Constructing multi-theory vignettes to measure the application of knowledge in ambivalent educational situations

Hendrik Lohse-Bossenz*, Christopher Bloss and Tobias Dörfler

Institute for Psychology, University of Education Heidelberg, Heidelberg, Germany

Research on evidence-based argumentation shows that (pre-service) teachers have difficulties in orienting their actions to existing theories and empirical evidence. This article addresses the knowledge content needed for this and presents a vignette-based procedure. Within each vignette, two different theoretical perspectives are addressed. The behavior of a teacher can be either suitable or unsuitable from both perspectives or more or less suitable depending on the perspective. In study 1, the procedure is piloted and in study 2, an intervention on a specific area of knowledge takes place. The results show that participants differentiate the vignettes as expected. The intervention leads to corresponding increases in knowledge, which likely relates to a change in the evaluations. The presented approach is discussed with regard to possible applications in the context of research on evidence-based argumentation.

Introduction

The decisions involved in planning, delivering, and evaluating school lessons are characterized by high levels of uncertainty (Floden and Buchmann, 1993). In the face of this uncertainty, teachers may rely on a variety of sources to make their pedagogical decisions: scientific theories, scientific evidence, subjective theories, beliefs, anecdotes, recipes, or even gut feelings (Stark, 2017; Kiemer and Kollar, 2021). Given that information linked to these specific sources is acquired *via* specific knowledge-building processes, its epistemic status varies, for example, with respect to trustworthiness and credibility (Fenstermacher, 1994). Although the idea that scientific evidence might be valuable in solving practical problems is controversial

(Brown and Rogers, 2015), the use of educational evidence to explicate the reasons for pedagogical judgments seems to be beneficial, at least in cases of classroom problems that appear to occur repeatedly.

Several sets of findings indicate that (pre-service) teachers encounter challenges on two levels of scientific reasoning (Csanadi et al., 2021). On the process level, they may struggle to engage in the inquiry process (Klahr and Dunbar, 1988) or to follow the trajectory of epistemic processes suggested by Fischer et al. (2014). This means that they might not collect enough evidence before engaging in evaluation of that evidence; as a result, the process is unsystematic and speculative. On the content level, they may not be able to relate scientific knowledge from relevant domains to actual classroom incidents, because they lack the requisite knowledge that would enable them to make this transfer, or to do so in a suitable manner (Brown and Rogers, 2015; Hetmanek et al., 2015; Hartmann et al., 2016).

To date, research has seldom addressed situations in which different lines of actions (in the sense of conflicting evidence) are available. Consider, for example, a typical classroom situation in which two students become angry with one other and are arguing at a time when all the students have been asked to work quietly on their worksheets. From a classroom management perspective, the teacher should intervene immediately to enforce the classroom rule that time on task should be maximized (see Lenske et al., 2016 for evidence on the influence of classroom management on students' learning gains). From the perspective of the development of a healthy classroom climate and peer relationships, the teacher may instead let the class stop working and use this situation to explicitly address productive ways to solve peer conflicts (Jennings and Greenberg, 2009). Here, two educational goals with unique theoretical and empirical backgrounds come into play and lead to divergence in the actions that the teacher could potentially take. Such situations are highly prevalent in educational settings, and teachers face the challenge of weighing the benefits of possible actions against one another and coming to a decision tailored to the specific situation.

Presumably, if teachers make decisions on the basis of the knowledge that is accessible to them, they may not even perceive certain cues in the situation that would have led to another decision (a reflection of insufficient knowledge). In other cases, in which the teacher does have access to knowledge from different fields, the process by which they evaluate strands of evidence which may lead to different decisions (fragmented knowledge) is not well understood.

To address these issues, this article presents the construction and validation of a vignette-based instrument, involving items presenting scenarios in which decisions may vary depending on the theoretical perspectives on which the decision is based. After describing the theoretical background and the process of constructing the instrument, we present two validation studies indicating that convergent and divergent theoretical perspectives lead to systematic differences in

decision-making (Study 1) and that knowledge input influences judgments in a manner that indicates deeper evaluation of the cues corresponding to the newly-acquired knowledge (Study 2).

Theoretical background and research aims

There seems to be an increasing demand for teachers and policymakers to orient their respective educational and political decisions more towards evidence rather than relying on other sources such as subjective theories or anecdotes (Davies, 1999; Bromme et al., 2014). Following Stark (2017) and other researchers, we consider *evidence* to broadly consist of both theories and obtained empirical results that are valued by an individual as being of high scientific quality. That means that evidence does not have an independent existence in an objective sense, outside the judgment of individuals who attribute to it the specific property of meeting scientific standards (Bromme et al., 2014).

Research in the domain of evidence-based education often makes reference to the medical profession, where the parallel term *evidence-based medicine* is employed (Sackett et al., 1996). Although the feasibility of transferring theoretical perspectives from the medical to the teaching profession is under debate (Stark, 2017), the basic idea that teachers use evidence in their argumentation for or against specific decisions seems plausible. In their description of evidence-based argumentation, Csanadi et al. (2021) differentiate between *content* and *process* levels. The *content level* relates to knowledge which is used for evidence-based argumentation. On this level, strands of knowledge with variable epistemic status (Fenstermacher, 1994) are brought to bear. In addition to scientific theories and empirical results, subjective theories or case knowledge can also be put to use as sources in the argumentation process (Kiemer and Kollar, 2021). The *process level* itself can be further subdivided into the selection and the use of specific sources (Kiemer and Kollar, 2021). In turn, the use of specific sources consists of further subprocesses, including problem identification, hypothesis generation, and drawing conclusions (Fischer et al., 2014).

Recent research on the process level has provided insight into the ways in which (pre-service) teachers use or do not use evidence. For instance, Hetmanek et al. (2015) have demonstrated that pre-service teachers – despite being provided with the necessary information – do not use scientific evidence in their case analysis. Concerning the content level, recent studies have directly compared types of source to explore their specific role in the argumentation process. For example, Kiemer and Kollar (2021) have demonstrated that scientific theories are used more often than anecdotes or subjective theories in case analysis.

Research gap

To date, there has been a paucity of research concerning the comparison of sources with comparable epistemic status, such as convergent or divergent scientific theories. In such situations, heuristics like 'scientific theories are more trustworthy than subjective theories' provide no value. Instead, the evidence has to be evaluated with respect to the specific situation at hand and different strands must potentially be weighted differently in order to arrive at a decision. To this end, relevant information in the scenario, typically referred to as cues, must be observed and ultimately taken into account in the argumentation process. Furthermore, research on evidence-based argumentation in the domain of education has mainly focused on generic issues in teaching, such as motivation or general instruction. Subject-specific theories are seldomly addressed as sources of evidence.

To address these gaps, we developed an approach using multi-theory vignettes. The basic idea is to present situations that can be perceived differently from different perspectives. By defining two perspectives and their related core principles *a priori* in the process of constructing the vignettes, we can explicitly model participants' decision-making processes and formulate hypotheses concerning their reactions to the situations depicted.

Construction of multi-theory vignettes

Vignettes as a test format are becoming increasingly popular in the field of teacher education (Brovelli et al., 2014). Under this approach, each vignette consists of a scenario that presents an authentic situation from a lesson in school involving specific issues which necessitate the activation of professional knowledge in order to address them, and they are considered to be a suitable tool to assess situational knowledge or the ability of participants to access their knowledge in specific situations. In particular, research in the field of professional vision regularly employs this approach (Santagata and Angelici, 2010; Meschede et al., 2017).

Under our multi-theory vignette (MTV) approach, we constructed a set of vignettes containing cues that would be relevant from two different perspectives: the first falling primarily under the scope of a specific model of teaching games in Physical Education (PE), and the second falling primarily under the scope of self-determination theory (SDT). The core principle from the former perspective is that of *complexity reduction*: most teaching approaches in the domain of PE agree that sporting games need to be reduced in complexity when they are integrated into a school's curriculum (Kolb, 2005). Therefore, the teachers' behaviors depicted in our vignettes can be considered suitable from a PE teaching perspective if they involve a cue indicating some kind of complexity reduction. The core principle from the latter perspective is the *fulfilment of basic psychological needs*. If students' basic psychological needs are fulfilled, this appears to

enhance their sense that their actions are self-determined and to increase their intrinsic motivation (Ryan and Deci, 2000). Research indicates that satisfaction of basic psychological needs is associated with self-determined motivation (Chen and Jang, 2010; Goldman et al., 2017; Hu and Zhang, 2017) and positive learning outcomes (Baeten et al., 2013; McEown et al., 2014; Salmi and Thuneberg, 2019). Therefore, teachers' behaviors depicted in our vignettes that address students' psychological needs can be considered suitable from the SDT perspective.

As complexity reduction and need satisfaction are conceptually unrelated and are principles that arise from different theoretical perspectives, we combined both perspectives with their core principles in our vignettes. A *convergent vignette* would depict a pedagogical situation in which the action of the fictitious teacher is either suitable (the core principles are fulfilled) or unsuitable (the core principles are not fulfilled) according to both perspectives. We adopted a labelling scheme in which convergent vignettes depicting suitable teacher behavior were labelled SS because they suggested a **s**uitable teacher action as seen from both perspectives. In contrast, convergent vignettes depicting unsuitable actions were labelled UU, as they suggested an **u**nsuitable teacher action from both perspectives. A *divergent vignette* would depict a teacher action that is suitable from one of the perspectives and unsuitable from the other. These vignettes were labelled U_gS_m if they depicted an action which could be considered **s**uitable or need-supporting from the perspective of **m**otivational psychology or SDT, but an **u**nsuitable action from the perspective of teaching **g**ames; or S_gU_m if they depicted an action which could be regarded as suitable from the perspective of teaching games but unsuitable from the perspective of SDT or motivational psychology.

A total of 10 experts in the field of sports science with a focus on teaching games and 11 experts in motivational psychology were asked to evaluate our categorizations of 26 drafted vignettes as illustrating suitable or unsuitable actions from their expert perspective. To this end, they were informed beforehand of which teacher actions we considered to be suitable or unsuitable in terms of complexity reduction and need satisfaction. The sports science experts were not informed of the SDT interpretation of the vignettes, nor were they asked to rate the vignettes with this perspective in mind, and vice versa. The experts were also asked to name possible alternative actions for the teacher in each vignette. In general, the experts considered the vignettes to be authentic and suitable for our research purposes. However, some disagreement emerged concerning the suitability of the actions described, and experts from both fields suggested alternative actions for the teacher in a number of cases. It became clear that sports science experts with a focus on teaching games weight motivational considerations more heavily than psychology experts weight sports science considerations. After discussing all the results, excluding 10 vignettes, and slightly reformulating some vignettes, we arrived at a final set of 16 vignettes, four of each type (UU, SS, U_gS_m, and S_gU_m).

Example multi-theory vignette

> There is not much activity happening on the field where 24 fifth-graders are playing dodgeball: only a few students are actively taking part by running, dodging the ball, trying to catch it and throwing it at their opponents. One of the less active players, who has already had to leave the active zone of the field, is now outside in the passive zone (from where it is possible to return to the active zone by successfully throwing the ball at an opponent). She is standing close to the teacher and says to him: "This game is sooo boring…," looking expectantly at the teacher.
>
> The teacher replies that it would not be so boring if she, the girl, took part in it more actively. When the first round of the game has finished and the second round is about to begin, the teacher reduces the size of the field.

This vignette was constructed and used in our test as a divergent vignette (S_gU_m). From the perspective of teaching games, the teacher reacts rather appropriately to the lack of activity among his students by reducing the field size (complexity reduction). This lack of activity is evident in the vignette through the descriptions of the many passive players on the field and also the girl's claim of boredom. Although it cannot be assumed that the teacher's response here represents the ideal reaction, it is certainly a possible solution to a lack of activity during a ball game. However, from the SDT perspective, the teacher's reaction to the girl's complaint is inappropriate because he does not address the basic psychological needs of the student in this situation (need satisfaction). His answer makes it clear that he would prefer the girl to eliminate her negative emotions as quickly as possible. Additionally, he gives an unclear instruction by telling the girl that she should take part more actively: it can be assumed that the girl does not know what 'taking part more actively' means. Therefore, the student's psychological needs are not satisfied.

General hypotheses on multi-theory vignette ratings

As described, each vignette contained a problem, a dilemma, or a challenge to which a fictitious teacher's reaction was depicted. Each ended with a description of the teacher's actions, which were generally verbal, but sometimes non-verbal. As part of the instrument, participants were then asked to rate the fictitious teacher's action in relation to the statement 'The teacher's action is suitable', with higher ratings indicating greater perceived suitability. Convergent vignettes (i.e., those in which the actions are either suitable or unsuitable from both perspectives) are rather clear, and we thus expected participants to provide polarized ratings: UU vignettes should receive the lowest rating and SS

vignettes the highest rating, indicating high unsuitability and high suitability, respectively. In contrast, we expected ratings for divergent vignettes (i.e., those in which the suitability of the actions varied depending on the perspective adopted) to be close to the middle of the scale, as participants should be undecided. An example train of thought for the participant might be: "This is an appropriate way of dealing with the issue [complexity reduction], but the way he talks to his students does not seem right… [no need satisfaction]." However, our objective was to establish a method of identifying the type of knowledge brought to bear by different participants in providing their ratings by investigating individual differences in the ratings of divergent vignettes. Specifically, if a participant judges the actions depicted in U_gS_m vignettes to be more suitable than those depicted in S_gU_m vignettes, it can be concluded that their knowledge of SDT seems to have been of greater importance in their decision; conversely, if a participant judges the actions depicted in S_gU_m vignettes to be more suitable than those depicted in U_gS_m vignettes, it can be concluded that cues relating to the perspective of PE teaching seem to have been more salient to them. Furthermore, by examining changes in these differences over time, we expected to be able to measure the effects of knowledge-building and application.

The present study

We conducted two studies to test the validity of the MTV instrument described above (Borsboom et al., 2004). In Study 1, we aimed to pilot the instrument with a sample of student teachers and a sample of sports science students. We expected the vignette ratings to exhibit the distribution described above, with SS vignettes receiving the highest ratings, UU vignettes the lowest, and US and SU vignettes receiving intermediate ratings. We also expected that the exact pattern would be dependent on the sample: specifically, we hypothesized that U_gS_m vignettes would be associated with lower suitability judgments than S_gU_m vignettes by sports science students and vice versa for student teachers. In Study 2, we tested the hypothesis that a knowledge intervention providing information on SDT would elicit an increase in the difference between participants' U_gS_m and S_gU_m ratings.

Study 1: Pilot

Methods

Student teachers (Sample 1)

Sample 1 consisted of 153 pre-service teachers (127 female) from a university of specializing in education studies. The mean age was 21.85 years ($SD = 3.06$); 78.5% were in semester 3 of their studies or below, and the remaining 22.5% were in semesters 4 to 13. The vast majority (79.7%) were working

towards a Bachelor of Arts in either primary or secondary education; a smaller number (18.3%) were working towards a master's degree in education.

Sports science students (Sample 2)

Sample 2 consisted of 48 sports science students (27 female), with a mean age of 21.10 years ($SD = 2.15$). Most were working towards a Bachelor of Science (87.5%); the remainder were working towards a Master of Science (12.5%). Approximately, 77% were in either their first or their third semester of study; only two (4.2%) had advanced beyond semester five.

Instruments

Participants completed the MTV instrument, in which they were presented with 16 MTVs (four in each condition, sample item in section 2.4) and were asked to rate the statement 'The teacher's reaction is suitable' in relation to each vignette. Ratings were given on a five-point Likert scale ranging from 1 (*I completely disagree*) to 5 (*I completely agree*). The instrument was administered online to both participant groups. Language complexity of vignettes (measured by the LIX-index, Lenhard and Lenhard, 2014-2022) seems comparable for all types of vignettes: $LIX_{UU} = 42.15$ ($SD = 6.35$); $LIX_{US} = 49.49$ ($SD = 3.20$); $LIX_{SU} = 45.94$ ($SD = 8.53$), $LIX_{SS} = 47.60$ ($SD = 6.48$); Kruskal-Wallis-H $= 2.54$, $df = 3$, $p = 0.47$; Bayes factor for ANOVA with H_1: $LIX_{UU} = LIX_{US} = LIX_{SU} = LIX_{SS}$ compared to unconstraint model H_u indicates moderate evidence for equality assumption ($BF_{iu} = 2.94$). Subsequently, each participant provided a self-assessment of their knowledge of both SDT and the teaching of games in PE on a four-point (sample 1) or five-point (sample 2) Likert scale. To compare these ratings between scales, we transformed individual ratings to scores on a scale ranging from 0 (*no knowledge*) to 1 (*advanced knowledge*).

Analyses

To explore whether participants' ratings followed the hypothesized patterns, mean ratings for each vignette were calculated. Next, a mean score was computed for each set of vignettes in the same condition (i.e., UU, SU, US, or SS); these scores can be interpreted as representing the mean rating for vignettes within each condition or cluster. Next, we conducted Welch's t-tests to compare the mean ratings given by participants in samples 1 and 2 for each condition. We were particularly interested in these comparisons for the conditions involving divergent vignettes. Finally, we calculated a Baysian analysis of variance with repeated measures (Gu et al., 2018; Hoijtink et al., 2019): mean scores for conditions represented the within-subject factor with four levels and sample was the between-subject factor with two levels (sample 1 and sample 2) and the following informed hypotheses:

H1: $\mu_{UU1} = \mu_{US1} = \mu_{SU1} = \mu_{SS1}$ (all means are equal in sample 1).

H2: $\mu_{UU2} = \mu_{US2} = \mu_{SU2} = \mu_{SS2}$ (all means are equal in sample 2).

H3: $\mu_{UU1} < \mu_{SU1} < \mu_{US1} < \mu_{SS1}$ (means are ordered with SU being lower than US in sample 1).

H4: $\mu_{UU1} < \mu_{US1} < \mu_{SU1} < \mu_{SS1}$ (means are ordered with US being lower than SU in sample 1).

H5: $\mu_{UU2} < \mu_{SU2} < \mu_{US2} < \mu_{SS2}$ (means are ordered with SU being lower than US in sample 2).

H6: $\mu_{UU2} < \mu_{US2} < \mu_{SU2} < \mu_{SS2}$ (means are ordered with US being lower than SU in sample 2).

H7: $\mu_{UU1} - \mu_{UU2} = \mu_{US1} - \mu_{US2} = \mu_{SU1} - \mu_{SU2} = \mu_{SS1} - \mu_{SS2}$ (mean differences for clusters are equal across samples indicating no interaction effect).

For each hypothesis, we calculated Bayes factors compared to the unconstrained hypotheses using the R-package bain (Hoijtink et al., 2019).

Results

Our research objective in Study 1 was to collect evidence on the validity of our proposed instrument by comparing participants' ratings of the suitability of the behaviors depicted in the convergent and divergent vignettes to our hypotheses regarding the expected pattern of ratings.

Preliminary analysis: Comparison of groups

As access to relevant knowledge was expected to influence evidence-based argumentation *via* its influence on the ability to identify relevant cues, participants were asked to rate their knowledge concerning the content of the vignettes; these ratings are summarized in Table 1. With respect to knowledge of SDT, both samples (i.e., both student teachers and sports science students) gave rather low self-reports (with mean ratings being 0.11 and 0.15, respectively); there was no significant difference between the groups, $t(82) = 0.94$, $p = 0.35$, Cohen's $d = 0.15$. However, as expected, sports science students reported having significantly more knowledge of teaching games ($M = 0.40$, $SD = 0.25$) than did student teachers ($M = 0.06$, $SD = 0.18$), $t(64) = 8.77$, $p < 0.001$, Cohen's $d = 1.69$.

Vignette ratings

Average ratings for each vignette are presented in Table 2. In line with the hypotheses, convergent vignettes of the UU type received the lowest ratings (sample 1: $M = 2.39$, $SD = 0.70$; sample 2: $M = 2.42$, $SD = 0.67$). In other words, both groups were in agreement on their judgments of teacher behaviors which we had constructed to represent unsuitable actions from both perspectives. Participants from each group rated the individual UU vignettes (UU1–UU4) slightly differently, but the groups were approximately in agreement on the overall ordering, with vignette UU3 receiving the lowest overall rating and vignettes UU1 and UU2 receiving the highest ratings within this condition. A similar pattern was observed for convergent vignettes of the SS type,

TABLE 1 Self-reported knowledge of self-determination theory (SDT) and teaching games in physical education.

Type of knowledge	Student teachers			Sports science students			Welch's t-test and effect size			
	N	M	SD	N	M	SD	t	d.f.	p	Cohen's d
SDT	150	0.11	0.21	48	0.15	0.21	0.94	82	0.35	0.15
Teaching games	152	0.06	0.18	48	0.40	0.25	8.77	64	<0.001	1.69

TABLE 2 Vignette ratings: descriptive statistics and group comparisons.

Vignettes*	Student teachers ($N=153$)		Sports science students ($N=48$)		Welch's t-test and effect size			
	M	SD	M	SD	t	d.f.	p	d
UU1	2.95	1.10	2.85	1.05				
UU2	2.91	1.19	3.13	1.05				
UU3	1.56	0.77	1.77	1.07				
UU4	2.12	1.15	1.91	1.01				
	2.39	0.70	2.42	0.67	0.27	81	0.79	0.04
US1	3.71	1.12	3.57	1.15				
US2	3.31	1.35	3.38	1.34				
US3	1.68	0.94	1.42	0.74				
US4	2.71	1.17	2.67	1.02				
	2.84	0.60	2.73	0.51	1.25	91	0.22	0.19
SU1	4.11	1.07	3.93	1.16				
SU2	2.60	1.31	3.30	1.23				
SU3	2.52	1.25	2.65	1.19				
SU4	3.62	1.13	3.56	1.33				
	3.23	0.73	3.35	0.83	0.90	71	0.37	0.16
SS1	4.01	1.15	3.64	1.26				
SS2	4.07	1.02	4.09	1.09				
SS3	3.81	1.15	3.75	1.26				
SS4	3.14	1.27	3.14	1.22				
	3.77	0.71	3.61	0.79	1.25	72	0.21	0.22

*U, unsuitable; S, suitable; first letter indicating sports science perspective, second letter indicating psychological perspective.

which received the highest overall ratings (sample 1: $M = 3.77$, $SD = 0.71$; sample 2: $M = 3.61$, $SD = 0.79$).

The conditions containing divergent vignettes (US and SU) received intermediate ratings from participants in both groups, a result that was also in line with the hypotheses. Additionally, participants in both groups judged the actions in US vignettes ($M_1 = 2.84$; $M_2 = 2.73$) to be slightly less suitable than those in SU vignettes ($M_1 = 3.23$; $M_2 = 3.35$). Within these conditions, the rank order of the suitability of individual vignettes was constant across both groups, although the mean ratings varied.

Repeated measures ANOVA indicated a significant main effect for the within-subject factor "vignette condition" with $F(3) = 119.334$, $p < 0.001$, partial $\eta^2 = 0.377$. Post-hoc comparisons with Bonferroni correction revealed significant differences ($p < 0.001$) between each condition (UU, US, SU, and SS).

Comparison of ratings by student teachers and sports science students

The results of an independent-samples Welch's t-test indicated that there was no significant difference between the two groups in their ratings of UU vignettes. The corresponding effect size (Cohen's $d = 0.04$) indicated that the difference was below the threshold to be considered even a small effect. Similarly, both groups gave comparable judgments in response to the SS vignettes, representing items in which the teacher action was intended to represent a suitable response from both perspectives. There was no statistically significant difference between the ratings given by each group on this condition, and the (statistically insignificant) standardized mean difference (Cohen's $d = 0.22$) was just above the threshold of what is considered to be a small effect.

Concerning the divergent vignette conditions, once again no significant effect of group was observed. A comparison on the descriptive level of the within-group difference between ratings of the SU and US vignettes across groups indicated that there was a larger difference in the case of sports science students, who self-reported having greater knowledge of PE teaching (ΔSU–US $= 0.61$, $SD = 0.87$), compared to pre-service teachers (ΔSU–US $= 0.38$, $SD = 0.88$). However, this difference in differences was not significant, $t(79) = 1.62$, $p = 0.11$, Cohen's $d = 0.27$.

In the repeated measures ANOVA results from the paired Welch-tests could be replicated by a non-significant main effect for sample $[F(1) = 0.158$, $p = 0.692]$. Further, a non-significant interaction between condition and sample $[F(3) = 1.518, p = 0.209]$ leads to the assumption that judgments did not depend on the sample. The results from the frequentist approach were supported by bayesian evaluation of informed hypotheses: Bayes factors indicated strong evidence for H4 (BF$_{iu}$ = 20.98, means are ordered with US being lower than SU in sample 1), H6 (BF$_{iu}$ = 21.52, means are ordered with US being lower than SU in sample 2), and H7 (BF$_{iu}$ = 13.49, mean differences for clusters are equal across samples indicating no interaction effect).

Overall, our results indicated that participants were able to identify relevant cues in the vignettes in judging the suitability of specific teacher actions, and this led to a pattern of ratings that conformed to the hypotheses. However, differences between the two groups in terms of the mean ratings they gave were observed only on the level of individual vignettes, with no differences observed in the groups' average ratings over any of the aggregated conditions (UU, US, SU, or SS). There was a tendency in the case of the divergent vignette conditions towards a difference between the groups, in the hypothesized direction, but this did not reach

statistical significance. This lack of systematic differences between the groups may be attributable to the fact that the participants had not had enough opportunities to build a sufficient knowledge base in their respective fields. To address this issue, we conducted Study 2, in which a specific knowledge intervention was implemented.

Study 2: Intervention

Methods

This study was carried out in the course of a unit of teaching taken by students as part of their degrees in education studies. We developed a short intervention and tested whether this changes how participants perceive the suitability of teacher actions in our MTVs.

Design and sample

To investigate whether knowledge input would change participants' judgments in relation to our MTVs, we employed a pre–post intervention design. At the beginning of the unit, a pre-test including a similar MTV instrument to the one used in Study 1 was administered to participants. After participating in the knowledge intervention, they also completed an identical post-test. The entire procedure, consisting of the pre-test, knowledge intervention, and post-test, took place during the regular 90-min session for delivery of the unit in question.

The 46 participants (72% female) were recruited from a single university specializing in education studies and had mean age of 24.59 years ($SD = 1.88$). All participants had already obtained a bachelor's degree in the field of education, either at the primary (65%) or the secondary (35%) level, and at the time of the study, they were working towards a master's degree in education. Most (approximately 85%) were in the first year of their master's studies; 36 had already completed an obligatory semester of practical teacher training in a school, and 10 had yet to do so. Of the 48 potential participants who initiated participation by beginning the pre-test, 46 (96%) completed the entire pre-test. Following the knowledge intervention, 42 participants began the post-test. Ultimately, full data (i.e., a linked pre-test and post-test) were available for 38 participants.

Intervention

The intervention administered in Study 2 aimed to enhance student teachers' knowledge of SDT, and specifically their understanding of the core principle of basic psychological needs and the ways in which teachers might foster need satisfaction in the classroom setting. The intervention was embedded in a seminar on learning and motivation theories, in the form of a unit which lasted approximately 60 min. The unit began with an overview of SDT, including presentations of cognitive evaluation theory and organismic integration theory as sub-theories (Ryan and Deci, 2000). Subsequently, the focus was on providing an explanation of competence, autonomy, and social relatedness as

basic psychological needs which foster self-determined forms of motivation. The key overall message, therefore, was that teachers design the motivational climate of their classrooms in such a way as to fulfil certain basic needs to variable extents. Finally, based on this theoretical perspective, various possible actions specifically linked to the satisfaction of basic needs were presented. Participants were free to ask questions during the unit and to make comments based on their own ideas or understanding. Questions and comments were handled discursively; nevertheless, this intervention overall can be considered to have been rather directive. The intervention was administered as an online course *via* the platform Zoom.

Instrument

The pre-test and post-test included an online questionnaire with several parts. In addition to collecting several demographic variables (e.g., gender and age), we asked participants to self-rate their knowledge of SDT on a scale from 1 (*no knowledge*) to 6 (*advanced knowledge*). Additionally, we constructed a knowledge test on the topic of SDT, consisting of nine multiple-choice items whose content was directly linked to the content of the knowledge intervention. Participants' responses to each item were coded as 0 = *incorrect* or 1 = *correct*, and overall test scores were calculated by summing these values, resulting in a range of possible test scores from 0 to 9.

The central component of both tests was the MTV instrument. Due to time restrictions, we divided the 16 vignettes into two comparable subsets, each containing eight items consisting of two from each condition (UU, SU, US, and SS). Participants were randomly assigned (*via* random assignment to breakout sessions in Zoom) to complete one of the subsets of items for the pre-test and each completed the same subset again in the post-test phase. In each case, participants rated the teacher's actions described in each vignette on a six-point Likert scale ranging from 1 (*The teacher's action is unsuitable*) to 6 (*The teacher's action is suitable*). The purpose of using a six-point scale, rather than a five-point scale as in Study 1, was to avoid the possibility of participants selecting the midpoint of the scale, thus encouraging them to choose at least a specific direction for their evaluation. We anticipated that this would allow us to observe any changes in their decision-making more clearly.

Analyses

To test our hypothesis that the knowledge intervention would lead to changes in participants' MTV ratings, we first examined whether knowledge gains had occurred using a paired-samples t-test to compare pre- and post-test scores on self-assessed knowledge and knowledge test scores. Subsequently, we computed mean scores for each condition of the vignette instrument (UU, SU, US, and SS) and carried out paired-samples t-tests comparing participants' pre- and post-test judgments for each condition. To quantify the effects, we also computed Cohen's d as a measure of effect size.

Results

Our research objective for Study 2 was to investigate whether an intervention involving knowledge input would alter participants' judgments of the suitability of teacher behaviors in the MTVs.

Effect of the intervention on self-determination theory knowledge

To test for an intervention effect independently of the MTV instrument, we analyzed participants' knowledge gains between the pre- and post-tests on the basis of their self-assessments and the more objective knowledge test. The results indicated that participants made sizeable knowledge gains: average self-assessed knowledge scores increased significantly from 2.31 ($SD = 0.80$) to 3.87 ($SD = 0.89$), $t(38) = 10.71$, $p < 0.001$, Cohen's $d = 1.60$; and average scores on the SDT knowledge test increased from 5.26 ($SD = 1.41$) to 6.46 ($SD = 1.07$), $t(38) = 5.45$, $p < 0.001$, Cohen's $d = 0.80$.

Effect of the intervention on MTV ratings

The intervention aimed to increase participants' knowledge of SDT, and more specifically their understanding of the core principle of the satisfaction of basic needs. Therefore, we expected that, following the intervention, participants would be better able to identify the SDT-related cues included in the MTVs, and thus that they may judge the actions depicted in UU and S_gU_m items to be less suitable, whereas they may judge those depicted in U_gS_m and SS items to be more suitable. Descriptive statistics (Table 3) indicated that participants' ratings of items in all conditions shifted in the expected directions; Figure 1 further illustrates the changes in average ratings and dispersion values between the pre- and post-test.

However, statistical comparisons of the mean differences indicated that these did not reach the level of statistical significance ($p < 0.05$), with the exception of condition UU, $t(37) = 2.30$, $p < 0.05$, Cohen's $d = -0.49$, $BF_{iu} = 1.98$ [weak evidence for H_i: $\mu_{Pre} > \mu_{Post}$]. According to the standardized mean differences (Cohen's d), the changes in rating for the US and SS conditions were rather small ($d = 0.24$ and $d = 0.13$, respectively),

TABLE 3 Vignette ratings in the intervention study and results of comparisons between pre- and post-test ratings.

Vignette*		Pre-test		Post-test		Paired t-test			
	N	M	SD	M	SD	t	d.f.	$p_{\text{one-sided}}$	$d_{\text{RM. pooled}}$
UUA1	19	1.84	0.83	1.74	0.99				
UUA2	19	2.26	1.05	2.16	1.07				
UUB1	19	2.84	1.26	2.42	1.17				
UUB2	19	2.21	1.27	1.79	1.18				
UU	38	2.29	0.89	2.03	0.96	3.02	37	0.01	−0.49
SUA1	19	3.11	1.49	3.05	1.31				
SUA2	19	2.42	1.02	2.58	1.07				
SUB1	20	2.90	1.29	3.05	1.54				
SUB2	19	4.32	1.16	4.11	1.29				
SU	39	3.22	0.98	3.18	1.13	0.25	38	0.40	−0.05
USA1	19	3.95	1.31	4.47	1.35				
USA2	19	3.47	1.26	3.74	1.56				
USB1	19	4.16	1.01	3.84	1.46				
USB2	20	3.40	1.27	3.85	1.31				
US	39	3.76	0.76	3.99	0.98	−1.56	38	0.06	0.24
SSA1	19	5.05	1.03	4.89	1.15				
SSA2	19	4.79	1.03	5.05	0.71				
SSB1	19	4.58	1.35	4.63	1.38				
SSB2	19	4.47	1.65	4.84	1.26				
SS	38	4.72	1.00	4.86	0.98	−0.89	37	0.19	0.13

*U, unsuitable; S, suitable, first letter indicating sports science perspective, second letter indicating psychological perspective, A and B indicate to which subset the respective vignettes belonged.
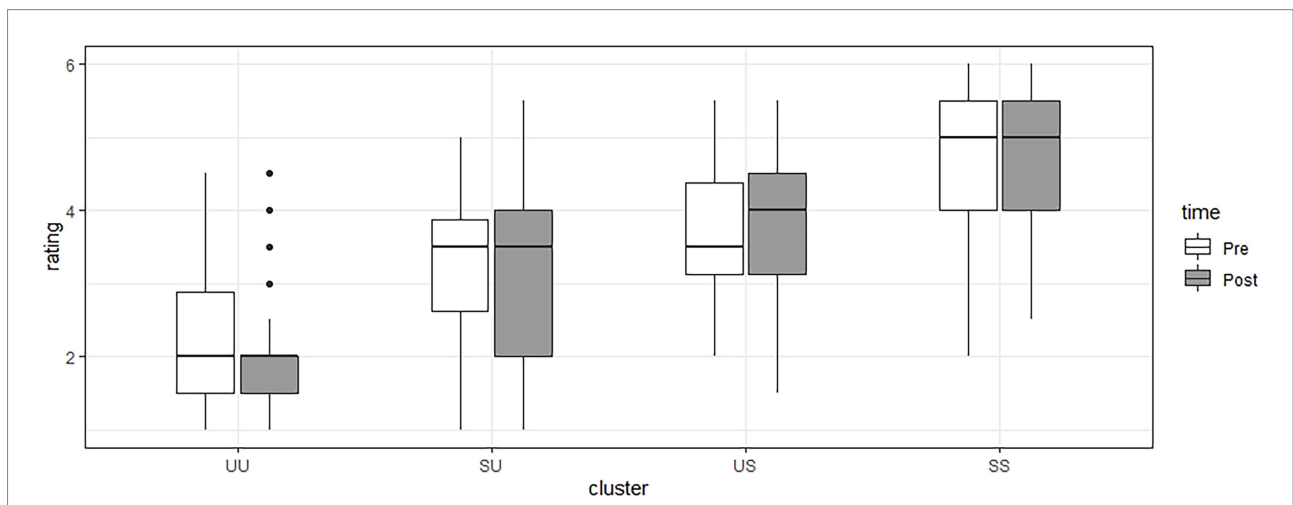

FIGURE 1
Comparisons of mean absolute ratings and mean difference in ratings for each condition between pre- and post-test.

while the change in rating for the SU condition was close to zero ($d = -0.05$). Figure 2 presents an illustration of Bayes factors for H1: $\mu_{Pre} > \mu_{Post}$, H2: $\mu_{Pre} = \mu_{Post}$, and H3: $\mu_{Pre} > \mu_{Post}$ against the unconstraint model.

Discussion

In this study, we tested MTVs as a tool for the presentation, in a measurement instrument, of authentic situations including content cues linked to two different theoretical perspectives. In particular, we constructed vignettes that could be evaluated from the perspective of teaching games in PE and from the perspective of self-determined motivation, which would involve application of the core principles of complexity reduction and need satisfaction, respectively. From each of these perspectives, the specific teacher action depicted in each vignette could be considered to be either suitable or unsuitable, producing four types of vignettes, two convergent and two divergent. In two studies, we demonstrated that participants' judgments of the suitability of the teacher's behaviors varied as expected according to vignette type. Furthermore, a brief knowledge intervention elicited change in participants' judgments between a pre-test and a post-test.
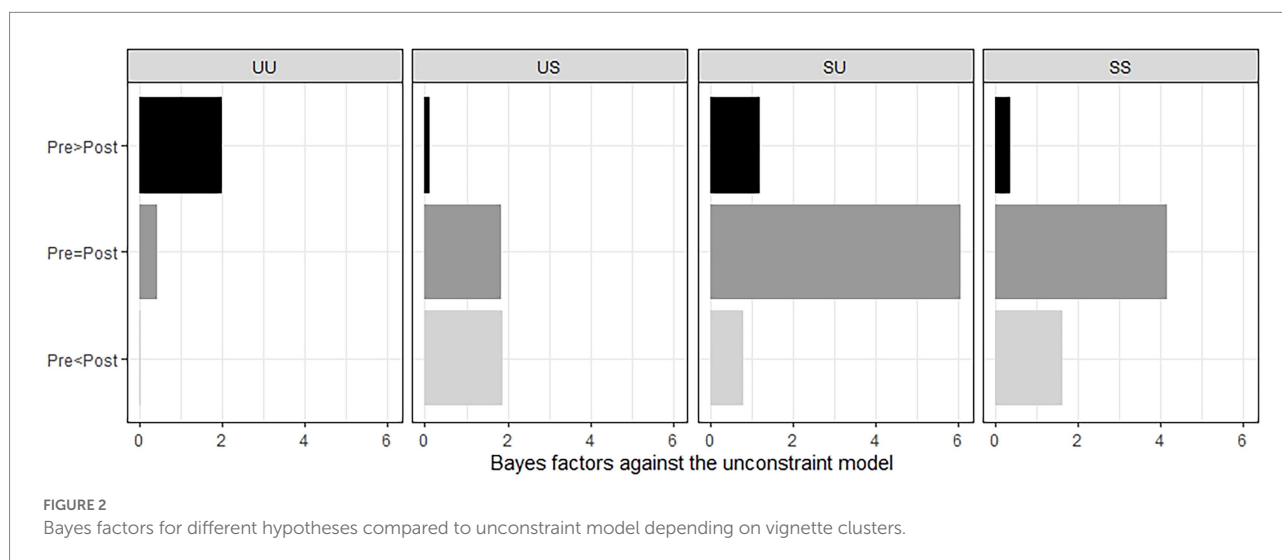
Multi-theory vignettes as a research tool

Evidence-based education as a field of research has gained in importance, with an increasing focus on the process level: that is, on evidence-based argumentation (Csanadi et al., 2021). In this domain, existing research indicates that (pre-service) teachers vary in their approaches to the selection and use of different sources of evidence (Kiemer and Kollar, 2021). Those sources can be considered to vary in their epistemic status (Fenstermacher,

1994) and utility value (Kiemer and Kollar, 2021). However, there is little existing research addressing the selection and use of competing strands of evidence that are of comparable epistemic status.

From the perspective of evidence-based argumentation, participants presented with our MTVs are confronted with a rather weakly-defined scenario: although some contextual information is provided, much other information has to be inferred. Nevertheless, the available information may lead to the generation of different hypotheses (step 3 in the process model proposed by Fischer et al., 2014) and, in the case of divergent MTVs, possible re-evaluation of one's thoughts (step 7). This specific step, in which two closely comparable hypotheses must be evaluated and weighed up to arrive at a decision or the solution to a problem, allows for a deeper exploration of the extent to which teachers make use of their knowledge of specific theories and empirical evidence. Here, a discrepancy between formal logic and participants' response becomes obvious: Formal logic would forecast that vignettes which contain at least one unsuitable action should be rated equally low – irrespective of other likely suitable actions. Instead, participants seem to apply a compensatory approach where suitable actions may compensate for unsuitable ones. In the present study, we further explored the influence of a specific knowledge intervention: specifically, we provided participants with information on a particular scientific theory as a source of evidence. However, future research may use other sources of evidence and explore the extent to which they can influence participants' judgments.

According to the perspective of professional vision as a knowledge-based ability (van Es and Sherin, 2002), participants presented with MTVs notice specific cues in the scenario, which then lead to a reasoning process. Both noticing and reasoning are considered to be knowledge-based processes, which means – drawing on process models of selective attention – that cues in the scenarios can only be noticed if the corresponding knowledge is

FIGURE 2
Bayes factors for different hypotheses compared to unconstraint model depending on vignette clusters.

represented in the cognitive system (see also Loibl et al., 2020). In Study 1, sports science students had higher levels of self-reported knowledge in the domain of teaching games and tended to orient their decisions to the suitability of the teacher's actions from that perspective to a greater extent. Furthermore, in Study 2, the intervention enhancing participants' SDT knowledge elicited an increase in their ratings of the suitability of the teacher's actions in divergent vignettes depicting behaviors which would be considered suitable from an SDT perspective, and a decrease in their ratings of the suitability of actions that would be considered unsuitable from that perspective. However, it remains unclear whether these results can be attributed to changes in their noticing or their reasoning processes. Nevertheless, MTVs may represent a potential tool for further differentiation between these processes, an issue which is seldom addressed in research in the domain of professional vision (Gold and Holodynski, 2017; Meschede et al., 2017).

Limitations

Despite the promising initial results, several limitations to the present study warrant further consideration. First, the participant samples recruited for the pilot (Study 1) were limited in scope. Both groups consisted of university students who were rather inexperienced in their fields of study. Therefore, the ratings they provided may have had a tendency to represent 'common sense' ratings rather than the results of systematic argumentation. Nevertheless, the mean ratings followed the hypothesized pattern. Additionally, we did not explicitly control for content, tone, and sentiments between and within vignette clusters. Further studies are needed to explore the extent to which specific knowledge may contribute to participants' judgments and to explore possible context factors.

Second, the MTVs that we constructed encompassed only two theoretical perspectives, namely teaching games in PE and SDT. Therefore, the results should be replicated across other theoretical perspectives in different content domains. However, we consider the instrument presented here to be a prototype, with reference to which many other MTVs can be generated in accordance with specific research questions. Additionally, the integration of multiple theoretical perspectives allows for research that crosses subject matter domains.

Third, in our analyses we theoretically assume that the judgments with a specific combination are uni-dimensional. Due to different knowledge domains (teaching ball games and self-determination theory) as well as different aspects within each vignette, this assumption could be questioned. However and as first evidence, dimensionality analysis within the structural equation modelling framework revealed that a four-dimensional model (UU, SU, US, and SS) fits better to our data of sample 1 than a one-dimensional model ($\Delta\chi2 = 28.051$, $\Delta df = 6$, $p < 0.001$). Nevertheless, testing the uni-dimensional assumption within each vignette and more complex latent structures across vignettes warrant further attention. Additionally, we calculated mean ratings within each vignette cluster. With this approach, we possibly reduced

heterogeneity on the level of vignette ratings. Future studies might inspect sources for differences on the vignette level and may relate them to knowledge differences, for instance.

Fourth, the results of the knowledge intervention study are limited in scope due to a deficiency in the design: specifically, we did not include a knowledge intervention focusing on knowledge about teaching games in PE (with the core principle of complexity reduction). Under the view we have presented here, we would hypothesize that gaining knowledge in this area would influence participants' judgments in the opposite direction. Furthermore, it would be of interest to investigate the outcome if participants are exposed to both interventions. This approach would allow for a deeper exploration of the processes involved in weighing up multiple hypotheses.

Contributions

Despite the limitations mentioned above, this study demonstrates that it is possible to construct vignettes that manipulate participants' judgments of the suitability of the actions of a teacher which can be considered under a specific theoretical perspective. Although this appears to be a rather research-oriented endeavor, the situations presented in the vignettes are of substantial importance to teachers' everyday experiences in the classroom, where it is very common for them to be confronted with situations in which there is no clearly correct or incorrect response, but rather competing solutions with comparable value. It therefore seems that it would be valuable to obtain further insight into the argumentation processes involved in such situations.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

HL-B wrote the manuscript, project management, and supervision. CB was responsible for vignette development and data collection. TD supervision and editing drafts of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2022.996029/full#supplementary-material

References

Baeten, M., Dochy, F., and Struyven, K. (2013). The effects of different learning environments on students' motivation for learning and their achievement. *Br. J. Educ. Psychol.* 83, 484–501. doi: 10.1111/j.2044-8279.2012.02076.x

Borsboom, D., Mellenbergh, G. J., and van Heerden, J. (2004). The concept of validity. *Psychol. Rev.* 111, 1061–1071. doi: 10.1037/0033-295X.111.4.1061

Bromme, R., Prenzel, M., and Jäger, M. (2014). Empirische Bildungsforschung und evidenzbasierte Bildungspolitik. *Z. Erzieh.* 17, 3–54. doi: 10.1007/s11618-014-0514-5

Brovelli, D., Bölsterli, K., Rehm, M., and Wilhelm, M. (2014). Using vignette testing to measure student science teachers' professional competencies. *Am. J. Educ. Res.* 2, 555–558. doi: 10.12691/education-2-7-20

Brown, C., and Rogers, S. (2015). Knowledge creation as an approach to facilitating evidence informed practice: examining ways to measure the success of using this method with early years practitioners in Camden (London). *J. Educ. Chang.* 16, 79–99. doi: 10.1007/s10833-014-9238-9

Chen, K.-C., and Jang, S.-J. (2010). Motivation in online learning: testing a model of self-determination theory. *Comput. Hum. Behav.* 26, 741–752. doi: 10.1016/j.chb.2010.01.011

Csanadi, A., Kollar, I., and Fischer, F. (2021). Pre-service teachers' evidence-based reasoning during pedagogical problem-solving: better together? *Eur. J. Psychol. Educ.* 36, 147–168. doi: 10.1007/s10212-020-00467-4

Davies, P. (1999). What is evidence-based education? *Br. J. Educ. Stud.* 47, 108–121. doi: 10.1111/1467-8527.00106

Fenstermacher, G. D. (1994). Chapter 1: the knower and the known: the nature of knowledge in research on teaching. *Rev. Res. Educ.* 20, 3–56. doi: 10.3102/0091732X020001003

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., et al. (2014). Scientific reasoning and argumentation: advancing an interdisciplinary research agenda in education. *Frontline Learn. Res.* 2, 28–45. doi: 10.14786/flr.v2i3.96

Floden, R. E., and Buchmann, M. (1993). Between routines and anarchy: preparing teachers for uncertainty. *Oxf. Rev. Educ.* 19, 373–382. doi: 10.1080/0305498930190308

Gold, B., and Holodynski, M. (2017). Using digital video to measure the professional vision of elementary classroom management: test validation and methodological challenges. *Comput. Educ.* 107, 13–30. doi: 10.1016/j.compedu.2016.12.012

Goldman, Z. W., Goodboy, A. K., and Weber, K. (2017). College students' psychological needs and intrinsic motivation to learn: an examination of self-determination theory. *Commun. Q.* 65, 167–191. doi: 10.1080/01463373.2016.1215338

Gu, X., Mulder, J., and Hoijtink, H. (2018). Approximated adjusted fractional Bayes factors: a general method for testing informative hypotheses. *Br. J. Math. Stat. Psychol.* 71, 229–261. doi: 10.1111/bmsp.12110

Hartmann, U., Decristan, J., and Klieme, E. (2016). Unterricht als Feld evidenzbasierter Bildungspraxis? *Z. Erzieh.* 19, 179–199. doi: 10.1007/s11618-016-0712-4

Hetmanek, A., Wecker, C., Kiesewetter, J., Trempler, K., Fischer, M. R., Gräsel, C., et al. (2015). Wozu nutzen Lehrkräfte welche Ressourcen? Eine Interviewstudie zur Schnittstelle zwischen bildungswissenschaftlicher Forschung und professionellem Handeln im Bildungsbereich. *Unterrichtswissenschaft* 43, 194–210.

Hoijtink, H., Gu, X., and Mulder, J. (2019). Bayesian evaluation of informative hypotheses for multiple populations. *Br. J. Math. Stat. Psychol.* 72, 219–243. doi: 10.1111/bmsp.12145

Hoijtink, H., Mulder, J., van Lissa, C., and Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychol. Methods* 24, 539–556. doi: 10.1037/met0000201

Hu, P., and Zhang, J. (2017). A pathway to learner autonomy: a self-determination theory perspective. *Asia Pac. Educ. Rev.* 18, 147–157. doi: 10.1007/s12564-016-9468-z

Jennings, P. A., and Greenberg, M. T. (2009). The prosocial classroom: teacher social and emotional competence in relation to student and classroom outcomes. *Rev. Educ. Res.* 79, 491–525. doi: 10.3102/0034654308325693

Kiemer, K., and Kollar, I. (2021). Source selection and source use as a basis for evidence-informed teaching. *Zeitschrift Für Pädagogische Psychologie* 35, 127–141. doi: 10.1024/1010-0652/a000302

Klahr, D., and Dunbar, K. (1988). Dual space search during scientific reasoning. *Cogn. Sci.* 12, 1–48. doi: 10.1207/s15516709cog1201_1

Kolb, M. (2005). "Sportspiel aus sportpädagogischer Sicht" in *Beiträge zur Lehre und Forschung im Sport: Band 147. Handbuch Sportspiel.* eds. A. Hohmann, M. Kolb and K. Roth (Schorndorf: Hofmann), 65–83.

Lenhard, W., and Lenhard, A. (2014-2022). Berechnung des Lesbarkeitsindex LIX nach Björnson. (Dettelbach: Psychometrica) Available at: http://www.psychometrica.de/lix.hml.

Lenske, G., Wagner, W., Wirth, J., Thillmann, H., Cauet, E., Liepertz, S., et al. (2016). Die Bedeutung der pädagogisch-psychologischen Wissens für die Qualität der Klassenführung und den Lernzuwachs der Schüler/innen im Physikunterricht. *Z. Erzieh.* 19, 211–233. doi: 10.1007/s11618-015-0659-x

Loibl, K., Leuders, T., and Dörfler, T. (2020). A framework for explaining teachers' diagnostic judgments by cognitive Modeling (DiaCoM). *Teach. Teach. Educ.* 91:103059. doi: 10.1016/j.tate.2020.103059

McEown, M. S., Noels, K. A., and Saumure, K. D. (2014). Students' self-determined and integrative orientations and teachers' motivational support in a Japanese as a foreign language context. *System* 45, 227–241. doi: 10.1016/j.system.2014.06.001

Meschede, N., Fiebranz, A., Möller, K., and Steffensky, M. (2017). Teachers' professional vision, pedagogical content knowledge and beliefs: on its relation and differences between pre-service and in-service teachers. *Teach. Teach. Educ.* 66, 158–170. doi: 10.1016/j.tate.2017.04.010

Ryan, R. M., and Deci, E. L. (2000). Self-determination and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.* 55, 68–78. doi: 10.1037/0003-066X.55.1.68

Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B., and Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *BMJ* 312, 71–72. doi: 10.1136/bmj.312.7023.71

Salmi, H., and Thuneberg, H. (2019). The role of self-determination in informal and formal science learning contexts. *Learn. Environ. Res.* 22, 43–63. doi: 10.1007/s10984-018-9266-0

Santagata, R., and Angelici, G. (2010). Studying the impact of the lesson analysis framework on Preservice teachers' abilities to reflect on videos of classroom teaching. *J. Teach. Educ.* 61, 339–349. doi: 10.1177/0022487110369555

Stark, R. (2017). Probleme evidenzbasierter bzw. -orientierter pädagogischer Praxis. *Zeitschrift Für Pädagogische Psychologie* 31, 99–110. doi: 10.1024/1010-0652/a000201

van Es, E., and Sherin, M. G. (2002). Learning to notice: scaffolding new teachers' interpretations of classroom interactions. *J. Technol. Teach. Educ.* 10, 571–596.