RedFame

Journal of Education and Training Studies
Vol. 1, No. 2; October 2013
ISSN 2324-805X E-ISSN 2324-8068
Published by Redfame Publishing
URL: http://jets.redfame.com

Validation of a Paper-and-Pencil Test Instrument Measuring Biology Teachers' Pedagogical Content Knowledge by Using Think-Aloud Interviews

Melanie Jüttner[1] & Birgit Jana Neuhaus[1]

[1]University of Munich, Germany

Correspondence: Melanie Jüttner, Institute for Biology Education, University of Munich, Winzererstrasse 45/II, 80797, Munich, Germany. Tel: 49-89-6492. E-mail: Melanie.juettner@lrz.uni-muenchen.de

Received: April 12, 2013 Accepted: May 2, 2013 Available online: June 14, 2013

doi:10.11114/jets.v1i2.126 URL: http://dx.doi.org/10.11114/jets.v1i2.126

Abstract

The topic of 'teacher professionalism' is one of the most crucial ones in quality education research. It has a potential to generate results that could inform and hence enhance the practice in classrooms. Thus, research in this field needs reliable instruments to measure the professional knowledge of our teachers to be able to generate reliable results for our research problems. Not many instruments have been developed with regard to this topic. At the same time, an adequate validation of the instrument developed is often missing (Schilling & Hill, 2007). Hence, in a bigger project *ProwiN* (German acronym for professional knowledge of science teachers), test instruments for measuring science teachers' pedagogical, pedagogical content and content knowledge (PK, PCK, and CK) were developed for the subjects biology, chemistry and physics. The present study tested the validity of some of these items which were used to measure the pedagogical content knowledge (PCK) of biology teachers. These items focused on measuring teachers' professional knowledge by analyzing 1) teachers' *knowledge about student understanding* (or lack of understanding) of several topics in biology and 2) *knowledge about instructional strategies* like the use of models or experiments. The content validity of these instruments was examined by think-aloud interviews with American and German Biology teachers (N=11). This study shows a high content validity for these items. Furthermore, this paper demonstrates the scope for adapting the conceptual framework of these items to measure biology teachers' PCK in other countries.

Keywords: Biology teachers, pedagogical content knowledge, professional knowledge, think-aloud interviews, validation

1. Introduction

In research about professional knowledge, there is broad agreement that teachers' domain-specific knowledge is most important for high-quality instruction (Krauss, Baumert, & Blum, 2008). Furthermore, there is a consensus about the fact that teachers' pedagogical content knowledge (PCK) and content knowledge (CK) are the two domain-specific and therefore subject-specific knowledge categories of teachers' professional knowledge (Shulman, 1986, 1987). However, claims about the conceptualization of PCK and CK are almost always theoretical and only a few aspects have been examined by a limited number of research groups (e.g., Gess-Newsome et al., 2012; Rowan et al., 2001; Schmidt et al., 2007). Currently, CK, PCK, and pedagogical knowledge (PK) are considered to be the main categories of teachers' professional knowledge (Baumert et al., 2010). Here, PCK is the one which describes the teachers' expertise in making the content comprehensible to students. It is often seen as the most specific type of knowledge for teachers and hence this work focuses on biology teachers' PCK. A variety of theoretical concepts about teachers' PCK currently exists (e.g., Baumert et al., 2010; Gess-Newsome, 1999; Hashweh, 2005; van Driel, Verloop, & de Vos, 1998). Until now, the main aim of researchers has usually been to identify the type of PCK that is necessary for high quality instruction and also about the way it correlates with teachers' performance in the classrooms (Park & Chen, 2012).

Several attempts have been made to analyze teachers' PCK but mostly certifications were used as an indirect indicator of teachers' knowledge (Abell, 2007; Baumert et al., 2010). Only a few studies have measured teachers' professional knowledge directly (e.g., *LMT, COACTIV, MT 21, TEDS-M*). For science, especially biology, such

representative studies are still missing (Abell, 2007). Furthermore, there are very few studies about test development and validation of tests that could be used to measure biology teachers' PCK (e.g., Schmelzing et al., 2013).

Also, such studies have usually analyzed teachers' written answers and have not combined them with their thinking. Additionally, Schmelzing et al. (2013) conducted a study on one specific topic (the cardiovascular system). Hence, there is a need to develop a generalized test instrument which includes different topics and also effectively measures biology teachers' PCK. We have therefore developed biology-specific test instruments measuring teachers' PCK. It includes items pertaining to the topics of *plants*, *vertebrates*, and *neurobiology* (Jüttner, Boone, Park, & Neuhaus, 2013; Jüttner & Neuhaus, 2012).

In this regard, we also need empirical evidences (in addition to teachers written answers) analyzing the different ways in which teachers think and their connection with real life practice in classrooms. Here, "test validation is almost universally viewed as the most unsatisfactory aspect of test development" (Schilling & Hill, 2007, p. 70) as there is a consistent disjunction between theoretical validity of a test and its application in the real scenario. This suggests that the validation of various test instruments need to be measured and discussed.

Thus, the proposed work focuses on a special method – "Think-Aloud Interviews" to measure the validity of various items that were used to measure the PCK test (Jüttner & Neuhaus, 2012; Jüttner, 2013).

Since the beginning of the 21st century, teachers' PCK has been measured by direct tests, prompting the need to investigate the validity of such tests because of their relevance to further research (Schilling & Hill, 2007). This paper, therefore, concentrates on the validation study of a developed and evaluated test instrument. To begin with, any test instrument should fulfill certain criteria so that the standardized tests possess adequate lucidity. There are three main test criteria: reliability, objectivity, and validity (Field, 2009). Herein, the instruments used for teachers' PCK research more often lack the validity tests (Hill, Dean, & Goffney, 2007). In 1985, the Standards for Educational and Psychological Testing noted that validity is "the most important consideration in test evaluation" (American Psychological Association, 1985, p. 9). Validity is very important because it indicates "whether the instrument provides information of interest to test consumers and whether scores generated by the test assist in making good decisions" (Schilling & Hill, 2007, p. 70).

In research literature, validity is categorized into content validity, criterion validity and construct validity (Field, 2009; Wright & Stone, 1999). Kane (2001, 2004) presented a historical overview about these different concepts of validity including the "toolbox" approach. Finally, Kane (2004) developed the argument-based approach based on construct validity to address difficulties concerning the gap between validity conceptualization and practical validation. Kane's approach consisted of two stages (the formative and summative stages) which might also be helpful for teachers PCK research concerning construct validity (cf. Schilling & Hill, 2007).

1.1 Methods for Analyzing Validity

In the field of teachers' professional knowledge, different methods have been used for validating the scores of the test instruments used in research studies. Measuring the construct validity (consisting of convergence and discriminate validity) of such scores was based on the idea that a correlation with other constructs or external criteria (e.g. education) could be examined. However, such construct validity test could not yield significant correlations in the *COACTIV*-study where they tried to examine the construct validity by investigating the correlation between teachers' beliefs and instruction behavior. Another method for testing the validity of test scores is the use of expert ratings (Carlson, 1990).

Furthermore, the construct validity (discriminant validity) can be examined by comparing the test scores of the teaching profession (here, teachers) to contrasting groups of other professions (Krauss et al., 2008). This validation study analyzes the structural assumption (cf. Schilling & Hill, 2007; Schmelzing et al., 2013). Another source of evidence could be cognitive interviews which help determine whether the teachers' check marked answers fit their individual way of thinking (Hill et al., 2007).

In this study, the developed PCK measurement instrument was validated by using all sources of evidence and also by conducting additional similar studies (see e.g., Jüttner, 2013; Jüttner & Neuhaus, 2012; Jüttner & Neuhaus, in press). The present paper highlights the use of think-aloud interviews to analyze the content validity of test instruments. Here, the validity is analyzed for each individual item of the instrument which is very different from the previous validity test where the entire test as a whole was tested for its construct validity (see Jüttner, 2013; Jüttner & Neuhaus, in press).

1.2 Item Format: An aspect influencing the choice of method that could be used to measure test validity

Especially for measuring teachers' PCK, much current discussion focuses on which item format might be best

used in standardized test instruments. More and more, studies are using open-ended questions because teachers have to construct their own answers instead of memorizing facts or theories (Krauss et al., 2011; Schmelzing et al., 2013). On the other hand, time limitation often causes the use of multiple-choice questions (Witner & Tepner, 2010). Both open-ended and multiple-choice items have many advantages and disadvantages, which is why the discussion about an ideal item format measuring teachers' PCK with paper-and-pencil tests still continues.

In this research work, several arguments for the use of open-ended and multiple-choice formats have been highlighted. First, the use of open-ended items will be discussed. Then the necessity of analysis of the test criteria—especially the validity of the paper-and-pencil tests will be briefly described.

Open-ended items, for example, are able to measure all cognitive levels that might be part of a competence model (Neumann, Kauertz, Lau, Notarp, & Fischer, 2007; Schecker & Parchmann, 2006). Complex open-ended items could be constructed for analyzing the development of understanding from the answers. Spontaneous ideas and associations could also be measured by open-ended items. This format does not limit the respondents' answers to the given possibilities. High content validity, their time-consuming nature, and a potentially poorer inter-rater objectivity are characteristics of open-ended items (Schmelzing et al., 2013). The potentially low objectivity could be improved by developing a detailed coding manual for the answers. Especially when measuring teachers' PCK, the decision if an answer is "right" or "wrong" is difficult (Krauss et al., 2011). In science education in particular, more than one theory exists regarding how someone might act pedagogically in a specific situation, which is a critical point concerning open-ended items measuring PCK (Ball, Hill, & Bass, 2005; Ma, 2000). A special way of coding might be a first step towards a solution to this problem. Until now, several ideas of how to code open-ended PCK items have been used: One could use empirical data (Jüttner & Neuhaus, 2012; Jüttner et al., 2013) or employ expert opinion while coding open-ended items (Carlson, 1990; Gardner & Gess-Newsome, 2011). Using experts is criticized because it is not clear who is an expert for which question. Often the opinions of experts are very heterogeneous, which might greatly influence the judgment of open-ended questions.

Currently, multiple-choice items are often used in large-scale studies about professional knowledge of teachers (e.g., *MT21, LMT, TEDS-M 2008*). They have clear advantages which include objectivity, economical use of time and the reduced cost for studies (Hill, Sleep, Lewis, & Loewenberg Ball, 2008). One major criticism of closed-ended items is that limited and predefined responses to multiple-choice items might not reveal how teachers might react in the actual situations. Also, the closed-item format might not be able to test all difficulty levels of cognitive thinking. However, Baumert and Köller (1998) assert that both of these item formats could be used for testing all levels of difficulty like autonomous thinking, methodological skills, and problem-based understanding (p. 15). In all, such multiple-choice items are more often preferred for their time related efficiency and acceptance level amongst teachers (Hill et al., 2008b; Rohaan, Taconis, & Jochems, 2009) and so are the Likert-scales (Rowan et al., 2001). Likert-scales are often used for measuring teacher beliefs and attitudes.

Nevertheless, if the item needs to measure the possible reaction or behavior in response to a specific situation, open-ended questions are a preferred format as they do not have normative "right" or "wrong" answers. Additionally, as mentioned earlier, multiple-choice answer options are not able to include all plausible and possible responses to pedagogical content knowledge (PCK) items (Hill, Loewenberg Ball, & Schilling, 2008; Hill et al., 2008b). Especially, knowledge about possible reactions could not be tested by multiple-choice items because the flexibility of teachers' reactions and their individual know-how might not be demonstrated (Baxter & Lederman, 1999). Finally, multiple-choice items have low validity because of their high probability of a correct answer choice (Hill et al., 2008a).

Such advantages and challenges in using each item format, as described above, call for the use of a blend of these formats, as recommended by previous researchers (Hill et al., 2008a; Hill et al., 2008b; Hill, Schilling, & Ball, 2004; Krauss et al., 2011). Additionally, intentional shift from the quantitative analysis of teachers' professional knowledge to an overall conceptual understanding of teachers on one hand and intense criticism of both item formats used in PCK research on the other hand has caused the researcher to use an appropriate mix of both item formats along with an additional description to justify the chosen answer (Abell, 2007; Hill et al., 2008a; Hill et al., 2008b; Park, Jang, & Chen, 2009).

Here, the proposed study used the test instrument that was developed as a part of the larger *ProwiN* study mostly included open-ended items as they were used to analyze teacher flexibility and responses in a specific situation (Jüttner & Neuhaus, 2012; Jüttner et al., 2013).

1.3 Research Aim

Due to the fact that conceptualization of teachers' PCK is yet not clear, this paper bases itself on a general assumption that the participant teachers might really need the intended knowledge type to answer PCK items used in the test instrument. According to Kane (2004) and Hill, Dean, and Goffney (2007), the fundamental assumption that must be addressed by validity test studies is the idea that a responses to any given test item actually reflect teachers' biological pedagogical content knowledge for teaching. Secondly, the present validity study will also address the question of whether the knowledge tested through each of the test items is actually relevant for teaching in real classroom situations.

2. Method

First, the sample size of the think-aloud interviews will be described. In the second part of this section, the material used for this study will be explained in detail. The idea is to provide readers with a lucid understanding of which of the items from the entire test instrument (Jüttner et al., 2013) were used for this particular study.

2.1 Sample Size

Eleven biology teachers were interviewed. They were on an average age about 42.1 years old (*SD* = 11.9, within a specific age range of 26–66 years) with average teaching experience of 14.1 years (*SD* = 12.6, within a specific range of 1–44 years). When the teachers were interviewed, they taught biology around 8.7 hours (*SD* = 8.3) per week. Six science teachers teaching biology in the United States and five German biology teachers were interviewed (cf. Tables 1 and 2). The names of the interviewed teachers were changed in the beginning of analysis.

Table 1. Background Information on the Interviewed Science Teachers from the U.S. (*N* = 6) (see Jüttner, 2013)

Name	Sarah	Erica	Andrew	Matthew	Pete	Sue
Gender	*female*	*female*	*male*	*male*	*male*	*female*
Education	Multi-disciplinary science	Master's in biology	Master of Biology Education	Master of Science Education	Bachelor of Science	Biological science education
School type	Middle school	Middle school	High school	College	High school	High school
Biology taught per week [hours]	0	0	25	6	25	20
Biology teaching experience [years]	11	4	17	1	44	11

All in all, seven teachers teaching students in grade 5 through 12 (in Germany, the so-called *Gymnasium*; in the US: high school) and four teaching students in grade 5 through 9 (in Germany, the so-called *Hauptschule*; in the US: middle school) were interviewed (see Tables 1 and 2 'schooltype'). One American teacher teaches Biology at the college.

Table 2. Background Information on the Interviewed Biology Teachers from Germany (*N* = 5) (see Jüttner, 2013)

Name	Michael	Brigitte	Kathrin	Inga	Ingrid
Gender	*male*	*female*	*female*	*female*	*female*
Education	Master of Science Education (biology, chemistry, sport)	Master of Science Education (biology, chemistry)	Master of Science Education (biology, chemistry)	Bachelor of Education(history, German, geography, art)	Bachelor of Education (mathe-matics, German, physics)
School type	Upper secondary (*Gymnasium*)	Upper secondary (*Gymnasium*)	Upper secondary (*Gymnasium*)	Lower secondary (*Hauptschule*)	Lower secondary (*Hauptschule*)
Biology taught per week [hours]	12	9	14	2	3
Biology teaching experience [years]	23	8	4	7	25

2.2 Material

The PCK test consists of 24 items on biological topics: *vertebrates, plants*, and *neurobiology* (Jüttner et al., 2013). Besides the biological topics, two more categories were used to measure PCK in the *ProwiN* project (see e.g., Jüttner et al., 2013; Tepner et al., 2012): three knowledge dimensions and three PCK-specific components. The knowledge dimensions were defined by psychological categories (Paris, Lipson, & Wixson, 1983): *declarative knowledge* (knowing What), *procedural knowledge* (knowing How), and *conditional knowledge* (knowing Why). Additionally, three PCK-specific components were defined based on the most frequently used PCK criteria in the literature. These components include: *knowledge of students' understanding*, *student errors* (SE) and *knowledge about instructional strategies* (Park & Oliver, 2008). Knowledge about instructional strategies was divided into *knowledge about models* (Mo) and *experiments* (Exp) as such knowledge is considered crucial for science teachers (Jüttner et al., 2013). Items per cell (per topic, per knowledge dimension, and per PCK-component) were developed according to the PCK blueprint (Jüttner et al., 2013).

As answering the entire PCK test (24 items; see Jüttner et al., 2013) takes a lot of time, only a few items were given to the teachers during think-aloud interviews. In addition, PCK items for various topics in biology were created using the same conceptual framework and thus all items measuring various knowledge dimensions pertaining to one same topic were given to teachers during these think-aloud interviews. The time fixed for these interviews was 40 minutes where 30 minutes were used for the actual item based think-aloud process and 10 minutes for a short follow-up interview. Due to this time limitation, various teachers could answer different number of items from the entire test instrument used for this study (see Table 3).

Here, according to the theoretical model , the validation study asked the biology teachers to answer items for all three knowledge dimensions (*declarative, procedural,* and *conditional knowledge*) as well as for different PCK components (*knowledge about students' understanding and errors* [SE] and *knowledge about instructional strategies concerning a special topic*). According to the Core Curriculum (2006) of the American State, American science teachers were given PCK items dealing with the topic of *plants* (see Table 3) as core topic of the study (i.e. *neurobiology*) is not included in high school biology. Furthermore, items from the topic *neurobiology* were given to German teachers as this was the main topic of the videotaping phase of the second phase of *ProwiN* (see Table 4). Hence, items from the main study (see Jüttner & Neuhaus, 2012; Jüttner et al., 2013) were used.

For the American test takers, the original test items were adapted to American curriculum but the knowledge dimensions (conceptual framework) remained same as in the original test.

Tables 3 and 4 summarize the items were given to teachers and also the item format used for each of these items.

Table 3. Overview of Parts of the Items Developed for the PCK Test Given to U.S. Biology Teachers for Think-Aloud Interviews (N = 6) (see Jüttner, 2013)

Content of item	**PCK components**	**Item format**	
		Open-ended	**MC**
Situation in the biology lesson; students' misunderstanding about the body of a blossom	Knowledge about possible reactions in a lesson situation with students' problems (SEp)	x	
Students' answer in a test about plants' anatomy	Knowledge about possible reasons for students' problems (SEc)	x	
The above-described experiment should be varied to be more student-involved	Knowledge about getting students more actively involved in experiments in school (Exp.p)	x	
An experiment in school about the analysis of photosynthesis products is described (same item stem than Exp.p)	The described experiment should be judged by the use of five different aspects (Likert scale) (Exp.c)		x
Learning about the structure of the blossom by use of a plastic model in a biology lesson	Dis-/advantages of the use of the model (Mo.d)	x	
Learning about the structure of the blossom by use of a plastic model in contrast to a real blossom (model vs. original object)	Criticism of the model in lessons (Mo.p)	x	

Note. The different items were created according to the theoretical model of the project *ProwiN*. The items dealt with the topic of plants, the knowledge dimensions (declarative, procedural, and conditional), and the PCK components SE, Experiments and Models. The grey written items were only given to teachers who had time left after answering the items before.

The items were developed in the biological part of the *ProwiN*-study (Jüttner & Neuhaus, 2012; Jüttner et al., 2013). The two last items (Mo.d; Mo.p) were used as backup items. Not all of the teachers answered these items due to the time limitation.

Table 4. Overview of Parts of the Items Developed for the PCK Test Given to German Biology Teachers for Think-Aloud Interviews (N = 5) (see Jüttner, 2013)

Content of Item	**PCK component**	**Item format**	
		Open-ended	**MC**
Answers of students to an achievement test about the knee-jerk used to show students' problems in understanding	Knowledge about how often such mistakes might arise in or after lessons (SEd)	x	
Situation in the biology lesson; students' misunderstanding about the knee-jerk in a written test	Knowledge about possible reactions in a lesson situation with students' problems (SEp)	x	
Students' answer in a test about knee-jerk	Knowledge about possible reasons for students' problems (SEc)	x	
Model of the human nervous system	Dis-/advantages of the use of the model (Mo.d)	x	
Model of the human nervous system vs. the real human nervous system	Criticism on the model in lessons (Mo.p)	x	
Different experiments about the topic of the ear	To know different possible school experiments for a given topic (Exp.d)	x	
Variation of the experiment on sound localization	Knowledge about how experiments could be changed to be more student-activating (Exp.p)	x	
Evaluation of a given experiment on sound localization	Rating 5 statements about the given experiment (Exp.c)		x

Note. The different items were created according to the theoretical model of the *ProwiN* project (Tepner et al., 2012). The items deal with the topic of neurobiology, the knowledge dimensions (declarative, procedural, and conditional), and the PCK components SE, Mo, and Exp.

2.2 Implementation of the Study

Think-aloud interviews were used to test content validity of newly-developed test instruments that are intended to measure biology teachers' PCK. Hill, Dean, and Goffney (2007) wrote that think-aloud interviews might be very useful during the development process of items (for an instrument). The teachers' thoughts and views about these items could indicate the scope for modification and also specific reasons for certain strange answers during observed during the written tests. Due to the same reason expert interviews with German Biology teachers were used during the development process of the test instrument for this study. Secondly, think-aloud interviews, allow for inferences about how teachers think with regard to certain situations and also why they answer the way they do (cf. Hamilton, Nussbaum, & Snow, 1997).

So far, this has been the only method proposed to understand whether the item constructed really calls for a teachers' PCK knowledge. It is the only method that allows the study of numerous cognitive processes that are involved in answering written open-ended questions. This was our key motivation for trying this method to study the validity of the proposed test instrument. The content validity test was conducted in spring 2011, where the whole test instrument was given to teachers as a part of the main study (see Jüttner et al., 2013). Also, considering the time constraints, a part of this instrument was also used for think-aloud validation test. Hence, as described earlier, the similar conceptual framework of items developed for each of the topics included in the test instrument (see Jüttner et al., 2013) allowed for a scope to choose specific topics (for think-aloud interviews) for each of the two samples (for teachers from America and Germany - see Tables 3 and 4). In addition, construct validity of the whole PCK instrument was also analyzed using Wright Maps (Wright & Stone, 1999) during the main study (see Jüttner et al., 2013) and in an additional study that compared this group with other experts such as biologists and pedagogues (Jüttner & Neuhaus, in press).

The think-aloud interviews took place in U.S. in January 2011 and in Germany in March 2011. The first part of this study was 'think-aloud interviews' where a simultaneous method was used: teachers had to answer an item by writing down their answer and by thinking aloud (Aitken & Mardegan, 2000; Ericsson, 2006; Leighton, 2009). After the think-aloud interviews, the teachers answered four follow-up questions about these items. These follow-up questions focused on the quality of items i.e. whether any of the items were confusing in their expression and also how far the situations used in these questions were realistic. Such questions help understand

whether teacher responses in interview are representative of how they might respond in actual classroom scenario. All in all, the interviews never exceeded 40 minutes.

The interview started with a short warm-up section in which 'thinking aloud' was briefly explained to participating teachers. The aim of this study was also described. The first page of this test requested teachers to provide information about their background. This phase was used as a practice for the 'think-aloud' interview process as such answers do not expect any cognitive activation from teachers. Also, during the whole think-aloud interview, the interviewer gave instructions, whenever some teacher forgot to say what they are thinking in that moment. In the five follow-up questions, the teachers were asked questions about the face validity of items answered.

2.3 Empirical Analysis

The think-aloud interviews were taped, transcribed verbally and analyzed for their content. Also, a theoretical model similar to the one used for item development was used for creating a coding manual and hence categorizing teachers' answers and thoughts. For e.g. to identify how often did teachers think about declarative CK while answering a question about possible reasons for students errors (*conditional PCK, knowledge about students' understanding*) (Van Someren, Barnard, & Sandberg, 1994). This allowed for direct comparison of the coded protocol and theoretical model underlying the item development (Van Someren, Barnard, & Sandberg, 1994). Two independent raters coded all teachers' interview transcripts ($N = 11$) as per the coding manual. The second coding was used for calculation of inter-rater reliability (ICC) and thus ensuring the objectivity of method used. Both raters coded each of the transcripts.

Here, due to the fact that theoretical model behind the item development process was already described in another paper (Jüttner & Neuhaus, 2012; Jüttner et al., 2013; Tepner et al., 2012), this model will not be described here in detail again. Furthermore, more about data collection of the main study and more information about the quality of the developed measurement instruments are provided there as well.

In the next section, short fragments of a verbatim protocol will be presented to make the coding process more clear.

2.4 Verbatim Protocol—Coding Scheme

The following short excerpt of the actual interview is provided to illustrate the way a think-aloud interview worked and also the way it was used to analyze different knowledge types used by a teacher while answering PCK items.

1	Teacher S.: (After reading aloud item SEc) "OK, so first of all, what I'm thinking is that I need to
2	identify that there's an error, but in my mind I want to make sure that I don't point it out so that that
3	student initially feels that I'm saying they're wrong because then their learning is going to shut down
4	immediately because they're wrong, and I think students react that way. So I need to come up with a
5	solution and try to maybe go around the fact that they're incorrect but let them know it's incorrect, umm,
6	by maybe addressing . . . using some specific examples, because the last thing I want to happen is for that
7	student to feel they're . . . pointed out in front of their peers, umm . . . because I want kids to continue to
8	answer and feel confident about their answers. So, if the student uses the term 'pollen sac' and they're
9	describing the female, umm . . . I probably would address, uhh . . . ask them specifically maybe, OK, so
10	in the pollen sac, what is the function of the pollen sac? And try to get them to focus on the word 'pollen'
11	rather than necessarily 'sac,' because I think that could be the confusing part and identifying what the
12	purpose of the pollen is in fertilization, in which then I would hopefully get them to come back and say,
13	'Oh, wait a second. The pollen sac is actually the male part rather than the female part' would be my
14	goal. Let me write that down."

Here, teacher had to answer a question about possible reactions when a student error arose in a sixth-grade classroom, while a teacher was presenting a lesson on the process of pollination. This error arose during the beginning of this particular lesson on pollination where a teacher asked a student to review previous day's lesson about various parts of a cherry blossom. The student, while naming the male and female parts, used the terms 'pollen sac' in his or her description of the female rather than the male part of this plant. The teacher being tested had to describe various ways in which this teacher could respond. The following sentences (Figure 1) show what Sue thought while answering this item about procedural knowledge about students' errors (see Table 3).

Figure 1: Extraction out of the think-aloud protocol from a biology teacher from the United States (Sue).

For this excerpt from Sue's think-aloud interview, the coder first read the passage. Then the coding manual was used to identify what knowledge type was used in each of her utterances to be able to successfully answer the question with regard to the situation given to her.

As can be seen in lines 2, 5–8, and 10–11, she talked or thought about how she would handle this situation. These different thoughts were coded singularly. In lines 11–12, she guessed that the word "sac" might be

misleading students. This was coded as knowledge about possible reasons for the student's error (conditional knowledge about students' errors). For the final calculation, one table was generated for each teacher where their uttered lines are noted with a special code.

In the end, the frequency at which a knowledge type was used by each of the teachers, while answering a given item, was counted. The mean (*M*) and the standard deviation (*SD*) of all the U.S. biology teachers' used PCK components and the mean of all German biology teachers' used PCK components were calculated and are summarized in tables (see Results section).

For an overview of how the different categories were coded, Table 5 summarizes the PCK categories and the conditions when they were coded. This coding scheme was based on the theoretical background model (see Jüttner et al., 2013). The coding manual for PCK was also developed based on the theoretical model for CK, as teachers in these interviews were often talking about the content while using their pedagogical content way of thinking.

Table 5. Overview of PCK Categories for Coding the Interview Transcripts. Bolded acronyms such as SEd used in results for the different knowledge types.

Components \ Knowledge	**Declarative knowledge**	**Procedural knowledge**	**Conditional knowledge**
Students' understanding	Knowledge about if and how often a given student's error could arise E.g., "This error arises often in the lessons about …" Not coded here: "This is a mistake."	Knowledge about how you could react in a special situation that confronts you with students' errors This includes reasons as well as intentions that lead to the reaction/enactment.	Knowledge about possible reasons for students' errors that might come up in lessons (coherence of reasons and acting of s.) Hypothetical statements about possible reasons out of lesson material—which is not available at the moment—are coded here as well.
	SEd	**SEp**	**SEc**
Models	Listed advantages and disadvantages of the shown model	Knowledge about when and how to act with the model in classroom situation; this includes knowledge about the framework of the usage as well as about how the model could be used.	Knowledge about possible student misunderstandings and/or errors coming up by using this model Important: Advantages and disadvantages must be strictly separated here.
	Mo.d	**Mo.p**	**Mo.c**
Experiments	Knowledge about numerous experiments dealing with one topic; factual knowledge about an experiment	Knowledge of many ways of changing an experiment to become as student-active as possible	Assessing of a described situation dealing with typical experiments used in biology lessons
	Exp.d	**Exp.p**	**Exp.c**

Furthermore, as think-aloud interviews were conducted to test whether the constructed items (in *ProwiN*) really measured respective knowledge categories, categories used for the item development will be presented in the results as well.

3. Results

The results are divided into three sections. In the first part, the results concerning the PCK items developed about the topic plants will be discussed, while in the second part the results concerning the neurobiology items will be presented. Thirdly, the inter-rater reliability for the objectivity of coding these topics will be reported. The outcomes of the study concerning content validity are presented (in tables below) as mean of the frequencies of each teacher knowledge type used to while answering the items in think-aloud interviews.

In Table 6, the content validity of each of the PCK items for the topic plants is shown as means of the frequencies of each knowledge type used while answering these questions.

Table 6. Results of the Coding of the Think-Aloud Interviews in the U.S. on the Topic of Plants (N = 6); Mean and (Standard deviation) are presented in % (see Jüttner, 2013)

Item	*N*	***PCK about SE***			***PCK about Exp.***			***PCK about Mo.***			***dCK***	***diSECK***
		SEd	*SEp*	*SEc*	*Exp.d*	*Exp.p*	*Exp.c*	*Mo.d*	*Mo.p*	*Mo.c*		
SEp	6		88.8 (12.5)	6.5 (6.5)							14 (0)	
SEc	6	19.5 (5.5)	29 (0)	82 (18.8)								40 (0)
Exp.p	5				50 (0)	70 (24.5)	50 (0)					
Exp.c	6						100 (0)					
Mo.d	3							85.3 (20.7)	22 (0)	22 (0)		
Mo.p	3							50 (0)	83.3 (23.6)			

Note. The empirical values compared to the intended PCK components of the items are marked in grey. dCK is declarative content knowledge; diSE means diagnosis of SE.

Table 6 shows that the intended PCK components (grey boxes) were used for more than 70% of all the answers of the interviewed teachers. Here, participating teachers were asked to answer the interview items in 30 minutes but some of them needed more time and hence did not answer all items (for example, the item about models). For item SEp where teachers had to think about possible reasons for student's errors, the intended procedural PCK about students' understanding was used by the six teachers in 88.8% (SD = 12.5) of their answers. For this item, teachers also used conditional PCK about students' understanding (M = 26.5%; SD = 6.5) and declarative CK (M = 14.0%; SD = 0).

For the topic of neurobiology, five biology teachers also used more than 63% of the intended knowledge when answering the items (see Table 7). Sometimes, as can be seen with the SEp item, one teacher talked a lot about almost everything he or she knew about students' problems in understanding facts; therefore, the conditional PCK about students' understanding was 100% (but only for one person) (SD = 0).

Table 7. Results of the Coding of the Think-Aloud Interviews in Germany on the Topic of Neurobiology (N = 5); Mean and (Standard deviation) are presented in % (see Jüttner & Neuhaus, 2012; Jüttner, 2013)

Item	*N*	***PCK about SE***			***PCK about Exp.***			***PCK about Mo.***			***dCK***	***diSE.CK***
		SEd	*SEp*	*SEc*	*Exp.d*	*Exp.p*	*Exp.c*	*Mo.d*	*Mo.p*	*Mo.c*		
SEd	5	63 (22)		41.3 (10.2)								20 (0)
SEp	5	50 (0)	71 (21.7)	100 (0)							33.3 (16.7)	
SEc	5	33 (0)		90.1 (13.2)							16.7 (0)	
Exp.d	5				100 (0)							
Exp.p	5				100 (0)	71.3 (6.1)	28.7 (6.1)				25 (0)	
Exp.c	5						87.5 (21.7)				50 (0)	
Mo.d	5							100 (0)				
Mo.p	5							70 (21.6)	66.3 (34.2)			

Note. The empirical values compared to the intended PCK components of the items are marked in grey. dCK is declarative content knowledge; diSE means diagnosis of SE.

The inter-rater reliability of the two independent raters for coding of think-aloud interview transcripts was found to be significantly high (ICC_{unjust} = .98; $F_{468,468}$ = 54.2; p < .001).

4. Discussion

The results of think-aloud interviews (tried on 11 biology teachers of two different countries) show that test items could most of the time draw intended knowledge type from participant teachers. This indicates that items in this test have adequate content validity to test the PCK of biology teachers.

Furthermore, analysis of curricula in the different federal states of Germany and U. S. was also conducted to ensure content validity for items used in think-aloud interviews. Additionally, 30 German biology teachers were asked to identify the most important topics from their curriculum which helped us in choosing the topics for item development (Jüttner, Spangler, & Neuhaus, 2009). In all, items pertaining to topics which were relevant to their curriculum and teaching also help ensure the content validity of the developed PCK test.

It is remarkable that the think-aloud interviews worked well both in Germany and the U.S. even while the items were developed originally for German biology teachers. Furthermore, the interviews were conducted in two different countries with different cultural background using two distinct languages; teachers in each of these countries could use the intended knowledge type to answer the questions about PCK. This result validates the first statement regarding the replication of these items for educational systems in other countries where relevant topic related items could be developed based on the proposed conceptual framework.

Here, it is noteworthy to know that teachers in the U.S. have very different cultural and educational backgrounds than Germany. But all in all, more than 60% of the intended knowledge type was used by all the teachers to answer these test items.

This might be initial evidence that the idea behind item development could be generalized and adapted for PCK item development in different countries. Moreover, for every change in the item in this instrument, a validation study must be conducted. The other test criteria must also be rechecked because of the changes (Schilling, 2007). The translation-replication of the items needs to adopt cultural characteristics of respective countries and perhaps alternative descriptions of classroom situations (Blömeke, König, Kaiser, & Suhl, 2010).

In particular, the items about students' errors might need special attention because the question about how often a student's error might arise in classroom situations is based on empirical data of German students (see Jüttner & Neuhaus, 2012). Thus, giving such test items to students in the U.S. might produce very different results. Additionally, the Core Curriculum in the U.S. did not include neurobiology and thus interviews in these countries dealt with different topics. This limits a potential generalizability. Also, due to time constraints, it was difficult to give more and different items to participating teachers.

Moreover, in Germany, the *neurobiology* topic was the focus of a second study concerning the relationship between the results of teachers' knowledge tests and their actual actions in (videotaped) lessons. Thus in future, the influence of use of various topics to develop items based on the proposed conceptual framework (Jüttner et al., 2013) should be investigated further with respect to the culture and curricula of the countries in question.

Furthermore, the results of the think-aloud interviews show that the development of PCK items that try to test special knowledge types according to a theoretical model (see Jüttner & Neuhaus, 2012; Jüttner et al., 2013) is possible only in a certain way. For almost all the items, teachers used more than 60% of the intended knowledge but they also needed other knowledge types as well. Such results demonstrate in a qualitative way that PCK and CK might be dependent on each other.

The follow-up questions after the think-aloud interviews were able to demonstrate face validity of this instrument. Here, for example, question about how far teachers' answers are able to demonstrate how they might or want to act in real-life classroom situations could be reaffirmed by their assertions like Sue said, "My answers represent what I strive to do". Moreover, each of the 11 participating teachers from both the countries suggested that the questions being very relevant to the teaching practice and real-life situations, the answers given by them might only demonstrate how they actually want to perform in real classroom scenarios. They speculated that any changes in their proposed action would depend on the way situations arise in the real-life. Thus theoretically, these items are able to demonstrate how teachers think about solving problems arising in lessons in general. This could be seen as one aspect of face validity, which is not described here in further detail as its criteria are very subjective (Krauss et al., 2011).

All in all, the results of the coded think-aloud interviews as well as their answers to the follow-up questions (about the item formats used) show that the presented test excerpt used a balanced mix of each of the item formats (see Tables 3 and 4). Here, teachers underlined this theory by pointing their views as (e.g. Sue pointed out directly: "I think it's a good mix."). According to the results of the think-aloud coding (see Tables 6, 7, and 8), all the different item formats were able to measure specific types of PCK with different knowledge dimensions (declarative, procedural, and conditional).

In summary, the method of think-aloud interviews could be used for content validation. The analysis of content validity of a newly developed test instrument for measuring different knowledge types especially needs a method that allows the acquisition of further information about their thinking process. The teachers' thoughts in the interviews compared to their written answers might also help to provide further idea about what might be difficult to write down even though it exists in their minds. For such a concrete analysis, fewer teachers were

interviewed. Finally, combination of teachers' written results of the whole test with their actions in lessons (video study) will help gather further information about the relation between knowledge (written) and action.

Acknowledgements

The present project was funded by the Federal Ministry of Education and Research in xxx; it is a cooperative project embedded in the framework program of "empirical research in education" (01JH0904). The full details of this study (including figures, tables, and diagrams) were reported in a German dissertation in Jüttner (2013).

References

Abell, S. K. (2007). Research on science teachers' knowledge. In S.K. Abell & N.G. Lederman (Eds.), *Handbook of research on science education* (pp. 1105–1149). Mahwah, NJ: Lawrence Erlbaum Associates.

Aitken, L. M., & Mardegan, K. J. (2000). "Thinking aloud": Data collection in the natural setting. *Western Journal of Nursing Research, 22*(7), 841–853.

American Psychological Association (APA), American Educational Research Association, and National Council on Measurement in Education (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Ball, D. L., Hill, H. H., & Bass, H. (2005). Knowing mathematics for teaching: Who knows mathematics well enough to teach third grade, and how can we decide? *American Educator, 29*(1), 14–46. http://hdl.handle.net/2027.42/65072

Baumert, J., & Köller, O. (1998). Nationale und internationale Schulleistungsstudien: Was können sie leisten, wo sind ihre Grenzen? [National and international assessment studies: What might they afford and what are the limitations?] *Pädagogik, 50*, 12–18.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., . . . Yi-Miau, T. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*(1), 133–180. http://dx.doi.org/10.3102/0002831209345157

Baxter, J. A., & Lederman, N. G. (1999). Assessment and measurement of pedagogical content knowledge. In J. Gess-Newsome & N. Lederman (Eds.), *Examining pedagogical content knowledge: The construct and its implications for science education* (pp. 147–161). Dordrecht, Netherlands: Kluwer Academic Publishers.

Blömeke, S., König, J., Kaiser, G., & Suhl, U. (2010). Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I. In S. Blömeke, G. Kaiser, & R. Lehmann (Eds.), *TEDS-M 2008. Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich* (pp. 97–136). Münster, Germany: Waxmann.

Carlson, R. E. (1990). Assessing teachers' pedagogical content knowledge: Item development issues. *Journal of Personnel Evaluation in Education, 4*, 157–163.

Ericsson, K. A. (2006). Protocol analysis and expert thought: Current verbalizations of thinking during experts' performance on representative tasks. In K. A. Ericsson, N. Charness, P.J. Feltovich, & R.R. Hoffmann (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 223–242). Cambridge, England: Cambridge University Press.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London, England: Sage.

Gardner, A. L., & Gess-Newsome, J. (2011, April). *A rubric to measure teachers' knowledge of inquiry-based instruction using three data sources*. Paper presented at the National Association for Research in Science Teaching Annual Meeting, Orlando, FL.

Gess-Newsome, J. (1999). Pedagogical content knowledge: An introduction and orientation. In J. Gess-Newsome & N.G. Lederman (Eds.), *Examining pedagogical content knowledge* (pp. 3–17). Dordrecht, Netherlands: Kluwer Academic Publishers.

Gess-Newsome, J., Taylor, J. A., Carlson, J., Gardner, A. L., Wilson, C. D., & Stuhlsatz, M. A. M. (2012). *Impact of educative materials and professional development on teachers' professional knowledge, practice, and student achievement.* Manuscript submitted for publication.

Hamilton, L. S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10*(2), 181–200. http://dx.doi.org/10.1207/s15324818ame1002_5

Hashweh, M. (2005). Teacher pedagogical constructions: A reconfiguration of PCK. *Teachers and Teaching: Theory and Practice, 11*, 237–292. http://dx.doi.org/10.1080/13450600500105502

Hill, H. C., Dean, C., & Goffney, I. M. (2007). Assessing elemental and structural validity: Data from teachers, non-teachers, and mathematics. *Measurement: Interdisciplinary Research and Perspectives, 5*(2–3), 81–92.

Hill, H. C., Loewenberg Ball, D., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education, 39*(4), 372–400.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal, 105*(1), 11–30. http://dx.doi.org/10501-000250500

Hill, H. C., Sleep, L., Lewis, J. M., & Ball, D. L. (2007). Assessing teachers' mathematical knowledge: What knowledge matters and what evidence counts? In F.K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 111-155). Charlotte, NC: Information Age Publishing.

Jüttner, M. (2013). Entwicklung, Evaluation und Validierung eines Fachwissenstests und eines fachdidaktischen Wissenstests für die Erfassung des Professionswissens von Biologielehrkräften. [Development, evaluation and validation of a content knowledge test and a pedagogical content knowledge test measuring biology teachers' knowledge]. (Doctoral Thesis, Biology Education, LMU Munich, Germany).

Jüttner, M., Boone, W., Park, S., & Neuhaus, B. J. (2013). Development of a test instrument to measure biology teachers' content knowledge (CK) and pedagogical content knowledge (PCK).*Educational Assessment, Evaluation and Accountability, 25*(1), 45–67. http://dx.doi.org/10.1007/s11092-013-9157-y

Jüttner, M., & Neuhaus, B. J. (2012). Development of items for a pedagogical content knowledge-test based on empirical analysis of pupils' errors. *International Journal of Science Education, 34*(7), 1125–1143. http://dx.doi.org/10.1080/09500693.2011.606511

Jüttner, M., & Neuhaus, B. J. (in press). Das Professionswissen von Biologielehrkräften. Ein Vergleich zwischen Biologielehrkräften, Biologen und Pädagogen. [Biology teachers' professional knowledge. A comparison between biology teachers, biologists and pedagogues.] *Zeitschrift für Didaktik der Naturwissenschaften.*

Jüttner, M., Spangler, M., & Neuhaus, B. J. (2009). Zusammenhänge zwischen den verschiedenen Bereichen des Professionswissens von Biologielehrkräften. [Connectednessbetween different categories of biologyteachers' professional knowledge.]In D. Krüger, A. zu UpmeierBelzen, S. Hof, K. Kremer, & J. Mayer (Hrsg.), *Erkenntnisweg Biologiedidaktik 8* (pp. 69–82). Gießen: Universitätsdruckerei Kassel.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319–342. http://dx.doi.org/10.1111/j.1745-3984.2001.tb01130.x.

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives, 2*(3), 135–170. http://dx.doi.org/10.1207/s15366359mea0203_1.

Krauss, S., Baumert, J., & Blum, W. (2008), Secondary mathematics teachers' pedagogical content knowledge and content knowledge: Validation of the COACTIV constructs. *International Journal of Mathematics Education, 40*(5), 873–892.

Krauss, S., Blum, W., Brunner, M., Neubrand, M., Baumert, J., Kunter, M., … Elsner, J. (2011). Konzeptualisierung und Testkonstruktion zum fachbezogenen Professionswissen von Mathematiklehrkräften. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV* (pp. 135–162). Münster, Germany: Waxmann.

Kromrey, J. D., & Renfrow, D. D. (1991). *Using multiple choice examination items to measure teachers' content specific pedagogical knowledge*. Paper presented at the Annual Meeting of the Eastern Educational Research Association, Boston, MA.

Leighton, J. P. (2009). *Two types of think aloud interviews for educational measurement protocol and verbal analysis*. Paper presented at the National Council on Measurement in Education (NCME), San Diego, CA.

Ma, L. (2000). *Knowing and teaching elementary mathematics: Teachers' understanding of fundamental mathematics in China and the United States*. Mahwah, NJ: Lawrence Erlbaum Associates.

Neumann, K., Kauertz, A., Lau, A., Notarp, H., & Fischer, H. E. (2007). Die Modellierung physikalischer Kompetenz und ihrer Entwicklung [Modelling of physical competency and its development]. *Zeitschrift für Didaktik der Natur-wissenschaften, 13*, 103–123.

Nielsen, J., Clemmensen, T., & Yssing, C. (2002). *Getting access to what goes on in people's heads?: Reflections on the think-aloud technique.* Proceedings of the NordiCHI 2002, Århus, Denmark.

Paris, S. G., Lipson, M. Y., & Wixson, K. K. (1983). Becoming a strategic reader. *Contemporary Educational Psychology*, *8*, 293–316. http://dx.doi.org/10.1016/0361-476X(83)90018-8

Park, S., & Chen, Y.-C. (2012). Mapping out the integration of the components of pedagogical content knowledge (PCK): Examples from high school biology classrooms. *Journal of Research Science Teaching, 49*(7), 922–941. http://dx.doi.org/10.1002/tea.21022

Park, S., Jang, J., & Chen, Y.-C. (2009, April). *Can we measure teachers' pedagogical content knowledge (PCK) using surveys?: Developing measures of PCK for teaching high school biology.* International Conference of the National Association for Research in Science Teaching, Garden Grove, CA.

Park, S., & Oliver, S. (2008). Revisiting the conceptualisation of pedagogical content knowledge (PCK): PCK as a conceptual tool to understand teachers as professionals. *Research in Science Education, 38*, 261–284. http://dx.doi.org/10.1007/s11165-007-9049-6

Rohaan, E. J., Taconis, R., & Jochems, W. M. (2009). Measuring teachers' pedagogical content knowledge in primary technology education. *Research in Science & Technological Education, 27*(3), 327–338. http://dx.doi.org/ 10.1080/02635140903162652

Rowan, B., Schilling, S. G., Ball, D. L., Miller, R., Atkins-Burnett, S., Camburn, E., … Geoff, P. (2001). *Measuring teachers' pedagogical content knowledge in surveys: An exploratory study.* State College, PA. Consortium for Policy Research in Education, Study of Instructional Improvement.

Schecker, H., & Parchmann, I. (2006). Modellierung naturwissenschaftlicher Kompetenz [Modelling of science competency]. *Zeitschrift für Didaktik der Naturwissenschaften, 12*, 45–66.

Schilling, S. G. (2007). The role of psychometric modeling in test validation: An application of multidimensional item response theory. *Measurement: Interdisciplinary Research and Perspectives, 5*(2), 93–106.

Schilling, S. G., & Hill, H. C. (2007). Assessing measures of mathematical knowledge for teaching: A validity argument approach. *Measurement: Interdisciplinary Research and Perspectives, 5*(2–3), 70–80.

Schmelzing, S., Van Driel, J., Jüttner, M., Brandenbusch, S., Sandmann, A., & Neuhaus, B. (2013). Development, evaluation, and validation of a paper-and-pencil test for measuring two components of biology teachers' pedagogical content knowledge concerning the "cardiovascular system." *International Journal of Science and Mathematics Education.* http://dx.doi.org/10.1007/s10763-012-9384-6

Schmelzing, S., Wüsten, S., Sandmann, A., & Neuhaus, B. (2010). Measuring declarative and reflective components of biology teachers' pedagogical content knowledge. In M.F. Tasar & G. Cakmakci (Eds.), *Contemporary science education research: Teaching* (pp. 71–77). Ankara, Turkey: Pegem Akademi.

Schmidt, W. H., Tatto, M. T., Bankov, K., Blömeke, S., Cedillo, T., Cogan, L., … Schwille, J. (2007). *The preparation gap: Teacher education for middle school mathematics in six countries. MT21 Report.* East Lansing, MI: Michigan State University.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*, 4–14.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*, 1–22.

Tepner, O., Borowski, A., Fischer, H. E., Jüttner, M., Kirschner, S., Leutner, D., …Wirth, J. (2012). Modell zur Entwicklung von Testitems zur Erfassung des Professionswissens von Lehrkräften in den Naturwissenschaften [Item development model for assessing professional knowledge of science teachers]. *Zeitschrift für Didaktik der Naturwissenschaften, 18*, 7–28.

van Driel, J. H., Verloop, N., & de Vos, W. (1998). Developing science teachers' pedagogical content knowledge. *Journal of Research in Science Teaching, 6*, 673–695. http://dx.doi.org/10.1002/(SICI)1098-2736(199808)35:6<673::AID-TEA5>3.0.CO;2-J.

Van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. (1994). *The think aloud method: A practical guide to modelling cognitive processes* (pp. 117–139). London, England: Academic Press.

Witner, S., & Tepner, O. (2010). Professional knowledge of chemistry teachers: Test development and evaluation. In M.F. Taşar & C. Çakmakci (Eds.), *Contemporary science education research: Preservice and inservice teacher education* (pp. 223–228). Ankara, Turkey: Pegem Akademi.

Wright, B., & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wilmington, DE: Wide Range.