Informatics in Education, 2017, Vol. 16, No. 2, 301–318

DOI: 10.15388/infedu.2017.15

dETECT: A Model for the Evaluation of Instructional Units for Teaching Computing in Middle School

Christiane G. von WANGENHEIM[1], Giani PETRI[1,2], André W. ZIBETTI[1],Adriano F. BORGATTO[1], Jean C.R. HAUCK[1] Fernando S. PACHECO[3], Raul Missfeldt FILHO[1]

[1]*Brazilian Institute for Digital Convergence (INCoD) – Department of Informatics and Statistics (INE) – Federal University of Santa Catarina (UFSC), Brazil*
[2]*Federal University of Santa Maria (UFSM), Brazil*
[3]*Department of Electronics – Campus Florianópolis Federal Institute of Santa Catarina (IFSC), Brazil*
e-mail: c.wangenheim@ufsc.br, gpetri@inf.ufsm.br, andre.zibetti@ufsc.br, adriano.borgatto@ufsc.br, jean.hauck@ufsc.br, fspacheco@ifsc.edu.br, raul.missfeldt.filho@grad.ufsc.br

Received: May 2017

Abstract. The objective of this article is to present the development and evaluation of dETECT (Evaluating TEaching CompuTing), a model for the evaluation of the quality of instructional units for teaching computing in middle school based on the students' perception collected through a measurement instrument. The dETECT model was systematically developed and evaluated based on data collected from 16 case studies in 13 different middle school institutions with responses from 477 students. Our results indicate that the dETECT model is acceptable in terms of reliability (Cronbach's alpha α=.787) and construct validity, demonstrating an acceptable degree of correlation found between almost all items of the dETECT measurement instrument. These results allow researchers and instructors to rely on the dETECT model in order to evaluate instructional units and, thus, contribute to their improvement and to direct an effective and efficient adoption of teaching computing in middle school.

Keywords: computing, evaluation, instructional unit, middle school.

1. Introduction

Teaching computing through summer camps, clubs or in family workshops is a worldwide trend (Gresse von Wangenheim and Wangenheim, 2014). There are several initiatives to teach computing such as Code.org (`http://www.code.org`), Code.club

(https://www.codeclubworld.org), Computing at Schools (http://www.computacaonaescola.ufsc.br/), among others. These initiatives are expected to contribute to the popularization of computing competencies as well as the awareness and interest of the students towards computing (Guzdial *et al.*, 2014; Garneli *et al.*, 2015).

Taking into consideration the growing number of alternative instructional units (IUs) for teaching computing, it is important to obtain evidence on the expected benefits as a basis for their systematic selection, adoption and improvement (Decker *et al.*, 2016). Following Guzdial (2004), a main contribution to this knowledge area is not necessarily the development of new programming environments or instructional units, but to find out how to study the existing ones. A more precise understanding of the results of using these instructional units would make it possible to know whether they contribute, in fact, positively to the achievement of the learning goals and compensate the cost involved in their adoption. However, although there is evidence that existing IUs can improve the teaching and learning process in middle school being used more widely in schools worldwide, there is little research on the analysis of the contribution that these IUs can bring to education (Decker *et al.*, 2016).

Currently, the evaluation of the quality of IUs is limited or even, sometimes, non-existent (Decker *et al.*, 2016; Garneli *et al.*, 2015). In many cases, a decision about the use of IUs is based on assumptions of their effectiveness (Gross and Powers, 2005; Wilson *et al.*, 2010). On the other hand, some studies focus on specific quality factors only, such as learning improvement (Gross and Powers, 2005; Kalelioğlu and Gülbahar, 2014). Other studies focus on the effectiveness of visual block-based programming languages (Weintrop and Wilensky, 2015; Grover *et al.*, 2014; Perdikuri, 2014). However, students' perceptions and intentions are also determining factors for successful learning (Giannakos *et al.*, 2013). Yet, few evaluations take into consideration aspects such as motivation and the students' experience during the instructional unit (Craig and Horton, 2009; Giannakos *et al.*, 2014), or students' attitudes toward technology acceptance (Giannakos *et al.*, 2013). In addition, studies that measure students' attitude toward computing are rather designed for higher education and seem to be outdated in the current context of teaching computing in schools (Garland and Noyes, 2008).

The measurements used to evaluate the quality of IUs to teach computing vary widely, ranging from generic scales of students' attitudes toward computing to measurement instruments developed in an ad-hoc way. Many measurements are developed without the definition of a model to derive the items of the measurement instrument based on theoretical constructs, which may make the validity of the results questionable. Thus, currently, there is a lack of systematically developed and evaluated evaluation models and/or measurement instruments that are widely accepted to evaluate the quality of IUs for teaching computing in schools. However, such evaluation models have to take into consideration the characteristics of such IUs typically performed more informally, for example, as programming workshops for parents and children outside the school environment. In such a context, it may be impracticable to carry out experiments that require pre-tests and inclusion of control groups, causing a major interruption and influencing the fun factor of the workshop. A more viable alternative may be the conduction of case studies, in which the evaluation of the IU is performed only at the end of the workshop/

course (post-test), typically through a questionnaire to obtain the students' perceptions (Wohlin *et al.*, 2012). An advantage of this study type is that evaluation can be performed with little effort and in a non-intrusive way at the end of the instructional unit. Studies based on the measurement of perceptions, using questionnaires, are conducted in a variety of different research areas providing reliable, valid and useful information (Devellis, 2016; Takatalo *et al.*, 2010, Sweetser and Wyeth, 2005; Poels *et al.*, 2007). Thus, the objective of this article is to present the development and evaluation of dETECT (*Evaluating TEaching CompuTing*), a model for the evaluation of the quality of instructional units for teaching of computing in schools based on the students' perception.

2. Research Method

In order to develop a model for the evaluation of instructional units for teaching computing, an applied research was carried out (Miller and Salkind, 2002), divided into four stages (Fig. 1):

- Stage 1. Literature review
- Stage 2. Developing of the dETECT Evaluation Model
- Stage 3. Design of the measurement instrument
- Stage 4. Application and evaluation of the measurement instrument

Stage 1. Literature review. In a first exploratory stage, we conducted a literature review on bibliography related to evaluation models of instructional units for teaching computing in schools.

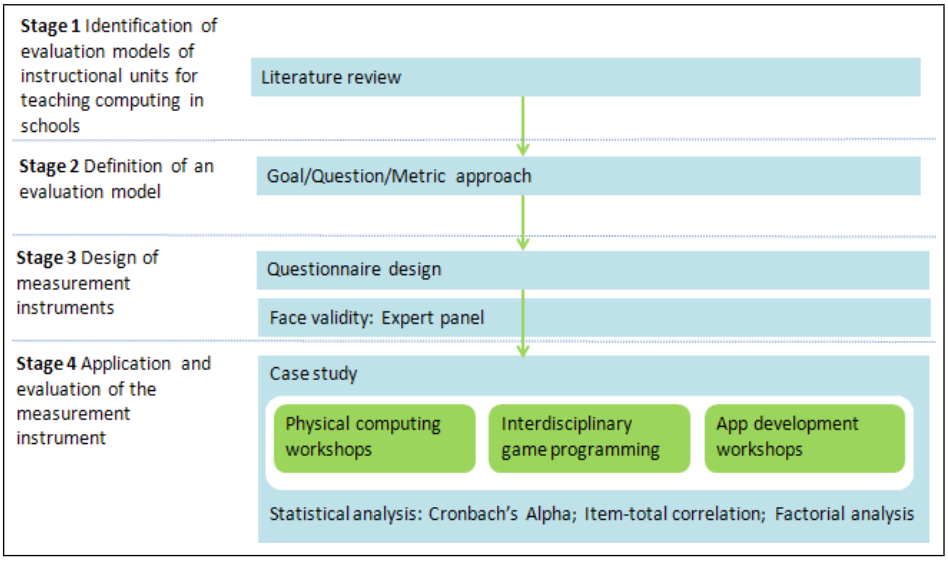


Fig. 1. Research method.

Stage 2. Developing of the dETECT Evaluation Model. Based on the results of the literature review, we systematically developed the dETECT evaluation model for measuring the quality of instructional units for teaching computing based on the perceptions of the students and their parents. Therefore, we used GQM – Goal Question Metric (Basili *et al.*, 1994), a popular approach to measure diverse quality attributes. Using GQM we systematically defined the evaluation objective(s) and decomposed the objective into analysis questions and measures.

Stage 3. Design of the measurement instrument. In order to operationalize the measurement, a questionnaire was developed by a multidisciplinary team, based on methods for scale and questionnaire development (Devellis, 2016; Krosnick and Presser; 2010; Malhotra, 2008; Kasunic, 2005). For each of the defined measure, questionnaire items have been defined also based on similar studies that were found in literature, considered adherent to the context of this study and to the defined measurement plan. The questionnaire has been revised and piloted with a small sample of the target audience.

Stage 4. Application and evaluation of the measurement instrument. A case study (Yin, 2009; Wohlin *et al.*, 2012) was conducted in order to evaluate the measurement instrument in terms of reliability and construct validity. For the definition of the evaluation, we used the GQM approach (Basili *et al.*, 1994). The objective of the study was decomposed into quality factors and analysis questions also in accordance with methods for scale development (Carmines and Zeller, 1979; Devellis, 2016; Trochim and Donnelly, 2008). During the case study, the dETECT model measuring instrument was applied as part of the evaluation of 16 courses/computing workshops carried out in different educational institutions collecting the required data. The pooled data was analyzed in order to answer our analysis questions, following the definition of Trochim and Donnelly (2008) and the scale development guide proposed by DeVellis (2016). In terms of reliability, internal consistency is typically measured based on the correlations between different items on the same measurement instrument (Carmines and Zeller, 1979; Trochim and Donnelly, 2008). Internal consistency is usually measured through Cronbach's alpha, a popular method to assess the reliability of the measurement instrument (Carmines and Zeller, 1979). In terms of construct validity, convergent and discriminant validity are the two subtypes of validity that make up construct validity (Trochim and Donnelly, 2008). Convergent validity refers to the degree to which two items of quality factors that theoretically should be related, are in fact related. In contrast, discriminant validity tests whether concepts or measurements that are supposed to be unrelated are in fact unrelated (Trochim and Donnelly, 2008). In order to analyze the convergent and discriminant validity of the dETECT measurement instrument, the intercorrelations of the items and item-total correlation are calculated (DeVellis, 2016). Intercorrelation refers to the degree of correlation between the items of a measurement instrument (Carmines and Zeller, 1979; DeVellis, 2016). The higher the correlations among items that measure the same quality factor, the higher the validity of individual items and, hence, the validity of the instrument as a whole. Item-total correlation is analyzed in order to check if any item in the measurement instrument is inconsistent with the averaged correlation of the others, and thus, can be discarded (Carmines and Zeller, 1979; DeVellis, 2016).

In addition, we used factor analysis to determinate how many factors underlie the set of items of the dETECT measurement instrument, following the analysis process proposed by Brown (2006). Each factor is defined by those items that are more highly correlated with each other than with other items. A statistical indication of the extent to which each item is correlated with each factor is given by the factor loading. Thus, the higher the factor loading, the more the particular item contributes to the given factor. Thus, factor analysis also explicitly takes into consideration the fact that the items measure a factor unequally (Carmines and Zeller, 1979).

This research was approved by the Ethics Committee of the Federal University of Santa Catarina (No. 1021541).

3. The Evaluation Model dETECT (Evaluating TEaching CompuTing)

The objective of the dETECT model is to analyze instructional units in order to evaluate the quality in terms of quality of the IUs, computing experience and the perception of learning, from the learners' perspective in the context of teaching computing in middle school. From this objective, the analysis questions and measures are derived based on literature (Fig. 2) (Keller, 1987; Sweetser and Wyeth, 2005; Poels *et al.*, 2007; Takatalo *et al.*, 2010; Ericson and McKlin, 2012; Tangney *et al.*, 2010; Wiebe *et al.*, 2003; Papastergiou, 2008; Sanchez-Franco, 2010; Giannakos *et al.*, 2013; Makris *et al.*, 2013; Shih, 2008; Sivilotti and Laugel, 2008; Lai and Lai, 2012; Lee *et al.*, 2009; Savi *et al.*, 2012; Kwon *et al.*, 2012).

In a general way, following the definition proposed by Wiggins and McTighe (2005), an IU is a set of lessons carefully designed to collectively achieve a selected group of learning objectives for a target audience. The unit consists of a coherent set of materials designed to support student learning in a specific educational context and offers goals, assessment tasks, instruction, implementation procedures, and resources. However, due to the lack of a definition of an IU for teaching computing in schools, based on the literature review (Keller, 1987; Sweetser and Wyeth, 2005; Poels *et al.*, 2007; Takatalo *et al.*, 2010; Ericson and McKlin, 2012; Tangney *et al.*, 2010; Wiebe *et al.*, 2003; Papastergiou, 2008; Sanchez-Franco, 2010; Giannakos *et al.*, 2013; Makris *et al.*, 2013; Shih, 2008;

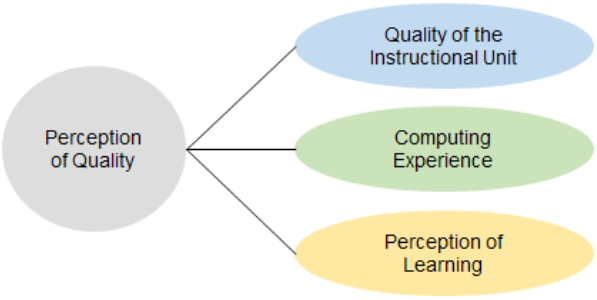


Fig. 2. Decomposition of the quality factors. Source: authors.

Sivilotti and Laugel, 2008; Lai and Lai, 2012; Lee *et al.*, 2009; Savi *et al.*, 2012; Kwon *et al.*, 2012) we consider, that an instructional unit (workshop, course, etc.) with quality achieves its learning objectives, promotes pleasant activities, facilitates learning, and that creates a positive perception and interest for computing.

Table 1
Measurement Instrument

No.	Description	Response Format
Quality of the Instructional Unit		
1	The workshop/course was:	(1) Lot of fun (2) Fun (3) Annoying (4) Very Annoying
2	The time of the workshop/course passed:	(1) Very quickly (2) Quickly (3) Slowly (4) Very slowly
3	The workshop/course was:	(1) Excellent (2) God (3) Regular (4) Bad
Computing Experience		
4	I will show my computer program to others:	(1) Yes (2) No
5	I want to learn more about how to make computer programs:	(1) Yes (2) No
6	Making a computer program is:	(1) Lot of fun (2) Fun (3) Annoying (4) Very Annoying
7	I like to make computer programs:	(1) Yes (2) No
8	Computing is useful in everyday life:	(1) Yes (2) No
9	I want to learn more about how to make computer programs:	(1) Yes (2) No
Perception of Learning		
10	The workshop/course was:	(1) Very easy (2) Easy (3) Difficult (4) Very Difficult
11	I can write computer programs:	(1) Yes (2) No
12	I can explain to a friend how to make a computer program:	(1) Yes (2) No
13	Making a computer program is:	(1) Very easy (2) Easy (3) Difficult (4) Very Difficult

The measurement is operationalized by the development of a questionnaire to be answered by the students at the end of the instructional unit, in order to obtain their perception about the quality of the instructional unit. The items that compose the questionnaire (Table 1) are defined for each of the measures derived from similar studies found in literature considered adherent to the context of this research (Keller, 1987; Sweetser and Wyeth, 2005; Poels *et al.*, 2007; Takatalo *et al.*, 2010; Ericson and McKlin, 2012; Tangney *et al.*, 2010; Wiebe *et al.*, 2003; Papastergiou, 2008; Sanchez-Franco, 2010; Giannakos *et al.*, 2013; Makris *et al.*, 2013; Shih, 2008; Sivilotti and Laugel, 2008; Lai and Lai, 2012; Lee *et al.*, 2009; Savi *et al.*, 2012; Kwon *et al.*, 2012).

The complete material of the dETECT model is available at: `http://www.computacaonaescola.ufsc.br/?page_id=45`.

4. Definition and Execution of the Evaluation of the dETECT Model

When developing evaluation models and questionnaires, it is fundamental to analyze whether they are measuring what is intended (construct validity) and whether the same measurement process produces the same results (reliability) (Carmines and Zeller, 1979). Therefore, we evaluated the measurement instrument of the dETECT model in terms of reliability and construct validity from the viewpoint of researchers in the context of instructional units for teaching computing in school. The following analysis questions are taken into consideration:

Reliability

AQ1: Is there evidence for internal consistency of the dETECT measurement instrument?

Construct Validity

AQ2: Is there evidence of the convergent and discriminant validity of the dETECT measurement instrument?
AQ3: How do underlying factors influence the responses on the items of the dETECT measurement instrument?

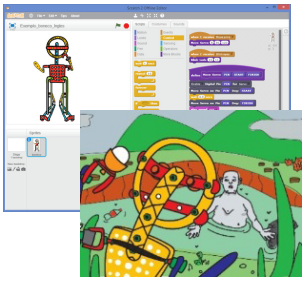
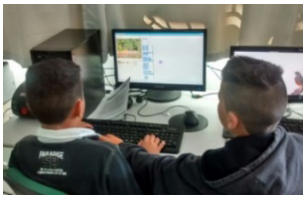
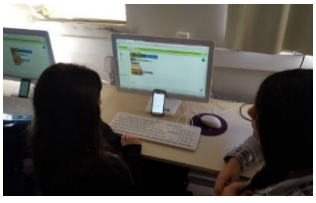
For the evaluation of the dETECT model, 16 case studies were performed applying three different instructional units in 13 different educational institutions between 2015 and 2016, involving a total of 477 students (Table 2). The measurement took place at the end of instructional units teaching computing, either in form of short 4-hours workshops or as part as interdisciplinary school units during 10–12 weeks (with 2 hours weekly). The units have been applied on the educational stage of middle school with children of age 10 to 14.

The target audience is middle school students including different types of activities during the regular school schedule as well as extracurricular workshops in Brazil. The instructional units aim at teaching computing focusing on programming and computational thinking (Table 3). More information on the instructional units is available at: http://www.computacaonaescola.ufsc.br.

Table 2
Summary of case studies

Instructional Unit	Institution/Date	Number of students
Physical Computing Workshop	INE/UFSC – Florianópolis – August 8, 2015	8
	Escola Hamônia – Ibirama/SC – August, 29, 2015	13
	INE/UFSC – Florianópolis – October 17, 2015	14
	IFSC – Gaspar/SC – October 20 and 22, 2015	32
	Escola Sabedoria Junior – Florianópolis/SC – November 4, 2015	22
	INE/UFSC – Florianópolis – November 7, 2015	16
	INE/UFSC – Florianópolis – November 14, 2015	15
Games with Scratch (Interdisciplinary Course)	Turmas 5Mat, 5 Vesp, 7A, 7B – Escola Autonomia, Florianópolis/SC – 2015	99
	Escola Básica Municipal Prefeito Reinaldo Weingartner, Palhoça/SC – 2015	25
	EEB Prof Vitorio Anacleto Cardoso, Gaspar/SC – 2015	43
	EEB Zenaide Schmitt Costa, Gaspar/SC – 2015	31
	EEB Luiz Franzoi, Gaspar/SC – 2015	15
	EEB Ferandino Dagnoni, Gaspar/SC – 2015	46
	EEB Prof Dolores dos Santos Krauss, Gaspar/SC – 2015	14
	EEB Norma Mônica Sabel, Gaspar/SC – 2015	49
App Inventor Workshop	Escola Básica Prof.ª Herondina Medeiros Zeferino – 2016	35
Total		**477**

Table 3
Overview of the instructional units applied

Physical Computing Workshop	Games with Scratch	App Inventor Workshop
		
Integrating Scratch/Snap! with Arduino and pieces of hardware in a low-cost solution, students learn to program an interactive robot.	In an interdisciplinary way students learn basic computer concepts by programming games involving different contents (e.g. history, Portuguese language, geography, etc.) using Scratch.	Student learn how to program a mobile app game using App Inventor.
		

4.1. *Analysis*

In order to obtain greater precision and statistical power through a larger sample size, the data collected in the 16 case studies were pooled to answer the defined analysis questions.

Reliability

AQ1: Is there evidence for internal consistency of the dETECT measurement instrument?

In order to answer this question, we evaluated the internal consistency of the dETECT measurement instrument through Cronbach's alpha coefficient (DeVellis, 2016; Trochim and Donnelly, 2008). Cronbach's alpha coefficient (Cronbach, 1951) indicates indirectly the degree to which a set of items measures a single quality factor. Thus, we want to know whether the dETECT measurement instrument measures the same quality factor, the perception of the quality of the instructional unit. Typically, values of Cronbach's alpha, ranging from 0.70 to 0.95 are considered acceptable (DeVellis, 2016), indicating an internal consistency of the instrument.

Analyzing the 13 items of the measuring instrument (Table 1), the value of Cronbach's alpha is acceptable ($\alpha = .787$). We, thus, can conclude that the answers to the items are consistent and precise, indicating the reliability of the measuring instrument items of the dETECT model.

Construct Validity

AQ2: Is there evidence of the convergent and discriminant validity of the dETECT measurement instrument?

Construct validity of a measurement instrument refers to the ability to actually measure what it purports to measure (Carmines and Zeller, 1979; Trochim and Donnelly, 2008). Convergent and discriminant validity are the two subtypes of validity that make up construct validity (Trochim and Donnelly, 2008). Convergent validity shows that the items that should be related are in reality related. On the other hand, discriminant validity shows that the items that should not be related are in reality not related (Carmines and Zeller, 1979; Trochim and Donnelly, 2008). In order to obtain evidence of the convergent and discriminant validity of the items of the dETECT measurement instrument, the *intercorrelations of the items* and *correlation item-total* are calculated (DeVellis, 2016).

Intercorrelations of the items. In order to analyze the intercorrelations between the items, we used the nonparametric Spearman correlation matrices (Table 4). The matrices show the Spearman correlation coefficient, indicating the degree of correlation between two items (item pairs). We used this correlation coefficient, as it is the most appropriate correlation analysis for Likert scales (Trochim and Donnelly, 2008). The correlation coefficients between the items within of the same dimension are colored. In accordance to Cohen (1988), a correlation between items is considered satisfactory, if the correlation coefficient is greater than 0.29, indicating that there is a medium or high correlation be-

Table 4
Spearman correlation coefficient

Item/Quality factor	1	2	3	4	5	6	7	8	9	10	11	12	13
	Quality of IU			Computing Experience						Perception of Learning			
1	1.000												
2	.268	1.000											
3	**.514**	**.294**	1.000										
4	.243	.090	.230	1.000									
5	**.291**	.137	**.295**	**.362**	1.000								
6	**.452**	.174	**.382**	.266	**.385**	1.000							
7	.211	.160	.181	**.325**	**.378**	**.395**	1.000						
8	.046	.028	.035	.240	**.327**	**.312**	**.345**	1.000					
9	.221	.107	.262	.203	**.324**	.226	.263	.197	1.000				
10	.174	.203	.153	.067	.133	.144	.022	-.086	.125	1.000			
11	.213	.059	.087	.273	**.320**	**.318**	.287	.239	.202	.157	1.000		
12	.162	.094	.135	**.312**	.246	.194	.266	.192	.247	.143	**.450**	1.000	
13	.156	.082	.155	.168	.266	**.325**	.225	.226	.199	**.391**	**.432**	.286	1,000

tween the items. Satisfactory correlations are marked in bold. The numbers of the items are related to the specification presented in Table 1.

Analyzing the interrelations between the items of the three quality factors (Table 4), we can observe that most of the item pairs have medium or high correlation regarding each quality factor. However, some item pairs have a low correlation (e.g., 1–2, 6–9, 10–11). Even so, the results indicate evidence of convergent validity.

On the other hand, some item pairs (e.g., 1–6, 3–6, 5–11) presented medium or high correlation with items of another quality factor. Thus, there is no evidence of discriminant validity. However, the lack of discriminant validity is acceptable, as, although the model is divided into three quality factors, all factors are also related to a single factor, which is the perception of the quality of the IU.

Item-total correlation. This method is complementary to the previous one in order to evaluate the correlation with all the other items. Each item of the instrument should have medium or high correlation with all the other items (DeVellis, 2016), as this indicates that the items present consistency in comparison to the other items. On the other hand, a low item-total correlation of an item undermines the validity of the scale, and, therefore, should be eliminated. Table 5 shows the correlation coefficients between a single item and the other items of the measurement instrument.

We used the method of corrected item-total correlation, which compares one item with every other one of the instrument, excluding itself. Reference values for the analysis are the same as presented in the previous section based on Cohen (1988), considering a correlation satisfactorily, if the correlation coefficient is greater than 0.29. Items with low correlation are marked in bold. In addition, Table 5 also shows the Cronbach's alpha if an item was deleted, expecting that no item elimination should cause a substantial decrease in the Cronbach's alpha (DeVellis, 2016).

Table 5
Corrected item-total correlation

Quality factor	No. Item	Corrected item-total correlation	Cronbach's alpha, if item was deleted
Quality of IU	1	.511	.764
	2	**.269**	.794
	3	.459	.769
Computing Experience	4	.431	.773
	5	.511	.769
	6	.594	.753
	7	.481	.771
	8	.338	.783
	9	.410	.774
Perception of Learning	10	**.280**	.787
	11	.470	.770
	12	.415	.773
	13	.474	.769

In general, item-total correlations are medium and high. Most items demonstrate acceptable item-total correlation and satisfactory values of Cronbach's alpha coefficient, if item was deleted, thus, indicating, the validity of the quality factors. Only the items 2 ("The time of the workshop passed:") and 10 ("The workshop was:") presented a low item-total correlation. In addition, item 2 presents a small increase in Cronbach's alpha if the item was deleted. Consequently, the results indicate that these items need to be reviewed.

AQ3: How do underlying factors influence the responses on the items of the dETECT measurement instrument?

In order to identify the number of factors (quality factors) that represents the responses of the set of the 13 items of the dETECT measurement instrument, we performed a factor analysis.

To analyze whether the items of the dETECT measurement instrument can be submitted to a factor analysis (Brown, 2006), we used the Kaiser-Meyer-Olkin (KMO) index. This method indicates how much the realization of the factor analysis is appropriate for a specific set of items (Brown, 2006). The KMO index measures the sampling adequacy with values between 0.0 and 1.0. An index value near 1.0 supports a factor analysis and anything less than 0.5 is probably not amenable to useful factor analysis (Dziuban and Shirkey, 1974). Analyzing the set of items of the dETECT measurement instrument, we obtained a KMO index of .827. Consequently, it indicates that factor analysis is appropriate in order to analyze the number of factors that represents the responses of the dETECT measurement instrument.

Running a factorial analysis, the number of factors retained in the analysis is decided (Glorfeld, 1995; Brown, 2006). Here we used the Kaiser-Guttman criterion for this decision, as it is the most commonly used method of determining the number of factors. This method states that the number of factors is equal to the number of eigenvalues greater than 1 (Glorfeld, 1995). The eigenvalue refers to the value of the variance of the all the

items which is explained by a factor (Glorfeld, 1995). Following the Kaiser-Guttman criterion, our results show that three factors should be retained in the analysis. Regarding the dETECT model, this means that the responses of the measuring instrument are representing three underlying factors, indicating a decomposition similar to the original definition of the model.

Once identified the number of underlying factors, another issue is to determine which items are loaded into which factor. In order to identify the factor loadings of the items, a rotation method is used (Brown, 2006; Tabachnick and Fidel, 2007). Here we used the Varimax with Kaiser Normalization rotation method being the most widely accepted and used rotation method (Tabachnick and Fidel, 2007). Table 6 shows the factor loadings of the items associated with the three retained factors. The highest factor loading of each item, indicating to which factor the item is most related, is marked in bold.

Analyzing the factor loadings of the items (Table 6), we can observe that, the first factor (factor 1), includes a set of 7 items (4, 5, 6, 7, 8, 9 and 12). Thus, this factor is directly related to the quality factor of the computing experience provided by the instructional unit (Table 1). With the exception of item 12, all items correspond to the referred quality factor in the original structure of the dETECT model. Although, item 12 has the highest factor loading on factor 1, it also presents a similar factor loading (.410) with respect to factor 3, thus, showing that this item contributes to both quality factors (computing experience and perception of learning). Regarding factor 2, a set of three items (1, 2 and 3) is considered. This result seems to suggest that these items are related to the factor related to the quality of the instructional unit of the dETECT model. In fact, these items correspond to the same quality factor (quality of the IU) in the original definition of the dETECT model (Table 1). Analyzing the results of factor 3, it includes a set of three items (10, 11 and 13), indicating that these items are related to a single quality factor (perception of learning).

Table 6
Factor loadings

Quality factor	Item no.	Description	Factors		
			1	2	3
Quality of IU	**1**	The workshop was:	.146	**.763**	-.018
	2	The time of the workshop passed:	-.096	**.619**	.043
	3	The workshop was:	.110	**.800**	-.078
Computing Experience	**4**	I will show my computer program to others:	**.591**	.101	.008
	5	I want to learn more about how to make computer programs:	**.571**	.217	.035
	6	Making a computer program is:	**.510**	.400	.055
	7	I like to make computer programs:	**.683**	.114	-.053
	8	Computing is useful in everyday life:	**.783**	-.213	-.090
	9	I want to learn more about how to make computer programs:	**.401**	.130	.165
Perception of Learning	**10**	The workshop was:	-.425	.239	**.823**
	11	I can write computer programs:	.415	-.177	**.546**
	12	I can explain to a friend how to make a computer program:	**.432**	-.139	.410
	13	Making a computer program is:	.230	-.102	**.720**

5. Discussion

The obtained results show sufficient evidence to consider the reliability and construct validity of dETECT as an acceptable model for the evaluation of instructional units for teaching computing in middle school.

In terms of reliability (AQ1), the results of the analysis indicate an acceptable Cronbach's alpha for all quality factors (Cronbach's alpha α=.787), indicating the internal consistency of the dETECT measurement instrument. Thus, it indicates that the items of dETECT measurement instrument are consistent and precise with respect to the evaluation of instructional units for teaching computing.

In terms of construct validity, with regard to convergent validity (AQ2), we identified that most items have medium and high correlation, mainly between items of the same quality factor (e.g., quality of IU, computing experience, and perception of learning). In this way, we can conclude that there is evidence of convergent validity considering the quality factors. This indicates that the items of the measuring instrument seem to be actually measuring what they intend to measure (e.g., quality of IU, computing experience, and perception of learning). However, some items have a low correlation, both within a single quality factor and in relation to the other factors (e.g., items 4–9). This may be due to the description of the items derived from the ones found in literature, and, thus, may indicate that these items need to be revised.

With respect to discriminant validity, in general, most of the items present a low correlation with items of other quality factors. However, some item pairs (e.g., 1–6, 5–11) have a medium or high correlation with items of another quality factor. Thus, the results do not indicate evidence of discriminant validity. However, in this case, the lack of discriminant validity is acceptable, because, although the model is divided into three quality factors, all factors are also related to a single factor, which is the perception of the quality of the instructional unit, as proposed in the original composition of the dETECT model (Fig. 2).

Analyzing the item-total correlation, again, the majority of the items presents a satisfactory correlation with the other items of the measuring instrument. Thus, indicating that the set of items of the measuring instrument of the dETECT model is related to measure what they propose to measure (perception of quality of an IU).

Based on the results of the factor analysis (AQ3), we identified that the data collected in the case studies are explained by three factors. This confirms the initial structure defined for the dETECT model, clearly grouping the items according to their defined quality factor (quality of IU, computing experience and perception of learning).

Threats to validity

Due to the characteristics of this type of research, this work is subject to various threats to validity. We, therefore, identified potential threats and applied mitigation strategies in order to minimize their impact on our research. Some threats are related to the design of the study. In order to mitigate this threat, we defined and documented a systematic methodology for our study. The dETECT model was defined based on the GQM approach, systematically decomposing the evaluation objective into analysis questions

and measures. The measuring instrument was developed following a scale and questionnaire development methods defined in literature and involving a multidisciplinary team of researchers. In addition, for the evaluation of the dETECT model measuring instrument, a case study was systematically defined and documented. Another risk refers to the quality of the data pooled into a single sample, in terms of standardization of data (response format) and adequacy to dETECT model. As our study is limited exclusively to evaluations that used the dETECT model this risk is minimized as in all studies the same data collection instrument has been used. Another issue refers to the pooled data from different contexts. To mitigate this threat we selected studies which considered only case studies of IUs for teaching computing in similar contexts.

In terms of external validity, a threat to the possibility to generalize the results is related to the sample size and diversity of the data used for the evaluation. In respect to sample size, our evaluation used data collected from 16 case studies evaluating three different instructional units, involving a population of 477 students. In terms of statistical significance, this is a satisfactory sample size allowing the generation of significant results (Clark and Watson, 1995; MacCallum *et al.*, 1999; Kasunic, 2005, Devellis, 2016).

In terms of reliability, a threat refers to what extent the data and the analysis are dependent on the specific researchers. In order to mitigate this threat, we systematically documented the development and evaluation of the dETECT model, defining clearly the study objective, the process of data collection, and the statistics methods used for data analysis. Another issue refers to the correct choice of statistical tests for data analysis. To minimize this threat, we performed a statistical evaluation based on the approach for the construction of measurement scales as proposed by DeVellis (2016), which is aligned with procedures for the evaluation of internal consistency and construct validity of a measurement instrument (Trochim and Donnelly, 2008).

6. Conclusion

Although the evaluation of instructional units for teaching computing is essential for their continuous improvement and effective and efficient application, few efforts are made for the development of evaluation models. In this context, this article presents a first step into this direction taking also into consideration practical limitations when running such evaluations in more informal outreach programs. Based on literature and practical experiences, the evaluation model dETECT and its 13-item measurement instrument have been developed systematically and applied at the end of 16 instructional units in middle school in Brazil.

Results from the analysis of the responses of 477 students indicate that the measurement instrument is acceptable in terms of reliability and construct validity. With respect to reliability, a Cronbach's alpha α=.787 indicates an acceptable internal consistency, which means that the responses between the items are consistent and precise. Our analysis also indicates convergent validity through an acceptable degree of correlation found between almost all items regarding the quality factors. Thus, it suggests

that the measurement instrument of the dETECT model can be a reliable and valid instrument for measuring the students' perception of instructional units for teaching computing. The results of the factorial analysis indicate that three underlying factors influence the responses of the items of the dETECT model measuring instrument confirming the original structure of the model, which defines three quality factors (quality of IU, computing experience and perception of learning) for the evaluation of instructional units.

Acknowledgments

This work was supported by the CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico – www.cnpq.br), an entity of the Brazilian government focused on scientific and technological development and the Google Rise Award.

References

Basili, V.R., Caldiera, G., Rombach, H.D. (1994). *Goal, Question Metric Paradigm*. In: Marciniak, J.J. (Ed.). *Encyclopedia of Software Engineering*. John Wiley & Sons, 528–532.

Brown, T.A. (2006). *Confirmatory Factor Analysis for Applied Research*. 1. ed. New York: The Guilford Press, 475.

Carmines, E.G., Zeller, R.A. (1979). *Reliability and Validity Assessment*. 1. ed. Beverly Hills: Sage Publications Inc., 75.

Clark, L.A., Watson, D. (1995). Construct validity: Basic Issues in objective scale development. *Psychological Assessment*, 7, 309–319.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* Routledge Academic.

Craig, M., Horton, D. (2009). Gr8 designs for Gr8 girls: a middle-school program and its evaluation. In: *40th ACM Technical Symposium on Computer Science Education.* Chattanooga, TN, USA, 221–225.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.

Decker, A., Mcgill, M.M., Settle, A. (2016). Towards a common framework for evaluating computing outreach activities. In: *ACM Technical Symposium on Computer Science Education*. Memphis, TN, USA, 627–632.

Devellis, R.F. (2016). *Scale Development: Theory and Applications*. 4. ed. SAGE Publications, 280.

Dziuban, C.D., Shirkey, E.C. (1974). When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychological Bulletin*, 81, 358–361.

Ericson, B., McKlin, T. (2012). Effective and Sustainable Computing Summer Camps. In: *43rd ACM Technical Symposium on Computer Science Education*. Raleigh, NC, USA, 289–294.

Garland, K.J., Noyes, J.M. (2008). Computer attitude scales: How relevant today? *Computers in Human Behavior*, 24(2), 563–575.

Garneli, V., Giannakos, M.N., Chorianopoulos, K. (2015). Computing education in K-12 schools: A review of the literature. In*: IEEE Global Engineering Education Conference*. Tallinn, Estonia, 543–551.

Giannakos, M.N., Hubwieser, P., Chrisochoides, N. (2013). How students estimate the effects of ICT and programming courses. In: *ACM Technical Symposium on Computer Science Education*. Denver, CO, USA, 717–722.

Giannakos, M.N., Jaccheri, L., Leftheriotis, I. (2014). Happy girls engaging with technology: assessing emotions and engagement related to programming activities. In: *International Conference on Learning and Collaboration Technologies*. Heraklion, Greece, 398–409.

Glorfeld, L.W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement*, 55(3), 377–393.

Gresse von Wangenheim, C., Wangenheim, A. (2014). teaching game programming in family workshops. *IEEE Computer Magazine*, 47(8), 84–87.

Gross, P., Powers, K. (2005). Evaluating assessments of novice programming environments. In: *International Workshop on Computing Education Research*. Seattle, WA, USA, 99–110.

Grover, S., Cooper, S., Pea, R. (2014). Assessing computational learning in K–12. In: *Conference on Innovation & Technology in Computer Science Education*. Uppsala, Sweden, 57–62.

Guzdial, M. (2004). Programming environments for novices. In: S. Fincher and M. Petre (Eds.), *Computer Science Education Research*. Swets and Zeitlinger. Chapter 3.

Guzdial, M., Ericson, B., Mcklin, T., Engelman, S. (2014). Georgia Computes! An intervention in a US state, with formal and informal education in a policy context. *ACM Transactions on Computing Education*, 14(2), article 13.

Kalelioğlu, F., Gülbahar, Y. (2014). The effects of teaching programming via Scratch on problem solving skills: A discussion from learners' perspective. *Informatics in Education*, 13(1), 33–50.

Kasunic, M. (2005). *Designing an Effective Survey*. Carnegie-Mellon University Pittsburgh Pa Software Engineering Inst.

Keller, J.M. (1987). Development and use of the ARCS model of motivational design. *Journal of Instructional Development*, 10(3), 2–10.

Krosnick, J.A., Presser, S. (2010). Questionnaire design.In: J.D. Wright & P.V. Marsden (Eds.), *Handbook of Survey Research*. 2. ed. West Yorkshire, England: Emerald Group.

Kwon, D.-Y., Kim, H.-S., Shim, J.-K., Lee, W.-G. (2012). Algorithmic bricks: a tangible robot programming tool for elementary school students. *IEEE Transactions on Education*, 55(4), 474–479.

Lai, C.S., Lai, M.H. (2012). Using computer programming to enhance science learning for 5th graders in Taipei. In: *International Symposium on Computer, Consumer and Control*. Taichung, Taiwan, 146–148.

Lee, B.C., Yoon, J.-O, Lee, I. (2009). Learners' acceptance of e-learning in South Korea: theories and results. *Computers and Education*, 53, 4, 1–44.

MacCallum, R.C., Widaman, K.F., Zhang, S., Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4, 84–99.

Makris, D., Euaggelopoulos, K., Chorianopoulos, K., Giannakos, M.N. (2013). Could you help me to change the variables? Comparing instruction to encouragement for teaching programming. In: *8th Workshop in Primary and Secondary Computing Education.* Aarhus, Denmark, 79–82.

Malhotra, N.K., Birks, D.F. (2008). *Marketing Research: An Applied Approach.* 3. ed. Trans-Atlantic Publications, 816.

Miller, D.C., Salkind, N.J. (2002). *Handbook of Research Design and Social Measurement*. 6. ed. SAGE Publications, 808.

Papastergiou, M. (2008). Are computer science and information technology still masculine fields? High school students' perceptions and career choices. *Computers and Education*, 51(2), 594–608.

Perdikuri, K. (2014). Students' Experiences from the use of MIT App Inventor in classroom. In: *Panhellenic Conference on Informatics*. Athens, Greece, 1–6.

Poels, K., Kort, Y.D., Ijsselsteijn, W. (2007). It is always a lot of fun!: exploring imensions of digital game experience using focus group methodology. In: *7th Conference on Future Play*. Toronto, 83–89.

Sanchez-Franco, M.J. (2010). WebCT – the quasimoderating effect of perceived affective quality on an extending technology acceptance model. *Computers and Education*, 54, 1, 37–46.

Savi, R. *et al.* (2012). *MEEGA – A Model for the Evaluation of Games for Teaching Software Engineering*. Technical Report INCoD – Nº 001/2012 – E – GQS, Federal University of Santa Catarina, Florianópolis/Brazil. `http://www.incod.ufsc.br/meega-a-model-for-the-evaluation-of-games-for-teaching-software-engineering`

Shih, H. (2008). Using a cognitive-motivation-control view to assess the adoption intention for Web-based learning. *Computer and Education*, 50, (1), 327–337.

Sivilotti, P.A.G., Laugel, S.A. (2008). Scratching the surface of advanced topics in software engineering: a workshop module for middle school students. In: *Technical Symposium on Computer Science Education. Portland*, OR, USA, 291–295.

Sweetser, P., Wyeth, P. (2005). GameFlow: a model for evaluating player enjoyment in games. *Computers in Entertainment*, 3(3), 1–24.

Tabachnick, B.G., Fidell, L.S. (2007). *Using Multivariate Statistics*. 5th. ed. Boston: Allyn and Bacon.

Tangney, B., Oldham, E., Conneely, C., Barrett, S., Lawlor, J. (2010). Pedagogy and Processes for a Computer Programming Outreach Workshop – The Bridge to College Model. *IEEE Transactions on Education*, 53(1).

Takatalo, J., Häkkinen, J., Kaistinen, J., Nyman, G. (2010). Presence, involvement, and flow in digital games. In: Bernhaupt, R. (Ed.). *Evaluating User Experience in Games: Concepts and Methods*, Springer, 23–46.

Trochim, W.M., Donnelly, J.P. (2008). *Research Methods Knowledge Base.* 3. ed. Mason, OH: Atomic Dog Publishing, 361.

Weintrop, D., Wilensky, U. (2015). To block or not to block, that is the question: students' perceptions of blocks-based programming. In: *Conference on Interaction Design and Children*. New York, NY, USA, 199–208.

Wiebe, E., Williams, L., Yang, K., Miller, C. (2003). *Computer Science Attitude Survey. Technical Report*. North Carolina State University at Raleigh, Raleigh, NC, USA.

Wiggins, G., McTighe, J. (2005). *Understanding by Design*, 2nd Edition. ASCD: Association for Supervision and Curriculum Development.

Wilson, A., Moffat, D.C. (2010). Evaluating Scratch to introduce younger schoolchildren to Programming. In: *Annual Workshop of the Psychology of Programming Interest Group*. Leganés, Spain

Wohlin, C., Runeson, P., Host, M., Ohlsson, M.C., Regnell, B., Wesslen, A. (2012). *Experimentation in Software Engineering*. Springer, 236.

Yin, R.K. (2009). *Case study research: design and methods*. 4. ed. Beverly Hills: Sage Publications, 312.

C.A.G. von Wangenheim, PMP is a professor at the Department of Informatics and Statistics (INE) of the Federal University of Santa Catarina (UFSC). Her main research interests are in the area of software quality evaluation and computing education. She received the title Diplom-Informatikerin (Master in Computer Science), with parallel qualification in Production Engineering at the University of Kaiserslautern (Germany) in 1995, the title Dr. Eng. at the Graduate Program in Production Engineering at the Federal University of Santa Catarina in 2000 and the title Dr. rer. nat. in Computer Science at the University of Kaiserslautern (Germany) in 2002. She is also PMP – Project Management Professional and Assessor and Implementor MPS.BR. She is also the coordinator of the initiative Computing at Schools visioning computing education at primary schools.

G. Petri is a professor at the Polytechnic School of the Federal University of Santa Maria (UFSM). Currently, he is a PhD. student in the Graduate Program in Computer Science (PPGCC) at the Federal University of Santa Catarina (UFSC). He received a bachelor's degree in Information Systems (2009) and a master's degree in Computer Science (2013). His main research interests are in the area of computing education and educational games.

A.W. Zibetti is a professor at the Department of Informatics and Statistics (INE) of the Federal University of Santa Catarina (UFSC). His main research interest focuses on the statistical and mathematical modelling, data analysis and stochastic processes.

A.F. Borgatto is a professor at the Department of Informatics and Statistics (INE) of the Federal University of Santa Catarina (UFSC). His main research interest focuses on the Item Response Theory applied in the area of education.

J. Hauck holds a PhD in Knowledge Engineering and a Master's Degree in Computer Science from the Federal University of Santa Catarina (UFSC) and a degree in Computer Science from the University of Vale do Itajaí (UNIVALI). He held several specialization courses in Software Engineering at Unisul, Univali, Uniplac, Uniasselvi, Sociesc and Uniarp. He was a visiting researcher at the Regulated Software Research Center – Dundalk Institute of Technology – Ireland. He is currently a Professor in the Department of Informatics and Statistics at the Federal University of Santa Catarina.

F.S. Pacheco received an undergraduate degree, a master's degree and a doctorate degree in Electrical Engineering at the Federal University of Santa Catarina. He is currently professor at the Federal Institute of Santa Catarina, where he works in technical and superior courses. In addition to experience in the area of Signal Processing, Prof. Fernando has worked in extension projects to disseminate computing and engineering knowledge in primary education.

R.M. Filho is an undergraduate student of the Computer Science course at the Federal University of Santa Catarina (UFSC) and a scholarship student at the initiative Computing at Schools.