JOURNAL OF MATERIALS EDUCATION (Electronic)

Original Research

Topic Priorities and Assessment Participation in Comparisons of Four Materials Teaching Approaches

Zora Goodwin[1]

[1]*University of Alaska Fairbanks, Fairbanks, AK 99775, USA*

Correspondence to:
dgoodwin4@alaska.edu

RECEIVED:
19/02/2026
ACCEPTED:
03/10/2026
AVAILABLE ONLINE:
31/10/2026

ABSTRACT

Comparative judgments about materials teaching depend on the knowledge assessed and the students represented at assessment. This study examines lecture, discussion, discussion with sketching, and discussion with sorting through five Materials Concept Inventory items in four undergraduate cohorts. The analysis calculates percentage-point changes, normalized gains, item-omission comparisons, curriculum-weight thresholds, and entering-class endpoint bounds. Mean normalized gains are 36.44%, 39.41%, 69.17%, and 44.80%, respectively. Sketching retains the largest mean after every single-item omission, although the ordering of the other approaches changes. When polymer melting receives increasing weight and the other topics share the remaining weight equally, lecture first overtakes sketching at 78.34% of total weight. Post-assessment coverage ranges from 74.51% to 85.00% of entering enrollment. Allowing unobserved students any endpoint score between zero and full correctness produces overlapping ranges for most comparisons. An apparent 0.68-point separation between sketching and sorting also disappears after allowance for whole-percentage rounding. These findings distinguish stable advantages within the selected assessment from conclusions that depend on curriculum emphasis or response completeness. Because each approach occurs in only one cohort and pre/post responses are unlinked, the results describe class differences rather than causal teaching effects. The practical implication is to evaluate teaching choices against explicit topic priorities and a clearly defined student denominator.

KEYWORDS

materials education; teaching comparison; sketching; discussion; assessment participation; normalized gain

© 2026 The Authors

This is an open access article published by InterPore under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0) (https://creativecommons.org/licenses/by-nc-nd/4.0/).

1. INTRODUCTION

Materials education requires students to connect processes at different scales with properties that can be observed or measured. Atomic motion must become intelligible within an apparently stationary solid; bonding must be related to melting without treating every bond as interchangeable; and alloy conductivity must be explained without assuming that a constituent property transfers unchanged into a mixture. These demands create a specific teaching problem. A student may recall terminology while applying an unsuitable mechanism to a new material. Teaching comparisons therefore need evidence about particular explanations and predictions, rather than an undifferentiated impression that one classroom is more active than another. Work on structure–property reasoning and materials characterization supports attention to the alignment between disciplinary tasks and the understanding elicited by assessment [1, 2].

Discussion, sketching, and sorting address this problem through different student actions. Discussion asks learners to make an explanation available to others and respond to competing interpretations. Sketching additionally requires a spatial or mechanistic representation whose elements must be selected and coordinated. Sorting asks students to distinguish examples, associate features, and justify category boundaries. None of these actions guarantees scientific accuracy. A fluent discussion can preserve a shared misconception, a clear drawing can encode an incorrect mechanism, and a correct match can reflect recognition without explanation. The educational question concerns what students actually reason about during the activity, including the feedback that follows, rather than the visible presence of a particular classroom format.

Evidence across undergraduate science and engineering supports active participation as an alternative to sustained information delivery [3]. However, broad evidence does not settle the relative value of two specific activities within a materials course. Observations across university classrooms show substantial variation in teaching practice [4]. That variation matters when a method name is treated as a complete description of instruction. Two classes described as discussion based can differ in question quality, preparation, response time, instructor intervention, and the requirement to defend an answer. A comparison that ignores those details risks assigning an entire class difference to a label that covers several distinct instructional decisions.

Learning outcomes and students' experiences also need separate treatment. Controlled work has shown that feeling successful during instruction can diverge from measured learning [5]. A materials demonstration may seem persuasive because an instructor supplies a smooth explanation, while a difficult drawing task may expose uncertainty that students find uncomfortable. Neither reaction alone determines which activity improves reasoning. This distinction is especially relevant when departments select teaching approaches through satisfaction ratings or informal impressions. Assessment should capture whether students can connect a material's internal organization with an observable response, while experience measures should address engagement, perceived difficulty, or confidence on their own terms.

The quality of participation has further consequences for who benefits. Active instruction can reduce achievement differences across student groups, but effects depend on implementation and context [6]. Evidence connecting active participation, self-efficacy, and attainment reinforces the need to consider more than a class average [7]. The percentages examined here do not support subgroup analysis. Nevertheless, this literature clarifies why participation at assessment is educationally consequential. A high endpoint average calculated from respondents may say little about students who left the course or did not complete the assessment. A fair comparison must make the represented population visible before interpreting its mean.

Research on engagement distinguishes receiving information from producing an explanation and responding substantively to another learner [8]. This distinction offers a useful account of why sketching might complement discussion. A drawing can make an implicit assumption inspectable: particles may be shown fixed when motion is required, or a macroscopic analogy may replace an atomic explanation. Research on learning through drawing also cautions that benefits depend on guidance and the comparison activity [9, 10]. Drawing should therefore be examined as a reasoning demand with particular constraints, rather than as an inherently superior medium. The same caution applies to visual maps, which organize relations differently from freehand representations [11].

Sequence and support can be as influential as activity type. Evidence concerning problem solving before instruction shows that the value of initial difficulty depends on subsequent teaching and the design of the preparatory task [12, 13]. In materials courses, students may first attempt to explain a property change and then compare their reasoning with a scientifically adequate account. Without that comparison, an incorrect explanation can persist. Consequently, a favorable result for a class using sketching should not automatically be attributed to drawing alone. Discussion, instructor feedback, and the order of representation and explanation may all contribute to the observed class profile.

Other instructional changes often redistribute time rather than introduce a single new cognitive action. Reviews of flipped learning identify generally favorable outcomes alongside variation across course designs [14–16]. These findings are relevant because a comparison between lecture and a participatory approach can involve changes in

preparation, practice, and contact time together. Likewise, retrieval opportunities can improve later performance, and transfer depends on the relation between practice and subsequent questions [17–19]. The present comparison does not estimate those contributions separately. It instead treats each class as an instructional combination whose assessment pattern can be examined without assigning its difference to an isolated component.

Feedback provides another reason to retain topic-level detail. A useful response to an incorrect explanation identifies the scientific relation that must change and gives the learner an opportunity to use it. Evidence on feedback demonstrates that its educational contribution varies with its content and conditions [20]. In a materials task, confirming an answer about melting is less informative than requiring the student to distinguish intrachain bonding from interactions between chains. A class average can conceal a topic where feedback did not produce the intended distinction. Interpretation should therefore preserve weak outcomes even when the overall result appears favorable.

Comparative analysis also depends on the mathematical definition of improvement. Absolute change asks how many additional percentage points separate the two assessments. Normalized gain divides that change by the available room for improvement. Debates about gain measures show that these quantities answer different questions and may interact with prior preparation and unmeasured differences between classes [21–23]. Neither transformation recovers individual learning from unmatched group percentages. The appropriate response is to present the carefully defined quantities transparently, explain their denominators, and examine whether a conclusion changes when a different defensible comparison is used.

The present study concentrates on the conditions under which a teaching ordering can be maintained. Its contribution is an arithmetic examination of curriculum priorities, selected-item influence, response coverage, and whole-percentage rounding within a direct materials-education comparison. Rather than treating a single mean as a verdict, the analysis asks which differences remain visible when those choices change. The research question is whether the leading approach and the ordering beneath it persist across these checks, and whether endpoint comparisons for entering classes support the same interpretation.

2. METHODS

2.1. Teaching Conditions and Assessment

Previous research investigated the effectiveness of different instructional approaches in introductory materials science education, comparing traditional lectures, classroom discussions, discussions supported by sketching, and discussions incorporating sorting activities [24]. Students' conceptual understanding was evaluated using selected items from the Materials Concept Inventory, with assessments conducted before and after instruction. The research provided insights into how different teaching strategies may influence students' understanding of fundamental materials science concepts. Although the same instructor taught the four classes, differences in student populations, classroom environments, and assessment conditions were present. These variations highlight the importance of considering instructional context when interpreting changes in student learning and comparing the effectiveness of different teaching approaches.

Table 1 presents the numerical inputs without reconstructing individual answer counts. Percentages correct are the measurement units for item performance; student counts describe enrollment and participation. The topics concern atomic motion in solids, polymer melting, nickel phases, saturation, and alloy conductivity. They represent selected questions rather than a complete examination of all course objectives. Identifiers are retained to connect the calculations with the assessment record, while the abbreviated topic names identify the scientific issue without reproducing question wording. The five items are given equal importance initially because an independent curriculum weighting is unavailable, and alternative weighting is then examined explicitly.

The enrollment and participation counts are treated as distinct quantities. Entering enrollment defines the population for the endpoint bounds; exiting enrollment describes continued enrollment; and endpoint respondents determine how much of the entering class is observed at post-assessment. Percentages attached to participation in the record are not substituted for calculations from these counts, because some printed percentages do not reconcile with the stated denominators. No assumption is made that the same students responded twice. This distinction

is necessary because research on missing assessment responses identifies potential bias when response availability depends on achievement or other student characteristics [25].

Table 1: Pre/post percentages correct used in the calculations. Each cell gives pre/post values. L denotes lecture, D discussion, K discussion with sketching, and S discussion with sorting.

Item	Topic	L	D	K	S
1	Atomic motion	29/65	24/68	16/100	0/63
5	Polymer melting	24/31	6/13	32/30	0/5
4	Nickel phases	45/51	47/53	55/93	50/73
16	Saturation	39/65	49/82	42/96	27/81
15	Alloy conductivity	20/75	12/61	13/75	14/45

2.2. Changes and topic-specific comparisons

Let the pre-assessment percentage for topic i in class j be denoted by p and its post-assessment percentage by q. Absolute change is expressed in percentage points, whereas normalized gain is expressed as a percentage of the available increase to full correctness. The two are calculated from the same printed percentages using Eq. (1). The class summaries are unweighted arithmetic means across the five selected items. They are not full-inventory scores, student-level gain averages, or standardized mean differences. No variance is inferred from the spread between topics because differences between scientific questions do not provide student sampling variation.

$$d_{ij} = q_{ij} - p_{ij}, \qquad g_{ij} = 100\frac{q_{ij} - p_{ij}}{100 - p_{ij}}, \qquad \overline{g}_j = \frac{1}{5}\sum_{i=1}^{5} g_{ij}. \tag{1}$$

Each pre/post pair is also displayed in a response plane in Figure 1. Both coordinates remain in percent correct, allowing an endpoint to be distinguished from the amount of change leading to it. The diagonal marks equal pre/post performance and has no causal meaning. Shape identifies topic and color identifies class. This presentation makes it possible to inspect whether a class begins at a different level, reaches a common endpoint by a different amount, or performs poorly on one topic despite improvement elsewhere. It complements the gain calculation rather than replacing it with a visual ranking.

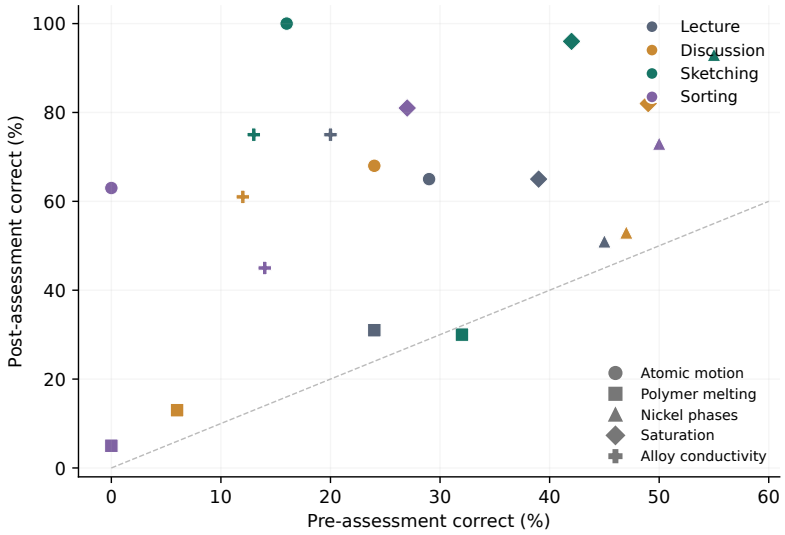


Figure 1: Pre/post response positions for the five materials topics in four teaching conditions. Points above the dashed equality line have higher post-assessment correctness. The coordinates describe group percentages and do not trace individual students.

2.3. Item omission and curriculum emphasis

To assess dependence on a selected question, the arithmetic mean is recalculated five times, omitting a different item on each occasion. The resulting values show how the class ordering changes when the assessed content changes. They are not bootstrap estimates and do not support a confidence interval. Every omission retains four observations about the same class rather than generating another teaching implementation. This distinction follows the more general requirement to recognize clustered educational observations and to align an analysis with the level at which instruction varies [26]. No inferential test treats topics as independent classes.

A separate calculation varies the emphasis placed on polymer melting, the only item on which sketching has a negative normalized gain. Its weight increases continuously from none to the entire assessment, while the other four topics divide the remaining weight equally. Eq. (2) defines the weighted comparison. Equal weighting of all five items occurs when the polymer weight is one fifth. Pairwise crossover weights are obtained by solving equality between the sketching class and each alternative. This restricted weighting exercise examines an explicit curriculum choice; it does not represent every possible set of topic weights or an estimate of how departments actually distribute instructional time.

$$G_j(w) = w g_{5j} + (1 - w)\frac{1}{4}\sum_{i \neq 5} g_{ij}, \qquad 0 \le w \le 1. \tag{2}$$

A further distinction is needed between equal residual weighting and unrestricted allocation across the remaining topics. The continuous polymer calculation fixes three degrees of freedom by requiring the other weights to be equal. To examine a conservative alternative, consider any nonnegative allocation of their remaining total. For each competitor, the smallest sketching advantage among those four items identifies the least favorable residual allocation. Combining that advantage with the polymer disadvantage yields a sufficient upper limit on polymer weight for sketching to remain ahead regardless of the residual allocation. This calculation is a property of the five gain vectors. It adds no observations and makes no claim about a preferred curriculum.

2.4. Rounding and unobserved responses

The input percentages are integers, so calculations at additional decimal places require a sensitivity check. Each value is allowed to vary by half a percentage point in either direction, with limits clipped at zero and full correctness. This is a nearest-whole-percentage rounding assumption, not a claim about the precision of the assessment instrument. Because normalized gain increases with q and decreases with p when q is below full correctness, the lower gain uses the smallest permitted q and largest permitted p; the upper gain uses the opposite corners. Means of these itemwise limits provide conservative envelopes for each class.

Entering-class endpoint bounds address a different quantity. Let N denote entering enrollment, n the number of post-assessment respondents, c their ratio, and the mean endpoint correctness among respondents be q-bar. The unobserved fraction is assigned a possible average u anywhere from zero to full correctness. Eq. (3) then gives the entering-class endpoint average under that assignment. The calculation assumes that endpoint respondents belong to the entering class and that the stated respondent count applies across the selected items. It does not impute student records, equate nonresponse with failure, or estimate the actual performance of absent students.

$$B_j(u) = c_j \bar{q}_j + (1 - c_j)u, \qquad c_j = \frac{n_j}{N_j}, \qquad 0 \le u \le 100. \tag{3}$$

The lower endpoint bound uses zero for u and the upper uses full correctness. Rounding-aware bounds additionally replace the observed endpoint mean with its permitted minimum or maximum. This produces a deterministic range compatible with both unknown outcomes and rounded inputs. These ranges are not confidence intervals: no sampling distribution, independence model, or response mechanism is estimated. The distinction is essential because overlapping ranges indicate unresolved ordering under the specified assumptions, while separated ranges

would support only that particular arithmetic comparison. They cannot establish an effect of instruction or a difference in individual improvement.

The analyses intentionally avoid significance testing, equivalence testing, and power calculations from the available aggregate percentages. Such procedures require additional information about variation, dependence, or meaningful effect thresholds that the four classes do not supply. Methodological guidance distinguishes an absence of statistical evidence from evidence that a difference is educationally negligible [27, 28]. Here, neither conclusion can be obtained from a small numerical contrast alone. The analysis therefore emphasizes the size and conditions of each descriptive difference, retaining the inferential limits throughout the interpretation rather than assigning certainty through a numerical threshold.

3. RESULTS AND DISCUSSION

3.1. The class average and the scientific topic

The five-item summaries give the sketching class the largest absolute change and normalized gain, followed by sorting, discussion, and lecture. Table 2 shows that both calculations preserve this ordering. However, endpoint correctness alone gives a different ordering beneath sketching: lecture finishes above discussion, and discussion finishes above sorting. This difference is important because an endpoint and an improvement answer different questions. The sorting class begins with the lowest mean pre-assessment correctness, so its larger improvement relative to lecture does not place it above lecture at the endpoint. A department must therefore state whether its immediate concern is final attainment, change, or improvement relative to remaining room.

Figure 2 separates the topic-specific gains that produce these averages. Atomic motion, nickel phases, and saturation contribute particularly strongly to the sketching profile. Alloy conductivity produces a much smaller separation between sketching and lecture, while polymer melting favors lecture over all the other classes under the normalized calculation. This pattern rejects an interpretation in which the leading class performs best on every scientific question. Instead, the advantage is concentrated in a group of topics, and its extent depends on whether those topics correspond to the educational priorities of the course being considered.

Table 2: Equal-weight class summaries across the five selected items. Percentage-point change is distinct from normalized gain.

Quantity	Lecture	Discussion	Sketching	Sorting
Mean pre correct (%)	31.40	27.60	31.60	18.20
Mean post correct (%)	57.40	55.40	78.80	53.40
Mean change (points)	26.00	27.80	47.20	35.20
Mean normalized gain (%)	36.44	39.41	69.17	44.80
Gain rounding lower (%)	35.20	38.22	68.16	43.79
Gain rounding upper (%)	37.66	40.58	70.05	45.67

Atomic motion illustrates why a large gain needs a precise interpretation. The sketching class moves from limited initial correctness to full endpoint correctness on the selected item. That result establishes a strong group-level response pattern for the question, but it does not show that every student developed a transferable account of diffusion. A correct response can reflect several kinds of knowledge, and a single selected-response item cannot distinguish them. A subsequent materials assessment would need a new task that asks students to explain movement within a solid under different conditions, rather than relying only on recognition of the earlier question.

Polymer melting presents the opposite pattern. The sketching class changes from 32% to 30% correct, while the other classes show small increases. The normalized value is therefore negative for sketching. This does not identify which students changed answers or demonstrate that drawing caused an error. It shows that the class-level endpoint did not improve on this question. For materials teaching, the result warrants close examination of whether students distinguish strong bonding along a polymer chain from interactions that permit motion between chains. An activity can be visually rich while leaving that particular distinction unresolved.

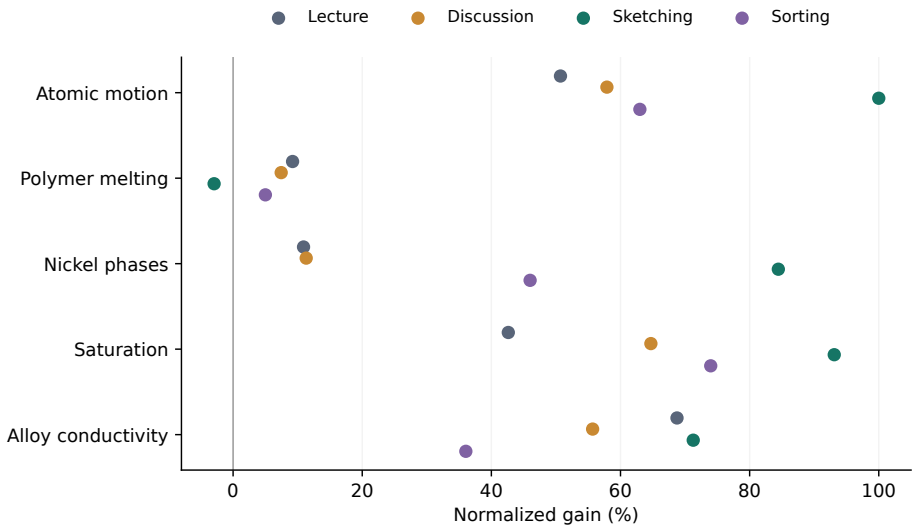


Figure 2: Normalized gains separated by materials topic. Values are recomputed from the pre/post percentages in Table 1. The zero line distinguishes a positive percentage-point change from a negative change; points do not carry student-level error bars.

The saturation item and the nickel-phase item show why the selected topics should remain separate. Although both support the leading class profile, they place different demands on learners. Saturation concerns a condition that must be interpreted without treating added material as automatically dissolved. Nickel phases require reasoning beyond a familiar everyday appearance. The appropriate follow-up teaching would differ between these topics. Combining them into a single gain is useful for describing the chosen assessment, but the combined value cannot specify which scientific explanation needs additional instruction or what student reasoning should be inspected next.

The conductivity item further constrains a broad recommendation for sketching. Lecture and sketching have the same endpoint correctness, although their pre-assessment percentages differ. Their normalized gains are consequently close without being identical. Figure 1 makes this distinction visible as a common vertical endpoint coordinate and different horizontal starting positions. The result exemplifies a general problem in pedagogical comparison: a higher transformed gain does not necessarily mean a higher final level on the measured question. Interpretation becomes clearer when the endpoint is retained beside the transformation instead of being replaced by it.

Taken together, these patterns favor a qualified reading of the sketching result. The leading average reflects substantial strengths on several assessed topics, alongside a persistent weakness and a near tie on another. The evidence does not support replacing every instructional activity with drawing. A defensible materials course would retain task-specific assessment and use the differences to identify where discussion, representation, or explanation should be strengthened. Research on teaching higher education connects achievement with multiple instructional and learner variables, which also argues against treating an activity name as a sufficient explanation for all observed variation [29].

Assessment validity introduces a related issue that percentage transformations cannot resolve. A selected response records whether an answer option was chosen, but the instructional aim may require an explanation, a prediction under changed conditions, or a distinction between competing mechanisms. Those are different evidential demands. A response about atomic motion can be correct without identifying how a vacancy participates in transport; an answer about conductivity can be correct without relating scattering to the direction of the property change. Consequently, increased correctness should be interpreted at the level actually measured. More elaborate disciplinary claims require assessment tasks that expose the necessary reasoning.

The precision of a question also affects how its instructional consequences should be read. A polymer prompt may ask about a class of materials whose behavior depends on chain structure, intermolecular interactions, and other conditions. Interpreting the answer requires knowing which simplifications the introductory course expects. The present calculations preserve the item identities and do not reinterpret them as comprehensive tests of polymer behavior. In applying the findings elsewhere, instructors should first inspect the question and its intended scientific distinction. Carrying over a numerical ordering without checking that distinction could compare classes against an objective that the new course does not share.

3.2. Removing a question changes the lower ordering

The single-item omissions retain sketching in the leading position in every calculation. Figure 3 displays the five means for each class, and Table 3 gives their values by omitted topic. The smallest separation between sketching and the next class is 18.92 normalized-gain points when nickel phases is removed. Thus, no one selected item is solely responsible for the leading average. At the same time, the span of sketching results is large because omitting polymer melting removes its only negative item. This illustrates why stability of position and stability of magnitude are different properties.

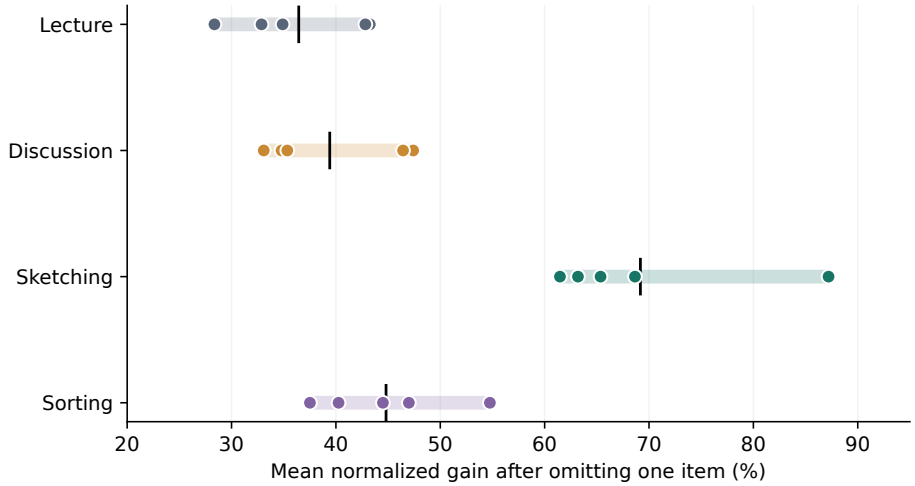


Figure 3: Class means after each single-item omission.

Table 3: Mean normalized gain after removing one assessed topic. All entries are percentages.

Omitted topic	Lecture	Discussion	Sketching	Sorting
Atomic motion	32.87	34.79	61.47	40.25
Polymer melting	43.25	47.40	87.20	54.75
Nickel phases	42.82	46.43	65.36	44.50
Saturation	34.89	33.09	63.19	37.51
Alloy conductivity	28.36	35.34	68.65	46.99

The lower positions are less stable. Removing nickel phases places discussion above sorting. Removing saturation places lecture above discussion. These changes do not arise from a different set of students or a new teaching condition; they arise solely from changing the scientific content included in the average. A curriculum committee that gives a high priority to phase reasoning may therefore reach a different interpretation from one emphasizing conductivity and saturation. Equal weighting is a transparent starting choice, but it cannot stand in for an explicit account of which materials knowledge the course is intended to develop.

The omission exercise also clarifies the limit of a robustness claim. It tests five nearby choices within a fixed set of selected questions. It does not test a new item about thermal treatment, a transfer task about an unfamiliar alloy, or performance after a delay. Stability across the omissions should consequently be described as local stability within this assessment. Extending the conclusion to the whole discipline would require wider topic sampling and

evidence that the questions elicit the intended reasoning. The retained leading position is informative precisely because its domain is stated, rather than expanded beyond the observations.

3.3. How much polymer emphasis changes the leading class

The continuous weighting calculation provides a more explicit curriculum comparison. At equal item weights, sketching has a clear advantage. As the share assigned to polymer melting increases, its weighted gain falls faster than those of the other classes because its polymer value is negative. Figure 4 shows the resulting intersections. Lecture first exceeds sketching at a polymer weight of 78.34%, discussion crosses sketching at 79.30%, and sorting crosses sketching at 80.34%. These thresholds apply only when the remaining topics share their residual weight equally; they are not universal limits for every possible curriculum.

The first crossover gives a concrete interpretation of the leading average. Within this weighting family, polymer melting must account for more than three quarters of the assessed importance before lecture overtakes sketching. The sketching advantage is therefore not fragile to a small increase in that topic's weight. Nevertheless, a specialized polymer course could reasonably reject an average dominated by diffusion, phases, saturation, and conductivity as an unsuitable decision criterion. The calculation does not prescribe a desirable weight. It identifies how a declared educational priority translates into a different arithmetic comparison.

The distinction between weighting and teaching time is also important. A question may receive greater importance because it represents a prerequisite for later reasoning, even if it occupies little contact time. Conversely, a lengthy activity may be a means of teaching several ideas without deserving a corresponding share of a final score. Weighting should therefore be justified by learning objectives and the decisions the assessment supports. Using the crossover as a scheduling recommendation would confuse evaluative importance with instructional duration. Its appropriate role is to make the decision rule explicit before selecting the apparent best teaching approach.

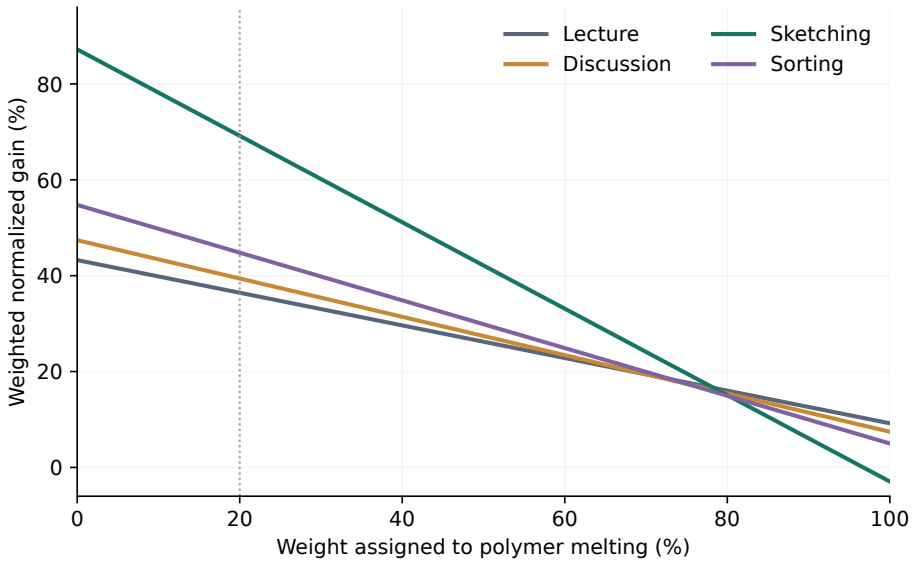


Figure 4: Weighted gains as the polymer-melting share increases.

Under unrestricted residual allocation, the lecture comparison is especially sensitive to the conductivity item, where the sketching advantage is small. Assigning all residual weight to that topic gives the earliest possible lecture crossover at a polymer share of approximately 17.14%. The contrast with the equal-residual threshold is substantial. It means that increasing polymer importance alone is not the only way to change a decision; concentrating the remaining importance on a near-tied topic also matters. A curriculum dominated by polymer melting and alloy conductivity can consequently favor lecture even though the broad five-item average strongly favors sketching. This conclusion concerns weighting, not a newly observed teaching effect.

The corresponding conservative limits against discussion and sorting are larger because sketching retains stronger advantages on their least favorable residual topics. However, it is unnecessary to assign those limits a universal educational interpretation. Their main purpose is to show why a single weighting curve must be read with its constraints. A curve that shares residual importance equally answers a particular curriculum question. A course organized around two highly specific objectives answers another. Neither calculation can establish which course objective is most valuable. That decision belongs to the disciplinary curriculum, and the arithmetic should follow it rather than quietly supply it.

3.4. Assessment participation changes the population represented

Figure 5 shows post-assessment coverage of the entering classes. Lecture has 49 respondents from 60 entering students, discussion 38 from 51, sketching 31 from 40, and sorting 34 from 40. These ratios range from roughly three quarters to slightly more than five sixths. Table 4 distinguishes them from course retention, which uses exiting enrollment instead. Sorting has the highest value under both denominators, but this coincidence does not make retention and assessment coverage interchangeable. A student can remain enrolled and still be absent from the endpoint assessment.

These differences matter even without evidence about the reasons for nonresponse. If absent students would tend to answer fewer questions correctly, the respondent average could exceed the entering-class average. If they would perform similarly, the difference could be small. Other patterns are also possible. The available counts cannot select among these possibilities. Evidence about missing educational responses supports examining the consequences of assumptions rather than automatically treating observed respondents as representative [25]. Here, the unrestricted endpoint bounds make those consequences visible without creating individual records or imposing an unsupported response model.

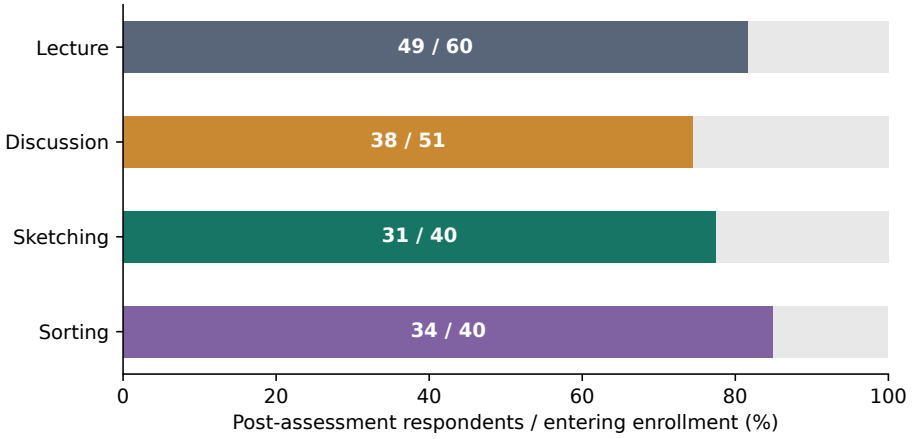


Figure 5: Post-assessment coverage with entering enrollment as the denominator.

Table 4: Enrollment and assessment counts.

Quantity	Lecture	Discussion	Sketching	Sorting
Entering enrollment	60	51	40	40
Exiting enrollment	54	48	33	38
Post-assessment respondents	49	38	31	34
Retention (%)	90.00	94.12	82.50	95.00
Coverage (%)	81.67	74.51	77.50	85.00

The bounds in Table 5 are wide enough to prevent a full ordering. Lecture, discussion, and sketching share substantial overlapping regions. Sorting initially appears separated from sketching by a small amount: the sketching lower limit exceeds the sorting upper limit by 0.68 percentage points when the printed values are taken exactly. That separation is much smaller than the observed difference between their respondent endpoint means. The contrast shows how a favorable comparison among assessed students can become uncertain when the comparison includes the whole entering class under unrestricted missing outcomes.

Table 5: Entering-class endpoint bounds in percent correct. The first pair treats the printed percentages as exact. The second pair also allows half-point rounding of each endpoint percentage.

	Printed values		Rounding allowance	
Class	Lower	Upper	Lower	Upper
Lecture	46.88	65.21	46.47	65.62
Discussion	41.28	66.77	40.91	67.14
Sketching	61.07	83.57	60.68	83.88
Sorting	45.39	60.39	44.97	60.82

Allowing whole-percentage rounding removes even that small separation. The sketching lower limit becomes 60.68%, while the sorting upper limit reaches 60.82%. Figure 6 displays both the exact-input intervals and their rounding envelopes. The overlap is narrow, but its width is not the principal point: the sign of the minimum separation changes. A claim of guaranteed sketching superiority for the entering classes would therefore depend on treating rounded percentages as exact. The defensible conclusion is that unrestricted missing outcomes and plausible rounding leave every pairwise ordering unresolved, even though the observed respondent means remain distinctly different.

An alternative assumption would assign the same unobserved average to every class. Under that restriction, the sketching entering-class endpoint exceeds each alternative throughout the permissible range because both extremes of each pairwise difference remain positive. This restricted result should not be confused with the unrestricted bounds, which permit missing students in different classes to have different averages. The shared-value assumption is easier to satisfy arithmetically but stronger educationally. Without information on absence, withdrawal, or prior achievement, the evidence cannot justify preferring it. Presenting both interpretations identifies exactly which assumption creates the stronger ordering.

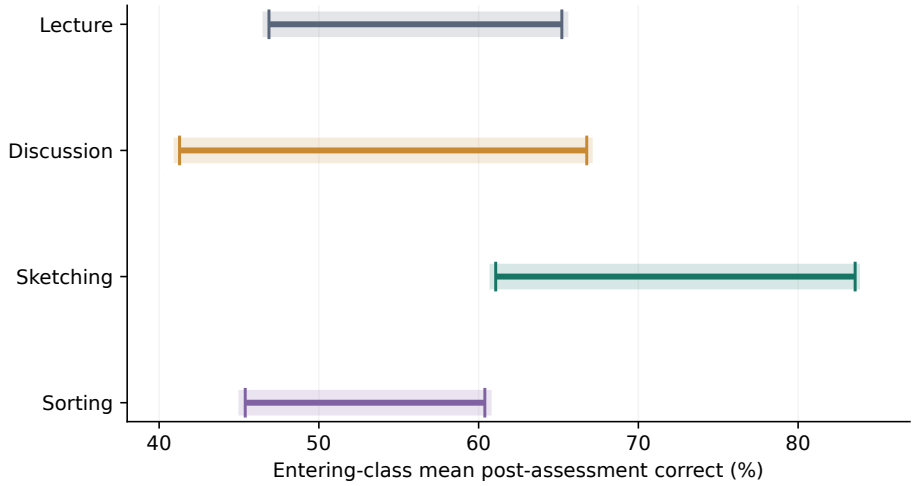


Figure 6: Entering-class endpoint ranges when unobserved correctness may lie anywhere from zero to full correctness. Thin dark segments use the printed percentages; pale envelopes additionally allow half-point rounding. The overlapping sketching and sorting envelopes prevent a guaranteed ordering under the combined allowance.

The entering-class denominator also deserves an educational justification. Departments often need to know what proportion of an initially enrolled group reaches an intended level, whereas instructors assessing a completed activity may need to know what responding students can do. Both questions are legitimate. Difficulty arises when the respondent average is used to answer the entering-class question without acknowledging absent outcomes. The endpoint bounds do not penalize nonrespondents or assume they learned nothing. They preserve the possibility that their performance differs in either direction, making the dependence on that unobserved portion explicit. A narrower range would require additional evidence or a more restrictive assumption.

Information about nonresponse could narrow interpretation without necessarily requiring a more complex statistical model. Reliable counts of assessment invitations, completed responses, withdrawals, and changes of enrollment

would establish the population represented at each occasion. Linked anonymous identifiers would establish whether pre/post comparisons involve the same people. Brief documentation of assessment incentives and administration would help distinguish a low-stakes response problem from course attrition. These practical records would be especially useful where paper and computer administration differ. Their value lies in clarifying what the percentages mean, rather than increasing the apparent numerical sophistication of the analysis.

3.5. Rounding preserves the observed leading gain

Rounding does not destabilize every conclusion. The normalized-gain envelope for sketching remains well above the other class envelopes in Table 2. Its lower limit is 68.16%, whereas sorting, the next class under equal weights, has an upper limit of 45.67%. This separation is too large to be explained by the half-point rounding allowance. The apparent contrast between this result and the endpoint-bound overlap is instructive: the calculations address different populations and different outcome definitions. Strong separation in one calculation does not transfer automatically into certainty about another.

The gain envelopes should still be interpreted conservatively. They allow each printed percentage to move to an extreme that minimizes or maximizes the class mean, even though the actual rounding errors may not align in that way. No probability is attached to the corners, and the envelope width is not an estimated standard error. The calculation addresses arithmetic precision only. It does not account for question validity, response composition, instructional differences, or selection of the five items. Keeping these sources of uncertainty separate prevents a narrow rounding range from being mistaken for a comprehensive statement of educational certainty.

The figures likewise serve distinct evidential purposes. The response plane preserves attainment and starting position simultaneously; the topic plot identifies heterogeneity within the selected assessment; and the omission plot shows sensitivity to a single question. The weighting curves make a curriculum decision explicit, the coverage bars identify the represented student fraction, and the endpoint intervals display the consequence of leaving the remainder unrestricted. None is a photograph of teaching or a depiction of unobserved student behavior. Their combined use avoids assigning one graph more interpretive work than it can support and lets the reader inspect each claim against its own numerical basis.

3.6. Teaching implications and limits of explanation

The evidence favors retaining drawing as a serious candidate for materials instruction while diagnosing what the drawing asks students to explain. The observed strengths on atomic motion, phases, and saturation suggest productive targets for activities that make an otherwise hidden process explicit. However, the polymer result warns against evaluating a drawing by visual completeness alone. Recent work on teachers' interpretation of chemistry drawings shows the importance of attending to disciplinary details in student representations [30]. In a materials course, the relevant details would include what entities move, what interactions are represented, and which changes account for the macroscopic property.

Sorting remains a useful comparison because it can require discrimination without demanding the same representational production. Its performance should be interpreted through the reasons students supply for a category assignment. A student who sorts several examples correctly may be using surface familiarity rather than a transferable relation, while another may use a scientifically adequate rule that needs refinement. Neither possibility can be distinguished by the aggregate percentages here. A future comparison should retain a brief written justification with each sorted group and assess that justification independently from the correctness of the final arrangement. Such evidence would clarify the reasoning elicited by the task.

Discussion also deserves a more precise description than its position in an overall ordering. It can expose competing explanations, but its effectiveness depends on who contributes, whether claims are challenged, and how the instructor responds. A few confident students may produce a strong group answer without all participants reaching the same understanding. Individual responses before and after discussion would distinguish collective completion from individual change. Research linking student acceptance of active instruction with engagement and course

performance further suggests that participation cannot simply be assumed from the assignment of a group task [31]. These considerations identify evidence to collect rather than effects demonstrated here.

Practical implementation requires attention to instructor constraints and student experience. Barriers to adopting active teaching have been measured in engineering contexts, indicating that decisions about class activities occur within institutional and personal conditions [32]. Anxiety can also vary across active-learning practices [33]. A materials department should consequently examine whether students have sufficient time, accessible drawing tools, clear expectations, and a reasonable way to contribute without being judged for artistic skill. These are implementation considerations, not explanations established by the comparison. They should be investigated directly before attributing differences in performance to motivation or comfort.

The most important explanatory limit is the complete alignment between teaching condition and cohort. A shared instructor reduces one source of variation, but it does not make the students, classroom, assessment conditions, or years exchangeable. With one class per approach, the analysis cannot separate teaching from those conditions. It would be misleading to treat the selected questions as independent replications and calculate a significance level for the approach. More generally, guidance on statistical interpretation emphasizes that a numerical threshold cannot replace design-based reasoning about what a comparison identifies [34]. The present arithmetic contrasts remain informative without being promoted into causal estimates.

A stronger comparison would require multiple classes using each approach and consistent assessment administration. Assignment should occur at a level compatible with classroom delivery, while the analysis should account for students nested within classes. Sample size should be justified by a meaningful difference and credible information about variation between classes, rather than by an arbitrary student total [35]. Because the current record lacks the necessary variation components, no numerical recruitment target is supplied. The appropriate next step is to obtain those quantities from suitable pilot work and specify the intended comparison before collecting confirmatory evidence.

Assessment content should also be designed before comparing teaching conditions. The selected materials topics would benefit from paired explanation and transfer tasks that distinguish recognition from mechanism-based reasoning. For polymer melting, this means determining whether students identify the relevant interactions; for conductivity, whether they connect an atomic-scale account with the direction of property change. Individual answers should be retained under anonymous identifiers and scored using the same rules across classes. A delayed assessment would test persistence without assuming that an immediate correct response represents lasting understanding. These requirements follow from the limitations of the available percentages, rather than from an untested claim that any new teaching sequence is effective.

Evidence from chemistry and broader science education is used here to explain relevant teaching demands, rather than to enlarge the empirical sample. Research about drawing, retrieval, feedback, or flipped instruction does not supply additional materials classes to the four-class comparison. Its role is interpretive: it helps identify why an activity can succeed under one design and fail under another. This boundary matters because a long reference list can otherwise give the impression of direct replication where only related evidence exists. The numerical conclusions remain restricted to the five items and four classes, regardless of how extensively the surrounding literature discusses active participation.

For classroom use, the findings support a disciplined cycle of selecting a scientific objective, choosing an activity that requires the intended explanation, and checking an individual response under comparable conditions. The checking task should be sufficiently different from the practice task to reveal whether the reasoning transfers, while remaining aligned with the same materials relation. A drawing activity could therefore be followed by a short written prediction rather than another copy of the same drawing. A sorting activity could be followed by justification of an unfamiliar example. These are design implications derived from the assessment problem, not additional outcomes measured in the classes.

Finally, the leading class should be treated as a reason for a better comparison rather than a definitive endpoint for instructional choice. Its strengths identify useful teaching candidates, while its weak item identifies an objective needing attention. Replicated classes could test whether those patterns recur when administration and student preparation are more comparable. If they do, the next question would concern the mechanism that distinguishes the activities. If they do not, the differences would still have served a useful purpose by exposing the assumptions hidden inside a simple ranking. In either case, progress depends on preserving the relation between a scientific question, the student response, and the population represented by the result.

The distinction between physical manipulation and explanatory construction is particularly important when interpreting sketching. Moving a pen is not the relevant achievement. The educational demand is to decide which entities, relations, and changes belong in the representation, then explain why they produce the property of interest. A teacher can inspect that reasoning by asking a student to account for one feature of the drawing and predict what would change if a stated condition changed. This approach also prevents artistic polish from becoming a substitute for scientific adequacy. The percentages cannot establish that these actions occurred, so they should be treated as questions for observation rather than an explanation already verified.

A similar distinction applies to the apparent weakness of lecture. The lecture class has the highest normalized gain on polymer melting and shares the conductivity endpoint with sketching. It would therefore be inaccurate to describe it as uniformly unsuccessful. Lectures can supply a carefully sequenced scientific explanation, while participatory tasks can provide opportunities to inspect and revise students' reasoning. The present evidence does not determine the best balance between those functions. It does show that a course-wide judgment can obscure a topic where the otherwise lower-ranked class has an advantage. A useful teaching comparison should preserve such exceptions because they direct attention toward the content and activity demands that may explain a difference.

The strongest defensible educational claim is consequently conditional. Sketching is the leading observed class when these five topics receive equal importance, and its lead survives the specified arithmetic checks. Statements about a specialized curriculum, an entering population, or a future teaching implementation require their own assumptions and evidence. Making those conditions explicit improves the usefulness of the comparison without enlarging its empirical reach.

4. CONCLUSION

The four classes display a stable leading mean for discussion with sketching across equal-weight gain calculations, all single-item omissions, and half-point rounding allowances. That stability is bounded by the selected scientific content: polymer melting favors other classes, and sufficiently strong emphasis on that topic changes the leading approach. Assessment participation creates a separate limitation. When entering-class endpoint averages allow unrestricted missing outcomes and rounding, no complete or guaranteed pairwise ordering remains. Materials teaching should therefore be evaluated through declared topic priorities, distinct attainment and change quantities, and an explicit student denominator. The evidence supports a disciplined comparison of these classes, while causal superiority and generalization require independently replicated instruction and linked assessments.

Conflict of Interests

The author declares that they have no conflicts of interest.

Funding Information

This research received no external funding.

Data Availability

No dataset are used in this research.

References

1. Underwood, S. M.; Kararo, A. T.; Gadia, G.. Investigating the impact of three-dimensional learning interventions on student understanding of structure–property relationships. *Chemistry Education Research and Practice* **2021**, 22(2), 247–262. https://doi.org/10.1039/d0rp00216j.
2. Chetehouna, M.; Viola, G.. A Teaching-Learning Framework for Materials Characterization. *Journal of Chemical Education* **2024**, 101, 1036–1048. https://doi.org/10.1021/acs.jchemed.3c00974.
3. Freeman, S.; Eddy, S. L.; McDonough, M.; Smith, M. K.; Okoroafor, N.; Jordt, H.; et al.. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences* **2014**, 111, 8410–8415. https://doi.org/10.1073/pnas.1319030111.
4. Stains, M.; Harshman, J.; Barker, M. K.; Chasteen, S. V.; Cole, R.; DeChenne-Peters, S. E.; et al.. Anatomy of STEM teaching in North American universities. *Science* **2018**, 359(6383), 1468–1470. https://doi.org/10.1126/science.aap8892.
5. Deslauriers, L.; McCarty, L. S.; Miller, K.; Callaghan, K.; Kestin, G.. Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences* **2019**, 116, 19251–19257. https://doi.org/10.1073/pnas.1821936116.
6. Theobald, E. J.; Hill, M. J.; Tran, E.; Agrawal, S.; Arroyo, E. N.; Behling, S.; et al.. Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. *Proceedings of the National Academy of Sciences* **2020**, 117(12), 6476–6483. https://doi.org/10.1073/pnas.1916903117.
7. Ballen, C. J.; Wieman, C.; Salehi, S.; Searle, J. B.; Zamudio, K. R.. Enhancing Diversity in Undergraduate Science: Self-Efficacy Drives Performance Gains with Active Learning. *CBE—Life Sciences Education* **2017**, 16(4), ar56. https://doi.org/10.1187/cbe.16-12-0344.
8. Chi, M. T. H.; Wylie, R.. The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist* **2014**, 49, 219–243. https://doi.org/10.1080/00461520.2014.965823.
9. Fiorella, L.; Zhang, Q.. Drawing Boundary Conditions for Learning by Drawing. *Educational Psychology Review* **2018**, 30(3), 1115–1137. https://doi.org/10.1007/s10648-018-9444-8.
10. Cromley, J. G.; Du, Y.; Dane, A. P.. Drawing-to-Learn: Does Meta-Analysis Show Differences Between Technology-Based Drawing and Paper-and-Pencil Drawing?. *Journal of Science Education and Technology* **2020**, 29, 216–229. https://doi.org/10.1007/s10956-019-09807-6.
11. Schroeder, N. L.; Nesbit, J. C.; Anguiano, C. J.; Adesope, O. O.. Studying and Constructing Concept Maps: a Meta-Analysis. *Educational Psychology Review* **2018**, 30(2), 431–455. https://doi.org/10.1007/s10648-017-9403-9.
12. Loibl, K.; Roll, I.; Rummel, N.. Towards a Theory of When and How Problem Solving Followed by Instruction Supports Learning. *Educational Psychology Review* **2017**, 29(4), 693–715. https://doi.org/10.1007/s10648-016-9379-x.
13. Sinha, T.; Kapur, M.. When Problem Solving Followed by Instruction Works: Evidence for Productive Failure. *Review of Educational Research* **2021**, 91(5), 761–798. https://doi.org/10.3102/00346543211019105.
14. Strelan, P.; Osborn, A.; Palmer, E.. The flipped classroom: A meta-analysis of effects on student performance across disciplines and education levels. *Educational Research Review* **2020**, 30, 100314. https://doi.org/10.1016/j.edurev.2020.100314.
15. van Alten, D. C.; Phielix, C.; Janssen, J.; Kester, L.. Effects of flipping the classroom on learning outcomes and satisfaction: A meta-analysis. *Educational Research Review* **2019**, 28, 100281. https://doi.org/10.1016/j.edurev.2019.05.003.
16. Bredow, C. A.; Roehling, P. V.; Knorp, A. J.; Sweet, A. M.. To Flip or Not to Flip? A Meta-Analysis of the Efficacy of Flipped Learning in Higher Education. *Review of Educational Research* **2021**, 91(6), 878–918. https://doi.org/10.3102/00346543211019122.
17. Adesope, O. O.; Trevisan, D. A.; Sundararajan, N.. Rethinking the Use of Tests: A Meta-Analysis of Practice Testing. *Review of Educational Research* **2017**, 87, 659–701. https://doi.org/10.3102/0034654316689306.
18. Pan, S. C.; Rickard, T. C.. Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin* **2018**, 144(7), 710–756. https://doi.org/10.1037/bul0000151.

19. Yang, C.; Luo, L.; Vadillo, M. A.; Yu, R.; Shanks, D. R.. Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin* **2021**, 147(4), 399–435. https://doi.org/10.1037/bul0000309.
20. Wisniewski, B.; Zierer, K.; Hattie, J.. The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research. *Frontiers in Psychology* **2020**, 10, 3087. https://doi.org/10.3389/fpsyg.2019.03087.
21. Nissen, J. M.; Talbot, R. M.; Nasim Thompson, A.; Van Dusen, B.. Comparison of normalized gain and Cohen's d for analyzing gains on concept inventories. *Physical Review Physics Education Research* **2018**, 14(1), 010115. https://doi.org/10.1103/physrevphyseducres.14.010115.
22. Coletta, V. P.; Steinert, J. J.. Why normalized gain should continue to be used in analyzing preinstruction and postinstruction scores on concept inventories. *Physical Review Physics Education Research* **2020**, 16, 010108. https://doi.org/10.1103/PhysRevPhysEducRes.16.010108.
23. Burkholder, E.; Walsh, C.; Holmes, N. G.. Examination of quantitative methods for analyzing data from concept inventories. *Physical Review Physics Education Research* **2020**, 16, 010141. https://doi.org/10.1103/PhysRevPhysEducRes.16.010141.
24. Krause, S.; Kelly, J.; Tasooji, A.; Corkins, J.; Baker, D.; Purzer, S.. Effect of pedagogy on conceptual change in an introductory materials science course. *International Journal of Engineering Education* **2010**, 26(4), 869–879. https://www.ijee.ie/articles/Vol26-4/13_Ijee2335.pdf.
25. Nissen, J.; Donatello, R.; Van Dusen, B.. Missing data and bias in physics education research: A case for using multiple imputation. *Physical Review Physics Education Research* **2019**, 15(2), 020106. https://doi.org/10.1103/physrevphyseducres.15.020106.
26. Van Dusen, B.; Nissen, J.. Modernizing use of regression models in physics education research: A review of hierarchical linear modeling. *Physical Review Physics Education Research* **2019**, 15(2), 020108. https://doi.org/10.1103/physrevphyseducres.15.020108.
27. Lakens, D.. Equivalence Tests. *Social Psychological and Personality Science* **2017**, 8(4), 355–362. https://doi.org/10.1177/1948550617697177.
28. Lakens, D.; Scheel, A. M.; Isager, P. M.. Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science* **2018**, 1(2), 259–269. https://doi.org/10.1177/2515245918770963.
29. Schneider, M.; Preckel, F.. Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin* **2017**, 143, 565–600. https://doi.org/10.1037/bul0000098.
30. Stammes, H.; de Putter-Smits, L.. Drawing meaning from student-generated drawings: exploring chemistry teachers' noticing. *Chemistry Education Research and Practice* **2025**, 26(2), 494–507. https://doi.org/10.1039/d3rp00253e.
31. Cavanagh, A. J.; Aragón, O. R.; Chen, X.; Couch, B. A.; Durham, M. F.; Bobrownicki, A.; et al.. Student Buy-In to Active Learning in a College Science Course. *CBE—Life Sciences Education* **2016**, 15(4), ar76. https://doi.org/10.1187/cbe.16-07-0212.
32. Carroll, L. J.; Reeping, D.; Finelli, C. J.; Prince, M. J.; Husman, J.; Graham, M.; et al.. Barriers instructors experience in adopting active learning: Instrument development. *Journal of Engineering Education* **2023**, 112(4), 1079–1108. https://doi.org/10.1002/jee.20557.
33. Cooper, K. M.; Downing, V. R.; Brownell, S. E.. The influence of active learning practices on student anxiety in large-enrollment college science classrooms. *International Journal of STEM Education* **2018**, 5(1), 23. https://doi.org/10.1186/s40594-018-0123-6.
34. Wasserstein, R. L.; Schirm, A. L.; Lazar, N. A.. Moving to a World Beyond p ¡ 0.05. *The American Statistician* **2019**, 73(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913.
35. Lakens, D.. Sample Size Justification. *Collabra: Psychology* **2022**, 8(1), 33267. https://doi.org/10.1525/collabra.33267.