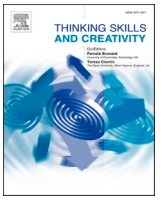
Thinking Skills and Creativity 42 (2021) 100954

Contents lists available at ScienceDirect

Thinking Skills and Creativity

journal homepage: www.elsevier.com/locate/tsc



Measuring primary school teachers' attitudes towards stimulating higher-order thinking (SHOT) in students: Development and validation of the SHOT questionnaire

Frances Wijnen [a,*], Juliette Walma van der Molen [a], Joke Voogt [b]

[a] ELAN, Department of teacher development, University of Twente, 7522 NB Enschede, The Netherlands
[b] Faculty of Social and Behavioural Sciences, University of Amsterdam, 1012 WX, Amsterdam. The Netherlands

ARTICLE INFO

Keywords:
Higher-order thinking
Attitude
Primary school teachers
Validation
Instrument development

ABSTRACT

This paper describes the development and validation of a new instrument to measure primary school teachers' attitudes towards stimulating higher-order thinking in students (SHOT questionnaire). It is believed that it is necessary to explicitly teach students to think, because it cannot be assumed that students will automatically become good thinkers. Therefore, teachers are expected to stimulate students to engage in higher-order thinking. However, we know little about teachers' attitudes towards teaching practices that engage students in higher-order thinking. Therefore, we need a valid and reliable measurement instrument that can be used to measure teachers' attitudes towards stimulating higher-order thinking (SHOT). Hence, we developed the SHOT questionnaire. Based on an earlier literature review, we identified four attitudinal factors that we aimed to measure with the SHOT questionnaire. In addition, we included a scale to measure teachers' behaviour aimed at stimulating higher-order thinking. Results of the exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) with 659 pre- and in-service primary school teachers' show that the requirements for construct validity were met. Furthermore, we found that in-service teachers, who are more positive about the relevance of stimulating higher-order thinking and their ability to do this, encourage students significantly more often to engage in higher-order thinking than pre-service teachers do.

1. Introduction

This paper presents a validation study of a questionnaire to measure primary school teachers' attitudes towards Stimulating Higher-Order Thinking in students (the SHOT questionnaire). Higher-order thinking skills, such as critical thinking, creative thinking, and problem solving are regarded as crucial, already at the primary school level, for students to develop in order to prepare them for their later (working) lives and are therefore mentioned in many models regarding 21st-century learning (OECD, 2018; Voogt & Pareja Roblin, 2012; World Economic Forum, 2016). Moreover, by engaging in higher-order thinking, students actively construct knowledge and engage in meaningful learning (Anderson et al., 2001). With the use of higher-order thinking, students are better able to make sense of what they learn, connect what they learn with previously acquired knowledge and store new knowledge in their long-term

* Corresponding author.
E-mail addresses: f.m.wijnen@utwente.nl (F. Wijnen), j.h.walmavandermolen@utwente.nl (J. Walma van der Molen), J.M.Voogt@uva.nl (J. Voogt).

https://doi.org/10.1016/j.tsc.2021.100954
Received 7 May 2021; Received in revised form 16 July 2021; Accepted 22 September 2021
Available online 24 September 2021

memory. This allows them to apply what they learned in new situations more easily (transfer of knowledge) (Anderson et al., 2001).

It is believed that it is necessary to explicitly teach students to think in such ways, because it cannot be assumed that students will automatically become good thinkers Elder (2003). Therefore, teachers are expected to stimulate students to engage in higher-order thinking. This means that teachers offer assignments in which students use complex cognitive skills (e.g., analysing, evaluating, creating) in order to find a solution or make a decision, prediction, judgement or product. However, primary school teachers mostly engage in teaching practices aimed at stimulating *lower*-order thinking skills, with an emphasis on knowledge transfer and there is little evidence that higher-order thinking is systematically stimulated and assessed in schools (Schulz & FitzPatrick 2016).

Previous work has shown that teachers' *attitude* towards specific teaching practices impact teachers' classroom behaviour (e.g., Stipek, Givvin, Salmon, & MacGyvers, 2001; Van Aalderen-Smeets & Walma van der Molen 2015). However, we know little about teachers' attitudes towards teaching practices that engage students in higher-order thinking, especially when it concerns primary school teachers (Schulz & FitzPatrick 2016). Do teachers believe it is important to stimulate higher-order thinking? Do teachers feel capable when stimulating such thinking? In order to support teachers, it is important that we gain an understanding of teachers' attitudes regarding this teaching behaviour. In order to do that, we need a valid and reliable measurement instrument that can be used to measure teachers' attitudes towards stimulating higher-order thinking. Furthermore, such an instrument can be used to investigate, monitor and evaluate the effects of teacher training programs aimed at developing more positive teacher attitudes. To our knowledge, however, no such instrument yet exists. Therefore, the goal of this study was to develop and validate such an instrument.

2. Theoretical underpinnings

2.1. Focus on attitude

Attitude is "probably one of the most important concepts in psychology" (Gawronski and Bodenhausen, 2007, p. 687), because (1) a person's attitude towards a particular object may impact a person's behaviour towards that object, (2) it influences how a person processes information regarding the attitude-object, and (3) it is possible to influence attitudes with educational efforts (Vogel & Wänke, 2016).

In this study, we used the well-known Theory of Planned Behavior (TPB) developed by Ajzen (1991; 2001) to define attitude. According to Ajzen (2001) "attitude represents a summary evaluation of a psychological object (the 'attitude-object'), captured in such attribute dimensions as good-bad, harmful-beneficial, pleasant-unpleasant, and likeable-dislikeable" (p. 28). The object about which an attitudinal evaluation is made is called an attitude-object. In this study, the attitude-object is 'stimulating higher-order thinking in students'.

Based on the TPB, we view attitude as an 'umbrella-term' consisting of three dimensions that, together, form a persons' attitude. The first dimension, *perceptions of behavioural attributes,* represents beliefs and feelings a person associates with a specific attitude-object. The second dimension, *perceptions of the social norm*, represents a person's perception of the social acceptability of the behaviour. The third dimension, *perceptions of behavioural control*, represents the person's perception of the level of control he/she has about enacting the behaviour. The latter perceptions can refer to external factors (e.g., availability of resources or time) or internal factors (e.g., perceived capability of enacting the behaviour, which is frequently defined as 'self-efficacy' based on Bandura's concept) (Armitage & Conner, 2001).

These three dimensions consist of subcomponents (i.e., attitudinal factors) that are specific for each attitude-object. A person's views with regard to each of these subcomponents may impact that person's *intention* to (not) enact a specific behaviour (Ajzen, 1991; Ajzen & Fishbein, 1980). It is assumed that the stronger an intention, the more likely it is that the person will enact the behaviour. In this context, this implies that the evaluation of the underlying attitudinal factors that constitute primary school teachers' attitudes towards stimulating higher-order thinking in students determines a teachers' intention to engage in teaching activities that are aimed at stimulating higher-order thinking in students.

2.2. Higher-order thinking

Definitions of higher-order thinking vary greatly. Cuban (1984, in Lewis & Smith, 1993) even referred to defining higher-order thinking as a 'conceptual swamp'. Labels such as critical thinking, problem solving, creative thinking, reasoning, metacognition, and reflective thinking are all used to refer to 'higher-order thinking'. Furthermore, perspectives on what higher-order thinking is may differ between disciplines. For example, philosophers generally view thinking as a means to decide what to do or to believe, whereas psychologists are more interested in the process of thinking and how this process can help people make sense of their experiences and surroundings by constructing meaning and imposing structure (Lewis & Smith, 1993; Ten Dam & Volman, 2004).

One of the best-known models addressing higher-order thinking is Bloom's cognitive taxonomy, first published in 1956. In this taxonomy, Bloom and his fellow authors described thinking skills as cognitive processes ranging from relatively simple to more complex. In 2001, Anderson and Krathwohl published a revision of Bloom's taxonomy in which the thinking skills of remembering, understanding, and applying were regarded lower-order thinking skills and analysing, evaluating, and creating were regarded higher-order thinking skills (Anderson et al., 2001).

Another description of higher-order thinking skills is provided by King, Goodson, and Rohani (1998) who define higher-order thinking as a set of skills that:

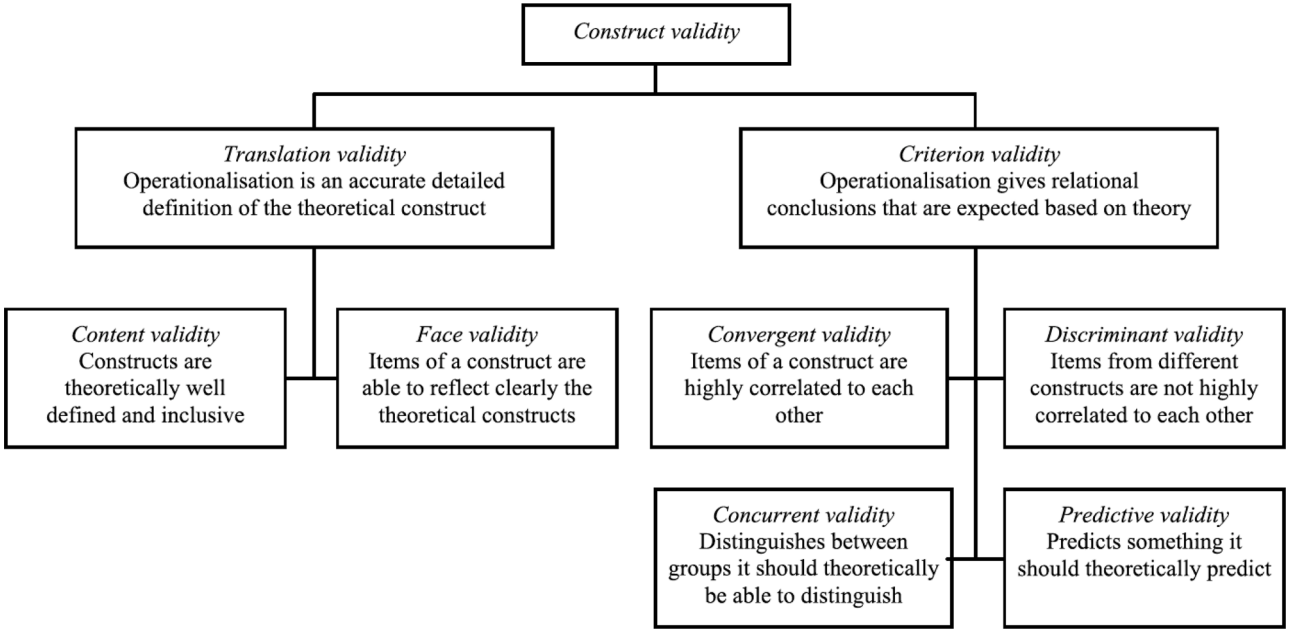


Fig. 1. Trochim and Donnelly's (2006) framework for construct validity.

> "... include critical, logical, reflective, metacognitive, and creative thinking. These skills are activated when individuals encounter unfamiliar problems, uncertainties, questions, or dilemmas. Successful application of the skills results in explanations, decisions, performances, and products that are valid within the context of available knowledge and experience and that promote continued growth in these and other intellectual skills." (p. 1)

In this study, we combined the definitions of King, Goodson, & Rohani (1998) and Bloom's revised taxonomy (Anderson et al., 2001) to define *stimulating* higher-order thinking, which is the attitude-object of this study, as follows: stimulating higher-order thinking in students means offering assignments, questions, problems or dilemmas where students need to use complex cognitive skills (such as analysing, evaluating and creating) in order to find a solution or make a decision, prediction, judgement, or product. Although stimulating higher-order thinking requires pedagogical knowledge in the context of HOT (i.e., knowledge about *how* to stimulate higher-order thinking), it also (and perhaps even more importantly), requires that teachers' have a positive *attitude* towards stimulating higher-order thinking, which is the focus of the measurement instrument that was developed in the current study.

2.3. Teachers' attitudes towards stimulating higher-order thinking

There is some research in which teachers' attitudes towards stimulating higher-order thinking is explored. Based on that research, we learn that primary teachers tend to see the relevance of stimulating students' higher-order thinking (e.g., AlJaafil, & Şahin, 2019; Ketelhut, Millis, Hestness, Cabrera, Plane & McGinnis, 2020; Tornero, 2017). However, not all teachers feel capable to engage students in this type of thinking (e.g., Cheeseman, 2018; Schulz and FitzPatrick, 2016) or they experience a lack of time and/or resources (Cheeseman, 2018; Hamdan Alghamdi & Saud Al-Salouli, 2013). Furthermore, a majority of teachers distinguishes between low- and high-achieving students and believes that low-achieving students are less able to engage in higher-order thinking (e.g., AlJaafil, & Şahin, 2019; Alwadai, 2014; Zohar, Degani, and Vaaknin, 2001). Although these studies have provided us with some insight in teachers' attitudes towards stimulating higher-order thinking, thus far studies have not combined different attitudinal factors, nor do we have a validated measurement instrument to assess teachers' attitudes in this context. Therefore, the goal of the present study was to develop and validate such an instrument.

3. Development of the SHOT questionnaire

We used Trochim and Donnelly's (2006) framework for construct validity to guide the validation of our questionnaire (see Velayutham, Aldridge, & Fraser, 2011 for a detailed description of the application of this framework). This framework (see Fig. 1) shows that an instrument has high construct validity if it can establish content, face, convergent, discriminant, concurrent and predictive validity.

3.1. Establishing translation validity

3.1.1. Content validity

In a previous literature study (Wijnen, Walma van der Molen, & Voogt, submitted for publication), we used the TPB as a guideline to structure the attitudinal factors that were previously studied as pertaining to primary school teachers' attitudes towards stimulating higher-order thinking in students. Based on our review of previous research, we were able to identify four attitudinal factors: perceived relevance, perceived students' ability, self-efficacy and context-dependency. In the present study, we used these four conceptual factors to develop the SHOT questionnaire.

3.1.1.1. Subscales and items. To our knowledge, no instrument yet exists that can be used to measure primary school teachers' attitudes towards stimulating higher-order thinking in students. However, there are some questionnaires that are used to measure teachers' attitudes in other contexts. We explored some of these questionnaires to determine whether they can be used as a basis for the development of the SHOT questionnaire. Van Aalderen-Smeets and Walma van der Molen (2013) developed the DAS-instrument to measure primary school teachers' attitudes towards (teaching) *science*. Three conceptual factors (perceived relevance, self-efficacy and context-dependency) that were developed for the SHOT questionnaire are also included in the DAS questionnaire and the wording of items used to measure these subcomponents seems appropriate. Furthermore, the validity and reliability of the DAS questionnaire has been thoroughly explored (Van Aalderen-Smeets & Walma van der Molen, 2013). We therefore used the DAS-questionnaire as a basis for the development of items for the perceived relevance, self-efficacy and context-dependency scales in the SHOT questionnaire.

Furthermore, Zohar and Schwartzer (2005) developed a scale to measure teachers' beliefs about teaching higher-order thinking to low-achieving students. We evaluated the items used by Zohar & Schwartzer (2005) to determine whether these items might be suitable for measuring the conceptual factor perceived student ability. However, we regarded the reliability (Cronbach's $\alpha = 0.59$) of that used instrument too low and therefore decided to develop new items for this scale. For the development of new survey items, we considered several important criteria, such as item wording, sequence and format Schwarz (2008).

Perceived relevance. Perceived relevance (PR) refers to teachers' beliefs about the importance of stimulating higher-order thinking for students' personal development. We adjusted the items used to measure the scale 'perceived relevance' from the DAS-instrument (Cronbach's $\alpha = 0.85$) for measuring PR in this study. A resulting item is: 'I think it is essential for the development of students to stimulate higher-order thinking'. We designed four items for this scale.

Perceived student ability. Perceived student ability (PSA) refers to teachers' beliefs about whether higher-order thinking is

Table 1
Factor structure for the SHOT questionnaire.

	Item	EFA PR	EFA PSA	EFA SE	EFA CD	CFA PR	CFA PSA	CFA SE	CFA CD
PR1	I think it is essential for the learning of students that they are encouraged to engage in higher-order thinking	0.780				0.892			
PR2	In order to stimulate students' development, I think that you cannot start early enough with offering assignments in which higher-order thinking comes into play	0.714				0.833			
PR3	I think it is essential for the development of students to stimulate higher-order thinking	0.878				0.882			
PR4	I think that stimulating higher-order thinking is so important, that all teachers should do this regularly in their lessons	0.896				0.850			
PSA1	I think that 'smart' students are much better at higher-order thinking than 'weak' students		0.717				0.676		
PSA2	I think that 'weak' students cannot handle assignments that require higher-order thinking		0.726				0.802		
PSA3	I think that most assignments that require higher-order thinking are too difficult for 'weak' students		0.864				0.824		
PSA4	I think that most assignments that require higher-order thinking are frustrating for 'weak' students		0.637				0.824		
PSA5	I think that assignments that require higher-order thinking are more appropriate for 'smart' students than for 'weak' students		0.751				0.802		
PSA6	I think that we cannot expect much higher-order thinking from 'weak' students		0.727				0.781		
SE1	I am well able to pose questions to my students that stimulate higher-order thinking			0.650				0.806	
SE2	I have enough skills to enrich my lessons with higher-order thinking assignments			0.764				0.818	
SE3	I am well able to guide students in doing assignments that stimulate them to engage in higher-order thinking			0.649				0.876	
SE4	I am well able to make-up assignments that stimulate my students to engage in higher-order thinking			0.890				0.820	
CD1	For me, extra time is decisive whether I will stimulate higher-order thinking in my students				0.548				0.623
CD2	For me, making higher-order thinking assignments is only possible when I have a method that describes how to do that				0.633				0.708
CD3	For me, a custom package with sample materials (for example Denksleutels) is conditional for stimulating higher-order thinking in my students				0.652				0.736
CD4	For me, the size of the group determines whether I will stimulate higher-order thinking in my students				0.642				0.493

Notes: a. Exploratory Factor Analysis (EFA) conducted with Maximum Likelihood (ML) estimation and Geomin rotation with subsample A, $N = 327$. Values represent factor loadings.
b. Confirmatory Factor Analysis (CFA) conducted with Maximum Likelihood (ML) estimation and Geomin rotation with subsample B, $N = 332$. Values represent factor loadings.
c. Factor loadings are only displayed for >.35
d. PR = Perceived Relevance, PSA = Perceived Student Ability, SE = Self-efficacy, CD = Context-dependency,
e. The items were originally developed in Dutch. This table presents the English translations.

suitable for both low- and high-achieving students. The items to measure this scale were all newly developed. An example item is: 'I think that most assignments that require higher-order thinking are frustrating for 'weak' students. We designed six items for this scale

Self-efficacy. Self-efficacy (SE) refers to teachers' self-perceived capability to stimulate higher-order thinking in students. We adjusted the items to measure the scale 'self-efficacy' (Cronbach's $\alpha = 0.90$) from the DAS-instrument for measuring SE in this study. A resulting item is: 'I am well able to pose questions to my students that stimulate higher-order thinking'. We designed four items for this scale.

Context-dependency. Context-dependency (CD) refers to teachers' perception that external factors, such as available time, or support are a *prerequisite* for them to be able to stimulate higher-order thinking in students. We adjusted the items to measure the scale 'context-dependency' (Cronbach's $\alpha = 0.74$) from the DAS-instrument for measuring CD in this study. We adjusted these items in order to fit the context of this study and to include external factors that are described as potential obstacles for teachers to stimulate higher-order thinking in students, such as time (e.g., Hamdan Alghamdi & Saud Al-Salouli, 2013) and teacher training (Al-Nouh, Abdul--Kareem, Taqi, 2014). A resulting item is: 'For me, extra time is decisive whether I will stimulate higher-order thinking in my students'. We designed six items for this scale.

Teaching behaviour. In order to determine whether teachers' attitudes towards stimulating higher-order thinking in students, impact teachers' actual teaching behaviour, we added a separate section to measure teachers' self-reported teaching behaviour related to stimulating higher-order thinking in students. This section was used as an outcome measure of the four factors measured with the SHOT questionnaire and contained eight items, such as: 'How often do you design a lesson that explicitly stimulates higher-order

thinking in students?' Responses were given on a 7-point Likert scale: (1) never, (2) a few times a year, (3) once a month, (4) a few times a month, (5) once a week, (6) several times a week, (7) every day (see Table 1).

3.1.1.2. Likert-scale. We chose to develop a Likert-scale instrument. Likert-scale instruments are suitable when items are organized in multiple subscales, they can easily be distributed among a large group of respondents, the questions are easy to answer, and Likert-scales enable parametric testing.

For the attitudinal scales of the SHOT questionnaire, we used a 5-point Likert-scale ranging from strongly disagree to strongly agree. Because it is not always clear how respondents interpret the midpoint of an uneven scale, some argue to only use an even number of response options (Kalton, Roberts & Holt, 1980). However, using an even number of response options forces respondents to make a choice, which may lead to skewed results. Therefore, we used a 5-point Likert-scale, where we did not label the middle response options, but only the two extremes (strongly agree- strongly disagree) thereby presenting the response options as a gliding scale.

3.2. Face validity

In extensive interviews, five primary school teachers evaluated the first version of the SHOT questionnaire (25 items). These teachers evaluated every item on comprehensibility and clarity. Furthermore, they indicated whether they thought the items were appropriate for measuring the underlying constructs and were asked whether any additional items were necessary. Based on these evaluations several changes were made. These changes included the addition of one more item to the scale PSA, which more explicitly addresses the belief that assignments that require students to engage in higher-order thinking are too difficult for 'weak' students, and the addition of two items to the CD subscale. These items represent two other external factors that teachers might perceive as pre-requisites for stimulating higher-order thinking: availability of teacher training, and size of the group of students.

4. Investigating criterion validity

4.1. Respondents

In order to meet the requirements for criterion validity, the questionnaire (28 items) was distributed among a large group of pre- and in-service primary school teachers ($N = 659$) in the Netherlands. This group consisted of 257 in-service primary school teachers and 402 third- and fourth-year pre-service teachers. The respondents were mostly females (79,4%), with a mean age of 30 years (range 18-65, $SD = 12.69$).

4.2. Procedure

The first author visited the primary schools (for in-service teachers) and the teacher education colleges (for pre-service teachers). After a short introduction and obtaining informed consent, respondents were directed to an online version of the questionnaire (84,2%), which they could fill in using their own devices, such as a smartphone, tablet or laptop or were given a paper-and-pencil version (15,8%). It took respondents approximately 10 min to complete the questionnaire. If a respondent did not understand a specific item, the researcher would provide clarification individually. When everyone completed the questionnaire, respondents got the opportunity to ask questions about the questionnaire and research.

In a few cases, it was not possible to agree on a specific time and date for the researcher to visit the school. Therefore, a small number of respondents (approximately 7,5% of the sample) received a link by email redirecting them to the online version of the questionnaire, which they completed on their own.

4.3. Data analysis

For our analyses regarding the construct validity and reliability of the SHOT questionnaire, we used similar steps as Post and Walma van der Molen (2019). We started our data analysis by checking the amount of missing data and calculated the standard deviation of respondents' scores on each item. Then, we used a random sampling procedure to extract two subsamples from the dataset, resulting in subsample A ($n = 327$) and subsample B ($n = 332$). In order to determine whether the subsamples were equivalent regarding the distribution of pre- and in-service teachers and regarding the sex of the respondents, Chi-square test were used. Results indicated that pre- and in-service teachers ($\chi^2 = .426$, $p = .514$) and males and females ($\chi^2 = .505$, $p = .477$) were equally distributed across both subsamples.

To explore discriminant and convergent validity, we used subsample A to conduct an exploratory factor analysis (EFA), with Maximum Likelihood estimation (ML) and, since we expected the subscales to correlate, Geomin oblique rotation was used Field (2009). This approach helped us identify latent factors underlying teachers' attitudes towards stimulating higher-order thinking in students. As a next step, we used subsample B to conduct a confirmatory factor analysis (CFA) with ML estimation. Unlike EFA, CFA allows for testing model fit, that is, how well the observed data fit a pre-defined hypothesized factor structure.

In addition to the factor analyses, we calculated the Average Variance Extracted (AVE), Maximum Shared Variance (MSV), Average Shared Variance (ASV) and Composite Reliability (CR) to further explore the convergent and discriminant validity of the subscales (Carter, 2016; Raykov, 1997). Discriminant validity is satisfactory when AVE is equal or greater than 0.50 and greater than the ASV

Table 2
Factor structure for the teaching behaviour scales

	Item	EFA TA	EFA ES	CFA TA	CFA ES
TA1	How often do you design a lesson that explicitly stimulates higher-order thinking in students?	0.927		0.836	
TA2	How often do you teach a lesson (self-designed or based on a teaching method) that explicitly stimulates higher-order thinking in students?	0.934		0.921	
TA3	How often do you give assignments to your students that require higher-order thinking?	0.720		0.888	
ES1	How often do you encourage your students to find more than one solution for a problem?		0.952		0.929
ES2	How often do you encourage your students to approach a subject from different perspectives (such as suggesting pro and counterarguments)?		0.777		0.851

Notes: a. Exploratory Factor Analysis (EFA) conducted with Maximum Likelihood (ML) estimation and Geomin rotation with subsample A, $N = 327$. Values represent factor loadings.
b. Confirmatory Factor Analysis (CFA) conducted with Maximum Likelihood (ML) estimation and Geomin rotation with subsample B, $N = 332$. Values represent factor loadings.
c. Factor loadings are only displayed for >.35
d. TA = Teaching Activities, ES = Encouraging Students
e. The items were originally developed in Dutch. This table presents the English translations.

and MSV of its factor. For convergent validity, the CR value should be equal or greater than 0.70 and be greater than the AVE of the subscale (Fornell & Larcker, 1981). Furthermore, we explored factor correlations. Results of these additional analyses are reported in Appendix A. For the analyses, we used the MPlus program (Muthén & Muthén, 1998-2015).

5. Results

5.1. Preliminary data checks

We started our data analysis by checking for missing data. In the online version of the questionnaire, we used a 'forced response' option. Therefore, respondents had to answer all items before being able to complete the questionnaire. This was not possible for the paper-and-pencil version of the questionnaire, but there were only a few respondents who did not answer one or more items. The percentage of missing data was 0.2% for subsample A and 0.4% for subsample B. We used the default option for handling missing data in MPlus.

Next, for each item, we calculated the standard deviation and checked whether each response option was used at least once. For a five-point Likert-scale, the standard deviation should approximate 1.0 (Coulson 1992). For the 5-point Likert scale standard deviations ranged from .839 to 1.089. For the 7-point Likert scale (teaching behaviour) standard deviations ranged from 1.387 to 1.873. All response options were used at least once. We concluded that the data were suitable for conducting factor analyses.

5.2. Exploratory factor analysis

We conducted iterative exploratory factor analyses with subsample A. Items with a factor loading below 0.35 or cross loadings less than 0.15 from an item's greatest factor loading (Worthington & Whittaker, 2006) were deleted from further EFA's.

Two items designed to measure context-dependency were deleted due to cross-loadings. The resulting EFA revealed a four-factor structure, as was anticipated (see Table 1). The four factors are: Perceived relevance (eigenvalue 5.487), perceived student ability (eigenvalue 4.404), self-efficacy (eigenvalue 1.915), and context-dependency (eigenvalue 1.169).

5.3. Confirmatory factor analysis

To explore how well the data from subsample B fitted the four-factor structure that we had identified with EFA, we performed a CFA. We used several goodness-of-fit indices to determine model fit. These indices are Standardized Root Mean Square Residual (SRMR), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and the Root Mean Square Error of Approximation (RMSEA). SRMR $\leq$0.08 indicates good fit Prudon (2015), CFI and TLI $\geq$.90 indicate acceptable fit and $\geq$.95 indicate good fit Brown (2006). RMSEA values $\leq$0.08 indicate adequate fit, and $\leq$0.05 indicate good fit Brown (2006). Since the χ^2 has some shortcomings (see Hooper et al., 2008; Prudon, 2015) we only report the χ^2 for completeness. The above thresholds are used as guidelines and not interpreted as strict rules Prudon (2015).

Results of the CFA for the SHOT questionnaire showed acceptable to good fit, SRMR = 0.067, RMSEA = 0.079, $\chi^2 = 389.203$, $df = 129$, $p < 0.001$, CFI = 0.929, TLI = 0.916. Table 1 presents the resulting factor structure of the SHOT questionnaire. As explained above, we calculated CR, AVE, ASV and MSV for each subscale as obtained by CFA as an additional check of convergent and discriminant validity. The results of these analyses indicated sufficient convergent and discriminant validity and are reported in Appendix A. Furthermore, we calculated the factor correlations to explore the relationships between the four attitudinal factors. The correlation matrix can also be found in Appendix A.

Table 3
Regression analyses per variable ($N = 659$).

	B	*SD*	*t*	*p*	R^2
Teacher activities					
Perceived relevance	0.418	0.036	11.709	<0.001	0.175*
Perceived student ability	-0.042	0.043	-0.976	0.329	0.002
Self-efficacy	0.608	0.029	21.126	<0.001	0.369*
Context dependency	-0.218	0.046	-4.790	<0.001	0.048*
Encouraging students					
Perceived relevance	0.577	0.030	19.134	<0.001	0.333*
Perceived student ability	0.071	0.043	1.623	0.103	0.005
Self-efficacy	0.482	0.038	12.573	<0.001	0.305*
Context dependency	0.038	0.048	0.781	0.435	0.001

*Significant at $\alpha < 0.05$

5.4. Teaching Behaviour

We analysed the teaching behaviour scale separately from the attitudinal factors. Before conducting our analyses, we first evaluated the items of the teaching behaviour scale to develop some hypotheses regarding the factor structure. Based on these evaluations we hypothesized that either a one-factor or two-factor structure might be appropriate. A one-factor structure might be appropriate because all items aim to measure the frequency of teachers' behaviour aimed at stimulating higher-order thinking. A two-factor structure might be appropriate because some items are focused on different activities a teacher can undertake to stimulate higher-order thinking (e.g., design a lesson, teach a lesson, give assignments), whereas other items are more focused on encouraging students to engage in different complex thinking processes (e.g., problem solving, creating new products). We therefore analysed both a one-factor and two-factor structure. Since the two-factor structure yielded better results, we will only report the results of the two-factor structure.

We conducted iterative EFA's using subsample A. Three items were deleted due to cross-loadings. Results of the EFA with a two-factor structure showed that three items related to teacher activities (TA) loaded highly together on one factor (eigenvalue 3.523) and two items related to encouraging students (ES) loaded highly together on another factor (eigenvalue 0.920) (see Table 2).

Next, we conducted a CFA using subsample B to further explore the two-factor structure. Results of the CFA are: $\chi^2 = 31.186$, $df = 4$, $p = <0.001$, RMSEA = 0.144, SRMR = 0.028, CFI = 0.977, TLI = 0.942. These results are confusing because CFI and SRMR indicate very good fit, TLI indicates reasonable fit, but RMSEA indicates very poor fit. A possible explanation for these findings could be that this model has small degrees of freedom ($df = 4$) and a relatively small sample size ($n = 327$) was used. According to Kenny, Kaniskan, and McCoach (2015)RMSEA can falsely indicate poor fit in models with small degrees of freedom. This effect is stronger when small sample sizes are used. Shi, DiStefano, Maydeu-Olivares and Lee (2021) recommend relying on SRMR and CFI in such situations. Since the SRMR and CFI indicate very good fit, we conclude that the two-factor structure is appropriate. The resulting factor loadings are reported in Table 2.

The factor encouraging students (ES) consists of two items. It is preferable that a factor is represented by at least three items. However, a factor with two items can be used when the items are strongly correlated with each other (Eisinga, Te Grotenhuis, & Pelzer, 2013) using the Spearman-Brown correlation. The two items for measuring ES were strongly correlated with each other, $\rho = 0.88$. Therefore, we decided to keep this factor. For teaching behaviour, we also calculated the CR, AVE, ASV and MSV for each subscale as obtained by CFA. The results of these analyses also indicated sufficient convergent and discriminant validity. Furthermore, we again calculated the factor correlations. The results of these analyses are also reported in Appendix A.

5.5. Predictive validity

We performed regression analyses (using all data) to investigate whether scores on the attitude subscales have predicative value for teachers' (self-reported) teaching behaviour, measured by the scales teacher activities (TA) and encouraging students (ES).

Results of the regression analyses (see Table 3) show that the attitude factors PR, SE and CD are significant predictors for TA. Furthermore, the attitude factors PR and SE are significant predictors for ES. Interestingly, PSA is not a significant predictor for TA or ES, indicating that teachers' beliefs about the capability of students to engage in higher-order thinking does not significantly impact their teaching behaviour. In addition, only PR and SE explain a reasonable amount of variance in TA and in ES, indicating that these factors have a stronger influence on teachers' teaching behaviour.

5.6. Concurrent validity

In order to evaluate concurrent validity of the SHOT questionnaire, we need to know whether we can expect differences in attitudes between different groups of respondents, based on theory (Trochim & Donnely, 2006). If such differences exist, one should be able to distinguish between these groups based on the scores of the respondents on the questionnaire.

In our study, we include two groups of teachers: pre-service teachers and in-service teachers. This indicates that if previous work shows that pre- and in-service teachers have different attitudes towards stimulating higher-order thinking in students, we should be

Table 4
Unweighted average scores for each attitudinal component.

	pre- and in-service teachers ($N = 659$)		in-service teachers ($N = 257$)		pre-service teachers ($N = 402$)	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
Perceived relevance	3.58	0.961	4.06	0.669	3.28	0.995
Perceived student ability*	2.28	0.737	2.36	0.772	2.22	0.708
Self-efficacy	3.04	0.813	3.21	0.692	2.92	0.864
Context-dependency	2.54	0.740	2.60	0.782	2.50	0.711

* a lower score indicates that the teacher believes HOT is suitable for low-achieving students

Table 5
Univariate post-hoc analyses

	pre-service vs in-service teachers	
	η2	*p*
Perceived relevance	0.159	**<0.001**
Perceived student ability	0.009	0.017
Self-efficacy	0.030	**<0.001**
Context-dependency	0.004	0.102

Note: $N = 402$ for pre-service teachers, $N = 257$ for in-service teachers

able to distinguish between pre- and in-service teachers based on their scores on the SHOT questionnaire.

However, to our knowledge no previous work exists in which pre- and in-service teachers' attitudes towards stimulating higher-order thinking in students are compared in order to evaluate whether they differ in their attitudes. Due to this lack of a theoretical basis, it is not possible to evaluate the concurrent validity of the SHOT questionnaire. However, we will evaluate whether pre- and in-service teachers' scores on the SHOT questionnaire differ significantly from each other. In order to do that, we first need to establish measurement invariance to ensure that the scores of the participants may be compared (i.e., whether both groups interpret the items similarly). We therefore evaluate whether the factor structure, factor loadings and factor intercepts are similar for both groups of respondents.

5.6.1. Measurement invariance

Detailed analyses of measurement invariance are reported in Appendix B. Results of these analyses showed that configural invariance (i.e., similarity of the factor structures) and metric invariance (i.e., similarity of factor loadings) could be established. However scalar invariance (i.e., similarity of factor intercepts) could not be established for item PSA1 ("I think that 'smart' students are much better at higher-order thinking than 'weak' students"). Marsh and Hocevar (1985) suggest that if the noninvariant items constitute only a small portion of the model, group comparisons may still be made, because the noninvariant items are not expected to affect the comparisons in a meaningful degree. We therefore decided to keep this item.

5.6.1.2. Teaching behaviour. Results of the measurement invariance analyses for the teaching behaviour scales indicated that configural and metric invariance could be established. However, scalar invariance could not be established for item TA1 ("How often do you design a lesson that explicitly stimulates higher-order thinking in students?"). Since omission of item TA1 would result in very few items for measuring teaching behaviour, we decided to keep this item. However, we conclude that caution should be used when comparing pre- and -in-service teachers' scores for the teacher activities scale.

5.6.2. Primary school teachers' attitudes towards stimulating higher-order thinking

Although the primary goal of this study was to develop and validate the SHOT questionnaire, the data also provide us with the opportunity to explore potential differences in pre- and in-service teachers' scores. To that end, we calculated the unweighted average scores of pre- and in-service primary teachers on the four attitudinal factors. Table 4 provides an overview of these scores.

The results indicate that pre- and in-service teachers believe stimulating higher-order thinking is relevant to support students in their development, believe that higher-order thinking is appropriate for low-achieving students, feel moderately capable in stimulating higher order thinking, and feel moderately dependent on context-factors.

To explore whether pre- and in-service teachers differ significantly in their attitudes, we performed a MANOVA with 'teacher type' as the between-subject factor, using SPSS version 24. As dependent variables we used the unweighted average scores, which represents the mean score for each participant for every attitudinal factor.

Results of the MANOVA resulted in a significant effect of 'teacher type', Wilks' Lamba $\Lambda = .811$, $F(4, 654) = 38.110$, $p = <0.001$, $\eta 2 = 0.189$. To determine on which factors pre- and in-service teachers score significantly different, we performed several post-hoc univariate analyses. Because we performed multiple ANOVA's, we used a Bonferroni correction to prevent incorrectly rejecting the null hypothesis (Type 1 error). This means that the *p*-value should be below 0.0125 to be significant. The results showed that (1) in-service teachers perceive stimulating higher-order thinking as more relevant and (2) in-service teachers perceive themselves significantly more capable to stimulate higher-order thinking in students (see Table 5).

Table 6
Frequency of stimulating higher-order thinking.

	pre- and in-service teachers ($N = 659$)		in-service teachers ($N = 257$)		pre-service teachers ($N = 402$)	
	N	%	*N*	%	*N*	%
Mean score TA						
1.00-1.99	112	17.0	70	27.2	42	10.5
2.00-2.99	250	38.0	81	31.4	169	42.1
3.00-3.99	122	18.6	40	15.6	82	20.3
4.00-4.99	101	15.3	40	15.6	61	15.2
5.00-5.99	45	6.8	13	5.1	32	7.9
6.00-7.00	24	3.6	11	4.3	13	3.2
Missing	5	0.8	2	0.8	3	0.7
Mean score ES						
1.00-1.99	43	6.6	17	6.7	26	6.5
2.00-2.99	168	25.5	37	14.4	131	32.6
3.00-3.99	98	14.9	37	14.4	61	15.2
4.00-4.99	117	17.7	49	19.1	68	16.9
5.00-5.99	98	14.9	52	20.2	46	11.4
6.00-7.00	129	19.6	62	24.1	67	16.7
Missing	6	0.9	3	1.2	3	0.7
	M	*SD*	*M*	*SD*	*M*	*SD*
Teaching Activities	2.87	1.359	2.74	1.414	2.95	1.319
Encouraging students	3.88	1.728	4.29	1.695	3.61	1.700

Note: 1 = never, 2 = a few times a year, 3 = once a month, 4 = a few times a month, 5 = once a week, 6 = a few times a week, and 7 = every day.

Table A.1
CR, AVE, MSV, ASV and factor correlations for the SHOT-questionnaire ($N = 332$).

	Perceived relevance	Perceived student ability	Self-efficacy	Context-dependency	Teacher Activities	Encouraging students
Perceived relevance		-0.084	0.790*	-0.091	0.473*	0.622*
Perceived student ability			-0.021	0.636*	-0.122*	0.053
Self-efficacy				-0.092	0.579*	0.542*
Context-dependency					-0.230*	0.072
Teacher Activities						0.680*
Composite Reliability (CR)	0.90	0.81	0.90	0.73	0.91	0.88
Average Variance Extracted (AVE)	0.85	0.72	0.80	0.61	0.87	0.88
Average Shared Variance (ASV)	0.21	0.14	0.21	0.14	0.46	0.46
Maximum Shared Variance (MSV)	0.62	0.40	0.62	0.40	0.46	0.46

* Factor correlation is significant at $p = <0.05$

Table B.1
Measurement invariance analysis for the attitudinal factors.

	χ^2	*df*	CFI	RMSEA	SRMR	$\triangle \chi^2$	*p*	△CFI	△RMSEA	△SRMR
Configural model	776.43	258	0.93	0.078	0.069	-	-	-	-	-
Metric model	812.79	272	0.93	0.078	0.078	36.36	<0.001	-0.00	0.000	0.009
Scalar model	953.80	286	0.91	0.084	0.085	141.01	<0.001	-0.02	0.006	0.007
Excluding PSA item 1										
Configural model	597.261	196	0.94	0.079	0.068	-	-	-	-	-
Metric model	632.38	208	0.93	0.079	0.078	35.11	<0.001	-0.01	0.000	0.010
Scalar model	711.31	220	0.92	0.082	0.083	78.934	<0.001	-0.01	0.003	0.005

Groups compared: pre-service teachers ($N = 402$) vs. in-service teachers ($N = 257$).

5.6.3. Teaching behaviour

On average, both pre- and in-service teachers engage in teaching activities aimed at stimulating higher-order thinking a bit more often than a few times a year ($M = 2.87$, $SD = 1.359$). Furthermore, they encourage students in higher-order thinking more often than once a month ($M = 3.88$ $SD = 1.728$). Table 6 gives an overview of how often teachers engage in teaching activities and how often they encourage students to engage in higher-order thinking.

In order to explore whether pre- and in-service teachers differ significantly in how often they stimulate higher-order thinking in

Table B.2
Measurement invariance analysis for teaching behaviour.

	χ^2	*df*	CFI	RMSEA	SRMR	$\triangle \chi^2$	*p*	△CFI	△RMSEA	△SRMR
Configural model	8.649	32	1.000	0.016	0.020	-	-	-	-	-
Metric model	9.304	29	1.000	0.000	0.023	0.655	0.884	0.000	0.016	0.003
Scalar model	18.060	26	0.998	0.030	0.036	8.756	0.033	0.002	0.030	0.013
Excluding TA1										
Configural model	1.618	26	1.000	0.000	0.004	-	-	-	-	-
Metric model	2.048	24	1.000	0.000	0.009	0.429	0.807	0.000	0.000	0.005
Scalar model	2.162	22	1.000	0.000	0.008	0.115	0.944	0.000	0.000	-0.001

students, we performed a MANOVA with 'teacher type' as the between-subjects variable. Results of the MANOVA resulted in a significant effect of 'teacher type', Wilks' Lamba $\Lambda = .908$, $F(2, 650) = 32.881$, $p = <0.001$, $\eta 2 = 0.092$. In order to determine on what factor pre- and in-service teachers score significantly different, we performed a post-hoc univariate analysis. Results showed that in-service teachers encourage students significantly more often to engage in higher-order thinking than pre-service teachers ($F(1, 651) = 24.642$, $p = <0.001$, $\eta 2 = 0.36$). There was no significant difference between pre- and-in-service teachers for how often they engage in teaching activities ($p = 0.055$, $\eta 2 = 0.006$).

6. Discussion

The goal of this study was to develop and establish the validity and internal consistency of the SHOT questionnaire. In an earlier conducted literature review (Wijnen, Walma van der Molen, & Voogt, submitted for publication), we used the TPB to extract four attitudinal factors that we aim to measure with the SHOT questionnaire. Results of our analyses showed that the construct validity for these four attitudinal factors could be supported.

Using the TPB to explore factors that may influence teachers' behaviour aimed at stimulating children's thinking is not entirely new. In a previous study, Lee (2018) used the TPB and Social Cognitive Theory from Bandura, to gain insight into what educators should do to help students become critical thinkers. Lee (2018) found, among other things, that positive (student) attitudes should be promoted by recognizing the positive consequences of critical thinking. Lee (2018) proposed to implement instructional strategies, such as role playing, student-initiated storytelling and problem-based learning. However, although the study by Lee provides suggestions for what teachers can do to stimulate critical thinking in students, it provides little insight into what factors account for teachers' *attitudes* towards stimulating higher-order thinking. After all, according to the TPB, a positive attitude towards stimulating higher-order thinking predicts teachers' intention to act accordingly. Therefore, we argue that insight in teachers' attitudes is important in training and coaching teachers in using the kind of instructional strategies Lee proposes. Furthermore, the SHOT questionnaire can be used to investigate, monitor and evaluate the effects of teacher training programs aimed at developing more positive teacher attitudes.

In the present study, we wished to explore whether we were able to measure the four attitudinal factors (PR, PSA, SE and CD) that we identified in our earlier conducted literature review. However, in that literature review we found that the amount of work on teachers' attitudes towards stimulating higher-order thinking was limited. It is therefore possible that other attitudinal factors might also be relevant when exploring teachers' attitudes towards stimulating higher-order thinking in students. Based on the TPB (Ajzen, 1991; 2001), we expect that such factors might include affective components (e.g., anxiety or enjoyment), teachers' views of the social norm regarding stimulating higher-order thinking (does a teacher believe that stimulating higher-order thinking is appreciated by important others, such as colleagues?), and teachers' perceived difficulty (does a teacher believe it is difficult or easy to engage students in higher-order thinking?). Inclusion of additional factors might also help in explaining more of the variance in the TA and ES scales. In the present study, we found that only PR and SE explain a reasonable amount of variance (R^2 for TA ranges between 0.175–0.369, R^2 for ES ranges between 0.305–0.333), but there is still variance that could not be explained by the four factors. We therefore recommend that researchers explore whether inclusion of additional factors to the SHOT questionnaire helps us better understand teachers' behaviour aimed at stimulating higher-order thinking.

As a starting point, we designed several potential items related to the affective component, the perceived social norm component, and the perceived difficulty component, which might be included in the SHOT questionnaire (see Appendix C). Furthermore, the teaching behaviour scale 'encouraging students' presently consists of two items only. It is preferable that a factor consists of at least three items (Eisinga et al., 2013) and we therefore propose another six items that might be used to expand this factor (see Appendix D).

In line with previous work (AlJaafil, & Şahin, 2019; Ketelhut et al., 2020; Tornero, 2017), we found that both pre- and in-service primary teachers think it is important to stimulate higher-order thinking in students. Also, in line with previous work, we found that most teachers do not feel capable to stimulate higher-order thinking. In addition, we found that both pre- and in-service teachers mostly agree that higher-order thinking is also suitable for low-achieving students. The latter finding is in contrast with previous work, which shows that most teachers believe that low-achieving students are unable to engage in higher-order thinking (e.g., Zohar et al., 2001). There could be a number of reasons for the differing results, such as differences in item wording or the fact that we did not measure how our respondents define 'smart' and 'weak' students. Further research, such as in-depth interviews might shed more light on this issue.

We found that the factors PR, SE and CD have a significant impact on teaching behaviour. This is valuable because, although

teachers' perceptions regarding the relevance of stimulating higher-order thinking, teachers' perceived ability to do this and teachers' perceived dependency on context-factors to stimulate higher-order thinking were explored in previous studies (e.g., Akinoglu & Karsantik, 2016; Baysal et al., 2010; AlJaafil, & Şahin, 2019; Cheeseman, 2018; Ketelhut et al., 2020), none of these studies investigated to what extent these factors impact actual teaching behaviour aimed at stimulating higher-order thinking.

Based on the TPB, we expected that a more positive attitude towards stimulating higher-order thinking would result in a higher frequency of teaching behaviour. In this study, we found that both pre- and in-service teachers are somewhat positive about stimulating higher-order thinking. Furthermore, we found, in line with the TPB, that in-service teachers, who are more positive about the relevance of stimulating higher-order thinking and their perceived ability to do this, encourage students significantly more often to engage in higher-order thinking than pre-service teachers do, but still not very often (approximately a few times a month).

One could argue that pre- and in-service teachers differ in the amount of overall teaching that they engage in. Unfortunately, we did not collect data on how much time the participating pre- and in-service teachers spent on teaching. However, we do know that in the Netherlands approximately 55% of in-service teachers work parttime (15% works less than 2,5 days and 40% works 2,5–4 days a week; Onderwijs, 2020). On the other hand, third- and fourth-year pre-service teachers teach at least one day a week with a maximum of 3 days a week. We therefore expect no large differences in teaching time that could account for the finding that in-service teachers stimulate higher-order thinking more often than pre-service teachers.

To conclude, the results of the current study show that the requirements for construct validity of the SHOT questionnaire could all be met. Although we carefully translated the items from Dutch to English for this article, cross-cultural validation should determine the construct validity of the SHOT questionnaire in different countries. In addition, we encourage researchers to extent the SHOT questionnaire with the additional factors that we suggested.

As a next step, we intend to explore whether the SHOT questionnaire can be used to identify different groups of teachers, based on their attitudes towards stimulating higher-order thinking in students. Identifying such groups would allow us to develop teacher training programs to support different groups of teachers in stimulating higher-order thinking in students .

CRediT authorship contribution statement

Frances Wijnen: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Visualization, Project administration, Funding acquisition. **Juliette Walma van der Molen:** Conceptualization, Methodology, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Joke Voogt:** Conceptualization, Methodology, Writing – review & editing, Visualization, Supervision, Funding acquisition.

Acknowledgements

The research described in this paper was funded by TechYourFuture, Centre of Expertise.

Appendix A: CR, AVE, MSV, ASV and factor correlations, for the SHOT-questionnaire

A.1. CR, AVE, MSV, and ASV for the attitudinal factors

We calculated CR, AVE, ASV, and MSV for each attitudinal subscale as obtained by CFA. The composite reliability scores indicated good internal consistency of the subscales. Scores for all factors were above the threshold of .70 and exceeding the AVE scores of each respective subscale. Furthermore, AVE scores indicated good discriminant power of the subscales, since all were ≥ 0.5 and exceeded the ASV and MSV scores for each respective subscale (see Table A.1).

A.2. CR, AVE, MSV, and ASV for the teaching behaviour scales

We calculated the AVE and CR score and explored the shared variance of the TA and ES scales. For both scales the AVE and CR scores were above the thresholds of 0.5 and 0.7 respectively (see Table A.1.), indicating good discriminating power and internal consistency of the subscales.

A.3. Factor correlations

We calculated the correlations to gain insight in the relationships between the four attitudinal factors. The correlation matrix (Table A.1.) shows that there is a significant positive correlation between perceived relevance and self-efficacy. It is possible that teachers who consider it important to stimulate higher-order thinking in students are more inclined to do this and because they have more experience have more positive views regarding their competency to engage students in higher-order thinking processes.

Furthermore, there is a significant positive correlation between perceived student ability and context-dependency. A higher score on perceived student ability indicates that teachers believe that higher-order thinking is not suitable for 'weak' students. This significant correlation might indicate that teachers who are unsure about whether higher-order thinking is suitable for 'weak' students, feel that they need additional support in order to be well able to engage students in higher-order thinking.

In addition, all four attitudinal factors (PR, PSA, SE and CD) are significantly correlated with the behavioural scale TA. This might indicate that these factors impact teachers' engagement in teaching activities aimed at stimulating higher-order thinking in students.

Also, the factors PR and SE are significantly correlated with the behavioural scale ES. This might indicate that teachers' beliefs about the importance of stimulating higher-order thinking and their perceived ability to do this impact teachers' behaviour to engage students in different higher-order thinking processes.

Appendix B: Detailed analyses of measurement invariance

B.1. Measurement invariance for the attitudinal factors

In order to establish measurement invariance, we conducted a multiple-group CFA (for a more in-depth explanation of measurement invariance, see Chen, 2007) including the factors PR, PSA, SE and CD. We explored configural invariance (i.e., similarity of the factor structures), metric invariance (i.e., similarity of factor loadings) and scalar invariance (i.e., similarity of factor intercepts) using all data ($N = 659$).

In order to determine invariance, differences in χ^2 (i.e., $\triangle\chi^{2)}$ between the different measurement models (configural, metric, and scalar) are explored. However, because χ^2 is sensitive to sample size and model assumptions (e.g., linearity, multivariate normality) (Byrne, Shavelson, & Muthén, 1989), we only report $\triangle\chi^2$ for completeness. Instead, we use $\triangle$CFI (Cheung & Rensvold, 2002), $\triangle$RMSEA and $\triangle$SRMR (Cheng, 2007).

For establishing metric invariance, a $\triangle$CFI of $\leq$-0.01, a $\triangle$RMSEA of $\leq$0.015 and a $\triangle$SRMR of $\leq$0.03 indicate invariance. For scalar invariance, $\triangle$SRMR should be $\leq$0.01. In order to establish metric invariance, fit indices of the metric model were compared with fit indices of the configural model. In order to establish scalar invariance, fit indices of the scalar model were compared with fit indices of the metric model.

As can be seen from Table B.1, fit indices for the configural model indicate acceptable fit. Furthermore, the $\triangle$-indices for the metric model were below the recommended thresholds, indicating metric invariance. However, $\triangle$CFI for the comparison between the metric and scalar model exceeded the recommended threshold of -0.01 and therefore scalar invariance could not be established.

In order to determine whether scalar invariance could not be established for a specific factor, several measurement invariance analyses were conducted. Results show that there was no invariance for the factor perceived student ability. More specifically, results showed that there was no invariance for the item 'I think that 'smart' students are much better at higher-order thinking than 'weak' students. Excluding this item resulted in scalar invariance. Since, Marsh and Hocevar (1985) suggest that if the noninvariant items constitute only a small portion of the model, then group comparisons may still be made, because the noninvariant items were not expected to affect the comparisons in a meaningful degree, we decided to keep this item.

B.2. Measurement invariance for the teaching behaviour scales

In order to test measurement invariance for teachers' behaviour aimed at stimulating higher-order thinking, we performed another multiple-group CFA for the TA and ES scales. Results, presented in Table B.2, show that configural invariance could be established. Furthermore, all indices indicate metric invariance, except $\triangle$RMSEA, which is slightly above the threshold of 0.015. Scalar invariance could not be established since both $\triangle$RMSEA and $\triangle$SRMR are above the recommended thresholds.

In order to explore whether scalar invariance could not be established for a specific item, multiple measurement invariance analyses were conducted. Results showed that there was no invariance for item TA1. Since omission of item TA1 would result in very few items for measuring teaching behaviour we decided to keep this item. However, caution should be used when comparing pre- and -in-service teachers' scores for the teacher activities scale.

Appendix C: Potential additional scales and items to expand the SHOT questionnaire

Enjoyment

- I feel enthusiastic when stimulating higher-order thinking in my students
- I feel satisfied when I stimulate higher-order thinking in my students.
- I enjoy stimulating higher-order thinking in my students
- I feel pleasure when I see my students engaging in higher-order thinking

Anxiety

- I get a sinking feeling when I have to design an assignment to engage my students in higher-order thinking
- I feel nervous about creating assignments to engage my students in higher-order thinking
- I feel anxious when I have to guide students through higher-order thinking assignments
- I feel tense when students pose unexpected questions
- I feel stressed about evaluating open ended solutions that my students come up with.

Social norm

- I think that my colleagues believe it is essential to engage students in higher-order thinking

- I think that my colleagues appreciate it when I design assignments to engage my students in higher-order thinking
- I think that, at our school, stimulating higher-order thinking in students is viewed as important
- I think that, at our school, it is expected of me that I engage my students in higher-order thinking

Perceived difficulty

- I think it is very difficult to pose questions that stimulate higher-order thinking in my students
- I think it requires a lot of effort to design assignments that stimulate my students to engage in higher-order thinking
- I think it is difficult to coach students when they engage in higher-order thinking

Appendix D: Potential additional items to expand the 'Encouraging Students' scale

Suggested additional items to the scale Encouraging Students

- How often do you encourage your students to approach a subject from different perspectives (such as suggesting pro and counterarguments)?
- How often do you encourage your students to find more than one solution for a problem?
- How often do you encourage your students compare different results?
- How often do you encourage your students to explain a finding?
- How often do you encourage your students to analyse information from different sources?
- How often do you encourage your students to evaluate a solution to a problem?

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179–211.

Ajzen, I. (2001). Nature and operation of attitudes. *Annual Review of Psychology, 52*, 27–58.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood-Cliffs, NJ: Prentice-Hall.

Akinoglu, O., & Karsantik, Y. (2016). Pre-service teachers' opinions on teaching thinking skills. *International Journal of Instruction, 9*(2), 61–76. https://doi.org/10.12973/iji.2016.925a

AlJaafil, E., & Şahin, M. (2019). Critical thinking skills for primary education: The case in Lebanon. *Turquoise International Journal of Educational Research and Social Studies, 1*(1), 1–7.

Al-Nouh, N. A., Abdul-Kareem, M. M., & Taqi, H. A. (2014). Primary school EFL teachers' attitudes towards creativity and their perceptions of practice. *English Language Teaching, 7*(9), 74–90. https://doi.org/10.5539/elt.v7n9p74

Alwadai, M. A. (2014). Islamic teachers' perceptions of improving critical thinking skills in Saudi Arabian elementary schools. *Journal of Education and Learning, 3*(4), 37–48. https://doi.org/10.5539/jel.v3n4p37

Anderson, L. W., Krathwohl, D., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., Raths, J., & Wittrock, M. (2001). *A taxonomy for learning, teaching and assessing*. United States: Longman.

Armitage, C. J., & Conner, M. (2001). Efficacy of the theory of planned behaviour: A meta-analytic review. *British Journal of Social Psychology, 40*, 471–499. https://doi.org/10.1348/014466601164939

Baysal, Z. N., Arkan, K., & Yildirim, A. (2010). Preservice elementary teachers' perceptions of their self-efficacy in teaching thinking skills. *Procedia - Social and Behavioral Sciences, 2*, 4250–4254. https://doi.org/10.1016/j.sbspro.2010.03.673

Brown, T. (2006). *Confirmatory factor analysis for applied research*. New York: The Guilford Press.

Byrne, B., Shavelson, R., & Muthén, B. (1989). Testing for equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456–466.

Carter, C. R. (2016). Using confirmatory factor analysis to manage discriminant validity issues in social pharmacy research. *International Journal of Clinical Pharmace, 38*, 731–737.

Cheeseman, J. (2018). Teachers' perceptions of obstacles to incorporating a problem solving style of mathematics into their teaching. In J Hunter, P Perger, & L Darragh (Eds.), *Proceedings of the 41st annual conference of the mathematics education research group of Australasia* (pp. 210–217). Auckland: Mathematics Education Research Group of Australasia.

Chen, F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*, 464–504.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255.

Coulson, R. (1992). Development of an instrument for measuring attitudes of early childhood educators towards science. *Research in Science Education, 22*, 101–105.

Eisinga, R., Te Grotenhuis, M., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach or Spearman-Brown? *International Journal of Public Health, 58*, 637–642.

Elder, L. (2003). *Teacher's manual: The miniature guide to critical thinking for children*. Dillon Beach, CA: Foundation for Critical Thinking Press.

Field, A. (2009). *Discovering statistics using SPSS*. New Delhi: Sage publications.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*, 39–50.

Gawronski, B., & Bodenhausen, G. V. (2007). Unraveling the process underlying evaluation: Attitudes from the perspective of the APE model. *Social Cognition, 25*, 687–717.

Hamdan Alghamdi, A. K., & Saud Al-Salouli,, M (2013). Saudi elementary school science teachers' beliefs: Teaching science in the new millennium. *International Journal of Science and Mathematics Education, 11*, 501–525.

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*, 53–60.

Kalton, G., Roberts, J., & Holt, D. (1980). The effects of offering a middle response option with opinion questions. *Journal of the Royal Statistical Society, 29*(1), 65–78.

Kenny, D., Kaniskan, B., & McCoach, D. (2015). The perfomance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research, 44*, 486–507.

Ketelhut, D. J., Mills, K., Hestness, E., Cabrera, L., Plane, J., & McGinnis, J. R. (2020). Teacher change following a professional development experience in integrating computational thinking into elementary science. *Journal of Science Education and Technology, 29*, 174–188.

King, F. J., Goodson, L., & Rohani, F. (1998). Higher order thinking skills: Definition, teaching strategies. *assessment. Center for Advancement of Learning and Assessment. Retrieved from*. http://www.cala.fsu.edu/files/higher_order_thinking_skills.pdf.

Lee, Y. L. (2018). Nurturing critical thinking for implementation beyond the classroom: Implications from social psychological theories of behavior change. *Thinking Skills and Creativity, 27*, 139–146.

Lewis, A., & Smith, D. (1993). Defining higher order thinking. *Theory Into Practice, 32*(3), 131–137.

Marsh, H., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First- and higher order factor models and their invariance across groups. *Psychological Bulletin, 97*(3), 562–582.

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus User's Guide. Seventh Edition*. Los Angeles, CA: Muthén & Muthén.

OECD. (2018). The future of education and skills: Education 2030. Retrieved from: http://www.oecd.org/education/2030/E2030 Position Paper (05.04.2018).pdf.

Onderwijs, Cultuur en Wetenschap [OCW]. (2020). Trendrapportage Arbeidsmarkt Leraren po, vo en mbo 2020. Retrieved from: https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/kamerstukken/2020/12/09/trendrapportage-arbeidsmarkt-leraren-po-vo-en-mbo-2020/trendrapportage-arbeidsmarkt-leraren-po-vo-en-mbo-2020.pdf.

Post, T., & Walma van der Molen, J. (2019). Development and validation of a questionnaire to measure primary school children's images of and attitudes towards curiosity (the CIAC questionnaire). *Motivation and Emotion, 43*, 159–178.

Prudon, P. (2015). Confirmatory factor analysis as a tool in research using questionnaires: A critique. *Comprehensive Psychology, 4*, 1–19.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*, 173–184.

Schulz, H. W., & FitzPatrick, B. (2016). Teachers' understandings of critical and higher order thinking and what this means for their teaching and assessments. *Alberta Journal of Educational Research, 62*(1), 61–86.

Schwarz, N. (2008). Attitude measurement. In W. Crano, & R. Prislin (Eds.), *Attitudes and attitude change* (pp. 41–60). Philadelphia, PA: Psychology Press.

Shi, D., DiStefano, C., Maydeu-Olivares, A., & Lee, T. (2021). Evaluating SEM model fit with small degrees of freedom. *Multivariate Behavioral Research*. https://doi.org/10.1080/00273171.2020.1868965

Stipek, D. J., Givvin, K., Salmon, J., & MacGyver, V. (2001). Teachers' beliefs and practices related to mathematics instruction. *Teaching and Teacher Education, 17*, 213–226.

Ten Dam, G., & Volman, M. L. L. (2004). Critical thinking as a citizenship competence: teaching strategies (IF: 1.3). *Learning and Instruction, 14*(4), 359–379. https://doi.org/10.1016/j.learninstruc.2004.01.005

Tornero, B. (2017). Prospective primary teachers' perceptions about their ability to use and teach thinking skills. *Journal of Constructivist Psychology, 30*(2), 127–145. https://doi.org/10.1080/10720537.2015.1120693

Trochim, W. M., & Donnelly, J. P. (2006). *The research methods knowledge base* (3rd ed.). Cincinnati, OH: Atomic Dog.

Van Aalderen-Smeets, S. I., & Walma van der Molen, J. H. (2015). Improving primary teachers' attitudes toward science by attitude-focused professional development. *Journal of Research in Science Teaching, 52*(5), 710–734. https://doi.org/10.1002/tea.21218

Van Aalderen-Smeets, S., & Walma van der Molen, J. (2013). Measuring primary teachers' attitudes towards teaching science: Development of the dimensions of attitude towards science (DAS) instrument. *International Journal of Science Education, 35*, 577–600.

Velayutham, S., Aldridge, J., & Fraser, B. (2011). Development and validation of an instrument to measure students' motivation and self-regulation in science learning. *International Journal of Science Education, 33*(15), 2159–2179.

Vogel, T., & Wänke, M. (2016). *Attitudes and attitude change*. New York, NY: Routledge.

Voogt, J., & Pareja Roblin, N. (2012). A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. *Journal of Curriculum Studies, 44*(3), 299–321. https://doi.org/10.1080/00220272.2012.668938

World Economic Forum. (2016). New vision for education: Fostering social and emotional learning through technology. Retrieved from http://www3.weforum.org/docs/WEF_New_Vision_for_Education.pdf.

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist, 34*, 806–838.

Wijnen, F.M., Walma van der Molen, J.H., & Voogt, J.M. Primary school teachers' attitudes towards technology and stimulating higher-order thinking in students: A review of the literature. (Paper submitted for publication).

Zohar, A., Degani, A., & Vaaknin, E. (2001). Teachers' beliefs about low-achieving students and higher-order thinking. *Teaching and Teacher Education, 17*, 469–485.

Zohar, A., & Schwartzer, N. (2005). Assessing teachers' pedagogical knowledge in the context of teaching higher-order thinking. *International Journal of Science Education, 27*, 1595–1620. https://doi.org/10.1080/09500690500186592