Appendix A:
Summary of Items on the ACAST

Below are tables that outline (A) I1-I12 of the generic formative assessment items, (B) the three phases of the gases scenario, and (C) the five phases of the stoichiometry scenario. In order to see the exact wording of the items, view the survey as it was administered online at:

http://tinyurl.com/otxc8sp
-OR-
https://miamioh.qualtrics.com/SE/?SID=SV_6Qe2fL8hzWFw1fv

Generic Formative Assessment Prompts

Item	Description (FA = formative assessment)	DDI*
I1a-c	Rank purpose of FA: Evaluate student understanding, feedback for teaching, & feedback for learning	G
I2	Percentage of FAs collected and/or evaluated	G, E
I3	Generality of knowing why students get items incorrect	E, C
I4a-c	Percentage use of FA: Evaluate student understanding, feedback for teaching, & feedback for learning	C
I5a-g	Frequency of use of FA to: prepare for exams, assess learning objectives, assess understanding, detect misconceptions, give practice, let students apply their knowledge, & assess knowledge of facts	G, C
I6	Percentage of FA that assess single vs. multiple concepts/skills	E
I7	Rank agreement to hypothetical conversations: A) Not enough practice, so give more. B) Students don't understand, so reteach using differing pedagogy. C) Students didn't get it, so recover at a slower pace.	C,A
I8a-d	Agreement with statements: Classroom evidence more reliable than FA results, FA solely for students' benefit, adequate time to analyze FA results, & FA results more reliable than teaching experience	G, E
I9a-d	Frequency aspects are considered when making FAs: What the item will measure, how well it aligns with goal(s), probability that students can respond correctly without understanding, & the format of FA items	G, E
I10a-c	Frequency in making conclusions about: student learning, teaching, & changes to teaching	C, A
I11	Check all sources of evidence used when analyzing FA results	E
I12	Check all determinations made by FA results	C

*Steps of DDI abbreviated: G=Goals, E=Evidence, C=Conclusions, and A=Actions

Gases Scenario

Phase I Most important content to assess on formative assessment about gases (5 options)	**Phase II** Choose *item(s) from below to include on a formative assessment that assess [goal selected previously] (8 options)	**Phase III** For each item selected previously, choose what else that item assesses (5 options)
1. Which gas laws describe which phenomenon 2. Problem-solving ability 3. Which students practice enough and which do not 4. Understanding of PVnT relationships 5. Particulate level understanding of PVnT relationships	❑ **Item 1:** Which gas law describes PV relationship? ❑ **Item 2:** According to Charles' Law, what happens to volume given temperature change? ❑ **Item 3:** How does change in pressure affect volume? ❑ **Item 4:** What affect would doubling pressure have on volume? ❑ **Item 5:** An ideal gas at x atm and y K decreased pressure to z atm, what is final temperature? ❑ **Item 6:** Will volume of an ideal gas be larger or smaller if temperature is raised from x K to y K? ❑ **Item 7:** Describe and draw a) gas molecules in a balloon and b) same molecules after decreased temperature. ❑ I would not pick any of these items.	❑ Algebra/Math skills ❑ Effort put into practicing problems ❑ Proficiency in solving gas law equations ❑ Proficiency in naming gas laws ❑ Understanding of relationship between PVnT ❑ Other:

*These are summaries of the actual items. For the exact wording, see online survey

Stoichiometry Scenario

Phase I Choose the item that best assesses mole-to-mole ratios (6 options)	**Phase II** Choose results you would look at to determine understanding (2 options)	**Phase III** Given item results below, determine understanding of mole-to-mole ratio, dimensional analysis, equation writing, and molar mass calculations (5 options for each determination)	**Phase IV** Choose action(s) based on conclusions (7 options)	**Phase V** Repeat Phase III and IV using...
Item 1: 1:1 ratio, equation not given - OR - *Items 1 or 3*	Raw score (% incorrect)	Assume 54% got *Item 1* incorrect	❑ Reteach as originally taught ❑ Give similar practice problems ❑ Ask students why they're struggling ❑ Alter teaching for current students ❑ Alter teaching for future students ❑ Cannot determine ❑ Other:	*Item 4*
	Individual students' work	Student consistently includes: $\ldots \times \frac{1molA}{1molB} =$ *answer*		
Item 3: 3:1 ratio, equation not given	Raw score (% incorrect)	Assume 54% got *Item 3* incorrect		
	Individual students' work	Student consistently includes: $\ldots \times \frac{1molA}{1molB} =$ *answer*		
Item 2: 1:1 ratio, equation given - OR - *Items 2 or 4*	Raw score (% incorrect)	Assume 54% got *Item 2* incorrect		
	Individual students' work	Student consistently includes: $\ldots \times \frac{1molA}{1molB} =$ *answer*		
Item 4: 3:1 ratio, equation given	Raw score (% incorrect)	Assume 54% got *Item 4* incorrect		*Item 1*
	Individual students' work	Student consistently includes: $\ldots \times \frac{1molA}{1molB} =$ *answer*		

Appendix B:

Summary of Participant and Expert Validation

Participant Validation

Item 1 – Rank of formative assessment purpose			
	Round 2		
		"Student evaluation of understanding" is ambiguous	Changed to "Evaluation of student understanding"
Item 5 – Frequency that formative assessment's purpose is...			
	Round 1		*Address*
		Frequency scale is not conducive to response	Changed frequency scale to something more gradient
		Would answer the question differently for different classes that I teach	Incorporated a class selection at the beginning of the survey
		Since assessments vary in length and scope, it's hard to pinpoint exactly	Consider in interpretation of results
		Some of these frequencies really depend where I am in the curriculum	Even though this was brought up, teachers expressed little trouble formulating a representative response
	Round 2		*Address*
		Measuring in between calendar days doesn't always mean as much to teachers in block schedules	"2 to 4 day" interval is large enough to encompass block scheduling while remaining meaningful
Item 8 – Agreement with 4 statements			
	Round 1		*Address*
		Ordering and wording of last two statements seemed like we were trying to "trick" participants	Spaced the statements apart and made one negatively worded and one positively worded
		Participants wanted a "neutral" option	While not their first choice, all participants gave adequate rationale to their ultimate decision; we wanted this question to force them to think in this way
Item 9 – What are you considering as you design formative assessments			
	Round 1		*Address*
		The phrase "as you design your formative assessments" excludes those formative assessments used but not designed by the teachers	Changed to "In making/choosing items..."
		Ambiguity in the term "goals" for formative assessment	Changed to "learning objectives"
	Round 2		*Address*
		One participant did not really think about this at all because she uses the same format of question for all	Consider in interpretation of results

		formative assessments	
Item 10 – What can you determine using formative assessments			
	Round 1		*Address*
		Phrase “formatting the results” was unclear	This phrasing was erroneous and the prompt was moved into Item 9 where and rephrased “format (multiple choice, free response, etc.) of the item...”
		Confusion about how to conclude that that students get the answer correct without understanding	Similar to the previous question, this prompt belongs in the design item (Item 9)
Item 11 – Check all evidence used to analyze data			
	Round 1		*Address*
		Higher occurrence of formative/summative assessment interchanging	Consider in interpretation of results
		Other parameters than the average (range, median, etc.) were used as evidence	Changed to “class average, median, and/or range of scores”
Gases Items			
	Round 1		*Addressed*
		Item #5 (mathematical gas law problem) would be without memorizing gas law formulas and instead use proportional reasoning	Consider in interpretation of results
Gases Assess			
	Round 1		*Addressed*
		Need to include “make sure they converted temperature units correctly”	Rejected this revision as it is included in “proficiency in solving gas laws”
Stoichiometry Item			
	Round 1		*Addressed*
		Would answer the question differently for different classes that I teach	Incorporated a class selection at the beginning of the survey
		One participant chose based on the “excitement of the reaction being portrayed”	Consider in interpretation of results
	Round 2		*Addressed*
		Participants do not see a difference between the 1:1 vs the 3:1 ratios	Included two additional options that allowed teachers pick both 1 & 3 or 2 & 4
		Participants were choosing based on how complicated the items looked due to format of item	Changed the longer items to make them analogous in appearance to “traditional stoichiometry” problems
Stoichiometry Conclusions			
	Round 1		*Addressed*
		Participants have to either assume arithmetic is correct or get a calculator and check it	Included wording “Assuming all arithmetic shown is correct...”

		Participants did not think that one could meaningfully conclude anything after seeing only one example problem	Included wording "...[assuming] this student consistently responds in this fashion..."
	Round 2		*Addressed*
		Some participants were wary of ever speaking in absolutes	Consider in interpretation of results
Stoichiometry Actions			
	Round 1		*Address*
		Participants wanted to see the previous question for reference	Placed the conclusions and actions items on the same page
		[For participants that were shown a single student's work] It's hard to determine an action based on one student	Consider in interpretation of results

Expert Validation

Item 1 Assertion: Preferred ranking is 1) Feedback for teacher, 2) Feedback for students, and 3) Evaluation of student understanding		
	General Education Experts	*Address*
	• Feedback to teachers requires evaluation of understanding, so those are equal • Issue with hierarchical ranking of purpose formative assessment • Formative assessment is for and about the learner • All are important and necessary	There is no "right" answer to this question. However, the results can speak to how teachers' view the purpose of formative assessment.
Item 2 Assertion: This item determines the ratio of formative assessments teachers collect and/or evaluate compared to those they do not.		
	General Education Experts	*Address*
	• Participants will have a difficult time estimating percentage in valid/reliable way • Unsure if this includes formal/informal formative assessments	It is important (and unavoidable) that teachers answer according to what they believe formative assessment is. Therefore we recognize that all types of "formative assessments" teachers recognize are included in this estimate.
Item 6 Assertion: Isolating one variable of interest generally yields more valid analyses and interpretations than assessing multiple variables.		
	General Education Experts	*Address*
	• By default, assessments will encompass multiple concepts/skills • Leave it for the teachers to decide what counts as a single versus a multiple skill/concept	We agree that assessments generally assess multiple versus single concepts/skills, but that is an important finding in and of itself. During participant validation interviews, we asked for an example of a single vs. multiple concept/skill. There was a high degree of alignment among the responses considering the chemistry curriculum.
Item 7 Assertion: Rank should be 1) Teacher B, 2) Teacher C, and lastly 3) Teacher A		
	General Education Experts	*Address*
	• Not enough information is provided in prompt to advise an action	Considering our qualitative data, whether or not teachers actually knew why students got an item incorrect didn't change their conclusions. What sought here was what action teachers found most appealing and for what reason.
Item 9 Assertion: Teachers should be thinking about all of these things on every assessment (10), but with all factors considered, frequencies will most likely be less than every assessment.		
	General Education Experts	*Address*
	• Should include readability of item and how many items are needed to effectively cover the learning objective	While these are good suggestions and would provide valuable insights, we chose to not include these options for

			consideration of the length of the survey. Other questions directly observed in qualitative data held priority over these suggestions.
Item 10 Assertion: Teachers should be determining all of these things on every assessment (10), but with all factors considered, frequencies will most likely be less than every assessment.			
	General Education Experts		*Address*
		• Should include "real-time" modifications to instruction	While this would provide valuable insight, the purpose of this survey was largely to reflect data-driven inquiry practices of *written* formative assessments as this was the main form teachers discussed in the qualitative interviews.
Item 11 Assertion: Use of scores should only be used with appropriate item design, Statements 4,6,7, and 9 are not validly assessed by most content-focused formative assessments, and student work is generally the most valid means of making analyses.			
	General Education Experts		*Address*
		• Most of these could be helpful if students show incorrect work	We agree that most would be helpful. However, the prompt stated that we didn't believe these were validly assessed by most content-focused formative assessments, a point this reviewer did not address.
Assertion about Gas Item 4: Only Item #4 assesses understanding of relationships in a reliable manner.			
	Chemistry Experts		*Address*
		• Question is not worded to give response desired, ambiguity in what is being asked, "what the gas molecules look like" has nothing to do with relationships, the main task of drawing the molecules doesn't get at the relationships, "contractable" is not a word, most students would only draw the "after" picture and not the before, which wouldn't get at their understanding as well • Items 3 and 5 could also assess this goal as they don't involve numbers • This assertion of the goal of "understand the relationships between pressure, volume, temperature, and/or moles" seems to assume *particulate* relationships when other exist • The word "reliable" in the assertion should be "valid" • Directions should say "that you would include" so that teachers don't think they had to have included that exact question in	Given the problems with the wording of item 4, we have revised it to incorporate all the suggestions to "Describe and draw a) gas molecules in a balloon and b) the same molecules after a decrease in temperature assuming constant pressure and moles." We believe (and further participant validation has confirmed) that teachers will much more readily accept the wording of this item in its current form. We have also sided with the experts and believe that Items 3,4, and 5 could all be seen as assessing student understanding in a valid manner. We also agree that we are assuming that "student understanding of relationships" refers to particulate, so we have included the option for teachers to specific that on the previous question in the survey. We have change language of our interpretations from "reliably" to "validly," as this is the accurate term and have also modified the

		order to check the box • The question and assertion equates "understanding" to the students not relying on memorization or formulas, which may not be the case	stem to include the word "would" per one of the suggestions. Lastly, we have evidence that teachers' will have widely varying definitions of what it means to "understand" something. One of the purposes of asking the question is to find out precisely what that means in the context of gas laws.
Assertion about Gas Items 3 and 5: Although these items assess understanding of relationships, students can memorize (#3) and solve mathematically using hypothetical numbers (#5), making these less reliable than Item #4			
	Chemistry Experts		*Address*
		• *[same critique as in previous question]* This assertion of the goal of "understand the relationships between pressure, volume, temperature, and/or moles" seems to assume *particulate* relationships when other exist • Does it matter how students come to know a relationship and can you differentiate between the ways students demonstrate their knowledge by Items 3 and 5?	We address this as we stated previously: by allowing teachers to specify whether they wish to look at the particulate level exclusively or not. To the second point, we acknowledge that you cannot determine how the student came to the answer s/he did, which is why we've made the assumption that Item #4 (as newly revised) is a more valid way of determining understanding. Additionally, 4/6 of the chemistry experts agreed that it was problematic that students could correctly answer these problems by not relying solely on understanding of the particulate level.
Assertion about assessing effort: Effort or motivation can never be reliably determined by these items.			
	Chemistry Experts		*Address*
		• This would depend on previous instruction because if a teacher demonstrates how to draw gas molecules, for example, it could imply that students would have had to at least try to pay attention and incorporate this into their work • Those who choose Items 6 or 7 might select this because if it's just about doing calculations, it could come down to how much effort they put into practicing	While it *could* depend on previous instruction, we believe that the majority of the time, it won't. Although the mathematical calculations generally imply that practice is required to complete them correctly, we still cannot validly determine effort/motivation because students will need varying levels of practice.
Assertion about stoichiometry item: Item #4 is the "best" item to exclusively assess molar ratios. Items 1 and 3 require students to write/balance equations, calculate molar masses, and convert grams to moles. While Item #2 similarly assesses only molar ratios, it is a 1:1 ratio, meaning that whether or not students consider this, they will get the same answer.			
	Chemistry Experts		*Address*
		• Item #4 is difficult to read, as such this might drawing teachers away from	Item #4, along with the rest of the items, have underwent revisions to make them

		choosing the problem • Item #4 gives too much direction and therefore you can't conclude anything from results • You could look at student work for Item #3 and assess mole-to-mole ratio understanding if you account for the other skills assessed • The stem of this question asks which problem teachers *would* use as opposed to which one *best* assesses mole-to-mole ratios • There should be a qualitative response option for teachers to state why they chose the item they did in order to get a better glimpse of the rationale instead of assuming what it is	appear more simplistic while retaining "similar structure" (meaning that one item does not stand out from the others because it clearly looks different, which could possibly draw teachers to choose it): "If 0.00788 mol of barium bromide reacts with excess lithium phosphate, how much (in moles) barium phosphate would be produced? Balanced equation is..." We acknowledge that you could look in the students' work in Item #3 to determine understanding, which is the primary reason we ask this in the following question. The rationale for excluding the word "best" from the stem is to ensure that teachers are answering this based on what they do in the class. If they are not originally thinking about the "best" item, we do not wish to impose that thought on them. While we agree that a qualitative response for why they chose the answer they did would provide valuable insights, we have evidence from the qualitative study to suggest that teachers do not always consider the content-specific nuances in selecting items for formative assessments. We also have additional items on this survey that will assess the degree to which they consider these nuances.
Assertion about what should be examined: Examining individual student work is always more valid/reliable than looking at aggregate scores. However, one could use scores to validly determine understanding from Items #4 and #2 to a degree.			
	Chemistry Experts		*Address*
		• Even in Item #4, students can make transcription and arithmetic errors, so teachers should always examine individual student work over scores • Examining percent of students that do certain things (i.e. 30% used the wrong molar ratio, 40% calculated molar mass incorrectly...) could be useful in determining understanding	We agree that students can still make errors that can't be determined by examining only raw scores. We will consider this in our results. While teachers can examine percent of students who did X incorrect, the prompt clearly states that the percent applies to the item as a whole. There was no confusion about this during the teacher validation interviews.
Assertion about what can be determined: For Item #4: 1) Student absolutely doesn't understand mole-to-mole ratios because the equation shows a 3:1 but the student puts a 1:1. 2) Absolutely understand dimensional analysis because the setup is correct and complete. 3) Cannot determine anything about understanding of balancing equations and calculating molar mass since			

these are given in the problem.		
	Chemistry Experts	*Address*
	• Just because it's shown correct does not mean that they "understand" dimensional analysis, students can "memorize" this format of converting units without knowing why they're doing what they're doing • Objection to the term "absolute"	We agree that students can "memorize" a pattern for dimensional analysis. Caution will be taken in the interpretation of this item. We also generally object to the term "absolute" as researchers. However, the confidence displayed by the teachers (initial survey data supports this as well) shows that teachers may not have a similar objection to the certitude expressed here.
Assertion about actions: There are no "right" answers as the action a teacher takes depends on a number of factors.		
	Chemistry Experts	*Address*
	• Agreeing with the assertion, why would you ask this question? • Cannot validly determine this based on a single student's selection • Changing instruction for future students is slightly worse than some of the others because formative assessment should be about the current students, not necessarily the future students	The purpose of asking this question is to determine what general outcomes the teachers favor (characterization). This will give us information about their beliefs on pedagogical response. We agree that that teachers should not be making pedagogical actions based on the response of one student and will exercise caution when interpreting the results of this question. Lastly, while we agree that there is a difference between the current and future students, if something was a problem for one group of students that causes a change, implementing that change for future "by default" is not really different than doing what was done previously. Either way, a teacher will end up doing something not knowing for sure it will work with a given population of students.

39
40

Appendix C:

Summary of Test-Retest Reliability

This document contains test-retest reliability for every item on the ACAST. Traditionally, a correlation is computed between scores or subscores on two administrations of the same instrument. However, due to limitations of the correlation calculation as well as the lack of score from the ACAST (it was not designed to produce a meaningful total score), we consider the consistency of student responses per item for continuous and ordinal level measures (Table 2) and nominal and dichotomous level measures (Table 2).

The zeta-range estimator shown in Table 1 is calculated by first allowing the participants to define their own error ("I can respond within +/- X of this value"). Then the proportion of those that fall outside of the error range is calculating, with lower scores indicating that more participants are able to respond within their measurement error, indicating better reliability. Instead of relying just on this one calculated proportion, 10,000 resamples are bootstrapped in order to make a confidence interval for the estimated proportion that will respond outside of the zeta-range.

The Chi g.o.f in Table 2 was determined by computing the total number of student who were consistent in their response (chose same response from test to retest administration) and comparing it against that proportion who were not consistent (chose different response from test to retest administration).

Table 1: Continuous and ordinal level measures

			Traditional Association				**P_ζ Estimator**	
Item	**Level**	**Description**	r	p (r)	ρ	p (ρ)	ζ	P_ζ CI (%)
I2	Continuous	FAs collected	0.46	<.001	-	-	15	33.9 – 59.7
I4A	Continuous	Design for feedback to learning	0.55	<.001	-	-	15	33.9 – 58.1
I4B	Continuous	Design for feedback to teaching	0.47	<.001	-	-	15	27.4 – 50.0
I4C	Continuous	Design for student evaluation	0.54	<.001	-	-	15	16.1 – 40.3
I6	Continuous	Assess single/multiple concepts	0.50	<.001	-	-	15	30.6 – 53.2
I9A	Ordinal	What the item will measure	0.18	0.168	0.45	0.002	2	4.8 – 19.4
I9B	Ordinal	Aligned with objectives	0.20	0.113	0.22	0.158	2	3.2 – 17.7
I9C	Ordinal	Respond correctly without understanding	0.50	<.001	0.58	<.001	2	19.4 – 41.9
I9D	Ordinal	The format of the item	0.46	<.001	0.38	0.010	2	21.0 – 43.5
I10A	Ordinal	Conclusions about student learning	0.38	0.002	0.36	0.017	2	3.2 – 17.7
I10B	Ordinal	Conclusions about teaching	0.41	<.001	0.39	0.009	2	1.6 – 12.9
I10C	Ordinal	Changes to teaching	0.48	<.001	0.42	0.005	2	0 – 11.3

84 Table 2: Nominal and dichotomous level measures

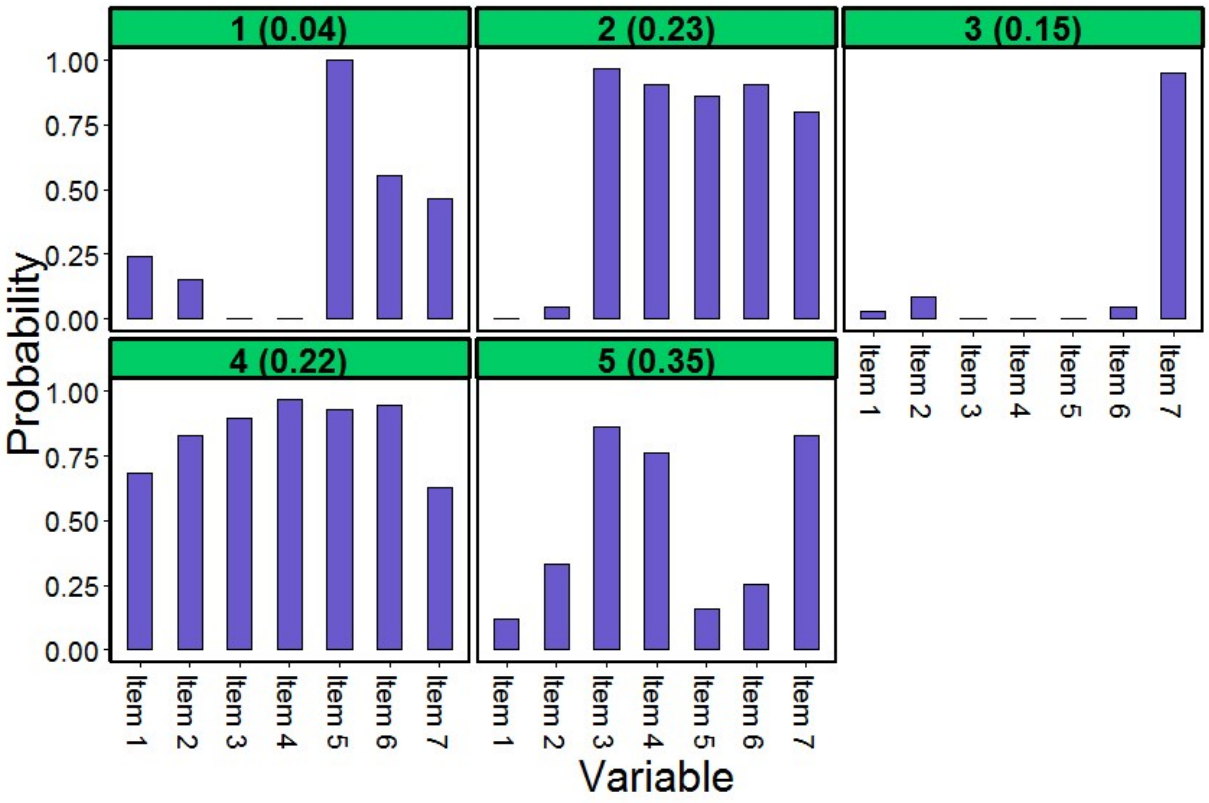
			Chi GOF		
Item	**Level**	**Description**	χ^2	p (χ^2)	w
I1A	Nominal	Evaluation of students	66.78	<.001	1.04
I1B	Nominal	Feedback to teacher	9.32	0.002	0.39
I1C	Nominal	Feedback to students	14.91	<.001	0.49
I3	Dichotomous	Why students are incorrect	31.23	<.001	0.71
I7	Nominal	Ranking three teachers	88.32	<.001	1.19
I11A	Dichotomous	Evidence: Score parameter	7.81	0.005	0.35
I11B	Dichotomous	Evidence: Work shown	25.81	<.001	0.65
I11C	Dichotomous	Evidence: Amount of content	7.81	0.005	0.35
I11D	Dichotomous	Evidence: Amount of practice	9.29	0.002	0.39
I11E	Dichotomous	Evidence: Familiarity with problem	0.26	0.612	0.06
I11F	Dichotomous	Evidence: Attention	14.52	<.001	0.48
I11G	Dichotomous	Evidence: Class Observations	3.16	0.075	0.23
I11H	Dichotomous	Evidence: Similar task performance	9.29	0.002	0.39
I11I	Dichotomous	Evidence: Motivation	12.65	<.001	0.45
I11J	Dichotomous	Evidence: Courses taken by students	25.81	<.001	0.65
I11K	Dichotomous	Evidence: Previous years' performance	58.06	<.001	0.97
I12A	Dichotomous	Conclusion: Problem-solving ability	7.81	0.005	0.35
I12B	Dichotomous	Conclusion: Grades	16.52	<.001	0.52
I12C	Dichotomous	Conclusion: Mathematic ability	10.90	<.001	0.42
I12D	Dichotomous	Conclusion: Understanding of content	20.90	<.001	0.58
I12E	Dichotomous	Conclusion: Chemistry in math tasks	50.58	<.001	0.90
I12F	Dichotomous	Conclusion: Motivation	7.81	0.005	0.35
I12G	Dichotomous	Conclusion: Teaching style	1.61	0.204	0.16
I12H	Dichotomous	Conclusion: Confidence	9.29	0.002	0.39
I12I	Dichotomous	Conclusion: Performance past years	20.90	<.001	0.58
G1	Nominal	Most important goal	90.86	<.001	1.21
G2A	Dichotomous	Gases: Item 1	28.45	<.001	0.68
G2B	Dichotomous	Gases: Item 2	7.81	0.005	0.35
G2C	Dichotomous	Gases: Item 3	12.65	<.001	0.45
G2D	Dichotomous	Gases: Item 4	18.65	<.001	0.55
G2E	Dichotomous	Gases: Item 5	7.81	0.005	0.35
G2F	Dichotomous	Gases: Item 6	0.26	0.612	0.06
G2G	Dichotomous	Gases: Item 7	31.23	<.001	0.71
S1	Nominal	Stoich item mole ratios	123.92	<.001	1.41
S2	Nominal	Results preferred	12.65	<.001	0.45
S3A	Nominal	Mole ratio conclusion	21.49	<.001	0.59
S3B	Nominal	Dimensional analysis conclusion	27.78	<.001	0.67
S3C	Nominal	Writing/Balancing conclusion	113.81	<.001	1.35
S3D	Nominal	Calculating molar mass	90.86	<.001	1.21

85
86

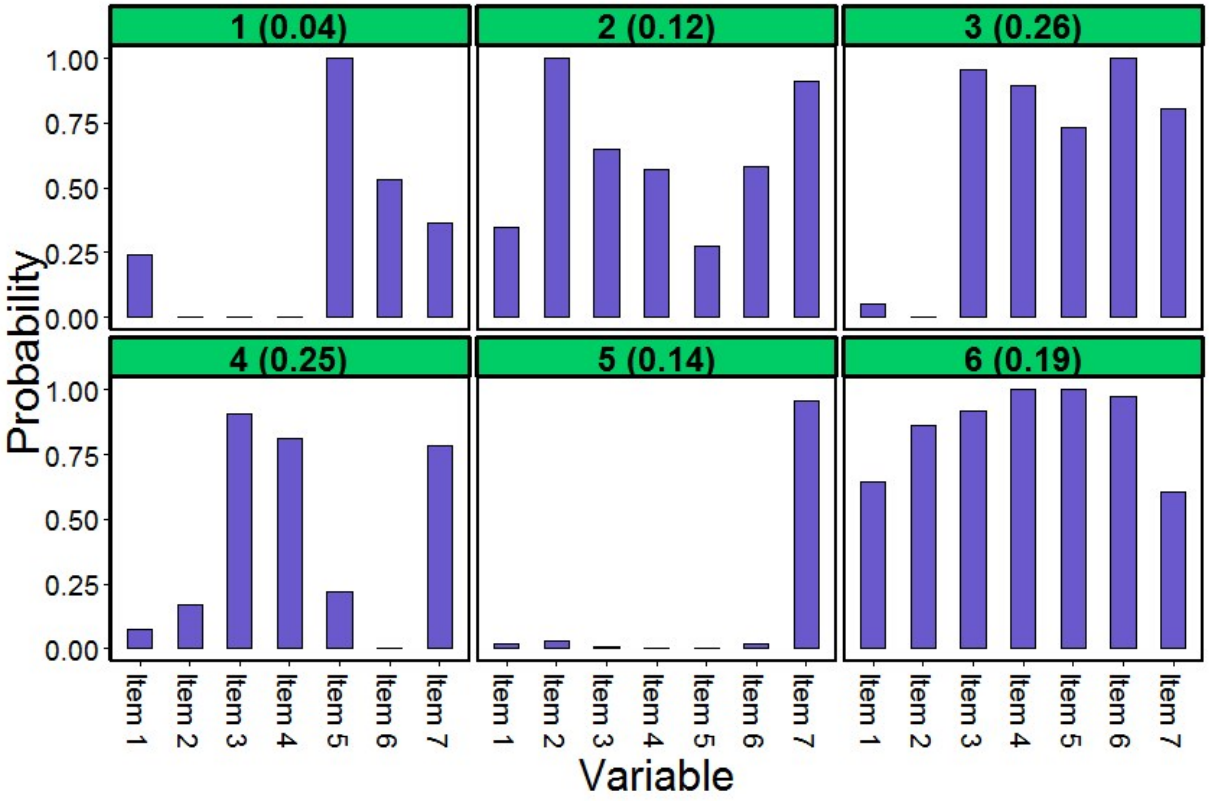
Supplemental Information B:
Models 1-6

Because only Models 4 and 6 are discussed in the manuscript, these and the other models not discussed are presented here.

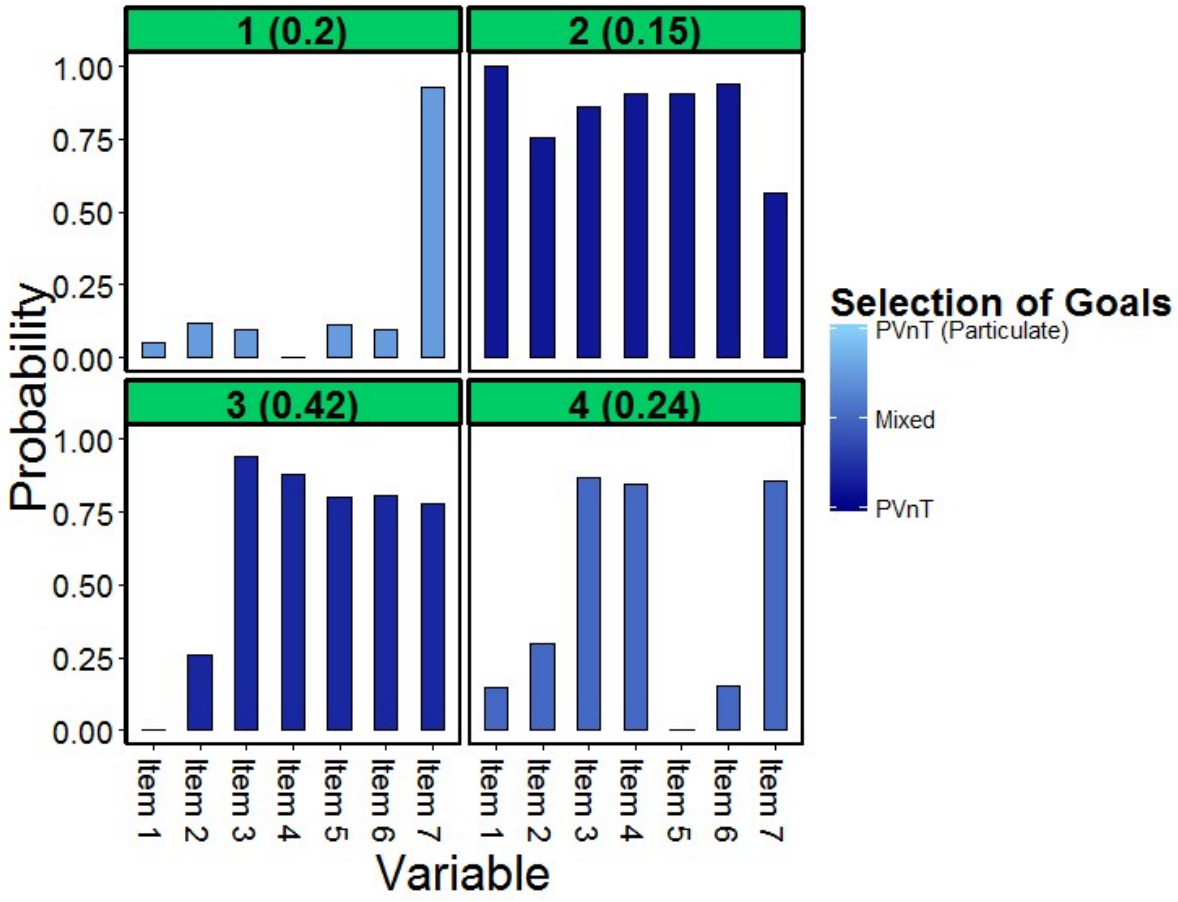
Model 1
Scenario: Gases
Models responses to: Selection of items only
Classes: 5



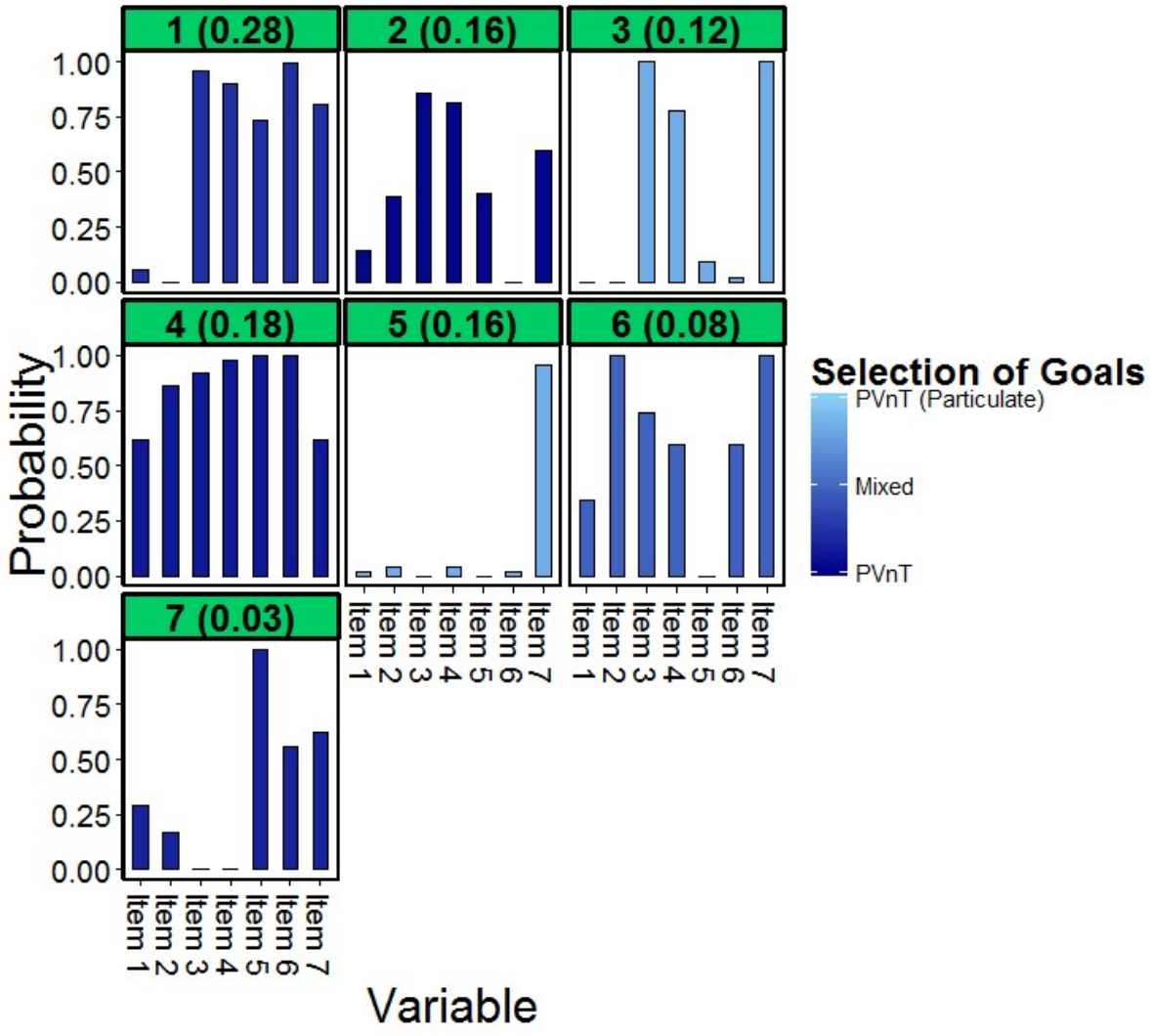
Model 2
Scenario: Gases
Models responses to: Selection of items only
Classes: 6



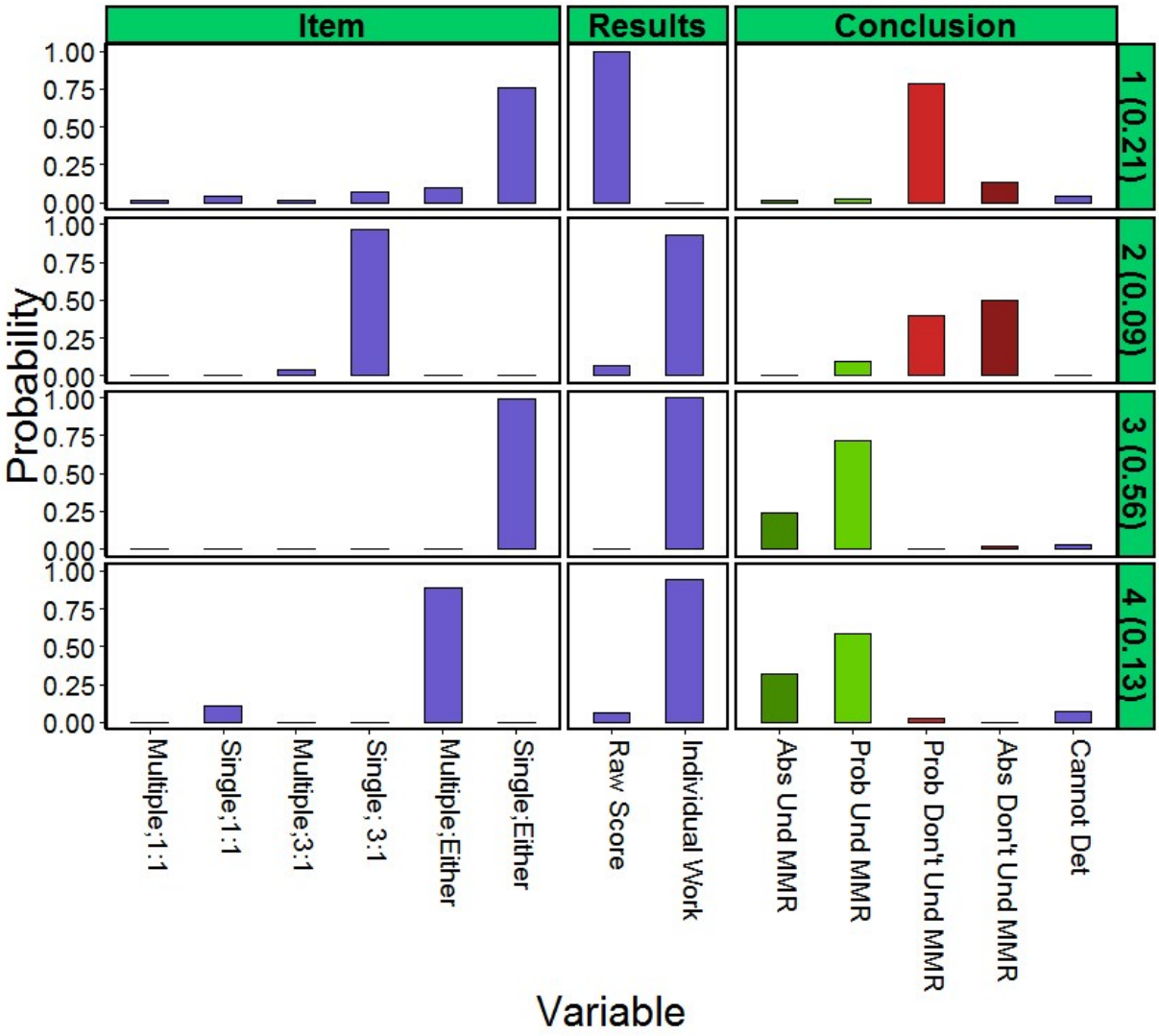
Model 3
Scenario: Gases
Models responses to: Selection of goals and items
Classes: 4



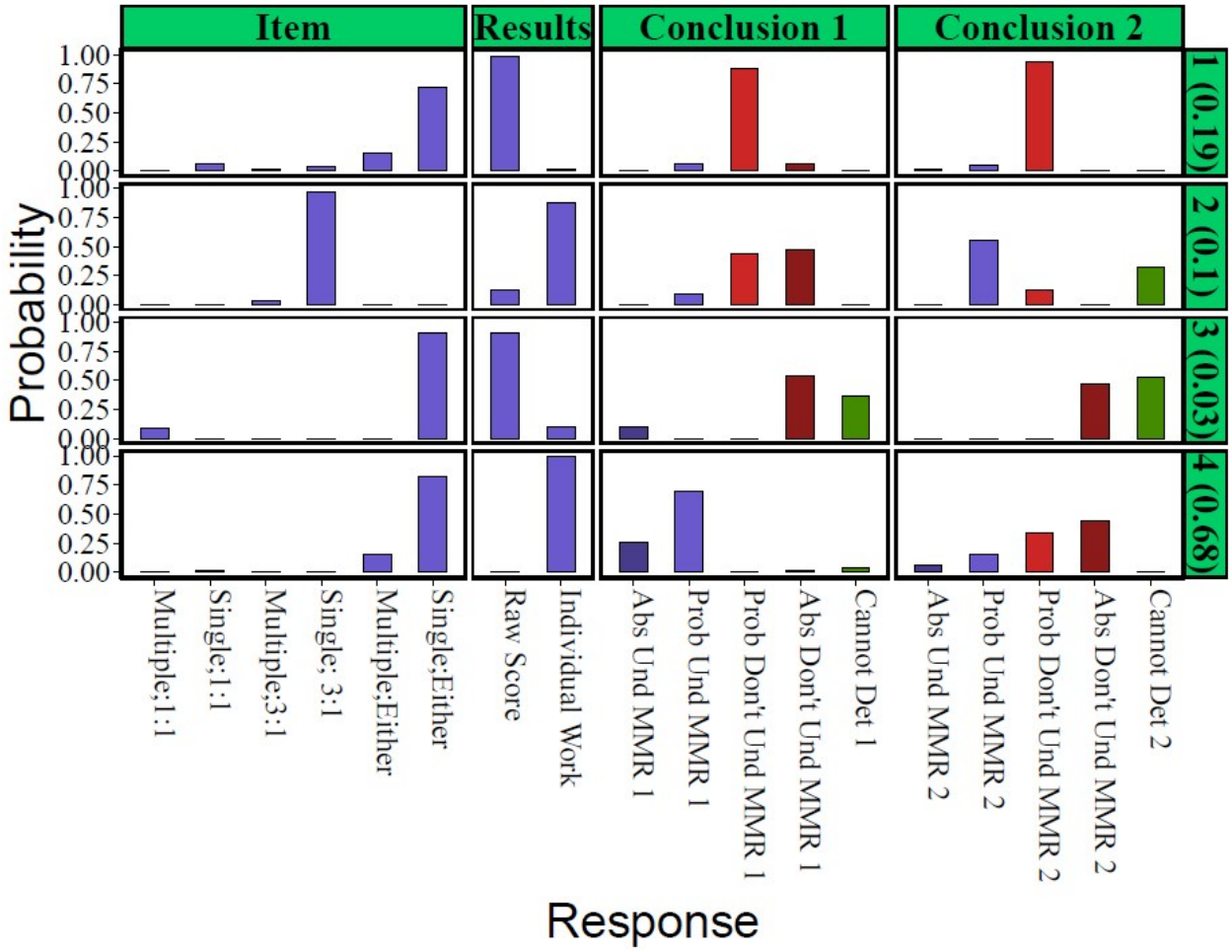
Model 4
Scenario: Gases
Models responses to: Selection of goals and items
Classes: 7



Model 5
Scenario: Stoichiometry
Models responses to: Selection of item, response format, and first iteration conclusions
Classes: 4



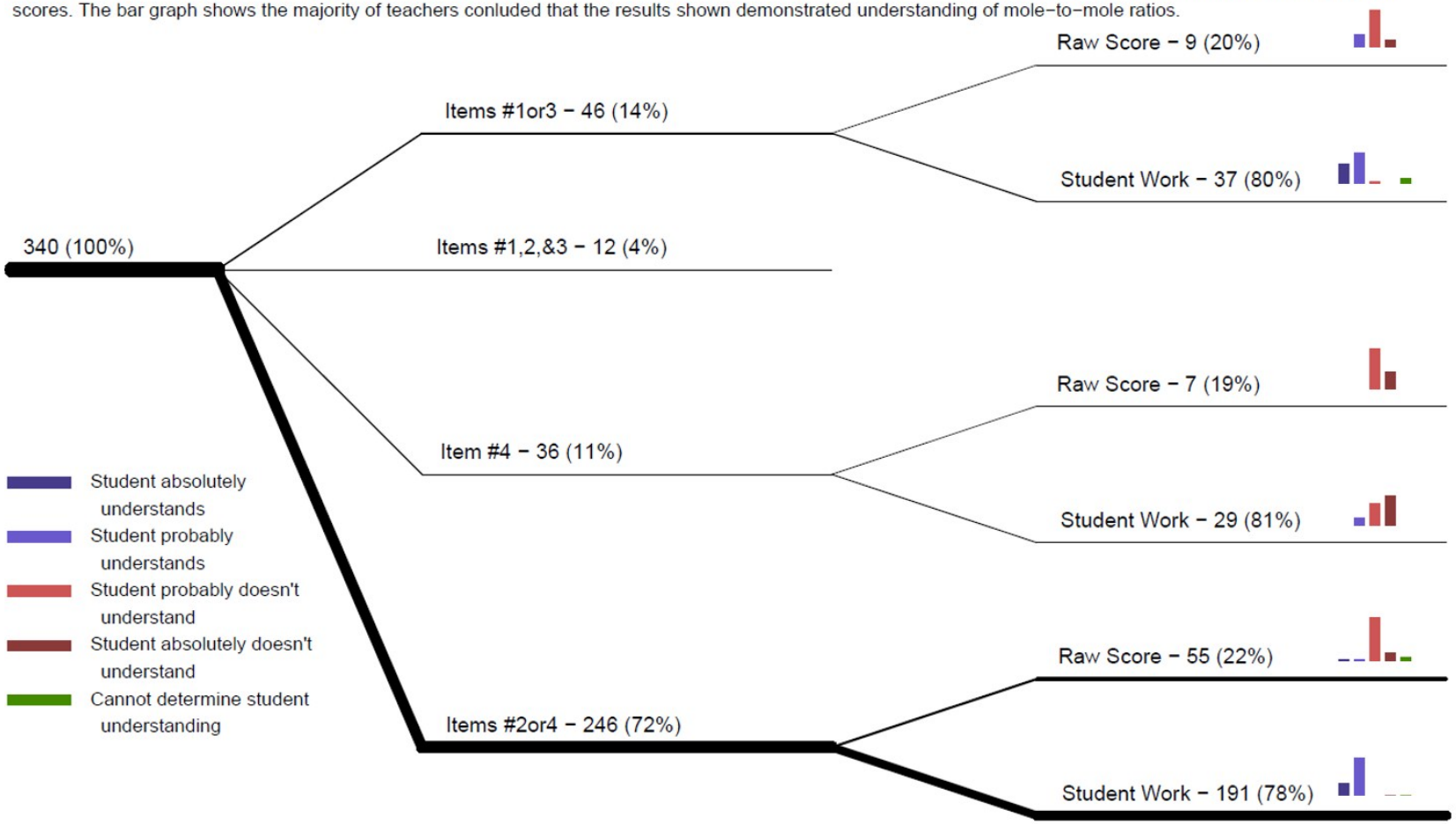
Model 6
Scenario: Gases
Models responses to: Selection of item, response format, and first and second iteration conclusions
Classes: 4



Supplemental Information E:
Stoichiometry Results

How to read chart

From left to right, frequencies (percent in parenthesis) of teachers that choose which item(s) on the first part of the stoichiometry scenario are given. Out of the number of teachers that chose each item(s) in the first part, the frequencies that chose to look at raw scores versus student work are shown next. Finally, the bar graph on the right displays the distribution of teachers that, based on (a) hypothetical student response(s), determined the students understood (blue), did not understand (red), or could not determine understanding (green) mole−to−mole ratios. An example interpretation of the thickest line (line width proportional to overall frequency) would be: Out of 340 teachers, 246 (72%) said that either Items 2 or 4 would best assess mole−to−mole ratios. Of these teachers, 191 (78%) preferred collection of student work over raw scores. The bar graph shows the majority of teachers conluded that the results shown demonstrated understanding of mole−to−mole ratios.



How to read chart

From left to right, frequencies (percent in parenthesis) of teachers that were given item(s) on the second part of the stoichiometry scenario are shown. Out of the number of teachers that were given each item in the second part, the frequencies that chose to look at raw scores versus student work are shown next. Finally, the bar graphs on the right display the distribution of teachers that, based on (a) hypothetical student response(s), determined the students understood (blue), did not understand (red), or could not determine understanding (green) of mole−to−mole ratios. An example interpretation of the thickest line (line width proportional to overall frequency) would be: Out of 340 teachers, 304 (89%) were asked to determine understanding of mole−to−mole ratios given student responses to Item 4. Of these teachers, 234 (69%) preferred collection of student work over raw scores. The bar graph shows the majority of teachers conluded that the results shown did not demonstrate understanding of mole−to−mole ratios, but a significant proportion concluded the opposite.