2026, Volume 58, Issue 2, 123-131, e-ISSN 1938-3800

Family Medicine

ORIGINAL ARTICLE

Conceptual Framework and Methods for Evaluating an Intervention to Improve Psychological Safety in Graduate Medical Education

Kate Rowland, MD, MS[a]; Lauren Anderson, PhD, MEd[a]; Elizabeth F. Avery, MS[a]; Cynthia Hays, PhD[b]

AUTHOR AFFILIATIONS:

[a]Department of Family and Preventive Medicine, Rush UniversityChicago, IL

[b]Rush Copley Family Medicine Residency, Aurora, IL, United States

CORRESPONDING AUTHOR:

Kate Rowland, Department of Family and Preventive Medicine, Rush University, Chicago, IL, Kathleen_rowland@rush.edu

HOW TO CITE: Rowland K, Anderson L, Avery EF, et al. Conceptual Framework and Methods for Evaluating an Intervention to Improve Psychological Safety in Graduate Medical Education. *Fam Med.* 2026;58(2):123-131. doi: 10.22454/FamMed.2026.110565

FIRST PUBLISHED: February 12, 2026

KEYWORDS: conceptual framework, graduate medical education, program evaluation, psychological safety

ABSTRACT

Background and Objectives: Psychological safety in graduate medical education influences key aspects of the clinical learning environment, including feedback and assessment. This study aimed to evaluate a 6-month, workshop-based intervention targeting four elements of psychological safety: culture, bias, power, and the hidden curriculum. The study used qualitative and quantitative methods to assess changes in knowledge, attitudes, and observable behaviors.

Methods: The participants were residents and faculty from two Chicago-area family medicine residency programs. The intervention was a series of four 2-hour workshops exploring different aspects of psychological safety and the clinical learning environment. Outcomes included change in the clinical learning environment, assessed by change in faculty feedback behavior on resident assessment forms and change in resident and faculty precepting room behavior. Additional outcomes included impact on individuals, assessed by responses to workshop surveys, pre- and poststudy surveys, and the standardized Postgraduate Hospital Educational Environment Measure. The study applied the Kirkpatrick Model to structure the evaluation of psychological safety interventions, focusing on reaction, learning, and behavior outcomes. A modified version of Miller's Pyramid of Clinical Competence informed the design of evaluation methods for behavior change. Data were analyzed using descriptive statistics and logistic regression to evaluate differences before and after the intervention.

Conclusions: Conceptual frameworks for program evaluation, clinical competence, and workshop content informed the design of this study. We evaluated outcomes for psychological safety, including observed change in teaching behaviors, observed change in written feedback behaviors, and participant self-report.

INTRODUCTION

Psychological safety is the concept that individuals are free to be themselves, including the freedom to share weaknesses without fear of shame or retribution.[1] Lower psychological safety in health care is associated with worse outcomes such as in patient safety, while better psychological safety is associated with outcomes such as clinician wellness.[2,3] In graduate medical education (GME), psychological safety is influenced by factors common to all teams, such as expectations and conscious and unconscious bias.[4-8] It is also affected by factors specific to GME, including graduated entrustment and inauthentic feedback and assessment.[9-12] The Accreditation Council for Medical Education (ACGME) has established a requirement for psychological safety in all GME programs, but few studies have examined interventions to improve psychological safety in GME.[13-15]

Psychological safety has few objective measures and is often studied using self-report or behavioral proxies; this approach has been noted as a barrier to research.[15] When considering the methods for evaluating a psychological safety or other health professions education

(HPE) intervention, researchers must ask the right questions, use valid measures, examine relevant outcomes, and ensure that methods are repeatable by independent investigators.[16-18] The results should be implementable and practical.[17] Identifying and implementing the best evaluation methods to achieve these objectives requires an understanding of the conceptual frameworks and educational theories underlying the research questions and study outcomes.[19-21]

The thoughtful application of a conceptual framework to a research question may result in a study design that leads to data from more than one source.[22] Data may also be in more than one format and can be integrated in many ways depending on the source and outcomes being studied.[23] The use of multiple sources of data to answer a single question improves the strength and validity of the findings through triangulation.[24,25] An effective research study must have a clear methodologic framework that aligns the data collection and analysis to answer the research question clearly.

This study sought to describe the methods used to determine the effects of a 6-month intervention addressing four aspects of psychological safety in GME: the culture of medical education, the hidden curriculum, power, and bias. We hypothesized that the intervention would improve individual participant self-report of knowledge, attitudes, and impression of the clinical learning environment (CLE). We also hypothesized that the intervention would improve observable feedback and teaching behaviors. We assessed changes in psychological safety in the CLE in two areas: self-reported impact on individual participants' knowledge, skills, and attitudes, and change in participant behavior based on resident assessment data and precepting room observations. (Table 1) This study used qualitative and quantitative methods to evaluate these changes.

METHODS

Participants

The participants were residents and faculty from two Chicago-area family medicine residency programs. Site one has four residents per class in a community-based university-affiliated setting; site two has eight residents per class in a federally qualified health center-based, university-affiliated teaching health center funded setting.

Intervention

The intervention was a series of four 2-hour workshops exploring different aspects of psychological safety and the clinical learning environment: psychological safety itself, the culture of medical education, expectations, and power and bias. The workshops were designed and delivered by subject-matter experts, with the goal of increasing understanding and fostering change in these topic areas. They were held during regularly scheduled residency didactic time, and faculty and residents attended together. The sessions were structured to encourage engagement, reflection, and action. Sessions were held in-person, virtual/hybrid, and in one instance via video recording of the presentation.

Ethical Considerations

All residents and faculty at both sites were invited to participate in the study, and all residents were required to attend the workshops. However, only those who provided consent to participate in research had data collected and included for analysis. The principal investigators (PIs) were a faculty member at one of the participating residencies and a senior department leader. Therefore, to mitigate the possibility of coercion, the study was introduced and informed consent obtained by a research team member unaffiliated with the residency programs. Invitations to complete study surveys were also sent by research team members unaffiliated with the residency programs. To avoid influencing the resident or faculty behavior, the PIs were not present during observation and data collection. The PIs did not have access to study data prior to de-identification during data analysis. This study was approved by the Institutional Review Boards at Rush University Medical Center and Rush Copley Medical Center.

Choice of Conceptual Frameworks

Conceptual frameworks clarify and define problems, develop potential solutions, and structure studies by guiding the formulation of research questions, choice of methods, and interpretation of findings. Conceptual frameworks use existing theories or models to create a shared vocabulary and structure for a study. Incorporating multiple frameworks offers complementary perspectives on both content and methods.[20]

The Kirkpatrick Model is a foundational framework for evaluating educational interventions in HPE. It comprises four levels: Reaction (learner satisfaction and engagement), Learning (knowledge acquisition, attitude, confidence and commitment), Behavior (application of skills), and Results (impact on organizational outcomes).[26] In HPE, this model has been applied to across a range of programs.[27] We chose to evaluate outcomes at several Kirkpatrick levels, including Reaction, Learning, and Behavior.

We used the Kirkpatrick model to identify an overall methodological approach. We drew on Miller's Pyramid of Clinical Competence to determine our approach to measuring the changes in behavior.[28] Miller's Pyramid provides a framework for assessing clinical competence, progressing from foundational knowledge ("knows") to performance in clinical settings ("does"). It reflects that true clinical competence must be demonstrated through observable behavior in real-world contexts, not solely through knowledge or simulation. Although psychological safety is not a traditional clinical skill, it is analogous in that it can be taught and measured across levels of mastery. We specifically drew on a modified framework (Miller's Prism) that emphasizes the assessment of knowledge, skills, and attitudes, as well as assessment methods of cognition and

TABLE 1. Study Framework and Research Plan

Research question	Hypothesis	Associated outcome(s)	Data source	Data detail
Does a 6-month intervention improve psychological safety in two family medicine graduate medical education programs?	Participation in the workshop intervention will improve feedback and increase resident requests for help, indicators of psychological safety. Participation in the intervention will decrease negative comments, power-associated comments, and biased comments, which are negatively associated with psychological safety.	Observed change in observed teaching/ learning behaviors	Direct observation in precepting room	• Teaching incidents • Feedback incidents • Help-seeking statements • Negative comments • Power comments • Bias statements
	Participation in the workshop intervention will improve the quantity, quality, and directness of written feedback.	Observed change in feedback and assessment behaviors	Resident end-of-rotation and work-based assessment forms	• Distribution of numeric scores (overall) • Distribution of numeric scores (by ACGME competency) • Relevance of written comments • Orientation of written comments • Politeness strategies of written comments
	Participation in the workshop intervention will improve participant self-report of knowledge and skill about factors associated with psychological safety.	Impact on individuals: study-long	• Before and after intervention survey • PHEEM	
	Participation in the workshop intervention will increase scores on a standardized scale of the clinical learning environment.	Impact on individuals: workshop	Workshop evaluations	• Knowledge • Attitudes • Satisfaction

Abbreviations: ACGME, Accreditation Council for Graduate Medical Education; PHEEM, Postgraduate Hospital Education Evaluation Measure

behavior for our study design and evaluation plan.[29] Thus, we used aspects of both the Kirkpatrick Model and Miller's Pyramid to choose rigorous methods of measuring study outcomes. The overlap between the two frameworks is imperfect (Table 2).

Data Collection and Analysis

To measure change in the psychological safety of the CLE using feedback and assessment, this study measured the impact of the intervention on two different outcomes: behavior change and individual participant reaction through self-report. Behavior change was measured objectively through observation of teaching and learning in the precepting room and through numeric and written comment data from resident assessment forms. Participant reaction was measured through before- and after-intervention surveys and individual post workshop evaluations. (Table 3).

Behavior Change: Precepting Observation Data

The observation data were collected by a trained research assistant. The research assistant spent at least four half-day sessions at each residency's family health center observing outpatient precepting interactions between residents and faculty. Observations occurred before and after the intervention. The research assistant used a data collection form for each precepting interaction. Data included the presence of teaching and feedback, help-seeking behaviors of residents, and positive and negative language used. Identifiers were not collected.

Precepting Observation Data Analysis

We used descriptive statistics to describe the presence and patterns of data collected. Differences before and after the intervention were examined using χ^2 tests. Due to the lack of identifiers, no sensitivity analyses related to faculty or resident clustering were performed.

Resident Assessment Data

We measured behavior change through resident assessment forms. Resident assessment forms completed by faculty were downloaded from the residency management software (MedHub [MedHub LLC]) at both programs. We analyzed both the written comments and numeric scores. During de-identification of the data, link identifiers were created to later adjust for clustering.

TABLE 2. Connection Between Miller's Pyramid of Clinical Competence and Kirkpatrick Model of Training Evaluation

Kirkpatrick level	Miller's Pyramid level	What it measures	How to measure
1: Reaction			
2: Learning	Knows/ knows how	Knowledge acquisition	Tests, surveys
3: Behavior	Shows how	Skills	Simulation, OSCE
	Does	Real-world performance and outcomes	Direct observation
4: Results	Does+	Mission-level results	Organization-level metrics

Note: Satisfaction (Kirkpatrick level 1) does not walk over to clinical competence.
Abbreviation: OSCE, objective structured clinical examination

TABLE 3. Conceptual Approach to Data and Analysis

	Data source	Analysis	Kirkpatrick model level	Miller's level of clinical competence
Change in learning environment		**Assessment forms**		
	Numeric scores	Distribution of numeric scores	Level 3: Change in behavior	Does
		Distribution of scores by ACGME competency	Level 3: Change in behavior	Does
	Written comments	Comment relevance and orientation categorized using a rubric	Level 3: Change in behavior	Does
		Presence of politeness strategies within comments	Level 3: Change in behavior	Does
		Observations		
	Precepting room observation form	Interactions between preceptors and residents	Level 3: Change in behavior	Does
Impact on individual		**Workshops**		
	Individual workshop evaluation	Satisfaction Attitudes Future changes/commitment Confidence	Level 1: Reaction Level 2: Learning	–
	Knowledge exam	Knowledge pre and post	Level 2: Learning	Knows
		Participant surveys		
	Modified Postgraduate Hospital Educational Environment Measure (PHEEM)	Perception of clinical learning environment (Likert data)		–
	Participants self-assessment (pre/post intervention)	Knowledge Perceived skills Attitudes	Level 2: Learning	Knows

Abbreviation: ACGME, Accreditation Council for Medical Education

Numeric Scores

We analyzed numeric scores within program because the sites used different assessment metrics. Where possible, assessments with similar metrics were combined. All metrics were ordinal Likert scales; one site used a 6-point scale for most items, and one site used a 5-point scale for most items. The number of items on the assessments ranged from 5 to 21. They were analyzed as both ordinal and nominal data using ordinal and multinomial logistic regression, respectively. We used descriptive statistics to describe the patterns of data collected. We assessed differences in the distribution of outcomes before and after intervention using logistic regression. We performed sensitivity analyses to explore possible data clustering by participant.

Written Comments

Each assessment form had at least one mandatory response field for written comments, but the total number of fields available to leave comments ranged from one to seven. Comments were extracted individually from the forms and analyzed independently of any other comments from the same assessment.

We analyzed the written comments using a previously published coding rubric for nature of feedback in narrative comments to examine before- and after-intervention changes in comment orientation (positive or critical feedback) and comment relevance[30-32] (Table 4). Prior to applying the rubric, the research team participated in a training session that involved reviewing operational definitions and multiple example comments to ensure a shared understanding of

each rating dimension. Two researchers then independently reviewed a sample of 32 comments and scored them for calibration. Each two-category scale was collapsed to a binary category (ie, relevant and irrelevant; critical or positive). The results of the sample were discussed, and consensus for each was reached. Both researchers independently scored the remaining comments for relevance and orientation.

The written comments initially were analyzed using a qualitative framework and transformed during the analysis into categorical data. This allowed the written comment data to be analyzed using descriptive statistics. We additionally looked at differences in distribution before and after the intervention using logistic regression. Data were not paired pre- and postintervention. Sensitivity analysis was performed adding a cluster adjustment for comments within the same assessment. A separate analysis was performed using faculty as the cluster adjustment.

Written comments were also coded using a similar two-reviewer method for the presence of two strategies of politeness: hedging and indirectness. These linguistic features are indicators of less direct and authentic feedback.[33]

Workshop Evaluations

After each session, participants completed a 6-item postevaluation assessing satisfaction (alignment with learning objectives, content, and pacing), attitudes toward the topics, confidence in applying changes, and personal commitment to future actions. The evaluation included both open-ended and multiple-choice questions. One workshop also included a brief pre- and postknowledge exam. Responses were analyzed using basic descriptive statistics along with simple coding for themes of open-ended responses.

Impact on Individual Participants

Survey instruments were comprised of two parts. Part one included a mix of Likert and open-ended response items created by the study team based on the content of the workshops and literature review. Part two was modified from the validated Post Graduate Hospital Educational Environment Measure (PHEEM), a 40-item inventory of the clinical learning environment.[34] Modifications updated language to reflect US training and vernacular. For example, an item originally written to ask about training hours compliance in alignment with the United Kingdom New Deal was revised to ask about compliance with ACGME requirements. Items on the PHEEM factor into domains including autonomy, social support, and teaching, among others. Each item on the PHEEM is scored on a 5-point Likert scale. This results in a final score of 0 to 160, with higher scores indicating a more positive educational environment.

Participants in this study were surveyed via REDCap (Research Electronic Data Capture) software before and after the intervention. To reduce participant burden, resident participants were offered time during scheduled educational activities to complete the surveys. Up to three reminders were sent via email. Before- and after-intervention changes in distribution of PHEEM scores were evaluated using an unpaired *t* test. Tests were not paired due to the decision not to collect identifiers on the surveys.

DISCUSSION

This project sought to measure resident and faculty response to a series of four workshops designed to improve psychological safety in graduate medical education. This study demonstrates the use of multiple data collection and analysis methods to measure behavior change and participant self-reported outcomes. It also demonstrates the transformation of qualitative, written comment data into categorical data.

Triangulation of Data

Triangulation is the use of multiple sources of data to converge on or approach a single result.[35] Triangulation assumes that one result exists that is best understood by being viewed from multiple lenses, or that the result is incompletely understood when viewed from a single lens. When a research outcome requires more than one method of evaluation to be fully understood, triangulation can be used to align multiple sources of data. Triangulation is used in mixed methods and multimethods study designs. Although some do not draw a distinction between these two designs, others indicate that mixed-methods designs must have substantial mixing of the quantitative and qualitative elements, whereby one component informs the other.[36-39]

To measure impact on assessment and feedback behaviors associated with psychological safety, the sources of data in this study included numeric scores and written comments from assessment forms, observation of teaching interactions, and participant report of knowledge, beliefs, and attitudes. Because the data varied in nature and included text, observation records, and Likert data from multiple sources, both qualitative and quantitative analysis were required to answer the research questions being asked. The data were collected in parallel, and the results of each analysis were used to compare, contrast, or reinforce results. In this study the qualitative and quantitative data triangulated to answer the same questions, but the qualitative arm did not inform the design of the quantitative arm.

Conceptual Frameworks

The use of different methods of data collection was informed by conceptual frameworks from education and training. This allowed for a more complete and robust exploration of the concept of psychological safety than a single measure or single framework would have allowed. We also used a conceptual framework for the content area of psychological safety, which is not discussed in detail here.[40] We measured impact on individual participants using self-report data. The use of self-report data is Kirkpatrick level two and evidence of "knows" or "knows how" on Miller's Pyramid. We also asked about satisfaction, commitment, and engagement, which are

TABLE 4. Rubric for Analysis of Written Comments

Comment relevance		
Level	**Definition**	**Example**
Highly relevant	Included specific items that could be used by the resident to improve or sustain practice, the CCC to make milestone placement or entrustment decisions, or by the program director or advisor to help with a learning plan	Documentation is excellent. I would recommend that [resident] work on billing. I am oftentimes seeing isolated billing codes for well-child visits/annual visits, without the appropriate E&M modifier. We walked through a few together; it can be hard if not done often.
Relevant	Helpful but might have lacked some of the specifics or action items; still contains observable behaviors and aligns with competencies or gives insight into progress	Resident continues to take constructive feedback and integrated into practice. IUD technique much improved this time.
Irrelevant	Lacked specific details, vague; offers impressions without actionable information or clear link to behaviors or milestones	Resident is very thoughtful in patient assessment.
Highly irrelevant	Very nonspecific, list of adjectives but not connected to anything specific; not useful in coaching or for program leadership	Open, humble, good work ethic
Comment orientation		
Level	**Definition**	**Example**
High praise	Multiple compliments with specific example(s); often aligned to competencies or demonstrating growth over time	Resident has shown tremendous growth over the rotation in clinical reasoning skills; she was able to independently and accurately complete a history and physical including differential for multiple complex patients.
Moderate praise	Generally positive, may mention improvement/growth and/or strengths but fewer examples	Doing well. She is reliable and always follows through on tasks without reminders.
Critical	Highlights areas for improvement using constructive language; may contain example(s)	Resident needs to work on staying organized when managing multiple patients; sometimes notes are incomplete or delayed when dealing with a schedule appropriate for PGY level.
Highly critical	Primarily negative in tone, may use judgmental language, often lacks constructive example	Resident seems disinterested and dismissive of patient concerns.
Politeness strategies		
Strategy	**Definition**	**Example**
Indirectness	Use of phrases and sentences that have contextually unambiguous meanings but avoid direct or specific critique—often relay on general or even neutral terms Common phrases: solid, sound	Resident demonstrated solid patient care.
Hedging	Lack of commitment to what was said—includes language that introduces doubt, minimizes critique, or shifts responsibility to others Common phrases: sort of, somewhat, a little, perhaps, I think, could have, a little more, as far as anyone could tell, clearly, apparently	Resident appears to have a strong understanding of CHF. Could perhaps benefit from a bit more confidence when leading rounds, but as far as anyone could tell she managed it well.

Note: Politeness strategies can serve to save face for both the feedback giver (faculty) and recipient (resident), maintaining social rapport. Hedging uses qualifiers or attributes statements to others. It softens statements or reduces the giver's perceived ownership or certainty. Similarly, indirectness uses common words (such as "meets expectations") that are understood to mean something else (in this case "below average").
Abbreviations: CCC, clinical competency committee; CHF, congestive heart failure; E&M, evaluation and management; IUD, intrauterine device; PGY, post-graduate year

Kirkpatrick level one results. We measured behavior change in multiple ways, including direct observation of behavior in the learning environment and analysis of resident assessment data. The use of resident assessment data is Kirkpatrick level three and evidence of faculty "does" on Miller's Pyramid. The use of preceptor room observations is also Kirkpatrick level three and evidence of "does" on Miller's Pyramid. These are more rigorous methods of measurement of both evaluation of a training program and of clinical competence compared with self-report.

Assessing multiple Kirkpatrick levels overcomes some of the historic critiques of the framework. These include the assumption that the four levels are progressive and causally linked.[41] Interestingly, the implication from assuming that the levels are linked causally is that at a certain degree of satisfaction or knowledge attainment, behavior would change. Existing literature suggests that satisfaction is not consistently associated with knowledge acquisition or future performance.[41,42] Kirkpatrick's model also has been criticized for the assumption that it is hierarchical, as implied by its pyramidal structure.[41] Assessing multiple outcomes also reduces this possible threat to validity. At the same time, most likely true in this case is that evidence of implementation is a better predictor of future culture than knowledge alone.[43]

Direct observation of participants allowed for measurement independent of participant self-report.[44] Self-reported behavior change is subject to biases, including recall, report, and social desirability biases.[45] Although observation is subject to other biases, including biases created simply by being observed, the behavior is directly witnessed and recorded.[46] The use of multiple forms of data also helps to overcome the possibility that participants are unaware of their own beliefs or their own skill level, which must be acknowledged with self-report data.[47]

Evaluating multiple Kirkpatrick or Miller's levels has challenges. Each step up the Kirkpatrick ladder is time-consuming and expensive.[48-50] Other fields have noted difficulty in accessing the correct tools to appropriately evaluate higher levels of the model.[51]

Generalizability

This study sought to measure change in psychological safety in GME after a brief intervention. Psychological safety and the clinical learning environment are both difficult to measure. Certain aspects of data analysis were limited by our decision to limit collection of identifiers; this decision was made to reduce the likelihood of resident identifiability. This study also prioritized real-world assessment and impact. Although faculty were aware of the intervention and what was being studied, they did not receive additional faculty development on completing assessments. Likewise, outside of the intervention workshops, faculty and residents were not encouraged to model a specific kind of teaching or learning during the precepting room observation. Reminders were not provided between workshops. This likely reduces the measured effect of the intervention in exchange for increasing the real-world applicability of any results.

Because assessment forms were not part of the intervention, they were not standardized pre-hoc. The items in the assessment forms varied by program and the setting being assessed (eg, inpatient vs ambulatory vs specialty rotation). The response options were not consistent between assessment forms and between programs. The number of assessments of a given resident varied within and between each time point.

Collaboration

This study reinforces the need for a comprehensive understanding of study design, research methods, conceptual frameworks, educational theory, and practical aspects of medical education to produce rigorous HPE research.[52] The research team for this study included team members with expertise in HPE, family medicine, psychology, and statistics. The availability of collaborators has been associated with research output in family medicine departments as well.[53]

CONCLUSIONS

Conceptual frameworks for program evaluation, clinical competence, and workshop content informed the design of this study. We evaluated outcomes for psychological safety, including observed change in teaching behaviors, observed change in written feedback behaviors, and participant self-report.

SUPPORT

This work was supported by a grant from the Josiah Macy, Jr. Foundation.

REFERENCES

1. Torralba KD, Jose D, Byrne J. Psychological safety, the hidden curriculum, and ambiguity in medicine. *Clin Rheumatol*. 2020;39(3):667–671. doi:10.1007/s10067-019-04889-4
2. de Lisser R, Dietrich MS, Spetz J, Ramanujam R, Lauderdale J, Stolldorf DP. Psychological safety is associated with better work environment and lower levels of clinician burnout. *Health Affairs Scholar*. 2024;2(7):qxae091. doi:10.1093/haschl/qxae091
3. Grailey KE, Murray E, Reader T, Brett SJ. The presence and potential impact of psychological safety in the healthcare setting: an evidence synthesis. *BMC Health Serv Res*. 2021;21(1). doi:10.1186/s12913-021-06740-6
4. Boysen PG II. Just culture: a foundation for balanced accountability and patient safety. *Ochsner J*. 2013;13(3):400–406.
5. Bing-You RG, Trowbridge RL. Why medical educators may be failing at feedback. *JAMA*. 2009;302(12):1330–1331. doi:10.1001/jama.2009.1393
6. McClintock AH, Fainstad TL, Jauregui J. Creating psychological safety in the learning environment: straightforward answers to a longstanding challenge. *Acad Med*. 2021;96(11S):S208–S209. doi:10.1097/ACM.0000000000004319
7. Sotto-Santiago S, Mac J, Slaven J, Maldonado M. A survey of internal medicine residents: their learning environments, bias and discrimination experiences, and their support structures. *Adv Med Educ Pract*. 2021;12:697–703. doi:10.2147/AMEP.S311543
8. Klein R, Julian KA, Snyder ED, et al. Gender bias in resident assessment in graduate medical education: review of the literature. *J Gen Intern Med*. 2019;34(5):712–719. doi:10.1007/s11606-019-04884-0
9. Torralba KD, Loo LK, Byrne JM, et al. Does psychological safety impact the clinical learning environment for resident physicians? Results from the va's learners' perceptions survey. *J Grad Med Educ*. 2016;8(5):699–707. doi:10.4300/JGME-D-15-00719.1
10. Ende J. Feedback in clinical medical education. *JAMA*. 1983;250(6):777–781. doi:10.1001/jama.1983.03340060055026
11. Ito A, Sato K, Yumoto Y, Sasaki M, Ogata Y. A concept analysis of psychological safety: Further understanding for application to health care. *Nurs Open*. 2022;9(1):467–489. doi:10.1002/nop2.1086
12. Anderson PAM. Giving feedback on clinical skills: are we starving our young? *J Grad Med Educ*. 2012;4(2):154–158. doi:10.4300/JGME-D-11-000295.1
13. Porter-Stransky KA, Horneffer-Ginter KJ, Bauler LD, Gibson KM, Haymaker CM, Rothney M. Improving departmental

psychological safety through a medical school-wide initiative. *BMC Med Educ*. 2024;24(1). doi:10.1186/s12909-024-05794-4
14. Accreditation Council for Graduate Medical Education. ACGME Common Program Requirements (Residency). *ACGME Common Program Requirements (Residency)*. Accessed December 3, 2025. https://www.acgme.org/globalassets/pfassets/programrequirements/2025-reformatted-requirements/cprresidency_2025_reformatted.pdf
15. O'Donovan R, McAuliffe E. A systematic review exploring the content and outcomes of interventions to improve psychological safety, speaking up and voice behaviour. *BMC Health Serv Res*. 2020;20(1):101. doi:10.1186/s12913-020-4931-2
16. Maudsley G, Taylor D. Analysing synthesis of evidence in a systematic review in health professions education: observations on struggling beyond Kirkpatrick. *Med Educ Online*. 2020;25(1). doi:10.1080/10872981.2020.1731278
17. Allen LM, Hay M, Palermo C. Evaluation in health professions education-Is measuring outcomes enough? *Med Educ*. 2022;56(1):127–136. doi:10.1111/medu.14654
18. Ellaway RH, O'Brien BC, Sherbino J, et al. Is there a problem with evidence in health professions education?. *Acad Med*. 2024;99(8):841–848. doi:10.1097/ACM.0000000000005730
19. Samuel A, Konopasky A, Schuwirth LWT, King SM, Durning SJ. Five principles for using educational theory: strategies for advancing health professions education research. *Acad Med*. 2020;95(4):518–522. doi:10.1097/ACM.0000000000003066
20. Bordage G. Conceptual frameworks to illuminate and magnify. *Med Educ*. 2009;43(4):312–319. doi:10.1111/j.1365-2923.2009.03295.x
21. Zackoff MW, Real FJ, Abramson EL, Li S-TT, Klein MD, Gusic ME. Enhancing educational scholarship through conceptual frameworks: a challenge and roadmap for medical educators. *Academic Pediatrics*. 2019;19(2):135–141. doi:10.1016/j.acap.2018.08.003
22. Kwok CS, Muntean EA, Mallen CD, Borovac JA. Data collection theory in healthcare research: the minimum dataset in quantitative studies. *Clin Pract*. 2022;12(6):832–844. doi:10.3390/clinpract12060088
23. Dill MJ, Yunker ED, Brandenburg K, Caulfield M. Secondary data sources for health professions education research: where to look and what you will find. *Acad Med*. 2016;91(12):e7. doi:10.1097/ACM.0000000000001437
24. Cristancho SM, Goldszmidt M, Lingard L, Watling C. Qualitative research essentials for medical education. *Singapore Med J*. 2018;59(12):622–627. doi:10.11622/smedj.2018093
25. Heale R, Forbes D. Understanding triangulation in research. *Evid Based Nurs*. 2013;16(4):98–98. doi:10.1136/eb-2013-101494
26. Kirkpatrick J, Kirkpatrick WK. *An Introduction to the New World Kirkpatrick® Model*. Kilpatrick Partners; 2022.
27. Anderson LN, Merkebu J. The Kirkpatrick model: a tool for evaluating educational research. *Fam Med*. 2024;56(6):403–403. doi:10.22454/FamMed.2024.161519
28. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med*. 1990;65(9 Suppl):S63–7. doi:10.1097/00001888-199009000-00045
29. Mehay R, ed. *The Essential Handbook for GP Training and Education*. CRC Press; 2012. 10.1201/9781846197918
30. Tekian A, Park YS, Tilton S, et al. Competencies and feedback on internal medicine residents' end-of-rotation assessments over time: qualitative and quantitative analyses. *Acad Med*. 2019;94(12):1961–1969. doi:10.1097/ACM.0000000000002821
31. Tekian A, Borhani M, Tilton S, Abasolo E, Park YS. What do quantitative ratings and qualitative comments tell us about general surgery residents' progress toward independent practice? Evidence from a 5-year longitudinal cohort. *Am J Surg*. 2019;217(2):288–295. doi:10.1016/j.amjsurg.2018.09.031
32. Anderson LM, Rowland K, Edberg D, Wright KM, Park YS, Tekian A. An analysis of written and numeric scores in end-of-rotation forms from three residency programs. *Perspect Med Educ*. 2023;12(1):497–506. doi:10.5334/pme.41
33. Ginsburg S, van der Vleuten C, Eva KW, Lingard L. Hedging to save face: a linguistic analysis of written comments on in-training evaluation reports. *Adv Health Sci Educ Theory Pract*. 2016;21(1):175–188. doi:10.1007/s10459-015-9622-0
34. Chan CYW, Sum MY, Lim WS, Chew NWM, Samarasekera DD, Sim K. Adoption and correlates of Postgraduate Hospital Educational Environment Measure (PHEEM) in the evaluation of learning environments - A systematic review. *Med Teach*. 2016;38(12):1248–1255. doi:10.1080/0142159X.2016.1210108
35. Battista A, Torre D, Konopasky A. Essential concepts for effective mixed methods research in the health professions: AMEE Guide No. 173. *Med Teach*. 2025;47(5):766–778. doi:10.1080/0142159X.2024.2401464
36. Creswell JW, Clar VLP. *Designing and Conducting Mixed Methods Research*.3rd ed. SAGE; 2017.
37. Stange KC, Miller WL, Etz RS. The role of primary care in improving population health. *Milbank Q*. 2023;101(S1):795–840. doi:10.1111/1468-0009.12638
38. Hesse-Biber S, Johnson RB, eds. *The Oxford Handbook of Multimethod and Mixed Methods Research Inquiry*. Oxford University Press; 2015. 10.1093/oxfordhb/9780199933624.001.0001
39. Anguera MT, Blanco-Villaseñor A, Losada JL, Sánchez-Algarra P, Onwuegbuzie AJ. Revisiting the difference between mixed methods and multimethods: Is it all in the name? *Qual Quant*. 2018;52(6):2757–2770. doi:10.1007/s11135-018-0700-2
40. Clark TR. *The 4 Stages of Psychological Safety: Defining the Path to Inclusion and Innovation*. Berrett-Koehler; 2020.
41. Reio TG, Rocco TS, Smith DH, Chang E. A critique of kirkpatrick's evaluation model. *New Horizons in Adult Education and Human Resource Development*. 2017;29(2):35–53. doi:10.1002/nha3.20178
42. Uttl B, White CA, Gonzalez DW. Meta-analysis of faculty's teaching effectiveness: student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*. 2017;54:22–42. doi:10.1016/j.stueduc.2016.08.007
43. Braithwaite J, Herkes J, Ludlow K, Testa L, Lamprell G. Association between organisational and workplace cultures, and patient outcomes: systematic review. *BMJ Open*. 2017;7(11). doi:10.1136/bmjopen-2017-017708

44. Mills AJ, Durepos G, Wiebe E, eds. Direct observation as evidence. In: *Encyclopedia of Case Study Research*. SAGE; 2010:302–304 Accessed May 21, 2025. https://methods.sagepub.com/ency/edvol/encyc-of-case-study-research/chpt/direct-observation-as-evidence
45. Choi BCK, Pak AWP. A catalog of biases in questionnaires. *Prev Chronic Dis*. 2005;2(1). https://www.cdc.gov/pcd/issues/2005/jan/04_0050.htm
46. Mahtani K, Spencer EA, Brassey J, Heneghan C. Catalogue of bias: observer bias. *BMJ Evid Based Med*. 2018;23(1):23–24. doi:10.1136/ebmed-2017-110884
47. Dang J, King KM, Inzlicht M. Why Are Self-Report and Behavioral Measures Weakly Correlated? *Trends Cogn Sci*. 2020;24(4):267–269. doi:10.1016/j.tics.2020.01.007
48. Rouse DN. Employing Kirkpatrick's evaluation framework to determine the effectiveness of health information management courses and programs. *Perspect Health Inf Manag*. 2011;8(Spring).
49. Jones C, Fraser J, Randall S. The evaluation of a home-based paediatric nursing service: concept and design development using the Kirkpatrick model. *J Res Nurs*. 2018;23(6):492–501. doi:10.1177/1744987118786019
50. Nestel D, Regan M, Vijayakumar P, et al. Implementation of a multi-level evaluation strategy: a case study on a program for international medical graduates. *J Educ Eval Health Prof*. 2011;8. doi:10.3352/jeehp.2011.8.13
51. Alsalamah A, Callinan C. Adaptation of Kirkpatrick's four-level model of training criteria to evaluate training programmes for head teachers. *Education Sciences*. 2021;11(3):116. doi:10.3390/educsci11030116
52. Sullivan GM. Deconstructing quality in education research. *J Grad Med Educ*. 2011;3(2):121–124. doi:10.4300/JGME-D-11-00083.1
53. Seehusen DA, Koopman RJ, Weidner A, Kulshreshtha A, Ledford CJW. infrastructure features associated with increased department research capacity. *Fam Med*. 2023;55(6):367–374. doi:10.22454/FamMed.2023.736543