Supporting Information for

Tracking Student Learning Across a Conceptual Landscape: Transitions and Differential Changes in Conceptual Modes During a Unit on Chemical Kinetics and Equilibrium

Jannik Lossjew[a] and Sascha Bernholt[a]

[a]Leibniz-Institute for Science and Mathematics Education, Olshausenstr. 62, D-24118 Kiel, Schleswig-Holstein, Germany

A. Competence goals of lower secondary education in both federal states (Schleswig-Holstein, Rhineland-Palatinate) with regard to the six developmental statements

Since we adhered to the exact wording of the respective curricula, in one case competence goals are formulated as operationalized learning objectives and in the second case more strongly as a description of content.

Table A1: Expected learning-related baseline based on the regional curricula with reference to the six formulated developmental statements.

Developmental Statement	Schleswig-Holstein	Rhineland-Palatinate
I	• explain the composition of substances and mixture using a particle model • describe and explain changes in states of matter using a particle model • apply their prior knowledge of the structure of matter to predict possible chemical reactions • explain specific properties of molecular substances using intermolecular interactions	• particle concepts are introduced and expanded: substances consist of particles (1), the particles of a pure substance are all identical (2), and particles are in motion (3) • polarity of compounds is determined by their molecular structure • electron pair bonds and intermolecular interactions determine the spatial structures of plastics
II	• describe activation energy as the energy required to bring substances into a reactive state	• -
III	• describe and explain chemical bonding in molecules • explain changes in chemical reactions at the atomic level • explain the energy balance of chemical reactions in simple terms with the breaking and formation of chemical bonds	• the bonds between atoms in hydrogen, oxygen, water and methane (as well as other hydrocarbons) are based on shared electron pairs • matter and energy transformations are modeled based on changes in molecules and electron pair bonds • matter transformations are modeled based on changes in particles and their bonds
IV	• explain different states of matter of a substance using the relationship between the kinetic energy of particles and temperature • describe activation energy as the energy required to bring substances into a reactive state • describe the effect of a catalyst on the activation energy	• the states of matter of water are explained using simple particle models • the change in energy carrier (energy release) [during chemical reactions] is noticeable through heating, movement or light • chemical reactions are controlled by varying the reaction conditions • catalysts are typically used to lower the activation energy
V	• -	• the formation and

		decomposition of a metal oxide are in principle reversible • chemical reactions in an accumulator are reversible
VI	• describe the effect of a catalyst on the activation energy	• catalysts are typically used to lower the activation energy

B. Coding Manual for assessing conceptual modes of the esterification reaction

Table A2: Overview of mode

Mode	Open Text Item*	Sub-modes	Description	Example	Cohens κ
Start (MS)	1) Describe at the submicroscopic level how acetic acid and ethanol molecules behave when the liquids are combined. 2) Explain why combining the two liquids directly causes the reaction to start.	Advanced Collision (**COLEXP**)	Students use the concept of random effective collisions and resulting (electronic) interactions to explain the reaction start.	The alcohol and acid molecules are always in motion. When these collide with sufficient kinetic energy, this initiates the start of the reaction. The fact that the activation energy of certain particle pairs can already by exceeded at room temperature also plays a role here.	0.71 (pre); 0.81 (post)
		Simple Collision (**COLSIMP**)	Students use the concept of random collisions without further interactions.	As soon as two molecules collide successfully, a reaction occurs.	
		Mutual Attractions (**ATTRAC**)	Students use the concept of mutual attraction to explain the start of reaction.	Polarization of the alcohol and acid molecules causes them to approach, leading to their reaction.	
		Initial Factor** (**INIT**)	Students use the concept of an initiating substance (reactivity) or name external factor as reason for the reaction start.	During an esterification, acid molecules a particularly reactive and start the reaction.	
		No Concept (**NOCONC**)	No concepts are used to explain the reaction start.	The two substances react to form an ester and water.	
Progress (MP)	3) Describe how you imagine the reaction at the particle level to proceed.	Advanced Collision (**COLEXP**)	Students use a combination of different interactions (continued collisions – including product particles, bond breaking/forming, escape/remaining of a substance in the system) to explain the progress.	Ethanol and acetic acid molecules continuously collide. If this happens (by chance) with sufficient kinetic energy, a bond is formed (between the hydroxyl group and the carboxyl carbon atom), followed by the cleavage of a stable water molecule. Product molecules become more.	0.81 (pre); 0.85 (post)
		Simple Collision (**COLSIMP**)	Students use the concept of perpetual collisions without describing further interactions.	The molecules continue to collide and thereby react, but they can also rebound after colliding. Product molecules become more.	
		Bond breaking and reformation (**BONDS**)	Students only use the concept of bond breaking/forming to explain the progress.	New bonds are formed between ethanol molecules and acetic acid molecules, followed by bond cleavage with the elimination of a water molecule.	
		Macroscopic change of substances (**MACRO**)	Students describe the course of the reaction from a substance-based perspective (e.g. increase in products/decrease in reactants)	More and more ester is formed.	
		No concept (**NOCONC**)	No concepts are used to explain the progress.	Acetic acid is more acidic than ethanol.	

Mode	Open Text Item	Sub-modes	Description	Example	Cohens κ
End (ME)	4) Describe how you imagine the end of the reaction at the particle level. Which particles are still present at this point? 5) Explain how the reaction ends at this point – are processes still detectable at the macroscopic or submicroscopic level?	Dynamic Equilibrium (**DYNAMIC**)	Students explain that reactions do not actually end, but instead reach a state of dynamic equilibrium.	The ratio between reactants and products no longer changes (macroscopic perspective). However, both the forward and reverse reactions continue to occur constantly at equal reaction rates, so that no net compositional change is observable.	0.85 (pre); 0.91 (post)
		Static Equilibrium (**STATIC**)	Students explain the and of the reaction by the achievement of a static equilibrium.	A state of equilibrium is established. Once this state is reached, neither the forward nor the reverse reaction occurs, which is why the ratio of reactants to products remains constant.	
		Limiting Component (**LIMIT**)	Students explain the reaction end by the presence of a limiting factor and/or by a complete reaction of the reactants.	The reaction is complete when all acetic acid and ethanol molecules have fully reacted.	
		No concept (**NOCONC**)	No concepts are used to explain the reaction end.	A colorless liquid is still visible in the flask.	

C1. Rationale for the Choice of the test for marginal homogeneity

Literature suggests, that the “classic” Chi^2 test for independence should not be used in paired samples (Agresti, 2002). We therefore discussed different test that would fit our data. In general, the McNemar-Bowker test replaces the "classic" Chi^2 test for contingency tables that exceed a 2×2 design. However, this test becomes unreliable when many cells have a frequency of zero, as (I) it cannot be computed globally in such cases, and (II) cell-level statements about symmetry should not be interpreted when cells in the pre-condition are unpopulated. For this reason, we applied a test of marginal homogeneity, which allows conclusions about overall changes in the marginal distributions of the transition matrix. This test, however, limits cellwise interpretations and – in our case – does not allow for conclusions about dependencies between pre- and post-sub-modes. To examine such dependencies, we used ordinal regression analysis.

C2. Theoretical Rationale for the Choice of the Additive Approach and Details on ATT Estimation

Applying the classical DiD framework to categorical outcomes such as conceptual modes presents challenges, as traditional approaches typically assume continuous outcomes relying on linear relationships (Graves et al., 2022). Graves et al. (2022) extend the classical DiD framework to research settings with categorical outcomes, explicitly addressing how treatment effects, i.e. taught contents, can be identified under different assumptions about transitions across outcome states and marginal outcome distributions.

A key distinction in their approach is the comparison between additive and multiplicative assumptions when modeling changes over time. The additive approach assumes that, in the absence of treatment, pre-existing differences between groups would remain constant over time – referred to as fixed group differences. In contrast, the multiplicative approach assumes that treatment and comparison group share a common transition structure. This means that while treatment may accelerate learning, both groups will eventually converge to the same long-term distribution, regardless of whether they receive treatment or not. Both scenarios are represented in figure 1, where the intervention effect of both approaches is shown as difference

($d_2 – d_3$) between group differences d_2 (expected group difference in the post-condition without intervention) and d_3 (expected group difference in the post-condition with intervention). Under the additive assumption, this intervention effect is defined as deviation from the fixed differences which the two groups would have maintained over time without any sort of intervention, i.e. any additional change in the treated group relative to the comparison group is attributed to the intervention. Under the multiplicative assumption, both groups follow the same transition patterns, progressing toward a shared stationary distribution. Without intervention, the gap between the groups would change over time (no fixed effects). The intervention, however, disrupts this natural convergence by accelerating the treated group's progress, thereby widening the gap. This distinction is crucial because it determines how treatment effects are identified and whether differences in outcome distributions are attributed to the intervention or to pre-existing structural differences.

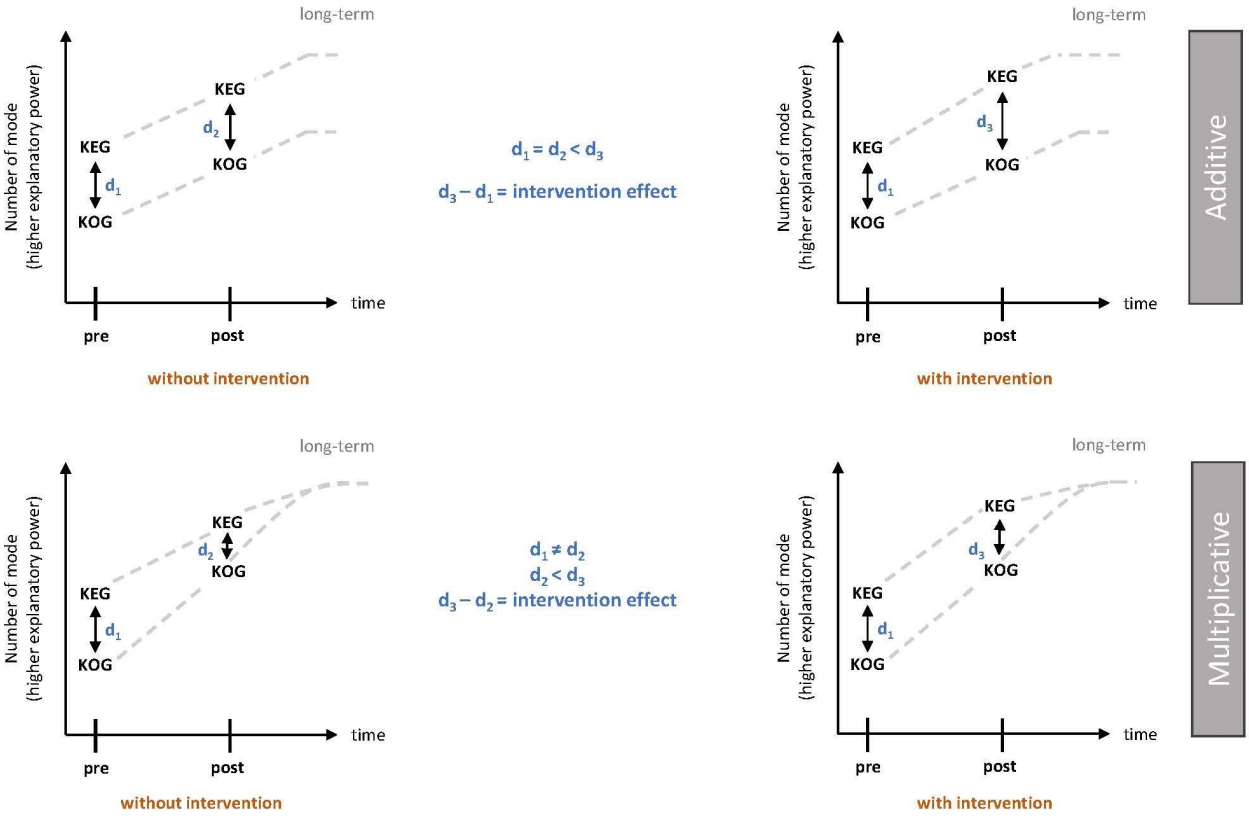

Figure A1: Comparison of theoretical developmental lines following either the additive approach (upper part of the figure) or the multiplicative approach (lower part of the figure).

In both cases, however, an ATT (Average Effect of the Treatment on the Treated) estimator is calculated, capturing the causal impact of the intervention by comparing the changes in

outcomes between the treatment and comparison groups over time. Following the framework of Graves et al. (2022), when treatment and comparison groups share the same baseline distributions, both the additive and multiplicative ATT estimators produce identical treatment effect estimates. However, when baseline distributions differ, the choice between these two options must be guided by theoretical considerations. In our study, students were naturally assigned to one of the two instructional conditions: (I) KOG received instruction only on chemical kinetics and (II) KEG received instruction on both chemical kinetics and chemical equilibrium. As in our case the two groups already exhibited differences in conceptual mode distributions at baseline, we explicitly choose the additive approach based on the following theory-driven reasoning: prior knowledge is a well-established predictor of learning outcomes in the sense of knowledge stability, with numerous studies indicating that differences in initial knowledge levels tend to remain stable over time when learning opportunities are held constant (Simonsmeier et al., 2022). Given that prior knowledge plays a key role in shaping individual progressions, we argue that pre-existing differences between groups would have remained constant in the absence of an intervention. Unlike more basic developmental processes, where individuals typically reach similar end states, conceptual learning processes in academic contexts often do not adhere to shared transition patterns or lead to stationary distributions. Thus, any deviation from this fixed difference should be attributed to the instructional intervention. Consequently, we adopt the additive approach, as it aligns with our theoretical expectations and empirical findings on the stability of knowledge differences over time. To quantify the effect of additional instruction on chemical equilibrium concretely, we measure the probability of students occupying each conceptual mode before and after instruction in both groups. This allows us to estimate how the probability of being in a given mode changes with each group over instructional time and then compare these changes across groups. Mathematically, the key idea that any change in the treatment group that exceeds the change in the control group can be attributed to intervention, is expressed as follows:

$$\pi = \left(P_{post}^{KEG} - P_{pre}^{KEG}\right) - \left(P_{post}^{KOG} - P_{pre}^{KOG}\right)$$

where P_{pre} and P_{post} represent the probabilities of a student being in a specific conceptual mode before and after instruction, for groups KOG and KEG, respectively. The estimator π corresponds to the aforementioned ATT, as it measures the additional effect of receiving instruction on chemical equilibrium for students in the treatment group (KEG) compared to the control group (KOG). One can see, that under these assumptions, fixed differences without intervention would lead to an ATT that equals 0, representing the stability of knowledge differences over time. Again, following Graves et al. (2022), we assess the statistical significance on the estimated ATT using permutation tests. Therefore, we implement a permutation-based inference procedure. The treatment condition (KOG, KEG) is randomly reassigned across students while preserving the original group sizes. For each permutation, the ATT estimator is recomputed, generating an empirical null distribution under the assumption of no treatment effect. The observed ATT is then compared to this distribution, and the proportion of permuted estimates that are equal to or exceed the observed ATT serves as the p-value. Since our hypotheses specify a directional effect, in most cases we compute one-sided p-values (see results for specifications). For a comprehensive description of the full ATT-procedure, interested readers are referred to Graves et al. (2022).

C3. Details on our ordinal regression analyses

Multicollinearity. To ensure statistical validity of our ordinal regression models, we first conducted a multicollinearity check for all predictors. We applied the Generalized Variance Inflation Factor (GVIF) for categorial variables and the traditional VIF for continuous predictors (Fox and Monette, 1992). The interpretation of VIF (or squared GVIF) values followed established threshold criteria, where values between 1 and 5 indicated mild but negligible multicollinearity, values between 5 and 10 suggested considerable multicollinearity that is noteworthy but not necessarily problematic and values exceeding 10 signaled severe multicollinearities, requiring intervention. In order to maintain a model's stability, we removed variables if their VIF/squared GVIF were substantially above 5 and suggested a strong redundancy with another predictor. In such cases, we identified the variable combination responsible for the collinearity and removed the less theoretically relevant predictor.

Model-building and selection. The model-building process started with a null model, which included only the intercept, and progressively incorporated relevant predictors in a predefined order: first, the last grade in chemistry was added to capture general prior knowledge, followed by learners' prior conceptual modes, representing specific prior knowledge. Finally, interest was included as an affective variable that is well known to be linked to academic achievement. At each step, we tested whether the inclusion of a new predictor significantly improved model fit by conducting likelihood-ratio tests, which conceptually resemble ANOVA as they assess model improvement by comparing changes in deviance. While a lower deviance generally indicates better model fit, improvements that are too small may not reach statistical significance, suggesting that an additional predictor does not substantially enhance explanatory power (Hosmer and Lemeshow, 2000). In this case the more parsimonious model was retained for the specific combination of main mode and reaction example.

Proportional Odds Assumption and Partial Proportional Odds Models. Following model selection, we assessed the proportional odds assumption for each model using the Brant test (Brant, 1990), which evaluates whether the assumption holds consistently across all levels of the ordinal dependent variable. In cases where the test indicated a potential violation, we estimated a partial proportional odds (PPO) model to examine whether the effects of the predictor in question varied substantially across cut points. If the PPO model revealed only minor deviations, with effects remaining in the same direction across thresholds, we followed Agresti (2002) in opting for the more parsimonious proportional odds model, as such small departures are often of limited practical relevance and may reflect sampling variability (in our case for example low frequencies in some outcome categories) rather than substantive differences. However, when the PPO model indicated substantial divergence in threshold-specific effects, suggesting meaningful differences how the respective predictor operates at different levels of the outcome, we retained the more flexible PPO model to adequately capture these variations (Christensen, 2019).

Statistical significance and Interpretation of Odd Ratios. Since our hypotheses were directional, we computed one-tailed p-values for all respective predictors. To enhance the

interpretability of our findings, we report Odds Ratios (ORs) alongside the regression coefficients. This allows for a more intuitive understanding of the magnitude of predictor effects in an ordinal regression model. For example, an OR of 2.5 for a continuous variable implies that for each unit increase in this predictor, the odds of being classified in a higher conceptual mode were 2.5 times greater, holding all other variables constant. For ordinal predictors, the OR reflected the odds of being in a higher outcome category compared to the respective reference category. ORs are particularly relevant when examining the report of all regression models provided in section E of this appendix.

D. Comparison of descriptive statistics of the predictor variable grade and subject-related interest (SRI) between the full and filtered (complete cases) dataset

Table A3: Descriptive summary of the predictors last chemistry grade and subject related interest, without restricting to complete cases.

Variable	Total (N)	Mean	SD	Min	Max
Grade	232	2.58	1.21	1.00	6.00
SRI	232	2.55	0.72	1.00	4.00

Of the 245 students from classes that participated in both, pre- and post-test, 232 provided data for both variables. The descriptive statistics indicate that the filtered dataset for ordinal regressions (N = 150) closely reflects the full sample, with only minimal differences in the mean value of grade (Δgrade = 0.06), suggesting no substantial bias toward selectively including only high-achieving learners.

E. Transition matrices

Reaction Start

Table A4: Pre- and post-distribution (marginals) and transitions between sub-modes from pre to post (transition-related and overall pre- and post-distribution percentages in parentheses) for the conceptual mode start of the reaction (MS).

Post / **Pre**	NOCONC	INIT	ATTRAC	COLSIMP	COLEXP	**Pre (total)**
COMBUSTION						
NOCONC	20 (23.3%)	37 (43.0%)	0 (0.0%)	10 (11.6%)	19 (22.1%)	**86 (47.0%)**
INIT	6 (6.4%)	37 (39.4%)	0 (0.0%)	11 (11.7%)	40 (42.5%)	**94 (51.4%)**
ATTRAC	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	**0 (0.0%)**
COLSIMP	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	3 (100.0%)	**3 (1.6%)**
COLEXP	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	**0 (0.0%)**
Post (total)	**26 (14.2%)**	**74 (40.4%)**	**0 (0.0%)**	**21 (11.5%)**	**62 (33.9%)**	**183 (100%)**
ESTERIFICATION						
NOCONC	62 (53.4%)	22 (19.0%)	8 (6.9%)	14 (12.1%)	10 (8.6%)	**116 (63.4%)**
INIT	14 (25.9%)	19 (35.1%)	1 (1.9%)	9 (16.7%)	11 (20.4%)	**54 (29.5%)**
ATTRAC	3 (23.1%)	2 (15.4%)	2 (15.4%)	2 (15.4%)	4 (30.7%)	**13 (7.1%)**
COLSIMP	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	**0 (0.0%)**
COLEXP	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	**0 (0.0%)**
Post (total)	**79 (43.1%)**	**43 (23.5%)**	**11 (6.0%)**	**25 (13.7%)**	**25 (13.7%)**	**183 (100%)**

Reaction Progress

Table A5: Pre- and post-distribution (marginals) and transitions between sub-modes from pre to post (transition-related and overall pre- and post-distribution percentages in parentheses) for the conceptual mode progress of the reaction (MP). All abbreviations refer to the sub-modes introduced in table 2.

Post / Pre	NOCONC	MACRO	BONDS	COLSIMP	COLEXP	**Pre (total)**
COMBUSTION						
NOCONC	27 (27.8%)	27 (27.8%)	9 (9.3%)	16 (16.5%)	18 (18.6%)	**97 (53.0%)**
MACRO	3 (9.4%)	6 (18.7%)	7 (21.9%)	8 (25.0%)	8 (25.0%)	**32 (17.5%)**
BONDS	4 (8.5%)	3 (6.4%)	7 (14.9%)	6 (12.8%)	27 (57.4%)	**47 (25.7%)**
COLSIMP	1 (33.3%)	1 (33.3%)	0 (0.0%)	1 (33.3%)	0 (0.0%)	**3 (1.6%)**
COLEXP	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (50.0%)	2 (50.0%)	**4 (2.2%)**
Post (total)	**35 (19.1%)**	**37 (20.2%)**	**23 (12.6%)**	**33 (18.0%)**	**55 (30.1%)**	**183 (100%)**
ESTERIFICATION						
NOCONC	45 (32.6%)	46 (33.3%)	22 (15.9%)	19 (13.8%)	6 (4.4%)	**138 (75.4%)**
MACRO	3 (23.1%)	5 (38.5%)	1 (7.7%)	3 (23.1%)	1 (7.7%)	**13 (7.1%)**
BONDS	5 (15.6%)	12 (37.5%)	5 (15.6%)	5 (15.6%)	5 (15.6%)	**32 (17.5%)**
COLSIMP	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	**0 (0.0%)**
COLEXP	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	**0 (0.0%)**
Post (total)	**53 (29.0%)**	**63 (34.3%)**	**28 (15.3%)**	**27 (14.8%)**	**12 (6.6%)**	**183 (100%)**

Reaction End

Table A6: Pre- and post-distribution (marginals) and transitions between sub-modes from pre to post (transition-related and overall pre- and post-distribution percentages in parentheses) for the conceptual mode end of the reaction (ME). Since one person in the full sample did not answer the last question, the sample size here is reduced to N = 182. All abbreviations refer to the sub-modes introduced in table 2.

Post / **Pre**	NOCONC	LIMIT	STATIC	DYNAMIC	**Pre (total)**
COMBUSTION					
NOCONC	40 (38.1%)	64 (61.0%)	1 (0.9%)	0 (0.0%)	**105 (57.8%)**
LIMIT	11 (14.7%)	58 (77.3%)	2 (2.7%)	4 (5.3%)	**75 (41.0%)**
STATIC	1 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	**1 (0.6%)**
DYNAMIC	0 (0.0%)	1 (100.0%)	0 (0.0%)	0 (0.0%)	**1 (0.6%)**
Post (total)	**52 (28.6%)**	**123 (67.6%)**	**3 (1.6%)**	**4 (2.2%)**	**182 (100%)**
ESTERIFICATION					
NOCONC	71 (48.3%)	37 (25.2%)	5 (3.4%)	34 (23.1%)	**147 (80.8%)**
LIMIT	3 (13.0%)	11 (47.8%)	2 (8.7%)	7 (30.5%)	**23 (12.6%)**
STATIC	0 (0.0%)	0 (0.0%)	1 (25.0%)	3 (75.0%)	**4 (2.2%)**
DYNAMIC	0 (0.0%)	0 (0.0%)	0 (0.0%)	8 (100.0%)	**8 (4.4%)**
Post (total)	**74 (40.6%)**	**48 (26.4%)**	**8 (4.4%)**	**52 (28.6%)**	**182 (100%)**

F. Full report of the tests for marginal homogeneity

Table A7: Overview of the results of the tests for marginal homogeneity.

	χ^2	df	p
Combustion			
Start (MS)	96.59	4	< .001
Progress (MP)	90.87	4	< .001
End (ME)	38.38	3	< .001
Esterification			
Start (MS)	52.72	4	< .001
Progress (MP)	90.42	4	< .001
End (ME)	73.13	3	< .001

G. Full report of ATT estimates

Table A8: Overview of ATT-estimators. The estimator values are interpreted for KEGG (whose has been provided extended with instructional options) in relation to KOG. A negative estimator indicates a lower probability of occupying the respective category, while a positive estimator suggests a higher proba

	Combustion Reaction					Esterification Reaction				
Start (MS)	**Sub-mode**	**ATT**	**P**	**Lower CI**	**Upper CI**	**Sub-mode**	**ATT**	**P**	**Lower CI**	**Upper CI**
	NOCONC	-0.032	.365	.359	.371	NOCONC	-0.008	.476	.472	.478
	INIT	-0,161	.068	.063	.073	INIT	-0.119	.105	.099	.111
	ATTRAC	-	-	-	-	ATTRAC	-0.030	.292	.285	.299
	COLSIMP	0.059	.135	.129	.141	COLSIMP	0.047	.192	.185	.199
	COLEXP	**0.137**	**.034**	**.030**	**.038**	**COLEXP**	**0.094**	**.038**	**.035**	**.042**
Progress (MP)	**Sub-mode**	**ATT**	**p**	**Lower CI**	**Upper CI**	**Sub-mode**	**ATT**	**p**	**Lower CI**	**Upper CI**
	NOCONC	-0.086	.195	.188	.202	NOCONC	-0.150	.061	.056	.067
	MACRO	0.123	.159	.149	.170	**MACRO**	**0.256**	**.001**	**< .001**	**.002**
	BONDS	-0,056	.265	-258	.272	**BONDS**	**-0.203**	**.007**	**.005**	**.009**
	COLCIMP	0.028	.334	.327	.340	COLCIMP	0.053	.169	.162	.175
	COLEXP	-0.001	.904	.895	.912	COLEXP	0.040	.153	.146	.159
End (ME)	**Sub-mode**	**ATT**	**p**	**Lower CI**	**Upper CI**	**Sub-mode**	**ATT**	**p**	**Lower CI**	**Upper CI**
	NOCONC	-0.137	.088	.082	.093	**NOCONC**	**-0.155**	**.054**	**.051**	**.059**
	LIMIT	0.133	.194	.183	.205	**LIMIT**	**-0.287**	**< .001**	**< .001**	**.001**
	STATIC	0.006	.822	.811	.832	**STATIC**	**-0.082**	**.018**	**.016**	**.021**
	DYNAMIC	-0.002	.867	.857	.876	**DYNAMIC**	**0.524**	**< .001**	**< .001**	**< .001**

H. Full report of all estimated models

MS Combustion

Table A9: Estimated coefficients and significance levels for the ordinal regression model for the start of the combustion reaction (MS Combustion).

Variable	Estimate	Std. Error	OR [95% one-sided CI]	OR [95% CI]	Wald χ^2
Variant (a) – NOCONC as reference category					
pre-INIT	**0.558 ***	**0.333**	**1.747 [1.010; ∞]**	**-**	**2.809**
Group	0.348 (n.s.)	0.340	1.417 [0.810; ∞]	-	1.046
Grade	**-0.742 *****	**0.160**	**0.476 [0; 0.619]**	**-**	**21.492**
Threshold Coefficients					
0|1	-3.760	0.623	-	-	36.350
1|3	-1.265	0.527	-	-	5.753
3|4	-0.679	0.519	-	-	1.701

As pre-ATTRAC contained no observations in the pre-condition and pre-COLSIMP had too few observations to yield reliable estimates, only pre-INIT was included as pre-sub-mode in the model. As only one sub-mode of the ordinally structured predictor (pre-INIT) is included in the model, variant (a) with NOCONC as the reference category and variant (b) with contrast coding yield identical results. Therefore, we only report variant (a). The inclusion of SRI did not enhance model fit, leading to its exclusion as an explanatory variable in the final reported model. Interaction terms were deemed unnecessary based on a pre-check of separate regressions for the two groups (KOG, KEG), which revealed no differences in the strength or direction of the predictor effects. Model fit indices were as follows: log-likelihood (LLH) = -165.48, null model log-likelihood (LLH_{Null}) = -186.42, and likelihood-ratio test $G^2 = 41.87$ ($p < .001$). Nagelkerke's pseudo-R^2 was 0.266. Significance levels: *** $p \leq .001$, ** $p \leq .01$, * $p \leq .05$.

MS Esterification

Table A10: Estimated coefficients and significance levels for the ordinal regression model for the start of the esterification reaction (MS Esterification).

Variable	Estimate	Std. Error	OR [95% one-sided CI]	OR [95% CI]	Wald χ^2
Variant (a) – NOCONC as reference category					
pre-INIT	**0.906 ****	**0.351**	**2.476 [1.391; ∞]**	**-**	**6.682**
pre-ATTRAC	**1.995 *****	**0.626**	**7.351 [2.627; ∞]**	**-**	**10.167**
SRI	**0.408 ***	**0.238**	**1.504 [1.016; ∞]**	**-**	**2.939**
Group	0.408 (n.s.)	0.367	1.504 [0.821; ∞]	-	1.231
Grade	**-0.391 ***	**0.183**	**0.676 [0; 0.913]**	**-**	**4.589**
SRI x Group	-0.595 (n.s.)	0.356	-	0.552 [0.275; 1.108]	2.791

Threshold Coefficients					
0|1	-0.809	0.504	-	-	2.572
1|2	0.249	0.502	-	-	0.243
2|3	0.559	0.506	-	-	1.219
3|4	1.457	0.525	-	-	7.590
Variant (b) – Sequential contrast coding					
pre-INIT	**0.906 ****	**0.351**	**2.476 [1.391; ∞]**	**-**	**6.682**
pre-ATTRAC	**1.088 ***	**0.650**	**2.969 [1.019; ∞]**	**-**	**2.799**
SRI	**0.408 ***	**0.238**	**1.504 [1.016; ∞]**	**-**	**2.939**
Group	0.408 (n.s.)	0.367	1.504 [0.821; ∞]	-	1.231
Grade	**-0.391 ***	**0.183**	**0.676 [0; 0.913]**	**-**	**4.589**
SRI x Group	-0.594 (n.s.)	0.355	-	0.552 [0.275; 1.108]	**2.791**
Threshold Coefficients					
0|1	-1.776	0.533	-	-	11.084
1|2	-0.718	0.517	-	-	1.928
2|3	-0.408	0.517	-	-	0.624
3|4	0.490	0.526	-	-	0.864

Model fit indices were as follows: log-likelihood (LLH) = -196.63, null model log-likelihood (LLH_{Null}) = -213.07, and likelihood-ratio test $G^2 = 32.86$ ($p < .001$). Nagelkerke's pseudo-R^2 was 0.209. Except for the interaction term, we have theory-driven directional assumptions for predictors, so one-sided CIs are reported. Significance levels: *** $p \leq .001$, ** $p \leq .01$, * $p \leq .05$.

MP Combustion

Table A11: Estimated coefficients and significance levels for the ordinal regression model for the progress of the combustion reaction (MP Combustion).

Variable	Estimate	Std. Error	OR [95% one-sided CI]	OR [95% CI]	Wald χ^2
Variant (a) – NOCONC as reference category					
pre-MACRO	0.580 (n.s)	0.435	1.786 [0.845; ∞]	-	1.623
pre-BONDS	**1.119 ****	**0.519**	**3.303 [1.406; ∞]**	**-**	**5.295**
Grade	**-0.428 *****	**0.152**	**0.652 [0; 0.838]**	**-**	**7.885**
pre-MARCO x Group	-0.523 (n.s.)	0.832	0.592 [0; 2.330]	-	0.396
pre-BONDS x Group	1.305 (n.s.)	0.883	3.687 [0.862; ∞]	-	2.181
Threshold Coefficients					
0|1	-1.932	0.535	-	-	13.068
1|2	-1.133	0.523	-	-	4.687
2|3	-0.594	0.524	-	-	1.284
3|4	0.243	0.531	-	-	0.209
0|1 Group	-0.803 (n.s.)	0.601	-	-	1.787
1|2 Group	0.229 (n.s.)	0.522	-	-	0.192
2|3 Group	0.419 (n.s.)	0.546	-	-	0.588
3|4 Group	0.706 (n.s.)	0.627	-	-	1.271
Variant (b) – Sequential contrast coding					
pre-MACRO	0.580 (n.s)	0.435	1.786 [0.845; ∞]	-	1.623
pre-BONDS	0.615 (n.s.)	0.527	1.849 [0.721; ∞]	-	1.153
Grade	**-0.428 *****	**0.152**	**0.652 [0; 0.838]**	**-**	**7.885**
pre-MARCO x Group	-0.523 (n.s.)	0.832	0.592 [0; 2.330]	-	0.396
pre-BONDS x Group	**1.828 ***	**1.000**	**6.224 [1.200; ∞]**	**-**	**3.338**
Threshold Coefficients					
0|1	-2.524	0.476	-	-	28.111
1|2	-1.725	0.451	-	-	14.608

2|3	-1.186	0.443	-	-	7.161
3|4	-0.349	0.439	-	-	0.632
0|1 Group	**-1.063 ***	**0.593**	**-**	**-**	**3.222**
1|2 Group	-0.032 (n.s.)	0.435	-	-	0.005
2|3 Group	0.159 (n.s.)	0.420	-	-	0.143
3|4 Group	0.445 (n.s.)	0.462	-	-	0.927

As pre-COLSIMP and pre-COLEXP showed only a few observations and were only present in the KEG subgroup, only pre-MACRO and pre-BONDS were included in the model as pre-sub-modes. The inclusion of SRI did not enhance model fit, leading to its exclusion as an explanatory variable in the final reported model. The variable Group violated a test of the Proportional Odds Assumption (POA) and also demonstrated that the effects at the individual thresholds indeed diverged substantially in a check of effects using a Partial Proportional Odds (PPO) model. Therefore, in this case, the PPO model was retained, in which the POA assumption for Group is relaxed. Accordingly, no global estimate for Group is reported, but rather threshold-specific estimates. These threshold-specific estimates should not be interpreted like global regression coefficients but rather as shifts of the respective threshold. A negative estimate corresponds to a downward shift of the threshold (which, in turn, facilitates the transition from the lower to the higher category). A positive estimate should be interpreted in the opposite way (Christensen, 2019). For example, it is shown that Group = 1 (which corresponds to KEG) significantly facilitates the transition from the lowest category (threshold 0|1). For the remaining thresholds, this effect reverses but does not reach significance in any of the cases. Furthermore, pre-checks of separate regressions for both groups (KEG, KOG) showed that the included pre-sub-modes pre-MARCO and pre-BONDS have substantially different effects in the groups. Therefore, we introduced an interaction term, which – due to the ordinal structure of the predictor pre – results in two interaction estimates. Previous analytical steps indicated that pre-MACRO is particularly unfavorable, while pre-BONDS is particularly favorable for development in KEG. Therefore, we also computed one-sided confidence intervals for the interaction terms at this stage. Regarding these latter, for example the odds ratio for pre-BONDS x Group indicates a positive effect in favor of Group = 1 (KEG). However, the 95% confidence interval is relatively wide, suggesting some uncertainty in the estimate, likely due to a rather small but still sufficient number of observations in this category. Model fit indices were as follows: log-likelihood (LLH) = -207.69, null model log-likelihood (LLH_{Null}) = -226.36, and likelihood-ratio test G^2 = 37.33 ($p < .001$). Nagelkerke's pseudo-R^2 was 0.236. Significance levels: *** $p \leq .001$, ** $p \leq .01$, * $p \leq .05$.

MP Esterification

Table A12: Estimated coefficients and significance levels for the ordinal regression model for the progress of the esterification reaction (MP Esterification).

Variable	Estimate	Std. Error	OR [95% one-sided CI]	OR [95% CI]	Wald χ^2
Variant (a) – NOCONC as reference category					
pre-MACRO	0.609 (n.s.)	0.602	1.839 [0.683; ∞]	-	1.022
pre-BONDS	0.032 (n.s.)	0.584	1.033 [0.395; ∞]	-	0.003
SRI	**0.388 ***	**0.205**	**1.473 [1.051; ∞]**	**-**	**3.561**
Grade	-0.239 (n.s.)	0.163	0.787 [0; 1.031]	-	2.134
pre-MACRO x Group	**-2.770 ***	**1.551**	**0.063 [0; 0.804]**	**-**	**3.186**
pre-BONDS x Group	1.130 (n.s.)	0.836	3.097 [0.783; ∞]	-	1.831
Threshold Coefficients					
0|1	-1.147	0.475	-	-	5.827
1|2	0.026	0.470	-	-	0.003
2|3	1.058	0.504	-	-	4.418
3|4	2.374	0.626	-	-	14.379
0|1 Group	-0.808 (n.s.)	0.507	-	-	2.538
1|2 Group	0.637 (n.s.)	0.476	-	-	1.790
2|3 Group	0.003 (n.s.)	0.520	-	-	0.0004
3|4 Group	0.038 (n.s.)	0.742	-	-	0.003
Variant (b) – Sequential contrast coding					
pre-MACRO	0.609 (n.s.)	0.602	1.839 [0.683; ∞]	-	1.022
pre-BONDS	-0.578 (n.s.)	0.791	-	0.562 [0.119; 2.648]	0.531
SRI	**0.388 ***	**0.205**	**1.473 [1.051; ∞]**	**-**	**3.561**
Grade	-0.239 (n.s.)	0.163	0.787 [∞; 1.031]	-	2.134
pre-MACRO x Group	**-2.770 ***	**1.551**	**0.063 [∞; 0.804]**	**-**	**3.186**
pre-BONDS x Group	**3.901**	**1.703**	**49.437 [3.000; ∞]**	**-**	**5.244**
Threshold Coefficients					

0|1	-1.361	0.495	-		7.573
1|2	-0.188	0.487	-		0.150
2|3	0.845	0.517	-	-	2.667
3|4	2.160	0.633	-	-	11.635
0|1 Group	-0.262	0.647	-	-	0.163
1|2 Group	**1.184 ***	**0.643**	**-**	**-**	**3.389**
2|3 Group	0.550	0.672	-	-	0.667
3|4 Group	0.584	0.849	-	-	0.446

As pre-COLSIMP and pre-COLEXP showed only a few observations and were only present in the KEG subgroup, only pre-MACRO and pre-BONDS were included in the model as pre-sub-modes. The variable Group violated a test of the Proportional Odds Assumption (POA) and also demonstrated that the effects at the individual thresholds indeed diverged substantially in a check of effects using a Partial Proportional Odds (PPO) model. Therefore, in this case, the PPO model was retained, in which the POA assumption for Group is relaxed. Accordingly, no global estimate for Group is reported, but rather threshold-specific estimates. Furthermore, pre-checks of separate regressions for both groups (KEG, KOG) showed that the included pre-sub-modes pre-MARCO and pre-BONDS have substantially different effects in the groups. Therefore, we introduced an interaction term, which – due to the ordinal structure of the predictor pre – results in two interaction estimates. Previous analytical steps indicated that pre-MACRO is particularly unfavorable, while pre-BONDS is particularly favorable for development in KEG. Therefore, we also computed one-sided confidence intervals for the interaction terms at this stage. However, in variant (b), pre-Bonds x Group exhibits an extremely large effect, drive by the combined influence of the more or less “harmful” neighboring pre-sub-mode pre-MARCRO (which seems to be very unfavorable in KEG) and the favoring effect of pre-BONDS. Given the relatively small number of cases in both pre-sub-modes, this effect should nevertheless be interpreted with caution. Model fit indices were as follows: log-likelihood (LLH) = -202.09, null model log-likelihood (LLH_{Null}) = -211.52, and likelihood-ratio test $G^2 = 18.86$ ($p < .05$). Nagelkerke’s pseudo-R^2 was 0.126. Significance levels: *** $p \leq .001$, ** $p \leq .01$, * $p \leq .05$.

ME Combustion

Table A13: Estimated coefficients and significance levels for the ordinal regression model for the end of the combustion reaction (ME Combustion).

Variable	Estimate	Std. Error	OR [95% one-sided CI]	OR [95% CI]	Wald χ^2
Variant (a) – NOCONC as reference category					
Group	**0.653 ***	**0.326**	**1.922 [1.124; ∞]**	**-**	**4.016**
Grade	**-0.300 ***	**0.138**	**0.740 [0; 0.929]**	**-**	**4.722**
Threshold Coefficients					
0|1	-1.530	0.445	-	-	11.842
1|2	-1.318	0.441	-	-	8.842

Neither the introduction of specific prior knowledge in the form of pre-sub-modes nor the SRI improved the model, so neither of these predictors was included in the final model. In this case, comparative contrast coding of the pre-sub-modes as an ordinal predictor is unnecessary, so only variant (a) is reported. Model fit indices were as follows: log-likelihood (LLH) = -106.40, null model log-likelihood (LLH_{Null}) = -110.07, and likelihood-ratio test $G^2 = 7.33$ ($p < .05$). Nagelkerke's pseudo-R^2 was 0.062. Significance levels: *** $p \leq .001$, ** $p \leq .01$, * $p \leq .05$.

ME Esterification

Table A14: Estimated coefficients and significance levels for the ordinal regression model for the end of the esterification reaction (ME Esterification).

Variable	Estimate	Std. Error	OR [95% one-sided CI]	OR [95% CI]	Wald χ^2
Variant (a) – NOCONC as reference category					
Group	**3.020**	**0.480**	**20.499 [9.301; ∞]**	**-**	**39.540**
Grade	**-0.854**	**0.209**	**0.426 [0; 0.601]**	**-**	**16.654**
Threshold Coefficients					
0|1	-1.514	0.495	-	-	9.370
1|2	0.255	0.498	-	-	0.263
2|3	0.586	0.511	-	-	1.316
0|1 pre-LIMIT	**-1.718 ***	**0.801**	**-**	**-**	**4.384**
1|2 pre-LIMIT	-0.381	0.638	-	-	0.356
2|3 pre-LIMIT	0.090	0.727	-	-	0.016

As only one sub-mode of the ordinally structured predictor (pre-LIMIT) is included in the model, variant (a) with NOCONC as the reference category and variant (b) with contrast coding yield identical results. Therefore, we only report variant (a). The inclusion of SRI did not enhance model fit, leading to its exclusion as an explanatory variable in the final reported model. The variable pre-LIMIT violated a test of the Proportional Odds Assumption (POA) and also demonstrated that the effects at the individual thresholds indeed diverged substantially in a check of effects using a Partial Proportional Odds (PPO) model. Therefore, in this case, the PPO model was retained, in which the POA assumption for pre-LIMIT is relaxed. Accordingly, no global estimate for pre-LIMIT is reported, but rather threshold-specific estimates. Model fit indices were as follows: log-likelihood (LLH) = -156.88, null model log-likelihood (LLH_{Null}) = -217.82, and likelihood-ratio test G^2 = 121.90 ($p < .001$). Nagelkerke's pseudo-R^2 was 0.531. Significance levels: *** $p \leq .001$, ** $p \leq .01$, * $p \leq .05$.

I. Developmental statements towards a learning progression

(I) Students explain the composition of matter and the structures that build it by describing the interaction of particles and emergent characteristics based on these interactions. Building on this, they explain that these interactions and thus the course of chemical reactions can be measured and influenced.

(II) On the basis of a permanent movement of particles, students describe the start of reactions by referring to random collisions of the reactants.

(III) Students explain how random collisions in the further course of reactions (i.e., progress of a reaction) can lead to bond breaking, which in turn leads to the reorganization of atoms and the formation of new bonds.

(IV) Regarding (II) and (III), students distinguish between different types of particle collisions (effective vs. ineffective) and explain how kinetic energy and environmental or even structural factors may influence the effectiveness of collisions.

(V) Students describe the end of the reaction based on systemic conditions and explain the state of dynamic equilibrium.

(VI) Overall, students explain how the time course of a reaction is influenced by applying knowledge listed under (I) to (V).

J. Integration of macro- and micro-adaptions in a (digital) learning environment

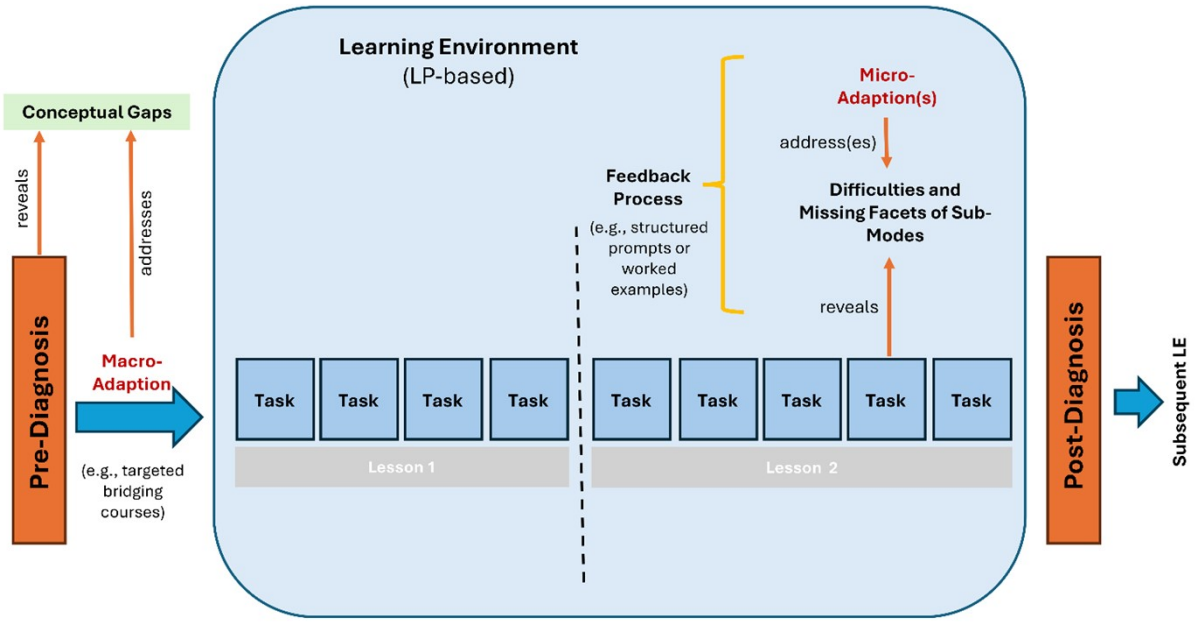

Figure A2: Diagnostic opportunities that can be utilized particularly in digitally supported learning environments, as well as the resulting opportunities for macro- and micro-level adaptations.

K. Methodological literature only cited in the supporting information

Agresti, A., 2002. Categorical Data Analysis, 1st ed, Wiley Series in Probability and Statistics. Wiley. https://doi.org/10.1002/0471249688

Brant, R., 1990. Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression. Biometrics 46, 1171. https://doi.org/10.2307/2532457

Fox, J., Monette, G., 1992. Generalized Collinearity Diagnostics. Journal of the American Statistical Association 87, 178–183. https://doi.org/10.1080/01621459.1992.10475190