12	18	7	4
2	3.9	11	15	6	4
3	3.5	9	12	5	3
4	3.1	7	10	4	3
5	2.8	6	8	3	2
6	2.5	5	7	2	3
7	2.3	4	6	2	2
8	2.0	4	5	1	2
9	1.8	3	4	1	1
10	1.5	2	3	0	1

Dyads ordered from highest to lowest written feedback quality. Written quality and sense-making rated on 1–5 scale.

Qualitative thematic findings:

Analysis of the 10 transcripts yielded three themes. The first- written comments as conversational scaffold – described how specific, behaviourally anchored written comments gave trainees named referent: something concrete enough to ask about without risking a challenge to the supervisor's overall judgment. The second evaluate closure, described the opposite: brief evaluative comments signaled that the assessment was complete, leaving no socially legitimate conversational opening. The third, hierarchy-mediated silence captured the way trainees in this cultural context actively suppressed questions when written comments lacked specificity – not from disinterest but from awareness that vague questioning could be read as

criticism of the supervisor. Several trainees articulated this directly this directly during post-observation informal discussions.

Discussion:

The central question this study asked was straightforward but its implications are not: does the quality of what a supervisor writes on a WBA form shape what happens educationally in the conversation that follows? Based on 120 WBA forms and direct observation of 10 trainee-supervisor pairs in peripheral FCPS training sites in KPK, the answer is clearly yes- and the relationship is stronger than expected.

The Gap between mean WBA performance scores (3.50) and mean narrative feedback quality (2.75), with further reductions in specificity (2.49) and actionability (2.52), is not incidental. It tells us that the supervisors in this setting regularly produce ratings that are more generous than the comments that accompany them. The numerical score offers a confident verdict; the written comment often provides little to support it developmentally. Pelgrim and colleagues documented essentially the same pattern in Dutch general practice training-narrative comments on Mini-CEX forms were predominantly evaluative and unspecific regardless of supervisor experience and feedback conversations were rarely sustained or dialogic [8,9]. The persistence of this finding across contexts and time reflects something structural: writing specific, behaviouraly anchored feedback requires skills that most clinical supervisors have never been formally taught and systems that rarely enforce and narrative standard.

The specificity-quality correlation (ρ =0.56) confirms the rater recognize specificity as the primary marker of a comment's worth. Shute's extensive review of formative feedback research identified specificity and elaboration as the dimensions most consistently associated with learning gains across educational domains [10]. And our data extend that finding into a direct observational context. A comment that names what was observed, in which patient encounter, with what consequence, gives the trainee a cognitive anchor. A comment that says only "good history taking" gives them nothing to work with.

The decoupling of WBA scores from feedback quality- correlations of only of ρ =0.32 and ρ =0.23- confirms an argument Holmboe and colleagues have long made: the educational valve of direct observation lies not in the rating but in the conversation the rating [11]. In or dataset, low specificity comments did not just fail to inform- they foreclosed the conversations. In dyad after dyad, brief written evaluations were followed by monologic encounters where the supervisor restated the score and the trainee offered no of response. The written comment had already ended the exchange before it began.

The strongest single finding is the association between written quality and trainee-initiated questions (ρ =0.86). This is not merely a statistical relationship; it describes a mechanism. A specific written comment-one that names a behavior in a clinical context – gives the trainee a bounded, legitimate object of inquiry. They can ask about the behavior without the question carrying an implied critique of the supervisor. Ajjawi and Boud's analysis of feedback dialogue argued that written artifacts shape the relational process that follows: what can be asked, how and by whom [12]. The present data provide observational evidence for this in a hierarchical, resource-constrained context that existing feedback frameworks have no adequately theorized.

The quality theme of hierarchy-mediated silence deserves direct attention. In KPK peripheral hospitals, the authority gradient between consultants and trainee is steep, culturally reinforced and – from trainee' perspective- not safely challengeable. What emerged from observations and informal post-session discussions was those trainees were not simply quiet; they are strategically quiet. They want to ask questions but in the absence of specific written comment, any question felt like it might be heard as "you didn't explain this well enough" – a social risk too large to take. Specific written feedback removed that risk. It created a named conversational object that both parties could examine without either party losing face. Fuentes-cimma et al. identified power asymmetry as an undertheorised moderator of feedback effectiveness in non- Western training

settings [6]; the present findings give that observation a specific mechanism.

The practical implication of the duration and dialogic turn correlations (ρ =0.72 and ρ =0.71) is important for administrators who read research about feedback and conclude that supervisors simply need more time. More time is not always the answer, and in high workload peripheral hospitals it is rarely available. The data suggest different intervention: the same 2 minutes of conversation can be far more educationally productive if the supervisor arrived at it having written one sentence that names what they observed and one sentence that says what to do differently. That is achievable within the existing time constraints. Watling and Lingard's work on feedback culture documented that trainee in programmes where written records were systematically substantive described feedback as generative and meaningful, not as a formality [13]. The written record shapes the culture of the conversation, not just its content.

The sense-making correlation (ρ =0.59), while the weakest in the dyad dataset, is the most educationally consequential. Sense-making is the mechanism through which feedback actually produces learning- the moment when the trainee connects external information to their own understanding of their performance and generates a plan [14]. That not a single dyad with written quality below 2.5 produced a sense-making rating above 2 suggests something close to a threshold effect: below a minimum standard of written specificity, the conditions for sense-making may simply not exist, regardless of how much goodwill exists on both sides.

Limitations of this study must be stated plainly. Ten dyads is a small qualitative sample; the cross-phase Spearman correlations describe pattern not causation, and their precision should not be overstated. Observed sense-making is a proxy indicator, not a longitudinal outcome. We cannot determine from this design whether supervisors who write better comments also tends to conduct better conversations as an expression of a shared underlying educational disposition, or whether the written quality enables the conversation. Both possibilities have practical implications and both warrant investigation with larger, prospective designs. The single setting nature of the study limits direct generalizability, though the contextual conditions – resources constraints, hierarchical culture, and limited teaching time – are widely shared across LMIC postgraduate training environments.

The practical recommendations from these findings are specific and do not require new resources. First, FCPS training programmes should revise WBA form design to require two structural elements in the narrative comment field before the form can be submitted as complete: one named observable behavior and one concrete next step. This is a form design change, not a training intervention and it make effect immediately. Second, faculty development should treat written and verbal feedback as a single communicative competency, not as separate administrative and educational tasks- because the data show they function as one. Third, narrative quality indicators- word count, presence of behavioural anchor, presence of action step – are monitorable through routine WBA data without additional data collection and can be reported to supervisors as developmental feedback on their feedback practice.

Conflict of interest:
No potential conflict of interest relevant to this article was reported

Funding:
No funding was received for this study.

Data availability:
The anonymised quantitative dataset (WBA form coding scores) and qualitative observation data (dyad profiles) supporting the findings of this study are available as supplementary files submitted with this manuscript. Data will be deposited in Harvard Dataverse upon acceptance. De-identified data are also available from the corresponding author on reasonable request, subject to institutional review board conditions.

Acknowledgments:
The author thanks all the trainee physicians and supervisors who participated in this study, and the administrations of the participating hospitals for facilitating access during routine clinical activity.

REFERENCES:

1. ten Cate O. Competency-based postgraduate medical education: past, present and future. GMS J Med Educ. 2017;34(5):Doc69. https://doi.org/10.3205/zma001146
2. Natesan S, Ayyar S, Lal S. Feedback in medical education: a narrative review of the evidence, gaps, and future directions. Med Teach. 2023;45(6):748-755. https://doi.org/10.1080/0142159X.2023.1864735
3. Leclair M, Mills K, Spencer C. Written feedback on workplace-based assessment forms: content, quality, and educational value. Med Teach. 2023;45(4):412-419. https://doi.org/10.1080/0142159X.2023.1892986
4. Weller JM, Morrison S, Pritchard T. Sociocultural perspectives on feedback in clinical training: implications for practice and research. Med Educ. 2023;57(9):982-993. https://doi.org/10.1111/medu.14812
5. Alam L, Khan ZM, Shafi MN. Workplace-based assessment in Pakistan: challenges, implementation, and opportunities for reform. Pak J Med Sci. 2022;38(5):1237-1243. https://doi.org/10.12669/pjms.38.5.6117
6. Fuentes-Cimma S, Carrasco R, Perez-Arauz M. Enhancing feedback practices in workplace-based assessments: a scoping review of barriers, enablers, and contextual factors. BMC Med Educ. 2024;24(1):43. https://doi.org/10.1186/s12909-024-05672-w
7. Creswell JW, Plano Clark VL. Designing and Conducting Mixed Methods Research. 3rd ed. Thousand Oaks (CA): SAGE Publications; 2018. 520 p.
8. Pelgrim EA, Kramer AW, Mokkink HG, van der Vleuten CP. Quality of written narrative feedback and reflection in a modified mini-clinical evaluation exercise: an observational study. BMC Med Educ. 2012;12:97. https://doi.org/10.1186/1472-6920-12-97
9. Pelgrim EA, Kramer AW, Mokkink HG, van der Vleuten CP. The process of feedback in workplace-based assessment: organisation, delivery, continuity. Med Educ. 2012;46(6):604-612. https://doi.org/10.1111/j.1365-2923.2012.04266.x
10. Shute VJ. Focus on formative feedback. Rev Educ Res. 2008;78(1):153-189. https://doi.org/10.3102/0034654307313795
11. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach. 2010;32(8):676-682. https://doi.org/10.3109/0142159X.2010.500704
12. Ajjawi R, Boud D. Researching feedback dialogue: an interactional analysis approach. Assess Eval High Educ. 2017;42(2):252-265. https://doi.org/10.1080/02602938.2015.1102863
13. Watling CJ, Lingard L. Toward meaningful evaluation of medical trainees: the influence of participants' perceptions of the process. Adv Health Sci Educ Theory Pract. 2012;17(2):183-194. https://doi.org/10.1007/s10459-011-9307-1
14. Gauthier S, Cavalcanti R, Nair V, Waddell A. Formative feedback in residency training: is it actually used for improvement? A scoping review. Acad Med. 2021;96(11):1579-1589. https://doi.org/10.1097/ACM.0000000000004218

15. Shah MI, Qayum I, Bilal N, Ahmed S. Evaluation of Mini-CEX as a workplace-based assessment tool in a resource-limited postgraduate training programme in Khyber Pakhtunkhwa. Prof Med J. 2021;28(9):1346-1350. https://doi.org/10.29309/TPMJ/2021.28.09.6120

IEER
Institute for Excellence in Education & Research