Chapter 13
Innovative Assessment Technologies in Educational Games Designed for Young Students

Benö Csapó, András Lörincz, and Gyöngyvér Molnár

13.1 Introduction

Feedback is an essential process in regulating complex systems, and it can be found at every level and unit of an efficient educational system, from macrolevels, including entire national education systems, to microlevels of learning processes, including computer games. Therefore, feedback is the overarching concept that helps to explain and interpret the role of assessment in educational games.

Feedback involves collecting and processing information about the actual state of a system represented by some key variables and comparing it to certain predefined standards or normative data. Collecting information may involve a number of means, but for assessing some key target variables of education (e.g., students' knowledge and skills), testing has been considered as the most objective and reliable way. Feedback which is used by the learner is considered the most important, distinctive attribute of formative assessment (Taras, 2005).

For almost a century, paper-and-pencil tests have been used for educational assessment, but since the emergence of the first computers, they have been used for testing students' knowledge as well. Currently, computerized testing, or more generally, technology-based assessment (TBA) is the most rapidly developing area of educational evaluation (Csapó, Ainley, Bennett, Latour, & Law, 2012).

Computerized educational games, on the other hand, focus on teaching, but to maximize their functionality, several assessment mechanisms are embedded in the games to control the learning processes and guide students through the learning

B. Csapó (✉) • G. Molnár
Institute of Education, University of Szeged, Petőfi sgt. 30-34, Szeged 6722, Hungary
e-mail: csapo@edpsy.u-szeged.hu; gymolnar@edpsy.u-szeged.hu

A. Lörincz
Department of Software Technology and Methodology, Eötvös Loránd University, Pázmány Péter sétány 1/C, Budapest 1117, Hungary
e-mail: andras.lorincz@elte.hu

D. Ifenthaler et al. (eds.), *Assessment in Game-Based Learning: Foundations, Innovations, and Perspectives*, DOI 10.1007/978-1-4614-3546-4_13,

tasks. In most serious games, one of the functions of assessment and feedback is to adapt the actual challenge to the cognitive level of the gamer. This procedure, the *dynamic difficulty adjustment* (see, e.g., Westra, Dignum, & Dignum, 2011), is based on monitoring cognitive processes, and matching the complexity of the tasks to the level of gamers ensures optimal learning. The novel feature of the assessment we are experimenting with is monitoring the affective states of the student while playing the game. The feedback gained in this way may be used to optimize the emotional aspects of the gaming process. Due to these similarities, there are areas where TBA and teaching games are converging, and a number of innovations, including detection of emotional states, may be applied in both fields in similar ways.

The project from which this study stems from has been dealing with the form of assessment which is considered very close to the educational games. An *Online Diagnostic Assessment System* (ODAS) is being devised, which, when fully developed, will be able to regularly assess students' cognitive development in three main domains, reading, mathematics, and science, in the first six grades of primary school. The aim of the diagnostic assessment is to identify students' developmental deficiencies and learning difficulties in order to help them to cope with the challenges and overcome difficulties. Diagnostics should be followed by intervention that helps mastering some key concepts, supports understanding, fosters students' skills, and accelerates the development of their abilities. The most obvious method of delivering intervention materials is the utilization of the same online technology which is used in the ODAS.

The main function of diagnostic assessment is to directly support teaching and learning; therefore, it is essentially embedded in instructional processes. The detailed student level feedback information provided by the online assessment can be used to tailor and customize intervention. Therefore, both pedagogical principles and technological conditions suggest the application of teaching games for individualized compensatory instruction.

The first phase of the project has focused on framework development, the establishment of an online platform, and the construction of assessment items. The next phase will aim at devising a number of educational games to compensate for students' learning deficiencies. However, several existing games have been explored, and some new ones have been devised and piloted in the first phase as well. One of such a piloting work is forming the empirical basis of the present study.

In the first part of this chapter, we outline a conceptual framework of assessment in which we describe the parallel functional and technological developments between educational assessment and teaching games. In the second part, we show how innovations are applied in these areas and how they improve the feedback in both systems. We describe the role of contextual information in providing better feedback and introduce the video-based analyses of facial expressions of subjects completing online tests or playing games. In the third part, the piloting work of a game-based training will be presented. This part illustrates how games can be applied for training students for whom ODAS indicates learning deficiencies. Finally, we outline how the elements presented in the first parts of the chapter can be integrated into a complex individualized teaching system in which technology supports both identifying and treating learning problems.

13.1.1 Feedback and Assessment in Education

In the past decades, most impressive developments in education can be attributed to the improved feedback built in several levels of the system. Education, as any other complex system cannot be improved without proper feedback mechanisms. Setting goals, carrying out interventions, assessing the results, and then comparing goals and results are the basic stages of control in any unit of an educational system. The most visible educational feedback systems are the large-scale international projects which assess global outcomes of the entire national educational systems. These international projects, like PISA, TIMSS, and PIRLS, generate feedback information for decision-makers at the national educational policy level.

The unified efforts of large international expert groups advanced educational assessment in a number of fields, such as setting goals (analysis of knowledge to be assessed and framework development); devising assessment instruments; sophisticated methods of data analysis, which include more contextual information (e.g., students' attitudes and their socioeconomic status) for presenting more functional and applicable feedback; and new reporting styles which include visual and innovative presentation of the results (see, e.g., the PISA reports).

There are two developments in international assessments which are closely related to the issues of teaching games:

1. The limitations of paper-based assessment have been reached, and the shift to TBA has been started.
2. New areas of assessment have been explored which include general thinking abilities. For example, problem solving was assessed in PISA 2003 (see OECD, 2004), and dynamic problem solving (Greiff & Funke 2009, 2010) will be assessed in PISA 2012.

International projects draw on the advances of educational measurement in some countries, and later the scientific and technical outcomes of the international projects are utilized in several other areas of assessment; for example, many developed countries introduced a national assessment system. The national assessment systems mostly provide school level feedback which then can be used for institutional improvement and accountability.

The third level of assessment provides feedback at student level and helps directly the teaching and learning processes. This level requires frequent and detailed assessment and rapid feedback. These requirements cannot be satisfied by the traditional paper-based testing.

13.1.2 Technology-Based Assessment

Due to the developments described in the previous section, there is a growing interest in developing TBA systems and making them available for broad everyday use.

Many international[1] and national[2] initiatives aim at developing new assessment systems utilizing the potential of information-communication technology.

There are a number of advantages technology offers for assessment (see Csapó et al., 2012). Traditional domains can be assessed with a greater precision and efficiency. By the means of technology, assessment can be extended to new domains which cannot be assessed by other means. These are domains where technology is essential for the definition of the construct (e.g., ICT literacy, problem solving in technology-rich environment, reading electronic texts, etc.) and domains where technology is instrumental for the assessment (e.g., assessing dynamics, teamwork through network connection; see Tzuriel, 1998).

Technology accelerates data collection, supports real-time automatic scoring, speeds up data processing, and allows immediate feedback. Technology improves the precision of measurements as well. A variety of instruments may be used for data entry and response capture (innovative use of traditional input instruments, touch screen, drawing, microphone with voice recognition, video camera with analysis software, specific interfaces for capturing complex movements), and in this way, large amounts of data can be collected within relatively short periods. Instead of providing single indicators, such as a test score, TBA may produce rich, contextualized, well-structured data sets. Assessment data can easily be stored and analyzed, and this possibility supports the transition from single testing to complex systems of assessments.

Technology revolutionizes the whole process of assessment, including item development (authoring software, automatic item generation). TBA supports item banking, storing of items, and item metadata in large databases. It also vitalizes testing situation, increases motivation, and may improve validity. TBA allows innovative task presentation, including multimedia (sounds, animation, video, simulation).

Technology supports adaptive testing which means that the actual item presented depends on the success of the student in solving the previous item. Therefore, in computerized adaptive testing (CAT), items are scored real time, and a decision about the next step is made depending on the result. In this feature, CAT is similar to the assessment embedded in teaching games.

13.1.3 Assessment for Learning: Integrating Assessment into Teaching

Large-scale assessment projects aim at assessing outcomes of usually at least one, but more frequently several years of learning. Therefore, the actual *summative tests* may cover only a small sample of the entire knowledge to be assessed.

[1] See the Assessment and Teaching of 21st Century Skills project, http://atc21s.org.

[2] At present, the US Race to the Top Assessment Program is the largest national initiative.

Student level feedback requires a different approach, and for this purpose, *formative* and *diagnostic* tests are applied (Ainsworth & Viegut, 2006; Black, Harrison, Lee, Marshall, & Wiliam, 2003; Clarke, 2001, 2005; Leighton & Gierl, 2007). As the feedback in this case is used to control learning processes and to adapt the next phase of learning to the actual needs of the student, the test should cover every relevant area of the students' knowledge which is an essential precondition for a later learning task.

There are several consequences of this requirement. First, formative and diagnostic tests should be built on a careful analysis of the domain (see, e.g., Seel, 2010; Spector, 2010; Strasser, 2010). A model of the structure of knowledge is needed to describe how the pieces of knowledge are related to each other (e.g., a hierarchy of skills and component skills). Second, a large number of items have to be constructed to cover the domain in sufficient details. Third, formative and diagnostic tests should be administered to students frequently enough, so that learning problems could be identified early enough, and the necessary interventions could be implemented. Frequent diagnostic assessment may prevent the accumulation of deficiencies.

There are several problems with this ideal model of applying formative tests. First, paper-based testing is expensive, scoring may require a lot of work, and the feedback may be too late for being efficient. Second, learning and assessment may compete for the same instructional time; too much time spent for testing may endanger learning. Third, if we want to administer paper-and-pencil tests matched to the individual needs of students in different phases of development, what is always the case in practice, complex logistics is required. Because of these difficulties, formative and diagnostic assessment may not be systematically implemented in the regular classrooms.

TBA may be a solution for these problems. A testing center serving a large student population may reduce developmental costs per student to a reasonable level. Online delivery reduces costs, and applying a certain level of adaptive testing (CAT or multistage testing) may help to adjust the actual assessment to the needs of individual students. Some of the testing time may be regained if feedback information is accompanied by some brief immediate customized (online) tutoring. Formative assessment which may be efficient in practice does not only provide students with feedback information, but at the same time, it promotes their learning as well.

13.1.4 Learning and Assessment in Educational Games

Teaching games represent the other side of the coin. They are designed to support students' learning (e.g., Meyer & Sorensen, 2009), but at certain points, they assess students' knowledge as well. To optimize the use of instructional time, the increase of the direct impact of the assessment on learning was proposed in the previous section. As far as optimizing the time from the perspective of teaching games, and multiple utilization of time spent playing games are concerned, a similar approach is proposed here. The information gained by the assessment within a teaching game should be made available outside of the game.

From this perspective, serious games and computerized diagnostic assessment systems may be considered as similar instruments. Both may be used for assessment and, although to different extents, for teaching as well. Beyond the assessment of the outcome variables, teaching games may be utilized to gather other types of information about the learner which in turn helps to further optimize the learning process.

Similar functions of serious games and TBA systems promote the convergence of the two systems. Research on serious games and on TBA may mutually fertilize the other fields.

The development of the ODAS and the utilization of teaching games for intervention aim at benefiting from these fertilizing effects. Given that the same infrastructure is utilized for both aims, further integration seems possible. Integrating the two systems may result in further positive effects in other domains as well. As for the affective aspects, games are associated with pleasure and enjoyment, while assessment is linked to stress and anxiety. Gaming is driven by intrinsic motivation, while in case of assessment, extrinsic motivation is dominating. In order to reduce anxiety and improve motivation, teaching games should be utilized more frequently for assessment purposes.

13.2 Innovative Assessment Technologies for Logging and Analyzing Metadata

13.2.1 Recording and Analyzing Contextual Information

As we mentioned earlier, contextual information is playing a growing role at every level of educational assessment. The information gathered this way contributes to understanding the examined phenomenon and helps to explain what influences the actual values of the observed variables. For example, students' socioeconomic status can be used to estimate the "added value" of the school, the proportion of variance that can be attributed to the school, if the results are controlled for students' social background.

Such contextual information may be essential for the type of assessments we discussed earlier: the online diagnostic assessments and the assessments related to educational games. The standard gaming and testing situation in these cases are very similar or identical: the gamer/testee sits in front of a computer which is equipped with several response capture or input instruments. With these instruments, virtually every controlled or unconscious reaction of testee can be recorded and logged and can be analyzed separately or in relationship to the targeted cognitive achievement variables.

There are data entry instruments which are standard equipment for every computer, such as keyboard and mouse. Microphone and webcam are also broadly available, while touch screen displays, including monitors with large touch screen surfaces, may be purchased at reasonable prices. Some other instruments, such as

gaze tracking equipment, are already in use in cognitive laboratories. If specific pieces of equipments routinely used in physiology and cognitive neuropsychology research are also considered, the possibilities are really unlimited; heartbeat rate, breathing rate, and brain activity can also be monitored.

In this chapter, we only consider those instruments which are broadly available and can be used in an average school computer laboratory, and present the analyses of data collected by webcam in details.

Logging keystrokes are the most common way of collecting metadata. Recording time between keystrokes allows analysis of students' reasoning speeds and thinking and test-taking strategies. Especially rich datasets can be collected if a testee may scroll up and down between items and may revise the solutions of the items. Guessing, e.g., can be identified this way, and it can be checked if solving one item can prompt the revision of the solution of another item. This logging is allowed by the platform used for ODAS, and the related analyses can be carried out any time.

13.2.2 Using Video for Detecting Head Movement and Facial Expression

Special tools using infrared light have been developed for gaze direction estimation and communication by gaze interaction.[3] They have been applied to assess what is salient (Itti, 2007) and what is relevant visual information under free viewing conditions (Peters, Iyer, Itti, & Koch, 2005) and in task-related behavior (Renninger, Verghese, & Coughlan, 2007) and what drives visual attention (Baluch & Itti, 2010). The cost of the infrared instruments, however, prohibits widespread utilization, e.g., in the classroom or in mobile phone applications. There is an ongoing and quick change that decreases the cost of the tools and will increase the number of participants by orders of magnitudes in the near future.

Advances of computational power and the availability of webcams on computers, e.g., laptops, their utilization in games and in assisting technologies (Hévízi, Gerőfi, Szendrő, & Lörincz, 2005), and their widespread use in video chats accelerated face-related evaluation technologies, such as facial expression of estimation and gaze direction estimation. Enabling technology components of webcam-based monitoring and automated annotation includes:

1. Efficient open source computer vision libraries in C and C++)[4]
2. An efficient face detector (Viola & Jones, 2001)
3. Two- and three-dimensional face landmark identification algorithms using either active appearance models (Matthews & Baker, 2004) or constrained local models (Cristinacce & Cootes, 2008; Saragih, Lucey, & Cohn, 2011)
4. Gaze direction estimations (Ishikawa, Baker, Matthews, & Kanade, 2004)

[3]COGAIN—Communication by Gaze Interaction. http://www.cogain.org/wiki/Main_Page.

[4]Open Source Computer Vision Library: http://en.wikipedia.org/wiki/OpenCV.

Annotated databases on faces and facial expressions are available from Carnegie Mellon University,[5] University of Basel,[6] HUMAINE (HUman-MAchine INteraction Network on Emotion),[7] the database of the Rensselaer Polytechnic Institute,[8] among many others.

High-quality facial expression estimation is in the focus of interest, and the main conferences as well as research networks are measuring progress at each possible occasion on different benchmarks (see, e.g., the CVPR 2011,[9] the ICCV 2011[10] conferences).

High-quality videos, off-line evaluations, and infrared light measuring techniques have shown the potentials on learning special individual facial gestures, predicting performance and attention levels, e.g., from blink rates (Chermahinia & Hommel, 2010) to mention only one example of the many behavioral signs. Interviews, avatars, and multiplayer games have been used to provoke, detect, measure, and evaluate intentional and unintentional facial expressions in social interactions, including intentional deception and subconscious emotions and microexpressions (see, e.g., Biland, Py, Allione, Demarchi, & Abric, 2008; Ekman, 2006; Porter & ten Brinke, 2011). It is expected that low-cost webcams will eventually enable real-time estimations of user intentions and other hidden parameters, including cognitive and emotional profiles through monitoring performance, development, and facial expressions during games, training sessions, and interactions with human partners and avatars in real situations. The main challenges include robust head pose independent facial expression estimation, robustness against light conditions, and subject to occlusions. Partial solutions to these challenges have been worked out in the literature (see, e.g., Gross, Matthews, Cohn, Kanade, & Baker, 2010; Jeni, Hashimoto, & Lörincz, 2011; Saragih, Lucey, & Cohn, 2011).

Keystrokes and mouse movement are already widely used during internet searches to characterize and predict users and their intentions during surfing the internet; if one types "(user monitoring) and (advertisement)" after 2009 into Google's Scholar, then a large number of patents appear as the most important hits.

Our intention is to include such innovative tools into education in order to better characterize the students and to improve personalization of the training materials for them. With regards to personalization, machine learning techniques supporting collaborative filtering and recommender systems have also undergone considerable developments in recent years. In this task, one assumes a large matrix containing scores of users on subject matters, e.g., grading of videos or grade points received at courses. These matrices are partially filled since only a small fraction of videos (courses) are seen (taken) by individuals. The question is what should be the next video or course that gives rise to the best grading or added value.

---

[5] Cohn-KanadeAU-CodedFacialExpressionDatabase: http://vasc.ri.cmu.edu/idb/html/face/facial_expression/.

[6] Basel Face Model: http://faces.cs.unibas.ch/bfm/main.php?nav=1-0&id=basel_face_model.

[7] http://humaine-emotion.net.

[8] RPI ISL FaceDatabase: http://www.ecse.rpi.edu/~cvrl/database/ISL_Face_Database.htm.

[9] http://clopinet.com/isabelle/Projects/CVPR2011/home0.html.

[10] http://fipa.cs.kit.edu/befit/workshop2011/.

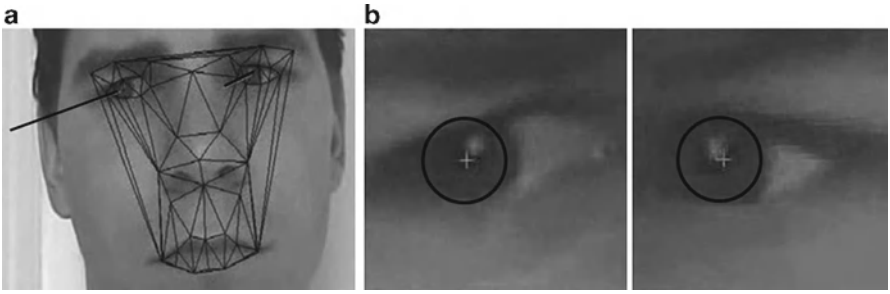

Fig. 13.1 Estimation of facial landmarks (**a**), iris borderlines (**b**), and center positions of the pupils (**c**). (**a**) Face with markers. Markers are connected with *lines*. Estimated gaze direction is shown by *straight lines*. (**b**) *Right and left* eyes with the iris. Estimated borderlines of the iris are shown by *circles*. *Plus* signs depict the estimated positions of the centers of the pupils

The problem is called matrix completion in mathematics. The problem is feasible if the matrix is of low rank. Recent developments showed that under certain, fairly restrictive, but still rather general conditions, the missing values can be filled in "exactly" (Candès & Recht, 2008; Candès & Tao, 2009; Chen, Xu, Caramanis, & Sanghavi, 2011). The method has been applied to a number of problems. For a fairly comprehensive list on methods and applications, see the Nuit-Blanche blogspot.[11] Recently, group-structured dictionary learning methods (Jenatton, Obozinski, & Bach, 2010) have been introduced to collaborative filtering since these methods search for the low-rank subspace *and* for more sophisticated structures (Szabó, Póczos, & Lörincz, 2011). The long-term goal is to keep the student in the zone of his/her proximal development as predicted by collected data on students' learning trajectories and teachers' experiences (Tudge, 1992). This problem, i.e., the task to make recommendations for users about subject matters as a function of time, has been approached in the literature recently (Gantner, Rendle, & Schmidt-Thieme, 2010; Thai-Nghe, Drumond, Horváth, Nanopoulos, & Schmidt-Thieme, 2011).

During the spring semester of 2010–2011, we collected video information during the pilot study that we describe in the next chapter. We also collected videos with nonspeaking, severely constrained, but speech understanding children. We are evaluating the collected materials. Using the experiences, we also utilize and develop tools for aiding automated annotations. At present, we can detect in many cases (1) if the student is present, (2) if she/he is engaged with the training material visible on the screen, or not, and (3) if she/he is talking (Fig. 13.1).

However, in other cases, our fits are not sufficiently precise, especially for large pose angles and untypical light conditions. For certain important cases, we develop special face model to improve tracking (Fig. 13.2). Model construction, however, is cumbersome. Collection of massive databases requires further improvement of our software.

[11] http://nuit-blanche.blogspot.com/.

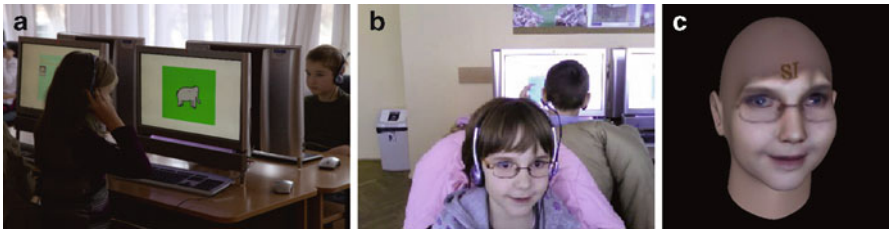

Fig. 13.2 Classroom, video recording, and a model of the user. (**a**) Working with the software in the classroom. (**b**) A video frame recorded by the webcam. (**c**) Model of the user (Model constructed with FaceGen Modeller: http://www.facegen.com/modeller.htm)

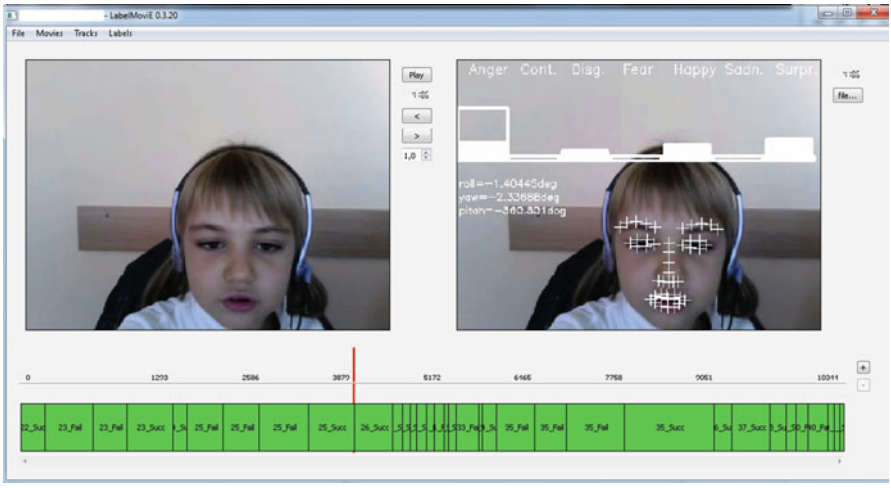

Fig. 13.3 Outputs of the facial expression analysis: *Left hand side* shows the original image. *Right hand side* shows the results of the evaluation. *Crosses*: marker points of the annotation. Roll, yaw, pitch: angle of head pose. Anger, etc.: emotions. *Solid white rectangles*: individual estimations of the different emotion sensors that may be misguided by similarities of the different emotions. *Larger open rectangle*: classification of the emotion

At present, we are evaluating the data and are looking for correlations with the performance measures that we collected. We are in the process of improving our face tracker and emotion estimation algorithms. We need to improve monitoring time since—at present—a relatively large portion of the data cannot be analyzed because of occlusions, e.g., if part of the face is out of the view of the camera or if it is covered by the hand. Similarly, improvements of robustness against light conditions and head pose angles are desired. These works are in progress at the moment.

We show two examples in Fig. 13.3. They exemplify additional research problems that automated annotation is facing and should solve. Notably, facial expressions during active participation (a) might need the analysis of a whole series of

images in the context of the task and (b) can be distorted by mouth movements related to the actual choice. Analysis says that the student in Fig. 13.3a is happy. However, she is uncertain about the solution of the problem and is about to take the risk. This can be inferred from the frame series (Fig. 13.3b) and the context of the task on the screen, but not from this single frame.

13.3 Game-Based Training of 6- to 8-Year-Old Students' General Thinking Abilities: A Pilot Study

The purpose of this pilot study is to investigate the opportunities and effectiveness of applying teaching games following the results of online diagnostic tests for compensating students' learning difficulties. As ODAS is still in experimental phase, no real feedback information is available yet at the main assessment domains (reading, mathematics, and science). However, several online assessments have been carried out to measure the achievements on an inductive reasoning test (see Csapó, Molnár, & Tóth, 2009). Therefore, the development of inductive reasoning by a teaching game was piloted, as a model for further similar computer games at other domains.

The training is based on Klauer's theory of inductive reasoning (Klauer, 1989; Klauer & Phye, 2008) and consists of 120 learning tasks integrated into a game, which can be solved through inductive reasoning. To verify the hypothetical assumptions, a 4-week pilot study was implemented. First and second grade students constituted the experimental and control group.

13.3.1 Methods

13.3.1.1 Participants

First and second grade students constituted the experimental group ($n=42$), who were diagnosed with developmental deficiencies and where it seemed essential to enhance the development of students' inductive reasoning skills. The performance of these students proved to be significantly lower than 50%. The control group consisted of students from the same grade in the same elementary school ($n=64$) with similar socioeconomic background (parents' education, number of owned books at home, own computer with internet connection, own room at home, etc.), but their achievement was 50% or above.

13.3.1.2 Instruments

The game-based inductive reasoning training consisted of 120 learning tasks integrated into a game, which can be solved through the application of appropriate inductive

reasoning processes. The games are designed for young children, which means that they have to meet several specific requirements compared to some traditional games: (1) the images, objects, and problems were fit into the program according to the interests of today's children and the stories they are familiar with; (2) touch screen computers were used during the study to eliminate the possible effect of mouse usage skills; (3) headsets were used to avoid the influential factor of reading skills by the training; and (4) special attention was paid to the task and help giving to ensure the interactivity of the games. Students perceived the training as playing games, not as learning. For a more detailed structure of the training, see Molnár (2011).

The effectiveness of the training was measured with a computer-based test of inductive reasoning (delivered by the ODAS), developed specifically for young learners. The test consisted of 37 items. When devising the items, special attention was paid to ensure the nonverbal character of the test. The reliability index of the whole test was Cronbach $\alpha = 0.87$.

The background questionnaires were filled out by the parents. By means of the paper-based parent questionnaire, we intended to gain information about students' socioeconomic background variables and motivation regarding the game-based training. A five-point Likert scale (1: strongly disagree…5: strongly agree) was used to explore students' attitude and motivation regarding the game-based training.

13.3.1.3 Procedures

In the first phase, the sample was divided in two groups according to students' inductive reasoning skill level. Students with lower skill level belonged to the experimental group, while the remaining part of the sample belonged to the control group. In the evaluation study, students were given the training individually. The time required for the work of development depended on the individual students. It was recommended that each session should last for 40 min and contain 20 tasks at most. This meant that the 120 tasks were divided into six sessions on average, depending on the students' skill level, ability to concentrate, motivation, and level of exhaustion. Every student received permanent feedback during the training after each game. This type of formative assessment, the real-time automatic scoring provided students not only with feedback, but it also supported their learning process directly. They could only get to/access the next game only if they managed to provide right solution/answer for the previous one. In other words, students had to repeat every game as long as they did not get the right solution.

The test-based data collections took place before and immediately after the training process. The interval between the pretest and the posttest was 1 month, the period during which the training was performed. To measure the stability of the training effect, a third data collection was conducted 1 month after the end of the training in the experimental group. All groups took the same reasoning test.

Besides the test-based data collection, innovative assessment technologies are explored by logging and analyzing metadata, such as keystrokes, mouse movement, head movement, and facial expressions. These data were collected by means of web cameras.

Table 13.1 Means and standard deviations of the inductive reasoning test (%)

	Pretest		Posttest		Follow-up test	
Group	*M*	SD	*M*	SD	*M*	SD
Experim. group ($n=42$)	28.3	7.9	43.2	9.9	43.7	12.5
Control group ($n=64$)	70.0	10.5	70.8	9.6	–	–

13.3.2 Results of Training

Significant differences were found between the performance of the experimental and the control group ($t=-21.1$, $p<0.00$) prior to the experiment. On the posttest, the control group still significantly outperformed the experimental group ($t=-13.1$, $p<0.00$); however, the differences were significantly lower (see Table 13.1).

There was no significant change in performance in the control group in this period of time ($t=-0.81$, $p=0.42$), while the experimental group managed to achieve significant development in the experimental period ($t=-9.4$, $p<0.00$). A month after the end of our training program, the follow-up study still indicated a significant ($p<0.001$) improvement in the inductive reasoning skills of the experimental group. The effect of the training proved to be stable over time.

In case of the experimental group, the comparison of the distribution curves for the pre- and posttest indicates that each member of the experimental group attained significant improvement in performance as a result of the training (see Fig. 13.4). However, despite the training, the distribution curve of the experimental group in the posttest still inclined to the left, indicating the need for more training. The control group has normal distribution curves in both of the pre- and posttest.

These results are supported by the two diagrams in Fig. 13.5 that show the changes in experimental and control group performance at student level. The performance levels recorded during the first and second data collection are projected onto each other. The abscissa shows comparative performance from the first data collection stage, and the ordinate displays this from the second. The symbols for students who performed identically in the two cases fall on the line. If a symbol is positioned above the line, it means that the given student showed a development between the two data collection points, while if it is below the line, it represents worse performance on the posttest than on the pretest. The broken lines indicate one standard deviation.

In case of the experimental group (see graph on the left), the symbols are distributed homogeneously around the mean line; i.e., the majority of these students performed better in the posttest than in the pretest. There were no students in the experimental group whose performance dropped significantly from pretest to posttest. Several students improved by more than one standard deviation; moreover, there was one participant who reflected a development of more than 40%. As the effect of the training, several students reached the developmental level of students in the control group, which consisted of students without diagnosed developmental deficiencies. A different tendency is displayed on the right-hand graph, showing the performance of the control group.

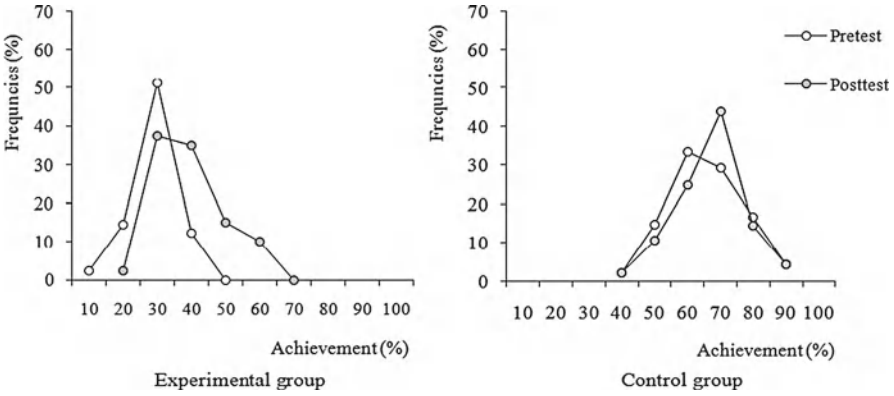

Fig. 13.4 Distribution curves of experimental and control groups in the pre- and posttest

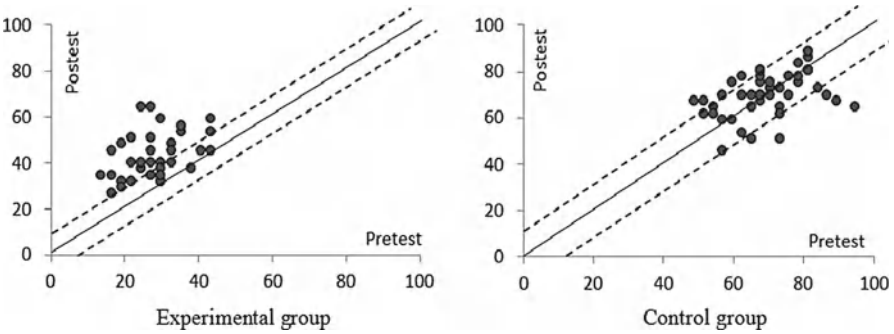

Fig. 13.5 Changes of the achievement of the experimental and control group from pretest to posttest

In case of the control group, the symbols are distributed homogeneously around the mean line; i.e., the majority of these students performed quite similarly in the two data collection phases.

Tables 13.2 and 13.3 show the mean performance of the experimental and control groups, by grade and gender. No differences are realized in the performance of first and second grade students in both the pre- and posttest in case of the experimental group. In the control group, first grade students achieved significantly higher on the pretest than second grade students, while their performance did not differ on the posttest.

No subsamples displayed significant differences in the relative performance of boys and girls in the experimental group, i.e., the training effect is not gender specific. Similarly, no gender-based differences were found in the control group either.

Table 13.2 Means and standard deviations of the inductive reasoning test by grade (%)

		Pretest		Sign.	Posttest		Sign.
Group	Grade	*M*	SD	*t*	*M*	SD	*t*
Exp.	1	26.6	7.1	n.s	40.5	8.1	n.s
Exp.	2	30.0	8.4		45.6	10.9	
Contr.	1	75.9	8.3	$t=5.1$; $p<0.001$	72.2	9.6	n.s
Contr.	2	63.3	8.7		69.5	9.5	

Table 13.3 Means and standard deviations of the inductive reasoning test by gender (%)

		Pretest		Sign.	Posttest		Sign.
Group	Gender	*M*	SD	*t*	*M*	SD	*t*
Exp.	Male	26.1	7.3	n.s	41.6	10.9	n.s
Exp.	Female	30.5	8.0		44.7	9.0	
Contr.	Male	66.2	10.5	n.s	68.2	12.8	n.s
Contr.	Female	71.3	10.3		72.1	7.4	

Table 13.4 Student and parental level of motivation and attitude towards game-based fostering

	Experimental group		Control group	
Variable	*M*	SD	*M*	SD
Students' attitude towards the training	4.93	0.27	–	–
Students' attitude towards computer-based games	–	–	4.44	0.94
Parents' attitude towards game-based training	4.43	0.55	4.13	0.59

The effect size of the training program was $d=1.66$ ($p<0.01$). Using Cohen's (1988) convention for describing the magnitude effect size, it is clearly a large effect.

Table 13.4 presents the mean results regarding motivation and attitude towards game-based fostering. Students' attitude towards the game-based training were absolutely positive, most of them chose the highest response category (5: I liked it very much.) on the Likert scale. This is supported by the low value of the standard deviation (0.27).

Parent's attitude is absolutely positive towards game-based fostering (see Table 13.5). There is no significant difference between the parental opinions of the experimental and control group students. However, 85% of the parents support both kinds of trainings, about 10% of the parents only prefer technology-based training to other kinds of fostering, and about 3.5% of the parents completely reject technology-based fostering. There are no parents in the sample who consider the need for any kind of training unnecessary.

Table 13.5 Parents' opinion about training with or without technology

Question	Parents' of the experimental group (%)	Parents' of the control group (%)
Training using different technology tools	12.5	10.4
Training without using different technology tools	2.5	4.2
Training with or without different technology tools	85.0	85.4
No training is needed	0.0	0.0

13.4 General Conclusions and Directions for Further Research

In this chapter, we placed innovative assessments applied in educational games into a broader theoretical framework. We consider feedback as the most important function of assessment; therefore, we showed how it works at the different levels of the education system. We described recent tendencies of TBA and highlighted the advantages it offers for improving feedback. We pointed out that feedback is most needed in everyday teaching processes where supporting students' learning requires reliable detailed and frequent feedback. This can best be done by diagnostic tests tailored to the actual developmental level and individual characteristics of students. We showed that this cannot be done by means of traditional tests; therefore, this is the context where TBA is most beneficial.

We analyzed the similarities and differences between online diagnostic assessment and educational games. As for functional similarities, the main mission of both diagnostic testing and assessment in educational games is guiding the learners through a series of learning tasks, so that they could always deal with learning tasks which are matched to their actual developmental level. The difference is that diagnostic assessment focuses on feedback and orients students' learning process towards the next learning tasks, while educational games support learning in a more direct way by presenting learning material and developmental stimuli. Educational games may play a complementary role to diagnostic assessment and can be used for compensating deficiencies identified by assessment. On the other hand, feedback cycles within educational games are even smaller; therefore, they should be more frequent than that of the diagnostic assessment. As for the technological aspects, similar or identical methods can be applied both in diagnostic assessment and in educational games. Therefore, innovations may be utilized in both areas in similar ways.

The empirical work presented in this chapter has been carried out in the framework of a project aiming at developing an ODAS. The experiment was designed so that it modeled the integration of educational games into (a renewed, assessment based) teaching process. One of the novel aspects of the study was that gaming took place in a school environment where students were together in the classroom space and at the same time played individually without disturbing each other. This shows how the types of teaching games we are experimenting with can later be embedded

into the regular teaching processes. In this experiment, student played with the same game, but in later practical implementations, students may play different games, according to their individual needs (identified by the diagnostic system).

A variety of assessments were carried out in the course of the experiment. Eseryel, Ifenthaler, and Ge (2011) distinguish internal and external assessments in the contexts of game-based learning. Internal assessment is part of the game, "Optimally assessment is part of the action or tasks within the game-based environment" (Eseryel et al., 2011, p. 166.). In our experiment, the real-time automated scoring of students' solutions played this role; it allowed an instant feedback and guided students through a series of games. This is not an original solution but can be applied in novel way in the future, when not only the difficulty level or complexity is taken into account, but, depending on the solutions students find or mistakes they make qualitatively, different types of tasks may be presented in the following steps. This development requires further research, especially on a better mapping of the construct defined in the framework into the gaming activity.

Two types of external assessment were applied. According to Eseryel et al. (2011), external assessment may take place before, during, or after playing the game, and as it is not part of the game, especially if applied during paying the game, it may disturb the player. We applied a pre- and posttest design, as proving the efficiency of the game was also part of the experiment. These assessments took place independently from the gaming sessions; therefore, they have not influenced students when playing the game. On the long run, only games with the proven efficiency will be introduced into the practice, so these assessments may be eliminated. On the other hand, when the games will be integrated into the diagnostic system, the information generated by the system may be utilized to monitor the effects of the games. While the way we have evaluated the effects of the game was not new, utilization of the diagnostic information for this purpose may result in novel solutions.

The most innovative aspect of the assessment we are dealing with is capturing contextual information while students play the game. This is also an external assessment, but it is carried out seamlessly, therefore, if it does not disturb the gaming process. The contextual information we focused on in this paper is related to one of the most rapidly developing areas of ICT, face recognition, and identification of emotional expressions. We have demonstrated that the automatic analysis of video data is accurate enough to provide significant feedback in the given contexts. It is also important to note that the precision of our face tracker—which was trained on databases of adults—is considerably worse on children than on adults. In turn, we need to retrain our system on facial expression databases of children. These works—which are in progress at the moment—further improve the precision of the system.

In the present phase of the research, the components examined here have not been completely integrated yet. However, the first results demonstrated the possibilities in each area. In the next phase of the project, more educational games will be developed: they will be connected to the results of the diagnostic assessments, and during their application, more contextual data will be collected. As for the identification of emotional expressions, the automated real-time evaluation of the

affective states and the results of cognitive processes may be connected. These assessments can be used themselves as feedback in the learning processes and can be utilized to improve educational games as well.

Acknowledgments The research reported in this chapter was funded by the European Union and co-funded by the European Social Fund and the European Regional Development Fund. Project ID numbers: TÁMOP 3.1.9.-08/1-2009-0001, TAMOP 4.2.1./B-09/KMR-2010-0003, KMOP-1.1.2-08/1-2008-0002. We are grateful to Brigitta Miksztai-Réthey for her thoughtful assistance during the experiments and in the ongoing evaluations.

References

Ainsworth, L., & Viegut, D. (2006). *Common formative assessments. How to connect standards-based instruction and assessment*. Thousand Oaks, CA: Corwin.

Baluch, F., & Itti, L. (2010). Training top-down attention improves performance on a triple conjunction search task. *PLoS One, 5*, e9127.

Biland, C., Py, J., Allione, J., Demarchi, S., & Abric, J.-C. (2008). The effect of lying on intentional versus unintentional facial expressions. *European Review of Applied Psychology, 58*(2), 65–73.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning. Putting it into practice*. Berkshire: Open University Press.

Candès, E. J., & Recht, B. (2008). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics, 9*, 717–772.

Candés, E. J., & Tao, T. (2009). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory, 56*(5), 2053–2080.

Chen, Y., Xu, H., Caramanis, C., & Sanghavi, S. (2011). Robust matrix completion with corrupted columns. arXiv http://arxiv.org/abs/1102.2254.

Chermahinia, S. A., & Hommel, B. (2010). The (b)link between creativity and dopamine: Spontaneous eye blink rates predict and dissociate divergent and convergent thinking. *Cognition, 115*(3), 458–465.

Clarke, S. (2001). *Unlocking formative assessment. Practical strategies for enhancing pupils learning in primary classroom*. London: Hodder Arnold.

Clarke, S. (2005). *Formative assessment in action. Weaving the elements together*. London: Hodder Murray.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cristinacce, D., & Cootes, T. (2008). Automatic feature localisation with constrained local models. *Journal of Pattern Recognition, 41*(10), 3054–3067.

Csapó, B., Ainley, J., Bennett, R., Latour, T., & Law, N. (2012). Technological issues of computer-based assessment of 21st century skills. In B. McGaw & P. Griffin (Eds.), *Assessment and teaching of 21st century skills* (pp. 143–230). New York: Springer.

Csapó, B., Molnár, Gy, & R. Tóth, K. (2009). Comparing paper-and-pencil and online assessment of reasoning skills. A pilot study for introducing electronic testing in large-scale assessment in Hungary. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing* (pp. 113–118). Luxemburg: Office for Official Publications of the European Communities.

Ekman, P. (2006). Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences, 1000*, 205–221.

Eseryel, D., Ifenthaler, D., & Ge, X. (2011). Alternative assessment strategies for complex problem solving in game-based learning environments. In D. Ifenthaler, K. P. Isaias, D. G. Sampson, & J. M. Spector (Eds.), *Multiple perspectives on problem solving and learning in the digital age* (pp. 159–178). New York: Springer.

Gantner, Z., Rendle, S., & Schmidt-Thieme, L. (2010). *Factorization models for context-/time-aware movie recommendations, in challenge on context-aware movie recommendation (CAMRa2010)*. Barcelona: ACM.

Greiff, S., & Funke, J. (2009). Measuring complex problem solving: The MicroDYN approach. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing* (pp. 157–163). Luxemburg: Office for Official Publications of the European Communities.

Greiff, S., & Funke, J. (2010). Systematische Erforschung komplexer Problemlösefähigkeit anhand minimal komplexer Systeme. In E. Klieme, D. Leutner, & M. Kenk (Eds.), *Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes* (Beiheft der Zeitschrift für Pädagogik, Vol. 56, pp. 216–227). Weinheim: Beltz.

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multi-PIE. Image and Vision. *Computing, 28*(5), 807–813.

Hévízi, G., Gerőfi, B., Szendrő, B., & Lörincz, A. (2005). Assisting robotic personal agent and cooperating alternative input devices for severely disabled children. *Lecture Notes in Artificial Intelligence, 3690*, 591–594.

Ishikawa, T., Baker, S., Matthews, I., & Kanade, T. (2004). Passive driver gaze tracking with active appearance models. In *Proceedings of the 11th world congress on intelligent transportation systems*. Nagoya: Japan.

Itti, L. (2007). Visual salience. *Scholarpedia, 2*(9):3327.

Jenatton, R., Obozinski, G., & Bach, F. (2010). Structured sparse principal component analysis. *JMLR Workshop and Conference Proceedings, 9*, 366–373.

Jeni, L., Hashimoto, H., & Lörincz, A. (2011). *Efficient, pose invariant facial emotion classification using 3D constrained local model and 2D shape information*. In Workshop on Gesture Recognition, Colorado Springs, USA. Retrieved June 20, 2011, from http://clopinet.com/isabelle/Projects/CVPR2011/posters/jeni_hashimoto_lorincz.pdf.

Klauer, K. J. (1989). *Denktraining für Kinder I*. Göttingen: Hogrefe.

Klauer, K. J., & Phye, G. D. (2008). Inductive reasoning. A training approach. *Review of Educational Research, 78*, 85–123.

Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment for education. Theory and applications*. Cambridge: Cambridge University Press.

Matthews, I., & Baker, S. (2004). Active appearance models revisited. *International Journal of Computer Vision, 60*(2), 135–164.

Meyer, B., & Sorensen, B. H. (2009). Designing serious games for computer assisted language learning—a framework for development and analysis. In M. Kankaaranta & P. Neittaanmaki (Eds.), *Design and use of serious games* (pp. 69–82). New York: Springer.

Molnar, G. (2011). Playful fostering of 6- to 8-year-old students' inductive reasoning. *Thinking Skills and Creativity, 6*(2), 91–99.

OECD. (2004). *Problem solving for tomorrow's world. First measures of cross-curricular competencies from PISA 2003*. Paris: OECD.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research, 45*(8), 2397–2416.

Porter, P., & ten Brinke, L. (2011). The truth about lies: What works in detecting high-stakes deception? *Legal and Criminological Psychology, 15*(1), 57–75.

Renninger, L. W., Verghese, P., & Coughlan, J. (2007). Where to look next? Eye movements reduce local uncertainty. *Journal of Vision, 7*(3), 1–17, http://journalofvision.org/7/3/6/, doi:10.1167/7.3.6.

Saragih, J. M., Lucey, S., & Cohn, J. F. (2011). Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision, 91*(2), 200–215.

Seel, N. M. (2010). Essentials of computer based diagnostics of learning and cognition. In D. Infenthaler, P. Pirnay-Dummer, & N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (pp. 3–14). New York: Springer.

Spector, J. M. (2010). Mental representations and their analysis: An epistemological perspective. In D. Infenthaler, P. Pirnay-Dummer, & N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (pp. 27–40). New York: Springer.

Strasser, A. (2010). A functional view toward mental representations. In D. Infenthaler, P. Pirnay-Dummer, & N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (pp. 15–25). New York: Springer.

Szabó, Z., Póczos, B., & Lörincz, A. (2012). *Collaborative filtering via group-structured dictionary learning*. *Lecture Notes in Computer Science, 7191*, 247–254.

Taras, M. (2005). Assessment—summative and formative—some theoretical reflections. *British Journal of Educational Studies, 53*(4), 466–478.

Thai-Nghe, N., Drumond, L., Horváth, T., Nanopoulos, A., & Schmidt-Thieme, L. (2011). Matrix and tensor factorization for predicting student performance. In *CSEDU 2011—Proceedings of the third international conference on computer supported education* (Vol. I, pp. 69–78). Noordwijkerhout: The Netherlands.

Tudge, J. (1992). Vygotsky and education: Instructional implications and applications of sociohistorical psychology. In L. C. Moll (Ed.), *Vygotsky and education: Instructional implications and applications of sociohistorical psychology* (pp. 155–172). New York, NY: Cambridge University Press.

Tzuriel, D. (1998). Dynamic assessment of preschool children: Characteristics and measures. In J. M. Martínez, J. Lebeer, & R. Garbo (Eds.), *Is intelligence modifiable?* (pp. 95–114). Madrid: Bruño.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01), 1*, 511. doi:ieeecomputersociety.org/10.1109/CVPR.2001.990517.

Westra, J., Dignum, F., & Dignum, V. (2011). Guiding user adaptations in serious games. In F. Didnum (Ed.), *Agents for games and simulations II. Trends in techniques, concepts and design* (pp. 117–131). Berlin: Springer.