Understanding Homeostatic Regulation: The Role of Relationships and Conditions in Feedback Loop Reasoning

Andrea Kiesewetter⬡* **and Philipp Schmiemann**⬡

Faculty of Biology, Biology Education Research and Learning Lab (BERLL), University of Duisburg-Essen (UDE), 45141 Essen, Germany

ABSTRACT

Understanding homeostasis is a goal of biology education curricula, as homeostasis is a core feature of living systems. Identifying and understanding the underlying molecular feedback mechanisms appear to be challenging for students. Understanding the properties and mechanisms of such complex homeostatic systems requires feedback loop reasoning, which is a part of systems thinking. Novices seem to struggle to 1) consider more than one initiating condition in cause–effect relationships and 2) track cause and effect across a sequence of processes. In this cross-sectional study, we analyzed how these factors impede feedback loop reasoning. High school and undergraduate students analyzed the organizational, behavioral, and modeling-related features of a homeostatic system (blood calcium regulation). Using multidimensional item response theory, we were able to confirm the three-dimensional structure of the theoretical systems-thinking model and to identify the factors causing item difficulty. As hypothesized, indirect relationships and derived inverse conditions are challenging factors for participants in the context of homeostasis across dimensions. Hence, we recommend paying special attention to these factors when teaching homeostasis as part of systems thinking. We assume that allowing students to reason from different initiating conditions in a learning setting may improve their systems-thinking skills.

INTRODUCTION

Homeostatic regulation is a key element of life science education that fosters higher-order thinking (Cary and Branchaw, 2017). In the Next Generation Science Standards for high school, homeostasis is embedded in the disciplinary core idea "from molecules to organisms: structures and processes" (NGSS Lead States, 2013). Furthermore, for undergraduate biology education, homeostatic regulation is crucial to various core concepts, including "information flow, exchange and storage" and "pathways and transformations of energy and matter" (American Association for the Advancement of Science, 2011). Finally, an understanding of homeostatic regulation is central, both as an expectation upon entering medical school and as a competency at the end of medical school (Association of American Medical Colleges and Howard Hughes Medical Institute, 2009). Previous studies on understanding homeostasis have shown that it is challenging to consider and relate elements at different levels of organization, particularly at the molecular level (Hmelo-Silver *et al.*, 2007; Ben Zvi Assaraf *et al.*, 2013). Students tend to describe structures of physiological systems at higher levels of organization (e.g., tissues and organs) and mechanisms of physiological systems at lower levels of organizations (e.g., molecules; Lira and Gardner, 2017). Linking the overall phenomenon of maintenance of homeostasis to the underlying mechanisms at the molecular level appears to be a key challenge (Snapir *et al.*, 2017; Tripto *et al.*, 2018). To gain further insights into this research field, it will be useful to apply another systems-thinking approach to this biological phenomenon.

Feedback loop reasoning is the ability to recognize, analyze, and model the structures, mechanisms, and functions of systems that are characterized by feedback. Consequently, this ability is seen as part of systems thinking (Wellmanns and Schmiemann, 2020). Feedback loops are mechanisms in which the effects of an emerging change in system states in turn affect the actual states (Sterman, 2000; Camazine *et al.*, 2003). Homeostatic regulation is a characteristic example of a biological phenomenon that consists of negative feedback loop mechanisms. As a result of a qualitative change in a regulated variable (e.g., increased blood glucose levels), processes are initiated (e.g., insulin secretion) that counteract the primary change (e.g., decreased blood glucose levels; Gatewood *et al.*, 1970). When attempting to assess feedback loop reasoning, it has been found that novice students do not refer to feedback mechanisms in their explanations (Batzri *et al.*, 2015). Instead, they refer to only one of the reciprocal relationships in a feedback loop, called open-loop causality (Booth Sweeney and Sterman, 2007). Novices explain that increased blood glucose levels cause increased insulin secretion. Thereby, they omit the feedback effect on blood glucose levels and outline an open loop. This tendency suggests that understanding feedback loop mechanisms is a particularly challenging task.

To explore these challenges in detail, this study uses a literature review to identify two potential factors that impede feedback loop reasoning in homeostatic regulation. The factors are the requirements to use more than one initiating condition of causal relationships (Chi *et al.*, 2012; Cho and Jonassen, 2012) and to apply the causal effects across more than one relationship (Mambrey *et al.*, 2020). To illustrate the two factors, one can consider the example of blood glucose levels. We may observe the effects of an increased insulin secretion but also the effects of a decreased insulin secretion as an initiating condition. The initiating condition that can be read directly from a representation (e.g., increased insulin secretion), we further call the "obvious condition"; the condition that can be inferred inversely (e.g., decreased insulin secretion), we call the derived "inverse condition." Moreover, on the one hand, the effects under investigation may be proximal and temporal (e.g., altered insulin secretion affects glucose uptake into cells), which we refer to as "direct relationships." On the other hand the effects may also be more distant (e.g., increased insulin secretion affects the blood glucose levels and thus insulin secretion in the long term) which we refer to as "indirect relationships." Thus far, there is missing evidence in systems-thinking literature regarding the extent to which these factors contribute to the requirements of systems-thinking tasks in homeostatic regulation. To operationalize feedback loop reasoning in this context, we apply a validated cross-contextual model of systems thinking (Mambrey *et al.*, 2020). We propose that understanding the more distant indirect cause–effect relationships (e.g., effects of altered insulin secretion on the insulin secretion itself) and the consideration of derived inverse initiating conditions (e.g., effects of decreased insulin secretion) is more difficult than understanding direct relationships and obvious conditions. Accordingly, we vary these factors in our design to compare and thus explain the requirements of homeostatic systems-thinking tasks.

THEORETICAL BACKGROUND

Homeostatic Systems

Homeostatic systems are characterized by the interactions of molecular processes to preserve the physical integrity of biological organisms (Camazine *et al.*, 2003). Maintaining physical integrity is defined as the primary function of a homeostatic system. The holistic view of systems theory does not focus on individual molecules and processes, but on the interplay of the organization of and the interaction between elements and processes that constitute the respective system (Boogerd *et al.*, 2007). Homeostatic systems are open systems that exchange substances, information, and energy with the environment. Complex nonlinear mechanisms have evolved to counteract external fluctuations and allow constant internal states (von Bertalanffy, 1973). Feedback loop mechanisms enable the system to respond to a wide array of factors that systems might experience from the outside. If disturbances from the external environment occur, processes are initiated that counteract the effect of the disturbance so that defined limits are not exceeded.

Systems-Thinking Skills

To operationalize the requirements of a system understanding of homeostasis and to promote it in the classroom, researchers and educators need knowledge about underlying skills. Many approaches have contributed to the conceptualization of general systems thinking, such as the structure–behavior–function theory (Hmelo-Silver *et al.*, 2007), and the systems-thinking hierarchical model (Ben Zvi Assaraf and Orion, 2005; Ben Zvi Assaraf *et al.*, 2013). These studies mapped and classified student behavior in educational settings. Accordingly, these models were empirically derived; however, verification of the structural design is mostly lacking. Based on a tested systems-thinking model from geography education (Mehren *et al.*, 2018), Mambrey *et al.* (2020) postulated a three-dimensional (3D) model consisting of the skills to identify system organization (SO) and to analyze the system behavior (SB) and system modeling (SM; see Table 1). The authors were able to validate the 3D structure of this systems-thinking model in ecology. According to Mambrey *et al.* (2020), SO is about identifying elements and relationships. The elements in homeostatic systems are regulated variables, glands and tissues. The relationships are usually the flow of matter and information and cause–effect relationships. SB addresses the analysis of dynamic developments within the system (Mehren *et al.*, 2018; Mambrey *et al.*, 2020). Analysis of dynamic developments includes the description of consequences triggered by changes in environmental or system-internal states (e.g., analysis of the consequences following a change in the regulated variable). Finally, SM describes the skill to weigh intentional actions and develop prognoses within the system (Mambrey *et al.*, 2020). A crucial part of this skill is the consideration of possible interventions to achieve specific target states (e.g., weighing measures that lead to an intended change in the regulated variable). The authors aimed to define a systems-thinking model that determines these requirements across different content areas. Thus far, this conceptualization has not been tested to determine whether it also reflects systems-thinking skills in homeostatic systems (i.e., feedback loop reasoning).

TABLE 1. Systems-thinking model, with questions in the context of calcium homeostasis (based on Mambrey et al., 2020)

	System organization	System behavior	System modeling[a]
Skill description	Students identify elements, processes, relationships and structures of complex systems.	Students describe system dynamic developments by analyzing the cause–effect relationships.	Students weigh actions to trigger the intended systems' states by analyzing the cause–effect relationships.
Overarching questions	Which elements affect blood calcium levels and how are they related?	How does the blood calcitonin level develop after calcium intake?	How or by which actions is an increased release of calcitonin triggered?

[a]In the context of our study, we refer to this dimension as regulatory measures (RM).

Requirements for Understanding of Homeostatic Regulation

Obstacles to building an understanding of homeostasis can arise from different perspectives. Students' struggle to grasp content-related aspects of homeostasis results in common misconceptions (Modell et al., 2015). The latest research indicates that knowing how the control center integrates incoming sensory information is particularly challenging (McFarland et al., 2017). In addition, system-related aspects are relevant to build an understanding of homeostasis. According to Lira and Gardner (2017), grasping the physiological concept of homeostasis requires explaining mechanisms and predicting future behavior, in other words, systems-thinking skills. Addressing mechanisms is a critical component; when teachers fail to present mechanisms in direct relation to structure and function, student learning is impeded (Lira and Gardner, 2017). To understand homeostatic mechanisms, it is necessary, albeit difficult, to analyze dynamic relationships (Ben Zvi Assaraf et al., 2013). Recognizing such relationships involves the description and analysis of changes in the states of elements and processes, as well as their effects in interaction with other elements and processes (Tripto et al., 2018). Dynamic relationships arise, for example, between water content in the blood, the amount of the antidiuretic hormone (ADH), and the reabsorption rate in the kidney. As a result of a low water content in the blood, the secretion of ADH from the pituitary gland increases, which causes the kidney to reabsorb more water, increasing the water volume in the blood (Freeman et al., 2017). In this example, there is a direct relationship between the change, "altered water volume in blood," and the consequence, "altered ADH secretion," as well as an indirect relationship between the change of state, "altered volume in blood," and the consequence, "altered water reabsorption," which is mediated by the changed ADH secretion. Students struggle to recognize the more distant, indirect consequences of changes; these effects can occur along a spatial and temporal spectrum, as demonstrated in ecological systems (Mambrey et al., 2022). When seeking a possible trigger for a given change in a system, students often consider only one triggering event, but it is usually necessary to consider a variety of processes that contribute to a particular event (Gilissen et al., 2020). For example, students attribute changes in blood glucose levels to changes in food intake, but not to the effect of the regulating hormones, insulin and glucagon (Gilissen et al., 2020; Wellmanns and Schmiemann, 2020). These previous findings contribute to the assumption that it is difficult to identify and apply indirect relationships.

Another challenge is to describe the different conditions related to the causal mechanisms. To describe the cause–effect relationship between two given elements or processes qualitatively, an analysis can start with two different initiating conditions (Chi et al., 2012). First, it is possible to describe the consequences of an increase in an element or process. To illustrate distinct initiating conditions, the blood volume example is used. "Increased ADH secretion" (i.e., increase in a process) leads to "increased water reabsorption" and "increased water volume in blood." Second, conversely, it is possible to describe the consequences of decreased elements or processes. "Decreased ADH secretion" (i.e., decrease in a process) leads to "decreased water reabsorption" and "decreased water volume in blood." Both explanations refer to the same sequence of cause–effect relationships. However, the initiating condition is different. One of these two initiating conditions is usually explicitly described in a text or representation of homeostatic mechanisms. We define this initiating condition as an obvious condition (in our example: increased ADH secretion) and the opposite initiating condition as a derived inverse condition (decreased ADH secretion). Cho and Jonassen (2012) found that most students explained the mechanisms in a homeostatic system by referring to only one of the two initiating conditions and not both. Accordingly, students concluded that either increased water intake leads to increased amount of urine or that decreased water intake leads to decreased amount of urine. Therefore, we assume that another challenge is the application of derived inverse conditions.

The factors we have listed may be used to explain what impedes students' reasoning surrounding homeostatic systems. We hypothesize that the need to recognize and analyze indirect cause–effect relationships ("relationships"; Mambrey et al., 2020) and the need to analyze derived inverse initiating conditions of cause–effect relationships ("conditions"; Cho and Jonassen, 2012) are factors that explain the increasing requirements of systems-thinking tasks. Thus far, little has been reported about the extent to which these factors impede students' reasoning in homeostatic regulation. To address this gap, we investigated the following research question: How do the theoretically assumed factors relationships and conditions of cause–effect relationships influence students reasoning in homeostatic regulation?

METHODS
Participants
To address our research question, we collected data from two different samples. We recruited students from high school biology courses (n = 77) and undergraduate students from an introductory university biology course (n = 136). We excluded three undergraduate students from further data analysis, because they did not respond to large parts of the assessment, which would have led to their results being systematically overestimated. This restriction resulted in a final sample of

210 students. The participating high school students averaged 15.8 (SD = 0.6) years of age, and the participating undergraduate students averaged 20.9 (SD = 3.6) years of age. As some students were minors, we informed teachers and parents in advance about the aim and procedure of the study according to the privacy regulations in our country. We emphasized that participation was voluntary. Parents of underage students provided written consent for students to participate in the study.

Test Design

Context and Instruction. Our purpose was to capture students' feedback loop reasoning related to a homeostatic system. Blood calcium control is an example of homeostatic regulation in the human body and emerges from the interplay of molecular feedback loop processes. We considered this context appropriate, because it had not been studied in our participants' previous biology classes. We deliberately excluded blood glucose control as a well-known example of homeostatic regulation, as we did not want to collect students' prior knowledge but rather their ability to use system representations and reason based on them. We consider our participants to be novices in terms of their knowledge of blood calcium control. Similar to other homeostatic processes, calcium control reveals the following mechanisms: The blood calcium level, as the regulated variable, is maintained within a certain range. Any change in calcium levels trigger processes that aim to oppose given changes. Sensors in the thyroid and parathyroid glands measure blood calcium levels and cause an adjustment in the secretion of the signaling molecules calcitonin and parathyrin. These processes result in 1) altered reabsorption of calcium from the tubular fluid into the blood, 2) altered calcium bone resorption, or building of bone, and finally, 3) altered calcium absorption into the blood (Modell *et al.*, 2015). These processes represent the respective effector responses, which in turn affect the regulated variable.

Because participants were presumed to have little prior content knowledge about calcium control in the human body, we initially provided them with an informational text explaining the relevant elements, processes, and functions of calcium regulation. We deliberately did not explain mechanisms regarding their underlying cause–effect relationships in the informational text (e.g., high blood calcium levels lead to increased secretion of calcitonin from thyroid cells), as we wanted participants to base their causal reasoning on insights gained from reading a representation. To illustrate the cause–effect relationships of calcium control in the human body, we used a representation showing 1) what sequence of processes is triggered when the calcium level deviates upward or downward, and 2) what effect this sequence of processes subsequently has on the calcium level as regulated variable. Such a representation can be used to promote understanding of the structure, relationships, and functions of a given system (Verhoeff *et al.*, 2018; Wilson *et al.*, 2020). We used a representation to avoid participants not showing feedback loop reasoning patterns at all because they did not remember underlying structures and relationships (Scott *et al.*, 2018). Furthermore, model-based reasoning is also a way of assessing students' ability to understand a homeostatic phenomenon's organization and temporal dynamics (Jansen *et al.*, 2019). Previous studies on model-based reasoning have shown that novices differ from experts, in that they address the

surface features of the models instead of underlying relationships, functions, and principles; in addition, novices tend to see models as static and fixed and overlook the possibility of applying these models as dynamic tools (Quillin and Thomas, 2015). Therefore, we assume that it is challenging for novices to use representations of systems models, although this capability is relevant for building more elaborate systems-thinking skills (Eilam and Poyas, 2010; NGSS Lead States, 2013). Informational text and representation of blood calcium control are provided in the Supplemental Material.

Conceptualization. To investigate feedback loop reasoning in a homeostatic context, we developed a test instrument characterized by the two hypothesized factors relationships and conditions and the three dimensions of systems thinking (Mambrey *et al.*, 2020; see Table 1). For the first dimension, SO, we identified and assessed students' understanding of the structure of a feedback loop mechanism. Students were asked to decide whether there was a relationship between two processes or elements. The second dimension, SB, includes the description and analysis of the consequences of changes in states. Students were asked to describe the effect of a certain change in one state on a second element or process in the system. The third dimension, SM, involves intentional selection of regulatory measures (RM) to achieve defined target states in homeostatic systems. RM addresses the evaluation of possible interventions to achieve the intended states. We gave students a target state of one element or process in the homeostatic system, and they had to choose a suitable measure to achieve this state. Both dimensions, SB and RM, relate to the analysis of system dynamics along a temporal axis. However, the dimensions differ with the overarching questions' directions (cf. Wellmanns and Schmiemann, 2020). Considering the relationship between the hormone calcitonin and the blood calcium level, an SB question could be "What consequences does an increased calcitonin level have on blood calcium levels?," whereas an RM question could be "What measures can be taken to lower blood calcium levels?" Essentially, both questions concern the same context, but are different types.

In the test instrument, we characterized direct relationships as being those in which cause and effect as elements are directly linked to each other via an arrow in the direction of action. In contrast, we defined indirect relationships between elements as those in which a chain of at least two consecutive arrows links cause and effect in the direction of action. Furthermore, we distinguish between obvious and derived inverse conditions. We define obvious conditions as the initiating conditions, which are explicitly presented in the given diagram, for example, increased blood calcium levels lead to an increased secretion of calcitonin (the diagram referred to can be found in the Supplemental Material). In contrast, we identify derived inverse conditions as the opposite initiating conditions that are not explicitly presented, but must be inferred, for example, decreased blood calcium levels lead to decreased secretion of calcitonin.

Assessment. To examine students' reasoning, we used 30 items with three different multiple-choice formats. All items were developed by science educators based on findings from a preliminary think-aloud study in the context of blood glucose regulation (Wellmanns and Schmiemann, 2020). All items are provided in the Supplemental Material. The formats were

TABLE 2. Number of items in SO, SB, and RM by relationships (direct vs. indirect) and conditions (obvious vs. inverse)

	Relationships	
Conditions	**Direct**	**Indirect**
SO		
Obvious	1	5
Inverse	0	0
SB		
Obvious	2	3
Inverse	2	5
RM		
Obvious	2	3
Inverse	2	5

adapted to either assess an understanding of the structure of a feedback loop (item format 1 for SO) or to assess the understanding of temporal dynamics (item formats 2.1, 2.2, both for SB and RM). The dimension SO was measured using two statements that described the reciprocal relationships between two processes in the blood calcium control (item format 1). Students had to decide whether both statements were correct (i.e., A affects B and B affects A), only one statement was correct (i.e., A affects B or B affects A), or neither statement was correct. Because the initiating conditions of the relationships are not examined, each of the six items for dimension SO refers to obvious conditions (see Table 2). Several items relate to reciprocal relationships in the feedback loops. As a result, even if one of the two relationships is direct (A→B), the second is indirect (B→…→A). Therefore, reciprocal relationships are rated as indirect relationships. Accordingly, for dimension SO, only one of the six items refers to a process from the external environment that has a direct effect on the regulated variable without any feedback effect, which makes it possible to illustrate a direct relationship (see Table 2).

To examine the reasoning about system dynamics, more specifically on the dimensions SB and RM, we used two different item formats. Item format 2.1 is two-tiered (Treagust, 1988). First, the prediction tier describes a system state and asks the students to predict consequences that will occur (SB) or actions that trigger the given system state (RM). Students had to choose one of two possible consequences or actions. Second, the justification tier offers students four possible explanations for their predictions. Students had to choose the explanation that best fits their predictions. There are eight items each for the SB and RM dimensions. In a balanced design, four items each were assigned to direct and indirect relationships and four items each for obvious and derived inverse conditions, leading to two direct/obvious, two indirect/obvious, two direct/inverse, and two indirect/inverse items.

Item format 2.2, for system dynamics (SB and RM), is a multiple-choice setting based on highly directed concept-mapping practices (Brandstädter *et al.*, 2012). The students were given a set of elements and processes (e.g., parathyrin concentration in the blood) and relations (e.g., X leads to more Y, X leads to less Y). The students were to trace the effects along a sequence of elements and relationships. SB tasks differ from RM tasks in that a state is given and participants must infer the resulting consequences. RM differ from SB tasks in that participants derive

purposeful actions to trigger a specified target state. There are four items each for the SB and RM dimensions. Because the requirement of these items is to track effects along a chain of cause–effect relationships, each item is rated as an indirect relationship. Three items each for dimensions SB and RM represent derived inverse conditions, and one item each represents obvious conditions (see Table 2). All item formats were dichotomously scored (i.e., right vs. wrong). Test booklets were developed, each containing the informational text, the diagram, and all items, so that we could generate complete data sets. To avoid item order effects, we used two different test booklets characterized by different sequences of items.

Data Analysis

Item Response Theory (IRT) to Measure and Investigate the Distribution of Item Difficulty. To investigate the influence of the item factors (i.e., relationships and conditions) on students' feedback loop reasoning, we needed a measure for task requirements. IRT-based item difficulty was used to operationalize requirements. IRT, like classical test theory, is a model-based measurement to analyze the relation of item properties to item responses (Yen and Fitzpatrick, 2006; Embretson and Reise, 2000). By transforming raw test scores, IRT maps and relates item properties and individual performance as latent traits on the same linear scale. While raw test scores are affected by measurement error, the latent traits are adjusted for these errors (Boone, 2020). IRT-based item difficulty as a latent trait is reported on a logit scale ranging from negative to positive values; the more positive an item measure is, the more difficult the item (Boone, 2020). After IRT analysis, measures of item difficulty are illustrated in a Wright map to investigate its distribution. In this study, Wright maps were used to check whether the items were arranged as expected (Boone, 2020).

Checking the Fit of Data to IRT Models. For IRT analyses, we used the R package TAM (Robitzsch *et al.*, 2020). Because we converted all student responses into dichotomous test scores, we applied IRT models that described dichotomous scores. As we only needed one estimator for item difficulty to address the research question, we further restricted the complexity of our model to one item measure describing item difficulty (1PL). We examined the quality of test items to detect potentially unsuitable items in advance. Weighted (infit) mean-square statistics indicate the association between the model and the data and thus can be used to test whether items fit the theoretical construct (Boone, 2020). Our analysis for the general unidimensional (1D) 1PL model revealed infit values ranging from minimum = 0.8 to maximum = 1.3, which is satisfactory (Bond and Fox, 2007, p. 243). The analysis of the 3D model showed item infits ranging from 0.87 (item rm1.3) to 1.18 (item so6), which is even more satisfactory and further supports the 3D model.[1] As the infit statistics were satisfactory, we included all items in the subsequent analyses. Further, to evaluate the reliability of the measurement instrument, we used expected a posteriori (EAP) reliability, a measure that can be interpreted similar to Cronbach's alpha (Bond and Fox, 2007, p. 59). The EAP reliability of 0.88 for the 1D model is seen as suitable.

---

[1]For interested readers, individual item fit statistics are listed in the Supplemental Material.

TABLE 3. Mean item difficulty (M), SD, and number of items (*n*) for the factors relationships (direct vs. indirect) and conditions (obvious vs. inverse)

Factor[a]	M	SD	*n*
Relationships			
Direct	−0.68	0.91	9
Indirect	0.10	0.96	21
Conditions			
Obvious	−0.67	0.92	16
Inverse	0.47	0.71	14

[a]The factors result from the main effects considered. Thus, the factors relationships and conditions are not independent.

Model Comparisons to Support the Assumed Structure of the Measuring Instrument. As there is no objective measure of model fit, we estimated a 3D model based on multidimensional item response theory (MIRT) in addition to the 1D model and compared their respective fits to assess construct validity. The 3D model, which distinguishes SO, SB, and RM as independent skills, was the theoretically supported model. Similar to confirmatory factor analysis, MIRT is a procedure used to evaluate the internal structure of a measurement instrument (Immekus *et al.*, 2019). It is advisable to also test the fit of models from MIRT to avoid the risk of underestimating the model, because "research experience so far indicates that overestimating the number of dimensions is less of a problem than underestimating [them]" (Hartig and Höhler, 2009, p. 60). To compare the goodness of fit of the 3D model and the general 1D model, we used deviance statistics and likelihood ratio tests. As the number of parameters affects the deviance statistics, we added the Akaike information criterion (AIC) and Bayes information criterion (BIC; Schwarz, 1978; Burnham and Anderson, 2004). AIC and BIC are motivated by different theoretical frameworks. We used BIC as the primary information criterion, as BIC is a conservative method that emphasizes parsimony of a model (Dziak *et al.*, 2020). In general, AIC and BIC are penalty scores, meaning that lower scores represent a better fit of the model. Furthermore, we examined the latent correlation between the dimensions to evaluate their independence (Hartig and Höhler, 2009). Latent correlations provide better estimates of the relations between dimensions than the correlations between manifest variables (Hartig and Höhler, 2009).

Analysis of the Role of the Factors Relationships and Conditions on Item Difficulty. To address the actual research question—to what extent the factors relationships and conditions affect the requirements of tasks—a two-way analysis of variance (ANOVA) with the factors relationships and conditions was computed. We examined the main effect models, as the number of items for the interaction groups was too small to make valid inferences about the interaction. Beforehand, we checked the data with respect to the assumptions of normality of residuals, homogeneity of variances, and independence of the observations (Luepsen, 2018). Because we applied an unbalanced design, we checked the data for positive pairing. There is a rather small positive correlation between the item number and SD for the main effect conditions (see Table 3). "The *F* test is conservative when sample sizes and variances are positively

related" (Hsiung *et al.*, 1994, p. 115). Accordingly, if there is a positive pairing, the probability of making a type I mistake is less than *p* = 0.05; however, the probability of making a type II mistake is increased (Field *et al.*, 2012). Thus, if a significant effect is identified based on the empirical data, it suggests that there is indeed an effect, but possibly with a biased effect size.

We conducted ANOVA tests using the R package car (Fox *et al.*, 2020). Because the item numbers per factor are unrelated to population size, we tested hypotheses associated with unweighted marginal means (type III sum of squares; Keselman *et al.*, 1995). We specified eta squared (η^2) as the effect size measure, indicating the amount of variance that can be explained by the factors relationships and conditions (Lakens, 2013).

RESULTS

Model Comparisons to Support the Assumed Structure of the Measuring Instrument

To provide empirical evidence for the hypothesized 3D structure, we compared the 1D between-item model with the 3D model.[2] Comparing the given goodness-of-fit parameters, the 3D model had data fit superior to the 1D model, as it had lower BIC scores (1D: 6947; 3D: 6904) as well as lower AIC scores (1D: 6843; 3D: 6783). The likelihood ratio test revealed that the 3D model was significantly better than the 1D model ($\chi^2(5) = 69.6$, $p < 0.001$). With EAP reliability ranging from 0.69 (SO) to 0.86 (RM), the three resulting dimensions exhibited acceptable values given the rather small number of items in SO. A further examination of the 3D model revealed that latent correlation coefficients ranged from 0.70 (SO related to SB) to 0.91 (SB related to RM). As expected, the last-mentioned was quite strong, because the items aimed at highly related sets of abilities. Latent correlations are corrected for measurement error, and consequently, those measurements are much higher than the correlations between the manifest variables. A measure of 0.9 indicates a strong interaction between the two dimensions, SB and RM, but both contextual reasons and the goodness-of-fit parameters indicated the superiority of the 3D model. Thus, we continue our analyses with measures for item difficulty resulting from the 3D model.

Factors Contributing to Requirements of Feedback Loop Reasoning in Homeostatic Systems

A Wright map provides a detailed overview of the distribution of item difficulty. Figure 1 presents the difficulties of test items representing obvious conditions (Figure 1a) and derived inverse conditions (Figure 1b) along the logit scale. By comparing the ordering of items from easy to hard and looking at the spacing of items, it is possible to assess whether the test measures feedback loop reasoning in a way that is consistent with prior definitions and theories. In general, items in all dimensions showed a wide range of empirical item difficulty: [−2.1–1.7] for SO, [−1.4–1.96] for SB, and [−1.2–0.8] for RM. Accordingly, there are rather easy and rather difficult items in all dimensions. Items for SO tended to be easier items, except for so6. A closer look at the task-specific demands showed that only so6, the

---

[2]For an additional overview of statistics comparing these and other combinations of dimensions, see the Supplemental Material.
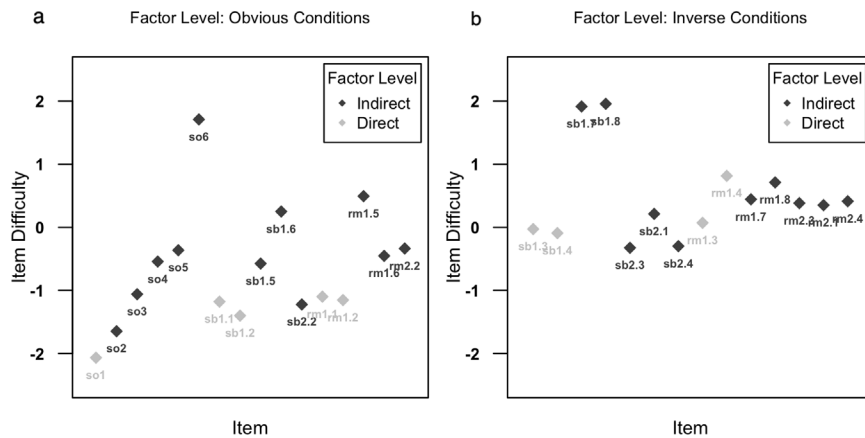
FIGURE 1. Wright maps showing the difficulty for items representing obvious conditions (a) and derived inverse conditions (b) on a logit scale (a high value indicates difficult items; a low value indicates easy items). The color of item measures represents the assigned level of the factor relationships (direct vs. indirect).

items related to direct relationships tend to be easier than items related to indirect relationships (cf., so1 vs. so2-so6, cf., sb1.1, sb1.2 vs. sb1.5, sb1.6, cf., rm1.1, rm1.2 vs. rm1.5, rm1.6). Within the items related to derived inverse conditions (see Figure 1b), items related to direct relationships tend to be easier than items related to indirect relationships (cf., sb1.3, sb1.4 vs. sb1.7, sb1.8). However, this tendency was not evident within items related to RM (cf., rm1.3, rm1.4 vs. rm1.7, rm1.8). Summarizing the differences between direct and indirect relationship across all items, the average item difficulty was higher for items regarding indirect relationships, whereas items regarding direct relationships had a lower average item difficulty (see Table 3). ANOVA revealed that the overall main effect of relationships was statistically significant, $F(1, 26) = 5.45$, $p = 0.03$, $\eta^2 = 0.11$. In summary, our graph-

most difficult, requires the recognition of a reciprocal relationship between two processes from separate loops (i.e., the release of signaling molecule A and the release of signaling molecule B). While the item's increased difficulty was as expected, it is surprising that only high-ability students possess a 50% probability of answering the item correctly. Items for SB showed a more prominent distribution along the poles of the logit scale compared with items for RM. Low-ability students possess a 50% probability of correctly answering items sb1.1 and sb1.2, which are expected to be the easiest ones. Only high-ability students possess a probability of 50% of correctly answering items sb1.7 and sb1.8, which are expected to be the most difficult ones. The remaining items for SB, including those in the concept-mapping format, showed only minor differences in item difficulty; hence mid-ability students possess a 50% probability of correctly answering these items. Similarly, for RM, low-ability students possess a 50% probability of correctly answering items rm1.1 and rm1.2, which are expected to be the easiest ones. In contrast, items rm1.7 and rm1.8 did not show such a high item difficulty, so upper mid-ability students possess a probability of 50% for correctly answering these items.

By comparing the average difficulty of items related to obvious conditions and derived inverse conditions, we obtained a qualitative overview of the influence of the factor conditions. Items related to obvious conditions showed a lower average item difficulty than items with derived inverse conditions (see Table 3). This trend was particularly evident for SB in the two-tier and concept-mapping item formats (cf., sb1.1, sb1.2, sb1.5, sb1.6 vs. sb1.3, sb1.4, sb1.7, sb1.8; sb2.2 vs. sb2.1, sb2.3, sb2.4). This trend was less evident in the dimension RM (cf., rm1.1, rm1.2, rm1.5, rm1.6 vs. rm1.3, rm1.4, rm1.7, rm1.8; rm2.2 vs. rm2.3, rm2.1, rm2.4). ANOVA revealed that the differences related to the main effect of conditions across all items were statistically significant, $F(1, 26) = 16.63$, $p < 0.001$, $\eta^2 = 0.34$.

Figure 1 also compares item difficulty related to direct relationships (light-colored thresholds) with the item difficulty related to indirect relationships (dark-colored thresholds). Within the items related to obvious conditions (see Figure 1a),

ical and statistical analyses support the hypothesis that the factors relationships and conditions help explain differences in item difficulty. Conditions, however, have a greater impact than relationships.

DISCUSSION

Based on findings from systems thinking in ecological systems (Mambrey *et al.*, 2020) and systems thinking in geography education (Mehren *et al.*, 2018), we applied a 3D model consisting of SO, SB, and SM (or RM[3]) to measure feedback loop reasoning in homeostatic systems. The theory-based 3D model indicated acceptable measures for model fit. Qualitative analyses and model comparisons suggest that the data on feedback loop reasoning in homeostatic systems fit to the 3D model. Consequently, this model can be applied to examine the research question. Regarding our research question, our results support the proposed hypotheses: tasks involving indirect relationships were more difficult than tasks involving direct relationships, and tasks involving derived inverse conditions were more difficult than tasks involving obvious conditions. Thus, we postulate that recognizing indirect relationships and using derived inverse conditions can serve as stepping-stones in constructing more elaborate feedback loop reasoning skills concerning homeostatic systems. These results complement other research that focused on content-related challenges of studying homeostasis (e.g., Westbrook and Marek, 1992; McFarland *et al.*, 2017). Modell *et al.* (2015) presented challenges at the level of undergraduate students from a physiologist's perspective. Accordingly, typical obstacles include the misconception that setpoints in physiological systems do not change over a lifetime; that "relatively constant" means that the regulated variable does not change over a period of an hour, a day, or week; or that homeostatic mechanisms operate like an on/off principle (Modell *et al.*, 2015). Each of these difficulties is apparently due to insufficient knowl-

---

[3]As noted earlier, we refer to the dimension SM within the context of homeostatic systems as RM, because it involves the intentional selection of regulatory measures to achieve intended target states.

edge of the dynamics of homeostatic systems. Analyzing underlying mechanisms, as provided by systems-thinking approaches, especially in the dimensions SB and RM, can offer descriptive insights into dynamic properties. Overall, mechanisms are crucial for linking structures and functions (Lira and Gardner, 2017). For example, analysis of various initiating conditions reveals that homeostatic systems do not behave according to an on/off principle. Instead, the analysis of mechanisms shows that flexible dynamics of elements and processes enable the constancy of the regulated variables as an emergent phenomenon. The factors considered in this research (i.e., relationships and conditions) refer to mechanisms of homeostatic feedback loop systems.

Factor Relationships

Direct and indirect relationships are a property of sequential processes, such as photosynthesis or blood glucose regulation (Chi *et al.*, 2012). The effects are indirect if the triggering event and the outcome are mediated by intermediate processes. Feedback effects, as a special case, are also indirect, because the cause and the actual effect on the preceding cause are mediated through a sequence of processes and events (Hokayem *et al.*, 2020). The ability to describe indirect, distant, or long-term ecological effects has been repeatedly examined (Grotzer and Basca, 2003; Ergazaki and Andriotou, 2010). Previous studies from geography education have demonstrated that difficulty arises from the type of networking, from monocausal (e.g., A acts on B) to complex (e.g., A acts on B, B acts on C, and C in turn acts on A; Mehren *et al.*, 2018). In ecological systems, Mambrey *et al.* (2020) also differentiated between direct cause–effect relationships and indirect effects. They concluded that analyzing indirect effects requires a higher skill level. Although the understanding of indirect relationships has been studied in ecological and geographic systems (Mehren *et al.*, 2018; Mambrey *et al.*, 2020), research on the role of the understanding of indirect relationships in homeostatic systems is still lacking. Our results illustrate this challenge in understanding homeostatic regulation.

Factor Conditions

Regarding the interpretation of qualitative dynamics of cause–effect relationships, it is necessary to distinguish between obvious and derived inverse conditions. Changes can run in opposite directions; a process can either increase or decrease in size or strength compared with previous states (Chi *et al.*, 2012). As a result, it is necessary to look at different conditions of relationships; causes and effects can be analyzed based on an increased process or state, but also on a decreased one. Cho and Jonassen (2012) have already indicated the challenge to apply distinct conditions of cause–effect relationships. They characterized low-quality responses by referring to one-sided explanations. In a physiological context, they were able to determine that students only focused on one of two possible conditions (i.e., either an increase or a decrease in a measure in the human body). Building on these findings, Wellmanns and Schmiemann (2020) found that some students intuitively applied derived inverse conditions when explaining scenarios in the context of blood glucose regulation. When asked what happens when glucose is ingested with food, some students explain that high glucose consumption causes blood glucose levels to increase signifi-

cantly, and that low consumption, conversely, causes blood glucose levels not to increase or to even decrease. Thus, these students considered two initiating conditions: increased and decreased consumption. Our research helped identify derived inverse conditions as a challenge for students describing homeostatic systems. Thus far, there has been no educational research on the role of conditions in systems thinking. We have now demonstrated that the analysis of derived inverse conditions, which is relevant for an analysis of systems dynamics, contributes significantly to the difficulty of test items. Thus, we would propose including the property of derived inverse conditions as a challenge in homeostatic system analysis.

Limitations

Because models for systems thinking are complex and must be adapted to the respective context and its particular features, there are some shortcomings with regard to the test instrument. We constructed our items according to the specific content requirements of each dimension, given that all items corresponded to a closed multiple-choice design. It turned out that the item formats for SB and RM were not suitable for assessing SO, which does not require analysis of dynamic developments. Therefore, the design of items on SO differs from the design of items on SB and RM, but those for SB and RM do not differ. To operationalize different skills, different task formats have already been used in previous studies on systems thinking (Ben Zvi Assaraf and Orion, 2005). Moreover, as per Dochy *et al.* (1999), it is necessary to use many different assessment tasks to obtain authentic measures for performance in different realms.

One clear limitation of this study is the rather small and limited sample of 210 students. The findings of this exploratory study, however, are plausible, as additional sources have already shown a 3D structure for systems-thinking skills (Mambrey *et al.*, 2020). Moreover, both, the 1D and 3D models were satisfactory in terms of item fit and reliability. Hence, the number of participants is still within a reasonable range to answer the research question, at least preliminarily. Further studies may evaluate to what extent our findings could be applied in other contexts and to other samples.

Another pitfall to be noted is that, due to the immanent characteristics of each skill and context, there is an unbalanced design regarding the assignment of items to factors explaining difficulty. Although the factorial ANOVA is based on an unbalanced design, and the results should therefore be interpreted with caution, the cross-dimensional ANOVA is still useful to provide an initial overview. Moreover, the graphical analysis of item difficulty indicates that the trends revealed by the factorial ANOVA are evident in each of the three skills. Therefore, we conclude that the main effects of the factors relationships and conditions on item difficulty were adequately assessed.

As the sample is composed of two cohorts, high school students and undergraduate students, measures of individual performance could have been affected. This should not be a concern, however, because the intention was not to provide reliable measures of individual performance but to provide empirical evidence for the hypothesized structure of feedback loop reasoning in homeostatic systems. This split sample allows for a higher variance in learning ability. In future work, the assessment could be used in larger samples to obtain an estimator for individual performance in feedback loop reasoning skills. In our data, there

are already initial indications that undergraduate students may perform better than high school students across all dimensions.

Finally, it should be noted that we examined the structure of feedback loop reasoning skills in homeostatic systems in only one context. We highlight what makes feedback loop reasoning challenging in this particular homeostatic system. Future work is necessary to confirm the findings in further homeostatic and non-homeostatic systems.

CONCLUSIONS AND EDUCATIONAL IMPLICATIONS

Our aim was to contribute to a better understanding of the requirements of systems-thinking tasks in the context of homeostasis. Our analyses confirmed that both direct and indirect relationships and obvious and derived inverse conditions are factors that pose challenges. We propose that curricula specify the use of indirect relationships and derived inverse conditions as learning objectives in teaching homeostasis. To learn to use dynamic relationships between elements in a homeostatic system, students need opportunities to begin their reasoning from different initiating conditions. Learning tasks may require students to describe the consequences starting from changes in an element or process in two opposite directions (Cho and Jonassen, 2012). For example, students could describe the consequences of increased reabsorption of calcium in the kidneys but also the consequences of decreased or pathologically missing reabsorption of calcium in the kidneys. Transferring this demand to a learning task addressing ecological relationships between predator and prey populations could target not only effects of an increasing but also effects of a decreasing predator population (Freeman *et al.*, 2017).

Furthermore, to learn to identify dynamic relationships, students need opportunities to explore relationships between elements and processes that are indirectly related to each other (Wellmanns and Schmiemann, 2020). An example of an indirect relationship is the long-term effect of calcium intake on blood calcium levels. Calcium intake increases the blood calcium level in the short term, but in the long term, the calcium decreases due to the effect of calcitonin release and the associated calcium incorporation into the bones (Modell *et al.*, 2015). Another, less obvious indirect relationship exists between hormones related to the regulated variable. For example, a change in the release of parathyrin causes a change in the release of calcitonin via a measurable change in the blood calcium level (Freeman *et al.*, 2017). In ecological contexts, long-term effects in learning tasks are even more illustrative. An increasing predator population affects the prey population in the short term. However, in the long-term, many actors in the ecosystem are affected by the broader consequences of the changes, including the predator population itself. Instructors may provide explicit instruction by presetting certain elements and processes that are indirectly related to each other and having students investigate the relationship. In ecological contexts, this could be the investigation of the relationship between the reintroduction of wolves and the increasing beaver population in Yellowstone National Park. In physiological contexts, this could be the investigation of the relationship between insulin injections and glucagon blood levels. We assume that engagement with such learning settings will help students recognize patterns so that they can identify and use derived inverse conditions and indirect relationships in their reasoning about complex systems.

REFERENCES
American Association for the Advancement of Science. (2011). *Vision and change in undergraduate biology education: A call to action*. Washington, DC.

Association of American Medical Colleges, & Howard Hughes Medical Institute. (2009). *Scientific foundations for future physicians*. Retrieved August 9, 2021, from www.aamc.org/system/files?file=2020-02/scientificfoundationsforfuturephysicians.pdf

Batzri, O., Ben Zvi Assaraf, O., Cohen, C., & Orion, N. (2015). Understanding the earth systems: Expressions of dynamic and cyclic thinking among university students. *Journal of Science Education and Technology*, *24*(6), 761–775. https://doi.org/10.1007/s10956-015-9562-8.

Ben Zvi Assaraf, O., Dodick, J., & Tripto, J. (2013). High school students' understanding of the human body system. *Research in Science Education*, *43*(1), 33–56. https://doi.org/10.1007/s11165-011-9245-2.

Ben Zvi Assaraf, O., & Orion, N. (2005). Development of system thinking skills in the context of earth system education. *Journal of Research in Science Teaching*, *42*(5), 518–560. https://doi.org/10.1002/tea.20061.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Boogerd, F. C., Bruggemann, F. J., Hofmeyr, J.-H. S., & Westerhoff, H. V. (2007). Towards philosophical foundations of systems biology: Introduction. In Boogerd F. C., Bruggemann F. J., Hofmeyr J.-H. S., & Westerhoff H. V., (Eds.), *Systems biology: Philosophical foundations* (pp. 3–19). Amsterdam, Netherlands: Elsevier.

Boone, W. J. (2020). Rasch basics for the novice. In Khine M. S., (Ed.), *Rasch measurement: Applications in quantitative educational research* (pp. 9–30). Singapore: Springer. https://doi.org/10.1007/978-981-15-1800-3_2

Booth Sweeney, L., & Sterman, J. D. (2007). Thinking about systems: Student and teacher conceptions of natural and social systems. *System Dynamics Review*, *23*(2–3), 285–311. https://doi.org/10.1002/sdr.366

Brandstädter, K., Harms, U., & Großschedl, J. (2012). Assessing system thinking through different concept-mapping practices. *International Journal of Science Education*, *34*(14), 2147–2170. https://doi.org/10.1080/09500693.2012.716549

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Camazine, S., Deneubourg, J.-L., Franks, N. R., Sneyd, J., Theraulaz, G., & Bonabeau, E. (2003). *Self-organization in biological systems*. Princeton, NJ: Princeton University Press.

Cary, T., & Branchaw, J. (2017). Conceptual elements: A detailed framework to support and assess student learning of biology core concepts. *CBE—Life Sciences Education*, *16*(2), 1–10. https://doi.org/10.1187/cbe.16-10-0300

Chi, M. T. H., Roscoe, R. D., Slotta, J. D., Roy, M., & Chase, C. C. (2012). Misconceived causal explanations for emergent processes. *Cognitive Science*, *36*, 1–61. https://doi.org/10.1111/j.1551-6709.2011.01207.x

Cho, Y. H., & Jonassen, D. H. (2012). Learning by self-explaining causal diagrams in high-school biology. *Asia Pacific Education Review*, *13*(1), 171–184. https://doi.org/10.1007/s12564-011-9187-4

Dochy, F., Segers, M., & Buehl, M. M. (1999). The relation between assessment practices and outcomes of studies: The case of research on prior knowledge. *Review of Educational Research*, *69*(2), 145–186. https://doi.org/10.3102/00346543069002145

Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R., & Jermiin, L. S. (2020). Sensitivity and specificity of information criteria. *Briefings in Bioinformatics*, *21*(2), 553–565. https://doi.org/10.1093/bib/bbz016

Eilam, B., & Poyas, Y. (2010). External visual representations in science learning: The case of relations among system components. *International*

Journal of Science Education, *32*(17), 2335–2366. https://doi.org/10.1080/09500690903503096

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum).

Ergazaki, M., & Andriotou, E. (2010). From "forest fires" and "hunting" to disturbing "habitats" and "food chains": Do young children come up with any ecological interpretations of human interventions within a forest? *Research in Science Education*, *40*(2), 187–201. https://doi.org/10.1007/s11165-008-9109-6

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Los Angeles, CA: Sage.

Fox, J., Weisberg, S., & Price, B. (2020). *car: Companion to applied regression*. Retrieved February 19, 2021, from https://CRAN.R-project.org/package=car

Freeman, S., Quillin, K., Allison, L., Black, M., Podgorski, G., Taylor, E., & Carmichael, J. (2017). *Biological science* (6th ed., Global ed.). Harlow, UK: Pearson.

Gatewood, L. C., Ackerman, E., Rosevear, J. W., & Molnar, G. D. (1970). Modeling blood glucose dynamics. *Behavioral Science*, *15*(1), 72–87. https://doi.org/10.1002/bs.3830150108

Gilissen, M. G. R., Knippels, M.-C. P. J., & van Joolingen, W. R (2020). Bringing systems thinking into the classroom. *International Journal of Science Education*, *42*(5), 1253–1280. https://doi.org/10.1080/09500693.2020.1755741

Grotzer, T. A., & Basca, B. B. (2003). How does grasping the underlying causal structures of ecosystems impact students' understanding? *Journal of Biological Education*, *38*(1), 16–29. https://doi.org/10.1080/00219266.2003.9655891

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation*, *35*(2–3), 57–63. https://doi.org/10.1016/j.stueduc.2009.10.002

Hmelo-Silver, C. E., Marathe, S., & Liu, L. (2007). Fish swim, rocks sit, and lungs breathe: Expert-novice understanding of complex systems. *Journal of the Learning Sciences*, *16*(3), 307–331. https://doi.org/10.1080/10508400701413401

Hokayem, H., Jin, H., & Yamaguchi, E. (2020). Feedback loop reasoning and knowledge sources for elementary students in three countries. *EURASIA Journal of Mathematics, Science & Technology Education*, *16*(2), 1–13. https://doi.org/10.29333/ejmste/112582

Hsiung, T.-H., Olejnik, S., & Huberty, C. J (1994). Comment on a Wilcox test statistic for comparing means when variances are unequal. *Journal of Educational Statistics*, *19*(2), 111–118.

Immekus, J. C., Snyder, K. E., & Ralston, P. A. (2019). Multidimensional item response theory for factor structure assessment in educational psychology research. *Frontiers in Education*, *4*, 1–15. https://doi.org/10.3389/feduc.2019.00045

Jansen, S., Knippels, M.-C. P. J., & van Joolingen, W. R. (2019). Assessing students' understanding of models of biological processes: A revised framework. *International Journal of Science Education*, *41*(8), 981–994. https://doi.org/10.1080/09500693.2019.1582821

Keselman, H. J., Carriere, K. C., & Lix, L. M. (1995). Robust and powerful nonorthogonal analyses. *Psychometrika*, *60*(3), 395–418.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*(863), 1–12. https://doi.org/10.3389/fpsyg.2013.00863

Lira, M. E., & Gardner, S. M. (2017). Structure-function relations in physiology education: Where's the mechanism? *Advances in Physiology Education*, *41*(2), 270–278. https://doi.org/10.1152/advan.00175.2016

Luepsen, H. (2018). Comparison of nonparametric analysis of variance methods: A vote for van der Waerden. *Communications in Statistics—Simulation and Computation*, *47*(9), 2547–2576. https://doi.org/10.1080/03610918.2017.1353613

Mambrey, S., Schreiber, N., & Schmiemann, P. (2022). Young students' reasoning about ecosystems: The role of systems thinking, knowledge,

conceptions, and representation. *Research in Science Education*, *52*, 79–98. https://doi.org/10.1007/s11165-020-09917-x

Mambrey, S., Timm, J., Landskron, J. J., & Schmiemann, P. (2020). The impact of system specifics on systems thinking. *Journal of Research in Science Teaching*, *57*(10), 1632–1651. https://doi.org/10.1002/tea.21649

McFarland, J. L., Price, R. M., Wenderoth, M. P., Martinková, P., Cliff, W., Michael, J., Modell, H., & Wright, A. (2017). Development and validation of the Homeostasis Concept Inventory. *CBE—Life Sciences Education*, *16*(2), 1–13. https://doi.org/10.1187/cbe.16-10-0305

Mehren, R., Rempfler, A., Buchholz, J., Hartig, J., & Ulrich-Riedhammer, E. M. (2018). System competence modelling: Theoretical foundation and empirical validation of a model involving natural, social and human-environment systems. *Journal of Research in Science Teaching*, *55*(5), 685–711. https://doi.org/10.1002/tea.21436

Modell, H., Cliff, W., Michael, J., McFarland, J., Wenderoth, M. P., & Wright, A. (2015). A physiologist's view of homeostasis. *Advances in Physiology Education*, *39*(4), 259–266. https://doi.org/10.1152/advan.00107.2015

NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states*. Washington, DC: National Academies Press.

Quillin, K., & Thomas, S. (2015). Drawing to learn: A framework for using drawings to promote model-based reasoning in biology. *CBE—Life Sciences Education*, *14*(1), 1–16. https://doi.org/10.1187/cbe.14-08-0128

Robitzsch, A., Kiefer, T., & Wu, M. (2020). *TAM: Test analysis modules*. Retrieved February 19, 2021, from https://CRAN.R-project.org/package=TAM

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Scott, E. E., Anderson, C. W., Mashood, K. K., Matz, R. L., Underwood, S. M., & Sawtelle, V. (2018). Developing an analytical framework to characterize student reasoning about complex processes. *CBE—Life Sciences Education*, *17*(3), 1–14. https://doi.org/10.1187/cbe.17-10-0225

Snapir, Z., Eberbach, C., Ben Zvi Assaraf, O., Hmelo-Silver, C. E., & Tripto, J. (2017). Characterising the development of the understanding of human body systems in high-school biology students—a longitudinal study. *International Journal of Science Education*, *39*(15), 2092–2127. https://doi.org/10.1080/09500693.2017.1364445

Sterman, J. D. (2000). *Business dynamics: Systems thinking and modeling for a complex world*. Boston, MA: McGraw-Hill.

Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. *International Journal of Science Education*, *10*(2), 159–169. https://doi.org/10.1080/0950069880100204

Tripto, J., Ben Zvi Assaraf, O., & Amit, M. (2018). Recurring patterns in the development of high school biology students' system thinking over time. *Instructional Science*, *46*(5), 639–680. https://doi.org/10.1007/s11251-018-9447-3

Verhoeff, R. P., Knippels, M.-C. P. J., Gilissen, M. G. R., & Boersma, K. T (2018). The theoretical nature of systems thinking. Perspectives on systems thinking in biology education. *Frontiers in Education*, *3*, 1–11. https://doi.org/10.3389/feduc.2018.00040

von Bertalanffy, L. (1973). *General system theory: Foundations, development, applications*. Harmondsworth, UK: Penguin.

Wellmanns, A., & Schmiemann, P. (2020). Feedback loop reasoning in physiological contexts. *Journal of Biological Education*, Advance online publication, 1–21. https://doi.org/10.1080/00219266.2020.1858929

Westbrook, S. L., & Marek, E. A. (1992). A cross-age study of student understanding of the concept of homeostasis. *Journal of Research in Science Teaching*, *29*(1), 51–61. https://doi.org/10.1002/tea.3660290106

Wilson, K. J., Long, T. M., Momsen, J. L., & Bray Speth, E. (2020). Modeling in the classroom: Making relationships and systems visible. *CBE—Life Sciences Education*, *19*(1), 1–5. https://doi.org/10.1187/cbe.19-11-0255

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In Brennan R. L., (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: Praeger.