

Article

Teaching Motivational Interviewing Skills Using Deliberate Practice with Artificial Intelligence Scoring and Feedback

Richard E. Heyman ^{1,*}, Alexandra K. Wojda-Burlij ¹, Jennifer Piscitello ¹, Kelly A. Daly ¹, Ana Ivic ¹, Anna Segura ^{1,2}, Jasara N. Hogan ¹, Kimberly A. Rhoades ¹, Danielle M. Mitnick ¹ and Amy M. Smith Slep ¹

¹ Family Translational Research Group, New York University, New York, NY 10012, USA; akw286@nyu.edu (A.K.W.-B.); jp4300@nyu.edu (J.P.); kelly.daly@nyu.edu (K.A.D.); ai679@nyu.edu (A.I.); as16200@nyu.edu (A.S.); jnh9335@nyu.edu (J.N.H.); kimberly.rhoades@nyu.edu (K.A.R.); dmm21@nyu.edu (D.M.M.); amy.slep@nyu.edu (A.M.S.S.)

² Salut Mental i Innovació Social, Universitat de Vic—Universitat Central de Catalunya, 08500 Vic, Spain

* Correspondence: richard.heyman@nyu.edu; Tel.: +1-212-998-9984

Highlights

What are the main findings?

- Dental students' initial attempts at motivational interviewing (MI) skills showed that 80% gave advice or tried to fix the patient's problem instead of drawing out the patient's own reasons for change; about one-third of students still struggled with this after instruction.
- After brief MI skill instruction, students fell into three groups. Two-thirds picked up the MI skills quickly. The remaining third split into two groups with distinct error patterns, and those patterns predicted summative performance.

What are the implications of the main findings?

- Artificial Intelligence (AI) personalized the MI portion of a behavior-change course for healthcare providers (four lectures plus four hours of small-group practice for 437 students) by transcribing handwritten MI enactment attempts; scoring them; delivering next-day individualized feedback; generating interactive AI-practice prompts tailored to individual students' needs; and coding nearly 7000 responses for research. AI can code at scale with high agreement with the gold standard.
- AI coding of student errors diagnosed specific changes needed in teaching MI, as some mistakes disappear with brief instruction, but others (the urge to give advice, responding to the sustain talk portion of patient statements, and giving deficient affirmations) require repeated, structured practice on individual skill components.



Academic Editors: Hessam Nowzari and Juan J. Segura-Egea

Received: 22 March 2026

Revised: 8 June 2026

Accepted: 27 July 2026

Published: 3 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Abstract

Background/Objectives: We analyzed 6555 written motivational interviewing (MI) skill attempts from 437 second-year dental students to (a) construct an inductive taxonomy of skill-enactment errors, (b) test whether error prevalence decreased after instruction, and (c) determine whether formative-error profiles predicted summative performance. **Methods:** We used a mixed-methods design. Students responded to patient prompts using specific MI skills (open-ended questions, affirmations, and reflections; 4 per skill) after brief instruction (T1) and during a midterm examination (T2; 3 items). Claude Artificial Intelligence (AI) performed initial coding using a qualitative constant-comparison protocol; the code registry reached saturation at $n = 200$. Five independent human coders validated the AI-generated codes. **Results:** Claude AI coded MI skill attempts with high agreement with the gold standard (linearly weighted kappa = 0.93). The two most prevalent (of seven

derived) errors at T1 were the fixing reflex (advising or correcting the patient; 79%) and responding to or eliciting sustain talk (75%). All error types decreased significantly by T2 ($p < 0.001$). Structural errors (e.g., wrong form) fell below 8%, whereas the fixing reflex (16%) and sustain-talk responding (52%) proved more durable. Latent profile analysis identified three groups: Attuned (64%; hallmark: successful enactment), Misaligned (17%; hallmark: fixing reflex), and Unanchored (18%; hallmark: generic responding). Misaligned students scored significantly lower at T2; Unanchored students matched the Attuned group. Conclusions: Structural errors yield to standard teaching, whereas the fixing reflex and sustain-talk responding require deliberate practice focused on subcomponents of affirming and reflecting. AI made it possible to provide individualized, criterion-referenced feedback to 437 students and to systematically code nearly 7000 responses.

Keywords: motivational interviewing; dental education; deliberate practice; artificial intelligence; constructivist grounded theory; error taxonomy; latent profile analysis; communication skills training; OARS skills; fixing reflex

1. Introduction

Healthcare providers have traditionally relied on information provision and clinical recommendations to modify patients' health-related behaviors [1]. Although providers are trying to be helpful, research indicates such offers are ineffective for most patients because (a) they presume that patients lack information they typically already possess, (b) longstanding behaviors are harder to change and rarely stem from information deficits anyway, (c) knowing and doing are entirely separate domains, and (d) health behavior change involves multiple biological, psychological, and social factors rather than a simple know–commit–change self-regulation process [2–5].

Even though clinicians typically try to confront, warn, or advise patients out of a sincere desire to improve their health, research documents that such attempts ironically backfire, making patients more likely to voice why change is undesirable, unnecessary, or impossible [6]. Motivational interviewing [7] (MI) was developed as a corrective to provider-directed change. MI is “a [collaborative] way of talking with people about change and growth to strengthen their own motivation and commitment” [7] by listening to and amplifying patients' own reasons for change in an atmosphere of partnership and acceptance. Meta-analyses have demonstrated that MI is effective across a wide range of health concerns, including oral health issues [8–11].

1.1. Training Healthcare Providers in MI

For healthcare professionals, MI training could borrow the tagline of 1970s advertisements for the board game Othello: “A minute to learn. . . a lifetime to master.” Understanding MI's purpose, concepts, and skill descriptions is straightforward. However, enacting them with both the overarching “MI spirit” elements and the skills' nuance requires considerable communication dexterity. Systematic reviews and meta-analyses [12,13] comprising studies across healthcare (e.g., psychology, medicine, dentistry, nursing) indicate that MI's basic communication skills (open-ended questions/prompts, affirming, reflecting, summarizing; OARS) are feasibly teachable in healthcare curricula. A systematic review [13] in a similar context to dental school (graduate medical education) reported that, compared with brief didactic instruction, intensive, experiential approaches (e.g., role plays, standardized patients, direct observation in clinical settings, objective structured clinical examinations) resulted in better learning and maintenance of MI skills.

In contrast, dental school curricula on clinical communication and the behavioral aspects of patient care are often limited to didactic lectures and multiple-choice examinations [14]. Some of these programs may include some behavioral science instruction; only one-third of U.S. dental schools offer classes in patient communication [15], despite longstanding recognition of the importance of “soft” skills in dentistry (including student competence requirements 2-16, 2-17, and 2-20 set by the Commission on Dental Accreditation). Providing opportunities for active practice of communication skills is increasingly recognized as both fundamental and neglected in dental education [14]. The extensive and expensive requirements of dental training preclude expanding pre-existing curricula to accommodate the development of soft skills [16].

Thus, dental educators have proposed time-limited, workshop-based interventions for communication skills practice [15,16]. Deliberate practice [17], a training model derived from research on the development of expertise [18], shows promise as an efficient approach to skill acquisition and is increasingly used in the training of behavioral intervention skills [19,20]. It involves operationalizing target skills (e.g., reflecting patient statements in an empathic manner that elicits change talk) into enumerated criteria across proficiency levels (e.g., beginner, intermediate, advanced) [21]. Individuals practice the target skill in response to a requisite patient prompt, receive real-time expert feedback aligned with the skill criteria, and repeat the attempt [21]. Unlike traditional training in healthcare communication, deliberate practice targets specific skills and refines individual performance through structured feedback and repetition until a trainee achieves a predefined level of competency [22]. Deliberate practice has been piloted as a means of teaching communication skills to medical residents [23] and MI skills to undergraduates [24]. Research suggests it can facilitate more efficient and effective learning of communication skills, as well as greater student self-efficacy, than typical training for healthcare professionals [23,25,26].

1.2. Study Aims and Research Questions

Deliberate practice requires both operationalizing key skills and a careful task analysis with enumerated criteria for beginner, intermediate, and advanced enactment. Although MI’s foundational skills (OARS) are enumerated, little work has examined the initial skill enactment of healthcare students, including the errors made and whether such errors are interrelated. The first critical question to guide better deliberate-practice instruction is, “When students attempt OARS skills, what do they produce?” Qualitative studies have examined neophytes’ self-reported difficulties with MI, such as confusing MI with persuasion [27] and clinging to practitioner-directed solutions [28–30]. Although useful as sensitizing concepts for a grounded-theory examination of error themes [31], these studies do not provide sufficient information to guide deliberate practice for beginners. A few studies have examined actual MI output (e.g., archetypal student responses [32], macro-level coding of dental student reflections on attempting MI skills [33]), but no research has systematically coded a large set of students’ formative and summative MI skill attempts to construct an empirically grounded classification of errors that in turn guides deliberate practice.

This study addresses this gap by using constructivist grounded theory [31,34] to analyze 6555 written MI skill responses to patient prompts produced by 437 dental students across two time points: (T1) 40-student small groups in which OARS skills were introduced and tried via deliberate practice, and (T2) a midterm examination several weeks later that served as the summative assessment. Given the number of responses to score, initial coding was AI-assisted under human oversight, with an independent human overlap sample; the human interpretive work was concentrated in protocol design, adjudication, and registry construction. The aim was to build an inductive taxonomy of skill-enactment errors. We then examined whether error prevalence shifted after initial instruction, whether students

clustered into identifiable profiles based on their pre-instruction error patterns, and whether those profiles differentially impacted summative performance. Thus, we sought to answer three research questions:

- RQ1: What types of errors do dental students produce in their initial attempt at using MI skills (OARS)? This question was addressed through Artificial Intelligence (AI)-assisted initial coding of all 6555 responses using a detailed qualitative protocol followed by human coder validation.
- RQ2: Does the prevalence of each error type change between T1 (initial instruction) and T2 (midterm summative examination several weeks later)? It was expected that, on average, students would demonstrate improvement over time, making significantly fewer errors by their midterm assessment.
- RQ3: Do students fall into distinct error profiles at T1, and do those profiles predict T2 performance? Latent profile analysis was used to identify subgroups of students defined by their constellations of T1 errors, then to test whether profile membership predicted midterm performance, whether each group's hallmark errors persisted or resolved by T2, and whether error rates at T2 continued to differentiate the profiles. Given the limited research in this area, no specific predictions were made regarding how many subgroups would emerge or how emergent groups would differ at T2. Thus, this aim was largely considered exploratory.

AI played two indispensable roles in this work: (a) it scored over 10,000 MI skill attempts with criterion-referenced, individualized feedback, and (b) it made feasible the systematic qualitative coding of 6555 of those attempts.

2. Materials and Methods

2.1. Study Design

This study used a mixed-methods design integrating qualitative and quantitative elements. The qualitative element employed constructivist grounded theory, an inductive qualitative analysis approach [31], to develop a descriptive taxonomy of MI skill-enactment errors among dental students. The analytic strategy followed an initial coding phase designed to remain close to the data (i.e., generating process codes focused on active, fluid experiences [31] that described what students did in their responses rather than applying predetermined categories). The taxonomy addresses a methodological gap in the MI training literature: although prior qualitative studies have examined practitioners' self-reported difficulties with MI [35,36], no published study has systematically analyzed actual skill output to construct an empirically grounded error classification. The quantitative element, described in Section 2.5, examined error prevalence across the two time points and used latent profile analysis to identify student subgroups.

2.2. Participants and Data Source

Participants comprised 437 second-year dental students enrolled in a course on the principles of behavior and behavior change; about a third of the course is focused on effective doctor–patient communication (including MI). The inclusion criterion was enrollment in the course (required of all second-year dental students). Thus, this anonymized, archival study included the entire student cohort ($N = 437$). Three students were absent from the midterm because of illness; they contributed no T2 data (T2 $N = 434$). T1 data were complete for all 437 students. The 15-lecture course includes four lectures on MI and two MI deliberate-practice small groups (about 36–40 students; 110 min each). This paper focuses on the OAR skills from the first small group. All students completed two assessments that asked them to provide a free response using specified MI skills in response to a patient prompt. Students completed a pre-instruction speed round (T1, administered

during Small Group 1) and a post-instruction midterm examination (T2). The average time between T1 and T2 was 26.92 days ($SD = 6.9$; range 15–36 days).

The small group comprised (a) a brief MI overview, (b) a brief introduction to a requisite skill (e.g., open-ended questions) including positive and negative examples, (c) a speed round comprising four instructor-read patient prompts to which students handwrote their attempt at the skill (e.g., open-ended question), with the instructor using a random-number list to call on two students to share responses aloud after the second and fourth attempts, (d) a recorded exemplar of a dentist–patient interaction modeling the skill (e.g., open-ended questions), and (e) a deliberate-practice round during which students would break into groups of three and take turns being the dentist, the patient, and an observer (to allow each student to practice the skill while enacting distributed clinical scenarios). The entire cycle was repeated over three rounds, facilitating student learning and practice of three separate MI skills (i.e., open-ended questions, affirming statements, and reflecting statements). The speed round was a low-stakes formative assessment (i.e., students received credit for good-faith responses). The midterm, in contrast, was a high-stakes summative assessment (i.e., graded on the quality of their response). Together, these assessments produced 15 free-response items per student (i.e., 12 at T1 and 3 at T2) for a total of 6555 responses.

Microsoft Copilot AI (‘Think Deeper’ mode) transcribed handwritten responses from small groups. Claude AI scored them using criterion-referenced rationales and generated individualized written feedback, which was emailed to students within 24 h (Sonnet 4.6 for small-group and practice-exam scoring and feedback; Opus 4.6 for midterm scoring, feedback, and the qualitative coding reported below). Transcription, scoring, individualized feedback, and qualitative coding thus constituted a single AI-assisted workflow; Section 4.1 describes how each function operated.

The two-model allocation within the Claude family was deliberate. Claude Opus 4.6 was used where instruction-following on long rubrics and stability mattered most: midterm scoring, midterm feedback, and the qualitative coding reported here. Claude Sonnet 4.6 (faster and less costly per call) handled the simpler T1 speed-round and practice-exam scoring tasks.

This study was conducted in full compliance with U.S. educational privacy regulations (i.e., The Family Educational Rights and Privacy Act) and was classified as exempt by the New York University Institutional Review Board. No personally identifying information appeared in any analytic file.

2.3. Instruments

2.3.1. Speed Round (T1)

The speed round comprised 12 items spanning three MI skill domains: open-ended questions (4 items), affirmations (4 items), and reflections (4 items). Each prompt represented a brief patient statement drawn from common patient–dentist interactions and asked the student to produce the specified MI skill. Patient scenarios varied in emotional content, the ratio of change talk to sustain talk, and clinical complexity. For example, open-ended question items ranged from a patient describing barriers to engaging in routine oral health care (“I know I should brush more, but I’m always running late in the morning”) to a patient declining necessary dental procedures (“I was supposed to get a root canal, but my tooth stopped hurting, so I don’t think I need it”). Supplementary Materials File S1 presents the full set of patient prompts.

2.3.2. Midterm Examination (T2)

Three midterm free-response items involved a clinical vignette for “Malia,” who tells the provider that (a) she cut down her soda consumption from three or four a day to one,

but (b) was unsure she wanted to stop altogether, rationalizing that daily gym visits and healthy eating offset the soda. Students responded to prompts asking for an open-ended question, an affirming statement, and a reflecting statement capturing Malia's ambivalence. The vignette had a 5:2 sustain talk/change talk ratio, which created a challenge for students to discern between the two and emphasize change talk.

2.3.3. Interrater Reliability Procedures

Six human coders (five small-group instructors, namely four Ph.D. clinical psychologists and one MSW, plus one additional Ph.D. clinical psychologist trained in MI) independently scored T1 and T2 subsets, with a total of 678 Claude AI–human scoring pairs (479 T1, 199 T2) across the three MI skill types. All six received the scoring manual used by the AI (see File S1).

Coders worked in pairs, one pair per skill, with responses presented in randomized order and no information about the score group. Sampling was stratified by the AI-assigned score to ensure every score level was represented; File S3 provides the per-coder quotas and the procedure for sparse strata.

The gold standard was derived from the AI and human scores rather than assigned independently of them. Where Claude AI and the human rater agreed on a score, that agreed-upon score became the gold standard. Where they disagreed, Dr. Heyman adjudicated the correct score by reviewing the student's response alongside both scores and their scoring rationales, applying the same scoring manual (File S1) used by the AI and the coders. Each ruling was recorded in a scoring decision log. Claude AI's self-audit identified 28 cases in which its rationale contradicted its assigned score; these internal inconsistencies were corrected before adjudication. Adjudication produced 50 corrections to Claude AI's original scores (8 T1, 42 T2). Reliability statistics were then computed among the human coders, Claude AI, and the gold standard (Table 1).

Table 1. Scoring Reliability: Linearly Weighted Kappa for Human Coders, AI, and Gold Standard.

	Comparison	N	κ_w	95% CI	Exact %	MAE
Pooled	Humans vs. Claude AI (Opus 4.6)	678	0.34	[0.28, 0.39]	42.5	0.236
	Humans vs. Gold Standard	678	0.38	[0.33, 0.44]	46.0	0.225
	Claude AI (Opus 4.6) vs. Gold Standard	678	0.93	[0.91, 0.95]	93.2	0.024
T1	Humans vs. Claude AI (Opus 4.6)	479	0.26	[0.19, 0.32]	38.6	0.263
	Humans vs. Gold Standard	479	0.28	[0.21, 0.34]	39.9	0.257
	Claude AI (Opus 4.6) vs. Gold Standard	479	0.97	[0.96, 0.99]	98.1	0.008
T2	Humans vs. Claude AI (Opus 4.6)	199	0.45	[0.36, 0.53]	51.8	0.170
	Humans vs. Gold Standard	199	0.53	[0.44, 0.62]	60.8	0.147
	Claude AI (Opus 4.6) vs. Gold Standard	199	0.82	[0.76, 0.87]	81.4	0.063
T1: Open-Ended Question	Humans vs. Claude AI (Opus 4.6)	159	0.06	[−0.04, 0.15]	30.8	0.302
	Humans vs. Gold Standard	159	0.06	[−0.04, 0.14]	30.8	0.302
	Claude AI (Opus 4.6) vs. Gold Standard	159	1.00	[1.00, 1.00]	100.0	0.000
T1: Affirming	Humans vs. Claude AI (Opus 4.6)	160	0.24	[0.13, 0.34]	35.0	0.286
	Humans vs. Gold Standard	160	0.24	[0.13, 0.34]	35.0	0.286
	Claude AI (Opus 4.6) vs. Gold Standard	160	1.00	[1.00, 1.00]	100.0	0.000
T1: Reflecting	Humans vs. Claude AI (Opus 4.6)	160	0.45	[0.34, 0.55]	50.0	0.202
	Humans vs. Gold Standard	160	0.51	[0.41, 0.61]	53.8	0.183
	Claude AI (Opus 4.6) vs. Gold Standard	160	0.92	[0.86, 0.97]	94.4	0.025
T2: Open-Ended Question	Humans vs. Claude AI (Opus 4.6)	60	0.62	[0.47, 0.76]	66.7	0.092
	Humans vs. Gold Standard	60	0.78	[0.66, 0.88]	78.3	0.058
	Claude AI (Opus 4.6) vs. Gold Standard	60	0.80	[0.68, 0.91]	85.0	0.050
T2: Affirming	Humans vs. Claude AI (Opus 4.6)	60	0.54	[0.37, 0.68]	60.0	0.138
	Humans vs. Gold Standard	60	0.59	[0.42, 0.74]	68.3	0.113
	Claude AI (Opus 4.6) vs. Gold Standard	60	0.73	[0.59, 0.85]	76.7	0.067

Table 1. *Cont.*

	Comparison	N	κ_w	95% CI	Exact %	MAE
T2: Reflecting	Humans vs. Claude AI (Opus 4.6)	79	0.32	[0.21, 0.42]	34.2	0.253
	Humans vs. Gold Standard	79	0.39	[0.26, 0.51]	41.8	0.241
	Claude AI (Opus 4.6) vs. Gold Standard	79	0.85	[0.77, 0.92]	82.3	0.070

Note. κ_w = linearly weighted Cohen's kappa. CI = bootstrap confidence interval (2000 iterations). Exact % = percentage of pairs with identical scores. MAE = mean absolute error. OEQ = open-ended question. Gold standard = Dr. Heyman's scoring. T1 = speed round (formative); T2 = midterm (summative). Six human coders independently scored subsets of student responses (two coders per skill domain at each time point); the AI scored all responses. T1 open-ended questions and affirming items had no gold-standard corrections (AI vs. Gold Standard = 1.00). Scores ranged from 0 to 1.0 on most items; the T2 reflecting item used a 0–1.25 scale, with the highest score indicating excellent reflections that landed on change talk.

2.4. Coding Procedure

2.4.1. Sensitizing Concepts

Initial coding was guided by sensitizing concepts drawn from MI [7] within a constructivist grounded theory framework [31]. Four MI conceptual domains structured the coding: MI Spirit, MI Technique, Selective Responding, and Successful Enactment. File S3 provides the domain definitions and explains how codes emerged from the data and were mapped to domains.

2.4.2. Artificial Intelligence (AI)-Assisted Coding and Human Coder Validation

AI-assisted coding was implemented with human oversight. Therefore, initial coding through AI did not replace human judgment. Initial coding was performed using Claude Opus 4.6 [37], a large language model, with a detailed qualitative coding protocol incorporating four foundational MI texts (see File S2). AI coding was used given the scale of qualitative coding (6555 items) and the need for consistent application of coding criteria across all items. The AI coder served as the initial analyst; human coders oversaw, calibrated, and validated its output throughout. We coded in nine batches, comparing each new response to the running registry of codes and reviewing the registry at the midpoint (Batch 5) and at completion (Batch 9). Saturation was reached after Batch 4. The remaining 237 students and 3555 responses produced no new codes, leaving a final registry of 52. Five independent coders at the master's or doctoral level then coded an overlap sample (22 students, 330 items) drawn by stratified random sampling across performance levels; discrepancies were resolved through a structured consensus process. Full procedural details, including the coding protocol, batch structure, quality gates, code registry, and decision log, are provided in File S3.

The scoring instructions used are provided in File S1 (S1.1 for the speed round and S1.2 for the midterm); the qualitative coding instructions are presented in File S2. The chat-message prompts used to invoke these instructions were brief operational directives (e.g., "Score the first 50 students on item 1" or "Attached are the speed round and midterm 1 responses") and carried no methodological content beyond what these files contain.

2.5. Quantitative Data Analytic Strategy

First, to address RQ1, count variables were created in SPSS Version 29 [38] for each error type within our final taxonomy, reflecting each error's frequency at T1 (speed rounds) and T2 (midterm exam). Specifically, simple percentages were used to calculate the proportion of students with at least one instance of each error at each time point. McNemar's test was then used to examine changes in error frequency between T1 and T2 (RQ2).

Next, to examine the heterogeneity of student errors at T1 (RQ3), a latent profile analysis (LPA [39]) was conducted in Mplus Version 8.11 [40] using maximum likelihood estimation with robust standard errors (MLR) to address missingness and non-normality of the data. Seven of the eight T1 final error types were included as continuous variables

in our LPA. This decision was guided by several factors. First, each error variable was examined prior to analysis. Zero proportions were not severe (<50%) for all variables except “Content Accuracy Errors” (90.4%). Second, visual inspection of variable distributions and skewness likewise suggested that the majority of error variables were sufficiently continuous in range with small-to-moderate skew (<1.00). Third, an LPA was originally estimated using a Poisson distribution for count variables but produced inadmissible model solutions (including class values that are theoretically impossible under a count distribution) and poor class separation. This suggests that model assumptions for count variables were poorly suited to our indicators and did not yield stable profile solutions. Thus, because continuous model specifications produced LPA solutions with greater computational stability and interpretation of class-specific means [41,42], all variables except “Content Accuracy Errors” were treated as continuous; “Content Accuracy Errors” was excluded given its low frequency and near-zero error variance (see Table 2), which also destabilized model estimation. Final model selection balanced model fit indices, theoretical interpretability, and profile stability [39]. Fit indices included log-likelihood, AIC, BIC, and SABIC, with lower values indicating better fit. Entropy, the Lo–Mendell–Rubin likelihood ratio test (LMR), and the Bootstrapped Likelihood Ratio Test (BLRT) were also evaluated. Higher entropy (>0.80) indicates greater classification certainty, and significant LMR and BLRT values indicate that the current model fit is better than that of a model with one fewer profile ($k - 1$).

Table 2. Final Error Taxonomy with Code Definitions and Prevalence at T1 and T2.

Domain and Error	Definition	T1 <i>n</i> (%)	T2 <i>n</i> (%)	χ^2	<i>p</i>
(D) Successful Enactment	Student deployed the requested MI skill with correct form, appropriate direction, and patient-specific content.	413 (94.4%)	396 (90.7%)	4.34	0.037
(B) Incomplete Execution	Student identified the correct skill and aimed it appropriately but omitted a structural element required for full MI consistency. Near-miss attempts.	351 (80.3%)	103 (23.6%)	216.34	<0.001
(A) Fixing Reflex	Student abandoned the MI skill to advise, educate, confront, or prescribe, inserting the clinician’s agenda in place of the patient’s.	345 (78.8%)	71 (16.2%)	248.43	<0.001
(C) Responding to/Eliciting Sustain Talk	Student engaged the wrong side of ambivalence, eliciting sustain talk through questions or reflecting barriers when change talk was available. Unified category spanning directional errors in open-ended questions and selective responding failures in reflections.	328 (74.9%)	225 (51.5%)	46.24	<0.001
(B) Form/Tool Error	Student deployed the wrong structural form: closed question for open, reassurance for affirmation, question for reflective statement, or multiple questions for one.	256 (58.6%)	25 (5.5%)	222.34	<0.001
(A) Provider-Centeredness	Student positioned the clinician as the reference point for the patient’s experience. The patient’s behavior or emotion is filtered through the provider’s evaluation or approval.	232 (53.1%)	26 (5.9%)	187.61	<0.001
(B) Generic Responding	Response contained no specific reference to the patient’s stated behavior, emotion, or situation. Could apply to any patient in any scenario.	173 (39.6%)	32 (7.3%)	129.80	<0.001
(B) Content Accuracy Error	Student misrepresented, fabricated, or failed to connect with what the patient said (i.e., change talk the patient never expressed, responding to the wrong scenario, reflecting content unrelated to the patient’s statement)	42 (9.6%)	8 (1.8%)	21.78	<0.001

Note. T1 = Time 1 (speed rounds); T2 = Time 2 (midterm).

After selecting a single optimal model, we assigned each participant to a latent class based on their highest posterior probability. To accurately describe each profile, we performed a post-hoc ANOVA to compare class means across the seven remaining T1 indicators using the latent class posterior probabilities identified during profile enumeration. Significant group differences across indicators were used to inform group description.

Second, we examined whether each group's error profile changed over time to determine whether "hallmark errors" during the speed round (T1) either resolved or persisted by the midterm (T2). To do so, we conducted a series of Wilcoxon signed-rank tests using the latent class posterior probabilities identified during profile enumeration to evaluate whether the prevalence of "hallmark errors" (e.g., proportion of fixing reflex responses) differed within group from T1 to T2. Finally, we examined whether T1 profile groups differed in their error rates at T2. To do so, Kruskal–Wallis and Mann–Whitney tests were used to compare class means across each of the seven T2 errors. (Wilcoxon signed-rank and Kruskal–Wallis/Mann–Whitney tests are nonparametric equivalents of the paired-samples *t*-test and one-way ANOVA, respectively. These were used due to the nonnormality of T2 data (higher zero counts for several errors)). All pairwise comparisons were examined across all groups using the Benjamini–Hochberg [43] correction to minimize Type I error. Cohen's *d* (or when applicable, *r*) was calculated to determine the effect size of each group difference.

3. Results

3.1. Interrater Agreement Among Humans, Claude AI (Opus 4.6), and Gold Standard

Pooled and item-level interrater agreement statistics are presented in Table 1 for human coders versus Claude AI (Opus 4.6), and for humans/Claude AI versus the gold standard. When aggregated across all items (i.e., pooled rows) using linearly weighted kappa (κ_w), Claude AI demonstrated extremely high consistency with the gold standard ($\kappa_w = 0.93$, 95% confidence interval (CI) [0.91, 0.95]), with 93.2% exact agreement and a mean absolute error (MAE) of 0.024. Human coders agreed much more modestly with both Claude AI ($\kappa_w = 0.34$, 95% CI [0.28, 0.39]) and the gold standard ($\kappa_w = 0.38$, 95% CI [0.33, 0.44]). The gap between AI and human reliability was due largely to the T2 reflection item. Claude AI did well on this item versus the gold standard ($\kappa_w = 0.85$, 95% CI [0.77, 0.92]), in contrast to human coders' agreement with the gold standard ($\kappa_w = 0.39$, 95% CI [0.26, 0.51]) and with Claude AI ($\kappa_w = 0.32$, 95% CI [0.21, 0.42]). For the T2 open-ended question and affirming items, which used simpler rubrics, human-to-gold-standard agreement was substantially higher ($\kappa_w = 0.78$ and 0.59 , respectively) but still lower than Claude AI's agreement ($\kappa_w = 0.80$ and 0.73).

3.2. Final Error Taxonomy

Each of the 6555 student-item units received a single primary code. When a response warranted multiple codes (e.g., an empathic lead-in followed by a fixing-reflex follow-up), the primary code was assigned using a hierarchical rule: Spirit Violations (A) superseded Selective Responding (C), which superseded Technical Errors (B), which superseded Successful Enactment (D). This hierarchy prioritized the most clinically consequential error, consistent with the study's focus on building a descriptive error taxonomy. The initial coding distributions were Successful Enactment (Domain D; 41.0%), Technical Errors (B; 39.7%), Spirit Violations (A; 12.5%), Selective Responding (C; 5.7%), and no-response items (1.1%). The ten highest-frequency codes accounted for 63.5% of all coded items.

The four broad domains were too general for interpretive work, and the 37 active sub-codes were too granular for cross-item comparison. Thus, we created a middle-tier taxonomy between them. The organizing question for each code was "What is the student fundamentally doing wrong (or right)?" This cross-cutting logic permitted categories

to reorganize across the original domain boundaries where the data warranted it. The resulting taxonomy contained eight categories: (1) Successful Enactment, in which the student deployed the requested skill with correct form, appropriate direction, and patient-specific content; (2) Fixing Reflex, in which the student abandoned the MI skill to advise, educate, confront, or prescribe; (3) Incomplete Execution, in which the student identified the correct skill and aimed it appropriately but omitted a structural element required for full MI consistency; (4) Provider-Centeredness, in which the student positioned the clinician, not the patient, as the evaluative reference point; (5) Responding to/Eliciting Sustain Talk, in which the student engaged the wrong side of ambivalence, whether by evoking sustain talk through questions or by reflecting barriers when change talk was available; (6) Form/Tool Error, in which the student deployed the wrong structural form for the requested skill; (7) Generic Responding, in which the response contained no specific reference to the patient's stated behavior, emotion, or situation; and (8) Content Accuracy Error, in which the student misrepresented or fabricated patient content. Blank or absent responses were coded as missing (999) and excluded from the taxonomy.

Several categories crossed the original domain boundaries. Fixing Reflex, for example, unified codes from Domain A (spirit violations, such as confronting patient reasoning) with those from Domain B (technical errors, such as embedding advice within an open question). In both cases, the underlying action (inserting the clinician's agenda) was the same, regardless of whether it manifested as an overt spirit violation or a subtler technical lapse. Provider-Centeredness similarly drew from Domain A (centering clinician approval as affirmation) and Domain B (reflecting with provider-centered framing). Domain C (Selective Responding) consolidated entirely into the Responding to/Eliciting Sustain Talk category, and Domain D (Successful Enactment) remained intact. The 37 active sub-codes were preserved as the granular coding layer beneath the eight middle-tier categories, enabling analysis at either level of abstraction.

3.3. Error Prevalence at T1 and T2

Frequencies of each error type at T1 and T2 are presented in Table 2. At both time points, most students (94.4% at T1, 90.7% at T2) demonstrated at least one successful MI skill enactment. In addition to these successes, three error types were most common. These included: (a) incomplete executions (T1 80.3%, T2 23.6%), (b) the fixing reflex (T1 78.8%, T2 16.2%), and (c) responding to/eliciting sustain talk (T1 74.9%, T2 51.5%). Two additional errors enacted by over half of students at T1 were (a) form/tool errors (T1 58.5%) and (b) provider-centered responding (T1 53.1%). By T2, these decreased substantially to 5.5% and 5.9%, respectively. Finally, the least frequent errors at T1 included (a) generic responding (T1 39.6%) and (b) content accuracy errors (T1 9.6%), each of which decreased by T2 (7.3% and 1.8%, respectively). Across all error types, McNemar's tests confirmed that reductions in error frequency from T1 to T2 were statistically significant (all $ps < 0.001$).

3.4. Modeling T1 Error Profiles

To select the optimal LPA model, we examined fit statistics for one- through six-class solutions (Table 3). Results supported either a three- or four-class solution. To make our final determination, we considered both statistical fit and theoretical interpretability of each model. Specifically, whereas the four-class solution demonstrated adequate model fit (lower LL, AIC, BIC, and SABIC values relative to the three-class model; entropy of 0.869; and significant LMR and BLRT values suggesting a better fit than a three-class model), it captured additional heterogeneity in student error profiles that did not appear to be theoretically distinct or meaningful (e.g., one group whose hallmark feature was higher-than-average provider-centered and generic responding, and another group whose

hallmark feature was provider-centered responding only). Therefore, we ultimately chose the three-class solution as our final model for its parsimony and adequate fit (lower LL, AIC, BIC, and SABIC values relative to the two-class solution; entropy of 0.882; and significant LMR and BLRT values, implying a better fit than the two-class model). Class-specific means for each error type are displayed in Table 4.

Table 3. Model Fit Indices for Latent Profile Analysis of T1 Error Types.

Class	LL	AIC	BIC	SABIC	Entropy	Smallest Class %	LMR <i>p</i>	BLRT <i>p</i>
1	−5259.111	10,546.223	10,603.342	10,558.913	--	--	--	--
2	−5114.322	10,272.644	10,362.402	10,292.586	0.962	18.50%	<0.001	<0.001
3	−5020.445	10,100.89	10,223.288	10,128.083	0.882	17.40%	<0.001	<0.001
4	−4946.928	9969.855	10,124.893	10,004.3	0.869	12.12%	0.007	<0.001
5	−4874.649	9841.299	10,028.976	9882.995	0.821	12.60%	0.005	<0.001
6	−4834.658	9777.317	9997.633	9826.265	0.827	7.30%	0.111	<0.001

Note. T1 = Time 1 (speed rounds). The bolded row represents the final class solution. LL = log-likelihood. AIC = Akaike Information Criteria. BIC = Bayesian Information Criteria. SABIC = Sample-adjusted Bayesian Information Criteria. LMR = Lo–Mendell–Rubin likelihood ratio test. BLRT = Bootstrapped likelihood ratio test.

Table 4. Means at T1 and T2 for T1 Error Profiles.

			Attuned (<i>n</i> = 281)			Misaligned (<i>n</i> = 76)			Unanchored (<i>n</i> = 80)		
Variables	Range	<i>M</i>	<i>SD</i>	% Total	<i>M</i>	<i>SD</i>	% Total	<i>M</i>	<i>SD</i>	% Total	
Latent profile indicators (T1)											
Successful Enactment	0–11	4.96	2.28	41.37%	1.92	1.28	16.01%	2.85	1.84	23.75%	
Fixing Reflex	0–9	1.27	1.02	10.59%	4.71	1.29	39.25%	1.64	1.11	13.65%	
Incomplete Execution	0–6	2.11	1.41	17.56%	1.49	1.21	12.39%	1.04	1.08	8.65%	
Provider-Centeredness	0–6	0.93	1.28	7.7%	1.09	1.21	9.10%	1.60	1.43	13.33%	
Responding to/Eliciting Sustain Talk	0–5	1.27	0.97	10.59%	0.93	0.90	7.79%	1.13	0.85	9.38%	
Form/Tool Error	0–6	0.94	1.07	7.8%	1.42	1.25	11.84%	0.91	1.10	7.6%	
Generic Responding	0–4	0.28	0.45	2.3%	0.21	0.44	1.75%	2.63	0.77	21.88%	
Midterm Performance (T2)											
Successful Enactment	0–3	1.93	0.85	64.41%	1.54	1.06	51.32%	1.45	0.97	48.33%	
Fixing Reflex	0–3	0.13	0.37	4.27%	0.29	0.56	9.65%	0.26	0.47	8.75%	
Incomplete Execution	0–3	0.23	0.44	7.83%	0.28	0.48	9.21%	0.25	0.46	8.33%	
Provider-Centeredness	0–3	0.05	0.22	1.66%	0.12	0.36	3.95%	0.05	0.22	1.67%	
Responding to/Eliciting Sustain Talk	0–3	0.53	0.57	17.56%	0.54	0.55	17.98%	0.66	0.59	22.08%	
Form/Tool Error	0–3	0.06	0.24	2.02%	0.08	0.27	2.63%	0.01	0.11	0.4%	
Generic Responding	0–3	0.01	0.12	0.5%	0.04	0.20	1.32%	0.31	0.47	10.42%	
Total Quality Score (sum 3 items)	0–3.25	2.54	0.42	--	2.30	0.59	--	2.63	0.38	--	

Note. T1 = Time 1 (speed rounds); T2 = Time 2 (midterm). “% total” are proportion scores adding to 100% of total errors at each time point, accounting for the number of available items on which students could have made errors during T1 (12 items) and T2 (3 items).

The largest (64.3%, *n* = 281) group, labeled the “Attuned,” is characterized by the highest rates of (a) successful enactment and (b) incomplete execution (i.e., appropriate skill identification and effort, with some missing components), suggesting that students in this group mostly demonstrated sufficient MI skill application even with minimal instruction. The second-largest group (18.3%, *n* = 80), labeled “Unanchored,” exhibited moderate rates of successful enactments, moderate use of the fixing reflex, and high levels of generic responding relative to the other groups, indicating inconsistent and unfocused skill execution. Finally, the smallest group (17.4%, *n* = 76), labeled “Misaligned,” exhibited higher rates of the fixing reflex and form/tool errors than those in the Attuned and Unanchored groups, suggesting a pre-existing orientation that deflected some key elements of brief MI instruction. Class comparisons on all error types are provided in Table 5.

Table 5. Group Comparisons Across T1 Error Profiles.

T1 Errors (Speed Round)	Attuned vs. Misaligned (T1)			Attuned vs. Unanchored (T1)			Misaligned vs. Unanchored (T1)		
	χ^2	Cohen's <i>d</i>	<i>p</i>	χ^2	Cohen's <i>d</i>	<i>p</i>	χ^2	Cohen's <i>d</i>	<i>p</i>
Successful Enactment	146.59	1.57	0.002	51.83	0.92	0.002	6.35	0.41	0.012
Fixing Reflex	116.98	1.4	0.003	6.28	0.32	0.012	58.26	1.23	0.002
Incomplete Execution	12.78	0.45	0.003	54.01	0.93	0.002	5.08	0.36	0.024
Provider-Centeredness	1.15	0.14	0.284	13.95	0.48	0.003	3.49	0.3	0.093
Responding to/Eliciting Sustain Talk	4.32	0.27	0.114	1.77	0.21	0.276	0.82	0.15	0.366
Form/Tool Error	5.28	0.30	0.022	0.001	0.004	0.977	4.28	0.33	0.038
Generic Responding	0.73	0.11	0.393	478.48	2.78	0.003	415.23	3.29	0.002
T2 Errors (Midterm)	Attuned vs. Misaligned (T2)			Attuned vs. Unanchored (T2)			Misaligned vs. Unanchored (T2)		
	<i>U</i>	<i>r</i>	<i>p</i>	<i>U</i>	<i>r</i>	<i>p</i>	<i>U</i>	<i>r</i>	<i>p</i>
Successful Enactment	3.00	0.16	0.003	3.99	0.22	<0.001	0.74	0.06	0.459
Fixing Reflex	−2.59	0.14	0.028	−2.78	0.15	0.016	−0.11	0.009	0.914
Incomplete Execution	−0.66	0.04	0.507	−0.20	0.01	0.839	0.38	0.03	0.707
Provider-Centeredness	−1.83	0.10	0.067	−0.01	0.00	0.995	1.47	0.12	0.140
Responding to/Eliciting Sustain Talk	−0.25	0.01	0.803	−1.83	0.09	0.067	−1.25	0.10	0.211
Form/Tool Error	−0.63	0.03	0.532	1.66	0.09	0.097	1.82	0.15	0.069
Generic Responding	−0.75	0.04	0.454	−9.03	0.49	<0.001	−6.54	0.52	<0.001
T1 vs. T2 Errors	Attuned T1 vs. T2			Misaligned T1 vs. T2			Unanchored T1 vs. T2		
	<i>W</i>	Cohen's <i>d</i>	<i>p</i>	<i>W</i>	Cohen's <i>d</i>	<i>p</i>	<i>W</i>	Cohen's <i>d</i>	<i>p</i>
Successful Enactment	−9.66	1.49	<0.001	−6.23	2.04	<0.001	−5.80	1.70	<0.001
Fixing Reflex	−6.65	0.90	<0.001	−6.77	2.47	<0.001	−2.67	0.63	0.007
Incomplete Execution	−7.57	1.06	<0.001	−1.50	0.35	0.132	−0.13	0.03	0.899
Provider-Centeredness	−7.22	0.99	<0.001	−3.12	0.77	0.002	−5.37	1.50	<0.001
Responding to/Eliciting Sustain Talk	−6.09	0.81	<0.001	−4.29	1.13	<0.001	−4.86	1.29	<0.001
Form/Tool Error	−7.99	1.14	<0.001	−5.07	1.43	<0.001	−5.87	1.74	<0.001
Generic Responding	−7.05	0.97	<0.001	−1.59	0.37	0.111	−5.75	1.68	<0.001

Note. T1 = Time 1 (speed rounds); T2 = Time 2 (midterm). Cohen's *d* and/or *r* of 0.2, 0.5, and 0.8 or higher are considered small, medium, and large effect sizes, respectively. *U* = the test statistic for Mann–Whitney U tests; *W* = test statistic for Wilcoxon signed-rank test. T1 vs. T2 comparisons were based on proportion scores because they are more directly comparable than raw error counts, with different total number of items across time points.

3.5. Group Comparisons on T2 (Midterm Exam) Performance

Overall quality of midterm performance at T2 is shown in Table 4. Students in the Misaligned group scored significantly lower ($M = 2.30$, $SD = 0.59$) on MI skill application than those in the Attuned ($M = 2.54$, $SD = 0.42$; $p = 0.009$) and Unanchored groups ($M = 2.63$, $SD = 0.38$; $p = 0.014$). Total quality scores for the latter two groups did not differ significantly.

Relatedly, the prevalence of certain error types changed significantly between T1 and T2 (Table 5), with each group showing significant improvements in successful enactments and decreases in hallmark errors. Specifically, by the midterm, students in the “Attuned” demonstrated significantly fewer incomplete executions. Following a similar trend, “Unanchored” students showed significantly less use of the fixing reflex and generic responses, and those in the “Misaligned” group demonstrated significant decreases in the fixing reflex, form/tool errors, and incomplete executions. Despite these general improvements, “Unanchored” students still had significantly higher rates of generic responding (10.42%) during the midterm than “Attuned” (0.5%) and “Misaligned” (1.42%) students (Table 5). Likewise, “Misaligned” students maintained significantly higher rates of the fixing reflex (9.65%) than “Attuned” (4.27%) and “Unanchored” (8.75%) students, highlighting areas in which some students may have had continued difficulty over time.

4. Discussion

Extensive research documents that patient-communication skills are essential in dentistry [44,45]. Good communication facilitates positive clinical outcomes and patient and provider satisfaction [45–48], whereas poor communication skills are the single most cited reason for both patient distrust and dentist relationship termination [14]. Students typically begin dental school with few of the manual “hand skills” or clinical knowledge required to practice. Curricula address these deficits via thousands of hours of supervised technical practice, including preclinical simulation with iterative feedback on technique. Preparation for the *human side* of dentistry receives comparatively minimal investment. Communication skills training, where it exists at all, occupies a fraction of the curriculum [10–16]. In our college (which graduates just under 10% of the dentists in the U.S. each year), MI deliberate practice constitutes just under four hours of individuals’ dental school training. Over two sessions, 36–40 students are scheduled with one instructor to learn the MI core skills. Under those constraints, providing deliberate practice as described in the expertise literature [17,21] (i.e., individualized performance, criterion-referenced feedback, repeated attempts at the same skill until proficient) to 437 students was utterly impossible until maturing AI arrived.

4.1. AI-Assisted MI Deliberate-Practice Implementation to 437 Students

AI facilitated five aspects of this course and made the research reported here possible. First, AI gave us the freedom to make a pedagogical choice that resulted in 1749 pages of handwritten text. Small-group sessions were made device-free, and students were required to handwrite their worksheets. One AI model (Microsoft Copilot, Think Deeper mode) transcribed 5244 speed-round responses and 874 deliberate practice worksheets, and another AI model (Claude) scored the speed-round and deliberate-practice responses. What began as an evaluative function (AI scoring for grading efficiency) revealed a didactic opportunity: the scoring rationales were detailed enough to serve as individualized instructional feedback, and we delivered next-day written feedback to every student after each small-group session. Second, we provided all students with AI prompts so they could practice responding interactively to patients’ statements, with the AI providing feedback (a form of deliberate practice), and the AI model drilling them on their demonstrated weaknesses. Third, we created a practice exam with AI scoring identical to that on the T2 midterm; students received detailed feedback and tailored AI prompts to practice key needs. Seventy-two students (16.5%) completed the practice exam, which also allowed us to tune the AI scoring instructions on actual student responses before the T2 midterm exam. Fourth, each student received detailed AI feedback on every T2 free-response answer, with extensive rationales explaining the scores they received. Thus, even with 6992 T2 responses to grade (comprising both MI and other content), AI could both score the responses and provide detailed feedback to every student, serving both evaluative and didactic functions. Fifth, AI scoring made this study’s thematic analysis possible: 6555 items were coded with consistent application of a 52-code registry across nine batches, a scale that would have taken human coders months versus Claude’s hours. In addition, because each AI-generated code was accompanied by an analytic memo citing specific language from both the patient prompt and the student’s response, there was a fully auditable coding record that enabled the human-validation protocol.

In summary, AI turned the irony of trying to simultaneously teach 437 students doctor-patient communication on its head: instead of the scale making personalized instruction impossible, it enabled the detection of student learning patterns and made next-day personalized feedback on every one of over 10,000 student responses possible. It also pointed to needed improvements in instruction. Education researcher Siegfried Engelmann famously

said, “If the student hasn’t learned, the teacher hasn’t taught” [49], indicating that student errors can be traced to instructional design flaws. Error-pattern analysis laid bare aspects of instruction that succeeded for two-thirds of students while leaving identifiable subgroups with distinct, correctable difficulties. As Engelmann’s work clearly documented, the error implies the remedy [49], guiding future improvements in MI instruction.

4.2. Interrater Agreement: AI as a Premier Rater

Claude AI rated the quality of MI skill attempts with high agreement with the gold standard ($\kappa_w = 0.93$, 93.2% exact agreement), in sharp contrast with the human MI instructors ($\kappa_w = 0.38$, 46.0% exact agreement). Unlike typical coding endeavors, the humans were given a manual and items to score but never met to calibrate. This underscores Claude AI’s advantage: it could code thousands of student MI skill attempts with high agreement with the gold standard in less time than it would have taken humans to code a small fraction and meet to calibrate, let alone iterate until the coding was complete. Human coders fared far better on open-ended questions and affirming than on the more complicated skill of reflecting. This study implemented a replicable, practicable, scalable strategy: AI applied a complex rubric with mechanical consistency at scale; human coders provided independent judgments, and an expert adjudicated disputes.

4.3. Research Question 1: Types of Errors

Another aim of this paper was to build a taxonomy of MI errors, informed by MI theory [7], using constructivist grounded theory. Eight categories (MI-consistent responding and seven error classes) were derived in our final coding taxonomy. These final categories (successful enactment, fixing reflex, incomplete execution, provider-centeredness, responding to/eliciting sustain talk, form/tool error, content accuracy, and generic responding) provided interpretable, actionable error categories to guide student feedback and to improve how OARS skills are taught and practiced. Several categories unified concepts across the original domain boundaries. For example, the fixing-reflex category unified occurrences classified as spirit violations (e.g., confronting patient reasoning) and technical errors (e.g., embedding advice within an open-ended question) because the behavior (i.e., inserting the provider’s agenda) was the same, regardless of the form. Thus, the taxonomy was an empirical reconceptualization of macro- and micro-elements.

It is both system-validating and theoretically and pedagogically significant that the two most frequent errors at T1, when students had just been introduced to OARS and MI concepts, were the fixing reflex (78.8%) (i.e., “the natural desire of helpers to prevent harm and promote a person’s welfare by trying to correct or repair perceived problems” [7]) and eliciting/responding to sustain talk (74.9%). The first error class, the fixing reflex, appears to be the default communicative stance of most people who become healthcare providers [7]: helpers want to help. This remained the signature problem of the Misaligned group. The widespread prevalence and subgroup “stickiness” of this error requires instruction for all and focused drilling for some. The second error, eliciting/responding to sustain talk, is a different problem. Students need to understand and recognize patient ambivalence (e.g., “I know getting the crown would benefit my oral health, but it’s very expensive”) and learn which side (e.g., health benefit) to respond to. This is a teachable listening and responding skill that was hardly covered before students began their small group 1 (OARS) because, due to scheduling, most students had not yet received the didactic lectures on ambivalence and selective responding. Thus, it is utterly unsurprising that students did not respond selectively before they had been taught the concepts underpinning this nuanced skill. AI made the discovery possible by classifying the nearly 7000 responses.

4.4. Research Question 2: Reduction from Formative to Summative Assessments

As anticipated, all error categories decreased from T1 to T2, indicating that MI instruction may have improved performance. However, the uneven reductions suggest that grasping certain MI elements requires additional focused work beyond what the course currently provides, something that only a data-driven approach can reveal. Three prevalent categories dropped from 40–59% to 6–7%: form/tool error, provider-centeredness, and generic responding. All can be remedied by instruction and learner inference (e.g., “a reflection is a statement, not a question;” “avoid first person in affirming statements”). Similarly, content accuracy errors dropped from 10% to 2%, which may be due to the stakes (in-class response vs. exam), format (oral prompt read once vs. written), and 12 T1 vs. 3 T2 responses (and thus more opportunities for error).

The prevalent fixing reflex dropped from 79% to 16%. This represents a notable win for teaching/learning, but one in six students still committed a cardinal MI violation (e.g., rewarding talk that keeps patients stuck or hopeless) on a summative assessment, an error that warrants attention. This issue is revisited when the profiles are examined in the next section (Research Question 3).

Helping students respond to sustain talk and elicit change talk in an MI-consistent fashion should receive additional focus in future years, as non-MI responding went from a prevalent 75% to a still-prevalent 52%. Successful enactment requires students to simultaneously (a) recognize patient change vs. sustain talk, (b) suppress the empathic pull of responding to or asking for elaboration of sustain talk, and (c) actively reflect, affirm, or elicit change talk. The T2 clinical scenario contained a 5:2 ratio of sustain to change talk. Furthermore, the reflecting exam item asked students to reflect the patient’s ambivalence. Although students were classified as successfully enacting if they either reflected both sides but landed on change talk or selectively reflected only change talk, the item construction was poor because it asked students to voice both sides of ambivalence rather than practice MI selective responding. Although future assessments will have to assess better whether this element might have dropped to a level closer to that of the fixing reflex, it is almost certain that responding to/eliciting sustain talk will require the most intensive task analysis, breaking the skill into subcomponents and, once firm, shaping complete responses until proficient.

4.5. Research Question 3: Error Profiles and Their Impact on Learning

Latent profile analysis identified three types of students as they begin their first MI trainings. The largest group (over two-thirds of students), Attuned, grasped MI concepts with very little instruction. Their hallmark T1 element was “successful enactment.” The other two groups equally split the remaining one-third of students. Misaligned students had the lowest success rate and the highest rate of their hallmark T1 element, the “fixing reflex.” Unanchored students fell in between, with a hallmark element of “generic responding.” At T1, they made significantly more errors than Attuned students on five of the seven error types; however, compared with the Misaligned group, they had higher rates of successful enactment and lower rates of the fixing reflex, incomplete execution, and form/tool errors.

The profiles did not differ on two error types: responding to/eliciting sustain talk and provider-centeredness (only 1 of 3 comparisons significant). These may suggest universal vulnerabilities at T1 that require universal attention and direct instruction.

If the profiles merely represented starting points that converged at T2 despite similar instruction, current instructional approaches could remain unaltered. Although Attuned and Unanchored students converged on MI quality at T2, the Misaligned group continued to trail. The fixing reflex, although reduced substantially, (a) still occurred twice as often (9.65% vs. 4.27%) compared with the Attuned group and (b) stands out as an early prognostic sign that particular students may have personalities, experiences, or cultural backgrounds

that lead them to approach doctor–patient communication with an MI-antithetical reflex. Students who show this early hallmark may benefit from focused, repeated practice to better align them with MI concepts. In contrast, the Unanchored group’s hallmark (generic responding) also remained elevated at T2 (10.42% vs. 0.5% for Attuned students), yet their overall MI quality scores converged with those of the Attuned group. Unanchored students appeared to compensate for T2 generic responses by increasing the quality of their responses, whereas the Misaligned group did not.

4.6. Implications: Using AI Error Scoring to Improve Teaching MI Skills Using Deliberate Practice

To improve instruction of MI skills based on these findings, the error data point to the value of a task analysis comparing the component skills each MI tool requires against what was actually taught during the small groups. That is, one-third of learners do not walk in the door ready to quickly absorb telegraphic microskills before full, didactic explanations. Decades of basic [50–53] and applied [54] learning research demonstrate that teaching to mastery from the outset is substantially faster and more resilient than allowing initial errors to consolidate and then having to extinguish them.

For all skills, the twin pillars of ambivalence (simultaneous sustain and change impulses) and the fixing reflex should be taught and tested via two sets of speed rounds before students attempt any OARS skills. For ambivalence recognition, the instructor would read a patient statement aloud once, and students would write the change talk and sustain talk they heard, building the auditory discrimination that selective responding depends on. For the fixing reflex, the instructor would read a patient statement designed to elicit advice-giving, and students would write their responses. Randomly selected students would read their responses in “bingo rounds” with the instructor providing corrective feedback.

With this foundation, the current open-ended question instruction could bring nearly all students to mastery, as the gap is conceptual sequencing rather than skill construction. Students can already produce open-ended questions; with prior instruction on ambivalence, they could now be asked to elicit change talk specifically, rather than merely asking a question that elicits any talk.

Affirming is a multi-component skill (form + specificity + character quality + patient-centering) that is currently taught as a single entity without specific instruction about each component. Reflecting also comprises multiple components (form + content accuracy + emotion labeling + directional choice + patient-centered framing). For both skills, each component should be drilled in speed rounds individually (see possible adaptations in Table 6). Only then should students attempt a multi-component response.

Table 6. Task-Analyzed Affirming and Reflecting Component Shaping.

Name	Procedure	Target Component
Affirming Subskill		
1. Identify Target Behavior	Instructor reads a patient statement. Students write the specific behavior that deserves recognition (e.g., “called the next day to reschedule”).	Specificity
2. Name Character Quality	Instructor says the behavior from #1 aloud. Students write a character quality it reflects.	Character quality inference
3. Build Affirming Statement	Instructor reads a patient statement aloud once. Students write an affirming statement using the scaffold: “You [behavior]. That shows [quality].”	Specificity + quality + centering
4. Error Discrimination	Instructor reads five sample affirmations aloud (not patient statements; finished affirmations). Three are correct; two are not (e.g., advice tacked on, provider-centered). After each, Students write “right” or “wrong.” For wrong examples, students each write corrected affirming statements.	Advice-spoiling + provider-centering (discrimination)

Table 6. *Cont.*

Name	Procedure	Target Component
Reflecting Subskill		
1. Paraphrase	Instructor reads a patient statement aloud once. Students write a paraphrase.	Content accuracy (foundation)
2. Identify patient emotion	Instructor reads a patient statement. Students write one or two emotion words on their worksheet.	Emotion labeling (isolation)
3. Identify Change Talk	Instructor reads an ambivalent patient statement containing both CT and ST. Students write CT and ST.	CT/ST discrimination (prerequisite)
4. Build a Reflection	Instructor reads a patient statement. Students write a complete simple reflection, combining content and emotion: “You’re [emotion] that [content].” (If the student reads their response in the form as a question, the instructor will correct it, indicating that, “A reflection is a statement. Your voice goes down at the end, not up.”)	Content + emotion + form (combination)
5. Two Landings	Instructor reads a patient statement with both CT and ST. Students write two reflections, one that lands on change talk, one that lands on sustain talk and labels each.	Directional choice (isolation)
6. Double-Sided Reflections, Landing on Change Talk	Instructor reads a patient statement with both CT and ST. Students write a double-sided reflection: “On one hand [ST], but on the other hand [CT].”	All components integrated (capstone)
7. Selective Responding	Instructor reads a patient statement with both CT and ST. Students write a selective reflection, responding only to the CT.	

Note. CT = change talk; ST = sustain talk.

Expansion of micro-skill building will require reducing the three patient–provider–observer practice sessions to one. As noted above, students will likely benefit from increasing the skills-firming components via more skill-building speed rounds, despite having less real-world analog practice. Mediocre practice produces mediocre performance, and the data imply that students would benefit far more from practice-to-firming before more unstructured practice.

4.7. Strengths and Limitations

4.7.1. Strengths

This study had several strengths. We scored 6555 MI skill-enactment attempts from 437 students at two time points (formative and summative assessments). We developed a grounded-theory-derived error taxonomy from actual student attempts (rather than self-reports) using AI-assisted coding, employing an eight-step protocol with sensitizing concepts, a decision log, seven quality gates, constant comparison, confirmed saturation, and human coder validation (using five independent coders and a structured consensus process; details in File S3). Claude AI (Opus 4.6) proved that it could code a large corpus of students’ attempts with high agreement with the gold standard. The latent profile analysis identified three student types with distinct learning and instructional implications. Finally, given the “if the student hasn’t learned, the teacher hasn’t taught” axiom [49], we used student error data to suggest teaching improvements targeted to specific learning gaps.

4.7.2. Limitations

There are also several notable limitations. First, the data are from a single U.S. dental school with a single student cohort. Despite the demographic and international diversity of NYU’s student body, the single source limits generalizability and demands replication. Second, because the deliberate practice strategies employed in our course were not formally evaluated in a randomized controlled trial, we cannot definitively state whether students’ improvement at T2 was attributable to our instructional practices or to other factors (e.g., practice, testing, maturation effects). Future research is needed to test empirically the

effectiveness of the educational practices used in our course against alternatives (e.g., a no-intervention control). Third, data are from students' written responses to oral or written patient prompts and thus might not generalize to actual provider behavior. Fourth, T1 formative assessments were administered under different conditions than T2 summative assessments (i.e., in T1, good-faith responding earned credit, patient prompts were given orally, and four prompts were used per OAR skill, whereas T2 exam responding earned grades, patient prompts were written, and one prompt was used per OAR skill). Fifth, the three latent profile analysis groups may not be stable and may reflect unmeasured differences among them; we modeled the seven error indicators as continuous, although they are counts (Section 2.5). Sixth, because of the scale of the necessary coding, the qualitative analysis used AI-generated initial codes (with a 5% human-overlap sample) rather than being coded entirely by humans. Although the size of the dataset made AI-thematic coding the only efficient, quick-turnaround approach possible, non-human coding deviates from typical constructivist grounded theory protocols, in which immersion in the data and code generation are themselves part of the interpretive process. Seventh, the T1 and T2 assessments were not psychometrically parallel, differing on format (oral vs. written prompts), stakes (non-graded formative vs. graded summative), and number of items (12 vs. 3, though similar in structure and content). Thus, T1–T2 comparisons should be interpreted with caution and replicated. Finally, this study used one AI-model family (Claude Sonnet 4.6 and Opus 4.6) in January and February 2026. Our findings reflect their capabilities then, and a replication a year from now would likely outperform what we report here; any replication would therefore need to re-establish reliability for its own model and version rather than assuming ours is invariantly established. AI scoring, like human scoring, costs money, which would have to be factored into any replication or dissemination plan. Claude's long context window accommodated instruction sets that exceeded the limits of other systems we tested (i.e., Gemini 3.5, Copilot). FERPA-protective guardrails were built into our instructions to prevent leakage of one student's response into the scoring of another; AI instructions must anticipate such AI-only problems. We encountered no hallucinations, but the possibility remains whenever AI systems are used. We also scored one item at a time for all students, which allowed retroactive recalibration when gray-area cases prompted refinements to the instructions; as with human scoring, an AI workflow that scored complete exams student-by-student would lock in inconsistencies that could otherwise be recognized and adjusted. Our gold standard was the consensus of two independent raters with expert adjudication of disagreements, the accepted method when no prior standard exists. Like any consensus-derived standard, it is not fully independent of the raters who produced it (human or AI), so Table 1 values report agreement with that adjudicated standard, as in any human interrater study.

4.8. Future Directions

First, in future years, we will greatly expand the didactic function of AI small-group feedback. AI can provide detailed, multi-page feedback on each student's responses within 24 h and suggest individualized AI prompts to drill on specific student error patterns. Second, our data suggested modifications to small groups, which should be tested to determine (a) if they improve student performance and (b) if other changes are needed to ensure student success. Finally, these methods should be replicated at other dental schools and with other health professions.

5. Conclusions

This study yielded four conclusions:

- A taxonomy of seven novice error types (and one success category) was derived from 6555 coded MI skill attempts by 437 second-year dental students.
- Many initial MI errors were highly remediable via practice, instruction, and study (e.g., form, provider-centering, generic responding), whereas responding to/eliciting sustain talk and, to a lesser extent, the fixing reflex were more durable, requiring a rethinking of how OARS are taught and practiced.
- Students entering training with an MI-antithetical orientation, marked by a high rate of the fixing reflex, trailed their peers on both formative and summative assessments.
- AI-supported feedback made deliberate practice with next-day, criterion-referenced feedback feasible at scale, supplementing expert instruction where course constraints have long limited it.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/oral6040097/s1>: File S1: Scoring Manual; File S2: Thematic Coding Manual; File S3: Qualitative Coding Procedures. Refs. [6,7,31,34,55–59] are cited in the Supplementary Materials.

Author Contributions: Conceptualization, R.E.H. and A.K.W.-B.; methodology, R.E.H. and A.K.W.-B.; software, R.E.H.; formal analysis, R.E.H. and A.K.W.-B.; investigation, R.E.H., J.P., K.A.D., A.I., A.S., J.N.H., K.A.R., D.M.M. and A.M.S.S.; data curation, R.E.H.; writing—original draft preparation, R.E.H., A.K.W.-B., J.P. and K.A.D.; writing—review and editing, R.E.H., A.K.W.-B., A.I., A.S., J.N.H., K.A.R., D.M.M. and A.M.S.S.; visualization, R.E.H. and A.K.W.-B.; supervision, R.E.H.; project administration, R.E.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study in accordance with U.S. research regulations (45 CFR 46.104(d)(1)), which exempts research involving normal educational practices (including evaluation of instructional techniques and curricula) that are not likely to adversely affect students' opportunity to learn required content. The study was determined exempt by the Institutional Review Board of New York University.

Informed Consent Statement: Informed consent was waived because the study was determined exempt under 45 CFR 46.104(d)(1). All data were derived from normal educational practices (routine coursework and examinations), no personally identifying information appeared in any analytic file, and individual participants cannot be identified from the published results.

Data Availability Statement: The data are not publicly available because student educational records are protected under the U.S. Family Educational Rights and Privacy Act.

Acknowledgments: During the preparation of this study and manuscript, the authors used the following AI tools. Microsoft Copilot (Think Deeper model) was used to transcribe 5244 handwritten speed-round responses and 874 deliberate-practice worksheets. Claude (Anthropic, San Francisco, CA, USA; Opus 4.6 except where noted) was used for (a) scoring speed-round, deliberate-practice, and midterm free-response items with criterion-referenced rationales; (b) generating individualized written feedback delivered to students after each small-group session (Sonnet 4.6) and the midterm examination; (c) generating interactive AI-practice prompts tailored to individual students' error patterns (Sonnet 4.6); and (d) initial qualitative coding of 6555 MI skill attempts following a detailed coding protocol with constant comparison, as described in the Method section. The authors reviewed and edited all output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Definition
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BLRT	Bootstrapped Likelihood Ratio Test
CI	Confidence Interval
FERPA	Family Educational Rights and Privacy Act
LMR	Lo–Mendell–Rubin likelihood ratio test
LPA	Latent profile analysis
MI	Motivational interviewing
MLR	Maximum likelihood estimation with robust standard errors
OARS	Open-ended questions, affirming, reflecting, summarizing
SABIC	Sample-adjusted Bayesian Information Criterion
T1	Time 1 (speed rounds; formative assessment)
T2	Time 2 (midterm examination; summative assessment)

References

1. Hoving, C.; Visser, A.; Mullen, P.D.; van den Borne, B. A history of patient education by health professionals in Europe and North America: From authority to shared decision making education. *Patient Educ. Couns.* **2010**, *78*, 275–281. [[CrossRef](#)] [[PubMed](#)]
2. Kelly, M.P.; Barker, M. Why is changing health-related behaviour so difficult? *Public Health* **2016**, *136*, 109–116. [[CrossRef](#)] [[PubMed](#)]
3. Hagger, M.S.; Hamilton, K. Progress on theory of planned behavior research: Advances in research synthesis and agenda for future research. *J. Behav. Med.* **2025**, *48*, 43–56. [[CrossRef](#)] [[PubMed](#)]
4. Hagger, M.S. Psychological Determinants of Health Behavior. *Annu. Rev. Psychol.* **2025**, *76*, 821–850. [[CrossRef](#)] [[PubMed](#)]
5. World Health Organization. *World Report on Social Determinants of Health Equity*; World Health Organization: Geneva, Switzerland, 2025; ISBN 978-92-4-010758-8.
6. Magill, M.; Apodaca, T.R.; Borsari, B.; Gaume, J.; Hoadley, A.; Gordon, R.E.F.; Tonigan, J.S.; Moyers, T. A Meta-Analysis of Motivational Interviewing Process: Technical, Relational, and Conditional Process Models of Change. *J. Consult. Clin. Psychol.* **2018**, *86*, 140–157. [[CrossRef](#)] [[PubMed](#)]
7. Miller, W.R.; Rollnick, S. *Motivational Interviewing: Helping People Change and Grow*, 4th ed.; Guilford Press: New York, NY, USA, 2023.
8. Lundahl, B.; Moleni, T.; Burke, B.L.; Butters, R.; Tollefson, D.; Butler, C.; Rollnick, S. Motivational interviewing in medical care settings: A systematic review and meta-analysis of randomized controlled trials. *Patient Educ. Couns.* **2013**, *93*, 157–168. [[CrossRef](#)] [[PubMed](#)]
9. Frost, H.; Campbell, P.; Maxwell, M.; O'Carroll, R.E.; Dombrowski, S.U.; Williams, B.; Cheyne, H.; Coles, E.; Pollock, A. Effectiveness of motivational interviewing on adult behaviour change in health and social care settings: A systematic review of reviews. *PLoS ONE* **2018**, *13*, e0204890. [[CrossRef](#)] [[PubMed](#)]
10. Gao, X.; Lo, E.C.M.; Kot, S.C.; Chan, K.C. Motivational interviewing in improving oral health: A systematic review of randomized controlled trials. *J. Periodontol.* **2014**, *85*, 426–437. [[CrossRef](#)] [[PubMed](#)]
11. Jahanshahi, R.; Amanzadeh, S.; Mirzaei, F.; Bagheri Moghadam, S. Does motivational interviewing prevent early childhood caries? A systematic review and meta-analysis. *J. Dent.* **2022**, *23*, 161–168. [[CrossRef](#)] [[PubMed](#)]
12. Maslowski, A.K.; Owens, R.L.; LaCaille, R.A.; Clinton-Lisell, V. A Systematic Review and Meta-Analysis of Motivational Interviewing Training Effectiveness Among Students-in-Training. *Train. Educ. Prof. Psychol.* **2022**, *16*, 354–361. [[CrossRef](#)]
13. Dunhill, D.; Schmidt, S.; Klein, R. Motivational interviewing interventions in graduate medical education: A systematic review of the evidence. *J. Grad. Med. Educ.* **2014**, *6*, 222–236. [[CrossRef](#)] [[PubMed](#)]
14. Moore, R. Maximizing Student Clinical Communication Skills in Dental Education—A Narrative Review. *Dent. J.* **2022**, *10*, 57. [[CrossRef](#)] [[PubMed](#)]
15. Alraiq, H.; Wolf, D.; Whalen, S.; Tepper, L. Communication Training for Dental Students. *JDR Clin. Trans. Res.* **2025**, *10*, 97S–103S. [[CrossRef](#)] [[PubMed](#)]
16. Alvarez, S.; Schultz, J.-H. A communication-focused curriculum for dental students—An experiential training approach. *BMC Med. Educ.* **2018**, *18*, 55. [[CrossRef](#)] [[PubMed](#)]

17. Ericsson, K.A. Deliberate Practice and the Acquisition and Maintenance of Expert Performance in Medicine and Related Domains. *Acad. Med.* **2004**, *79*, S70–S81. [CrossRef] [PubMed]
18. Ericsson, K.A.; Pool, R. *Peak: Secrets from the New Science of Expertise*; Houghton Mifflin Harcourt: Boston, MA, USA, 2016.
19. Chow, D.; Lu, S.H.X.; Kwek, T.; Miller, S.D.; Jones, A.; Hubble, M.A.; Tan, G.C.Y. Improving responses to challenging scenarios in therapy: A randomized controlled trial of a deliberate practice training program. *Train. Educ. Prof. Psychol.* **2025**, *19*, 1–13. [CrossRef]
20. Mahon, D. A scoping review of deliberate practice in the acquisition of therapeutic skills and practices. *Couns. Psychother. Res.* **2023**, *23*, 965–981. [CrossRef]
21. Rousmaniere, T. *Deliberate Practice for Psychotherapists: A Guide to Improving Clinical Effectiveness*, 2nd ed.; Routledge: New York, NY, USA, 2024.
22. Miller, S.D.; Chow, D.; Malins, S.; Hubble, M.A. (Eds.) *The Field Guide to Better Results: Evidence-Based Exercises to Improve Therapeutic Effectiveness*; American Psychological Association: Washington, DC, USA, 2023.
23. Koubek, R.; Jarecke, J.L.T.; Haidet, P. Merging the curriculum with the clinic: An intervention to foster deliberate practice of communication skills. *Patient Educ. Couns.* **2025**, *138*, 108824. [CrossRef] [PubMed]
24. Vega, A.L.; Olsen, J.; Ogles, B.M. Deliberate practice with motivational interviewing: Basic listening skills for undergraduates. *Teach. Psychol.* **2026**, *53*, 37–43. [CrossRef]
25. Li, J.; Li, X.; Gu, L.; Zhang, R.; Zhao, R.; Cai, Q.; Lu, Y.; Wang, H.; Meng, Q.; Wei, H. Effects of simulation-based deliberate practice on nursing students' communication, empathy, and self-efficacy. *J. Nurs. Educ.* **2019**, *58*, 681–689. [CrossRef] [PubMed]
26. Marsh, M.; Lauden, S.M.; Mahan, J.D.; Schneider, L.; Saldivar, L.; Hill, N.; Diaz, C.; Abdel-Rasoul, M.; Reed, S. Family-centered communication: A pilot educational intervention using deliberate practice and patient feedback. *Patient Educ. Couns.* **2021**, *104*, 1200–1205. [CrossRef] [PubMed]
27. Aujoulat, P.; Manac'h, A.; Le Reste, C.; Le Goff, D.; Le Reste, J.Y.; Barais, M. Investigating assumptions in motivational interviewing among general practitioners: A qualitative study. *BMC Prim. Care* **2025**, *26*, 15. [CrossRef] [PubMed]
28. Bell, D.L.; Roomaney, R. Exploring the barriers that prevent practitioners from implementing motivational interviewing in their work with clients. *Soc. Work/Maatskaplike Werk.* **2020**, *56*, 416. [CrossRef]
29. Langlois, S.; Goudreau, J. "From Health Experts to Health Guides": Motivational interviewing learning processes and influencing factors. *Health Educ. Behav.* **2024**, *51*, 251–259. [CrossRef] [PubMed]
30. Schoo, A.M.; Lawn, S.; Rudnik, E.; Lindsay, J.C. Teaching health science students foundation motivational interviewing skills: Use of motivational interviewing treatment integrity and self-reflection to approach transformative learning. *BMC Med. Educ.* **2015**, *15*, 228. [CrossRef] [PubMed]
31. Charmaz, K. *Constructing Grounded Theory*, 3rd ed.; Sage: Thousand Oaks, CA, USA, 2025.
32. Bray, K.K.; Catley, D.; Voelker, M.A.; Liston, R.; Williams, K.B. Motivational interviewing in dental hygiene education: Curriculum modification and evaluation. *J. Dent. Educ.* **2013**, *77*, 1662–1669. [CrossRef]
33. Hinz, J.G. Teaching dental students motivational interviewing techniques: Analysis of a third-year class assignment. *J. Dent. Educ.* **2010**, *74*, 1351–1356. [CrossRef]
34. Glaser, B.G.; Strauss, A.L. *The Discovery of Grounded Theory: Strategies for Qualitative Research*; Aldine: Chicago, IL, USA, 1967.
35. Barwick, M.A.; Bennett, L.M.; Johnson, S.N.; McGowan, J.; Moore, J.E. Training health and mental health professionals in motivational interviewing: A systematic review. *Child. Youth Serv. Rev.* **2012**, *34*, 1786–1795. [CrossRef]
36. Imel, Z.E.; Baldwin, S.A.; Baer, J.S.; Hartzler, B.; Dunn, C.; Rosengren, D.B.; Atkins, D.C. Evaluating therapist adherence in motivational interviewing by comparing performance with standardized and real patients. *J. Consult. Clin. Psychol.* **2014**, *82*, 472–481. [CrossRef] [PubMed]
37. Anthropic. *Claude Opus 4.6: Model Release Notes and System Documentation*; Anthropic: San Francisco, CA, USA, 2026. Available online: <https://www.anthropic.com/claude/opus> (accessed on 10 March 2026).
38. IBM Corp. *IBM SPSS Statistics*, version 29.0; IBM: Armonk, NY, USA, 2023.
39. Ferguson, S.L.; Moore, E.W.G.; Hull, D.M. Finding latent groups in observed data: A primer on latent profile analysis in Mplus for applied researchers. *Int. J. Behav. Dev.* **2020**, *44*, 458–468. [CrossRef]
40. Muthén, L.K.; Muthén, B.O. *Mplus [Computer Software]*; Muthén & Muthén: Los Angeles, CA, USA, 2024.
41. Bauer, D.J.; Curran, P.J. The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychol. Methods* **2004**, *9*, 3–29. [CrossRef] [PubMed]
42. Masyn, K.E. Latent class analysis and finite mixture modeling. In *The Oxford Handbook of Quantitative Methods*, 2nd ed.; Little, T.D., Ed.; Oxford University Press: New York, NY, USA, 2013; Volume 2, pp. 551–611. [CrossRef]
43. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **1995**, *57*, 289–300. [CrossRef]
44. Pilgrim, C.; Catunda, R.; Major, P.; Perez-Garcia, A.; Flores-Mir, C. Patient-provider communication during consultations for elective dental procedures: A scoping review. *Am. J. Orthod. Dentofac. Orthop.* **2024**, *166*, 413–422.e6. [CrossRef] [PubMed]

45. Ho, J.C.Y.; Chai, H.H.; Luo, B.W.; Lo, E.C.M.; Huang, M.Z.; Chu, C.H. An Overview of Dentist–Patient Communication in Quality Dental Care. *Dent. J.* **2025**, *13*, 31. [[CrossRef](#)] [[PubMed](#)]
46. Street, R.L., Jr.; Makoul, G.; Arora, N.K.; Epstein, R.M. How does communication heal? Pathways linking clinician–patient communication to health outcomes. *Patient Educ. Couns.* **2009**, *74*, 295–301. [[CrossRef](#)] [[PubMed](#)]
47. Bomhof-Roordink, H.; Gärtner, F.R.; Stiggelbout, A.M.; Pieterse, A.H. Key components of shared decision making models: A systematic review. *BMJ Open* **2019**, *9*, e031763. [[CrossRef](#)] [[PubMed](#)]
48. Becker, C.; Zumbunn, S.; Beck, K.; Vincent, A.; Loretz, N.; Müller, J.; Amacher, S.A.; Schaefert, R.; Hunziker, S. Interventions to improve communication at hospital discharge and rates of readmission: A systematic review and meta-analysis. *JAMA Netw. Open* **2021**, *4*, e2119346. [[CrossRef](#)] [[PubMed](#)]
49. Barbash, S. *Clear Teaching: With Direct Instruction, Siegfried Engelmann Discovered a Better Way of Teaching*; Education Consumers Foundation: Knoxville, TN, USA, 2012.
50. Terrace, H.S. Discrimination Learning with and without “Errors”. *J. Exp. Anal. Behav.* **1963**, *6*, 1–27. [[CrossRef](#)] [[PubMed](#)]
51. Bloom, B.S. Time and Learning. *Am. Psychol.* **1974**, *29*, 682–688. [[CrossRef](#)]
52. Kulik, J.A.; Kulik, C.-L.C. Timing of Feedback and Verbal Learning. *Rev. Educ. Res.* **1988**, *58*, 79–97. [[CrossRef](#)]
53. Bouton, M.E. Context and behavioral processes in extinction. *Learn. Mem.* **2004**, *11*, 485–494. [[CrossRef](#)] [[PubMed](#)]
54. Stockard, J.; Wood, T.W.; Coughlin, C.; Rasplia Khoury, C. The Effectiveness of Direct Instruction Curricula: A Meta-Analysis of a Half Century of Research. *Rev. Educ. Res.* **2018**, *88*, 479–507. [[CrossRef](#)]
55. Rollnick, S.; Miller, W.R.; Butler, C.C. *Motivational Interviewing in Health Care: Helping Patients Change Behavior*, 2nd ed.; The Guilford Press: New York, NY, USA, 2022.
56. Manuel, J.K.; Ernst, D.; Vaz, A.; Rousmaniere, T. *Deliberate Practice in Motivational Interviewing*; American Psychological Association: Washington, DC, USA, 2022.
57. Rosengren, D.B. *Building Motivational Interviewing Skills: A Practitioner Workbook*; Guilford: New York, NY, USA, 2009.
58. Moyers, T.B.; Manuel, J.K.; Ernst, D. *Motivational Interviewing Treatment Integrity Coding Manual 4.2.1.*; University of New Mexico Center on Alcoholism, Substance Abuse, and Addictions: Albuquerque, NM, USA, 2015. Available online: https://motivationalinterviewing.org/sites/default/files/miti4_2.pdf (accessed on 26 May 2026).
59. Martino, S.; Ball, S.A.; Gallon, S.L.; Hall, D.; Garcia, M.; Ceperich, S.; Farentinos, C.; Hamilton, J.; Hausotter, W. *Motivational Interviewing Assessment: Supervisory Tools for Enhancing Proficiency (MIA:STEP)*; Northwest Frontier Addiction Technology Transfer Center, Department of Public Health and Preventive Medicine, Oregon Health and Science University: Portland, OR, USA, 2006. Available online: <https://motivationalinterviewing.org/sites/default/files/mia-step.pdf> (accessed on 26 May 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.