Transfer of Training of An Educational Serious Game: The Effectiveness of the CASHIER TRAINER

E. A. P. B (Esther) Oprins and J. E (Hans) Korteling

Abstract We investigated the transfer of training of a stand-alone educational serious game supported with automated feedback and instruction compared to conventional on-the-job training (OJT). If transfer of training is sufficiently high, these types of games could reduce time-consuming and expensive OJT saving instructional personnel. In a case study with such a game, the CASHIER TRAINER supported with an ITS, we compared performance and competence of new cashier employees at the workplace in two matched groups: (a) experimental group: acquiring cashier skills in the game CASHIER TRAINER; (b) control group: acquiring cashier skills in conventional OJT. Performance and competence were measured by observation and by self-assessment. The results showed that performance and competence were at least as high and even higher on certain aspects for the CASHIER TRAINER than for the OJT condition. Thus, transfer of training with the CASHIER TRAINER was positive in comparison with OJT. Specifically, non-regular tasks which could be trained with gaming more often and structured than in OJT were performed better. It seems that automated feedback in stand-alone educational serious games for procedural tasks, if well-designed, could replace human tutoring to a certain extent.

Keywords Evaluation · Educational games · Instructional games · Serious games · Intelligent tutoring · Training effectiveness · Instructional effectiveness · Training simulation · Transfer of trainingTransfer of training

E. A. P. B (Esther) Oprins (✉) · J. E (Hans) Korteling
TNO, Delft, The Netherlands
e-mail: esther.oprins@tno.nl

J. E (Hans) Korteling
e-mail: hans.korteling@tno.nl

Y. Cai and S. L. Goei (eds.), *Simulations, Serious Games and Their Applications*,
Gaming Media and Social Effects, DOI: 10.1007/978-981-4560-32-0_15,

1 Introduction

The education and training community is increasingly accepting serious games as a potentially valuable, efficient, and effective alternative for conventional forms of education, training, or other applications. Many different types of games exist and the literature does not yet show complete consensus on its definition (e.g., Crookall 2010). Most popular definitions agree, however, that serious games are reality based (i.e., simulations), entertaining, interactive, rule-governed, goal-focused, and competitive (Bell et al. 2008; Tobias and Fletcher 2007; Vogel et al. 2006). They serve other goals than entertainment such as personnel recruitment, change of attitudes, research, selection, concept development, education, and training (e.g., Harteveld 2011, 2012). Within the broader category of serious games, this chapter focuses on the effectiveness of *educational serious games*, which are here referred to as interactive representations of (aspects) of reality intended for learning purposes that deliberately include elements of play or entertainment.

The reviews on studies on transfer of training that have been carried out so far consistently show that evidence is fragmented. Well-designed research on learning effects of educational games is scarce while it is necessary for designing games that are more effective for learning (e.g., Akl et al. 2010; Bekebrede et al. 2011; Connolly et al. 2012; Egenfeldt-Nielsen 2006; Fletcher and Tobias 2008; Freitas 2006; Girard et al. 2012; Hays 2005; Ke 2009; Lee 1999; Leemkuil et al. 2000; O'Neil et al. 2005; Randel et al. 1992; Sitzmann 2011; Vogel et al. 2006; Wouters et al. 2009; Young et al. 2012). Evidence on 'what works' is needed, that is, for which target groups, types of tasks and competences are educational games the optimal alternative and what are the conditions and other factors determining their learning effects? This kind of knowledge will help designers to improve their games (Bedwell et al. 2012). Moreover, for schools, companies, and training organizations it is important to get insight into the factors determining educational effectiveness of games that may be applied in their organizations. One of the applications of educational serious games, as discussed in this article, is to replace (parts of) on-the-job training (OJT). If evidence on transfer of training has been proved, this replacement of OJT will be more easily done by organizations.

Educational serious games are especially useful when conventional OJT does not provide suitable opportunities because of risks caused by errors, costs of using the real system, legislation, sustainability, motivation problems of trainees or limited availability of the operational environment (Korteling et al. 2012). Unlike learning at the workplace, the simulated environment offers the opportunity to train and practice tasks when suitable training opportunities are lacking (Farmer et al. 1999; Oprins 2008). This provides the opportunity to be confronted with learning tasks that do not frequently occur in practice and that can be built up in complexity gradually (Van Merrienboer and Kirschner 2007). In OJT an instructor or supervisor usually coaches the learner while performing tasks in the real-life setting (Oprins et al. 2011). It depends on situations that will happen which learning tasks the learners are confronted with. Educational games are supposed to be an effective alternative for OJT also because they can (partly) replace this staff and/or give them a more coaching or supervisory role. This is especially possible if games are used that are supported with a well-designed intelligent tutoring system (ITS; Polson and Richardson 1988), providing automated feedback on the actions of the learners. The learners can acquire basic knowledge and skills by themselves to a certain extent and have dealt with learning tasks that do not occur frequently in practice before they enter OJT. This saves instructional personnel at the workplace who supervise learners.

This chapter focuses on transfer of training of educational serious games with artificial instructional tutoring and feedback functions in comparison to OJT. We present a case study with a stand-alone game called CASHIER TRAINER supported with an ITS for new cashier personnel to train their skills on the cash desk in a simulated environment before they enter the supermarket as a possible replacement for (parts of) OJT.

1.1 Transfer of Training

Games may require substantial investments, not only in the design and development of the product itself, but also in the implementation and in education and training of personnel. Therefore, an important question concerns the degree to which educational games are really useful and effective for practical purposes such as job training. Educational games are likely to prove to be a good alternative for OJT if, and only if, they are able to achieve similar or higher learning outcomes at similar or lower expenses. This raises the question of what and how much, relatively speaking, is actually transferred from acquired competence in the game or in OJT to the workplace. This issue is typically captured under the notion of *transfer of training*, that is, skills acquired in a specific training environment may transfer to situations in a different (real world) task environment. After definitions provided by Baldwin and Ford (1988) and Gielen (1995) we define transfer of training in the context of educational gaming as the degree to which knowledge, skills and, attitudes that are acquired by playing a game can be used effectively in real-life (operational, professional) situations. With that, transfer results in a reduction of the amount of OJT required to obtain the training goals (Caro 1977). This definition also refers to the notion of competence that we define as the application of the total combination of acquired knowledge, skills and attitudes effectively in the job (Oprins 2008). This type of research can also be referred to as educational validity (Stainton et al. 2010). External educational validity, as opposed to intern educational validity, involves transfer of training by the game or simulation to the real world.

Educational serious games are supposed to be able to have a relatively high transfer of training in comparison with OJT because of their didactical advantages. From the point of learning theory, a surplus value of educational games is that competences can be acquired in a realistic, attractive, and challenging manner (Bedwell et al. 2012; Gee 2007; Korteling et al. 2012; Shaffer 2006; Squire 2003). Learners learn actively in authentic (realistic, practical, job-related) and flexible learning environments. Educational games are also assumed to be intrinsically motivating and engaging (Csikszentmihalyi 1990; Malone 1981) and give responsibility to the learner. We suppose that this latter aspect may enhance the learner's self-efficacy (Bandura 1997) and may facilitate the development of self-directed learning (Percival 1996; Stubbé and Theunissen 2008).

These potential didactical and motivational advantages of educational games are in line with modern general learning approaches. In the traditional instructor-centered situation the instructor is dominant. The instructors stand in the center of the attention whereas the learner more or less passively absorbs (lean back) the information that is provided. Contemporary theories of training and instruction, however, conjecture that learners should participate more actively during classroom lessons in a more 'lean-forward' style. Learners should have an active, central role while the instructors should be supportive rather than directive (Johnston and McCormack 1996; Petraglia 1998). Many contemporary constructivist visions of learning and instruction promote these active forms of learning through experience such as discovery learning (Steffe and Gale 1995), action learning (Smith and O'Neill 2003), and experiential learning (Jiusto and DiBiaso 2006). Although constructivism may take many forms (Petraglia 1998; Philips 1998), an underlying premise is that learning is an active process of sense making in which learners seek to build coherent and organized knowledge. Educational gaming may be brought to fit very well to these constructivist notions since it can be used to train relevant competences in a realistic, attractive, and challenging manner, using authentic (realistic, practical, job-related) and flexible learning environments in an active, self-directed way.

1.2 Measurement of Transfer of Training

Empirical studies to get evidence on transfer of training with educational games based on their motivational and didactical features are, however, scarce. Studies on transfer of training are difficult to perform in practice. The degree to which training results in behavioral change on the job is the gold standard of training. It refers to Kirkpatrick's third level of evaluation concerning the enhancement of actual behavior on-the-job based on the training activities (Kirkpatrick 1998). But many studies fall short in measuring training effects beyond Kirkpatrick level 1 (reactions) and 2 (declarative learning; Cohn et al. 2009). Transfer is difficult to measure because it requires research at the workplace that cannot easily be carried out and controlled (Bedwell et al. 2012; Korteling et al. 2013; Salas et al. 2009; Veldhuis and Theunissen 2009).

The previously mentioned reviews of effectiveness studies for educational games that have been done so far (see for a brief qualitative meta-review Harteveld

2012) have shown a large diversity in outcomes and quality of research. The effectiveness of educational games appears to be moderated by the didactical and technical characteristics of the game design as well as the instructional context in which the game is embedded (Sitzmann 2011). More in specific, this refers to, for example, different target groups, tasks, competences, and domains involved as well as types of games. To get insight into these complexities, it is very important to measure process variables related to the learning process itself instead of only measuring outcome variables (Bedwell et al. 2012; Salas et al. 2003, 2009). Process measures examine the manner in which a task is performed by the trainee, whereas outcome measures focus on how well a trainee accomplishes the overall task. Process measures can be a useful diagnostic tools explaining certain outcomes, thus why it happened, illustrating strengths and weaknesses of training program or game that should be either maintained, improved, or further developed to ensure that training goals are met (Cohn et al. 2009; Fowlkes et al. 1999). For practical reasons especially process measures usually have to be based on opinions, questionnaires, ratings, and (self-) evaluations. Direct quantitative measurements in terms of speed, time, or error are mostly more feasible for outcome variables (Korteling et al. 2013), which was also the case in the present experiment.

The reviews also indicate that some studies even do not measure learning effectiveness at all. Some limit themselves to Kirkpatrick's Level 1 (reactions of trainees and experts (e.g., Stehouwer et al. 2005) and others measure other aspects, such as the fidelity of the simulation (Allan et al. 1986). Sometimes they do not measure learning effectiveness appropriately, for instance, they do not use control groups and/or pretests (Sitzmann 2011). Other studies may be hampered by differences between experimental and control groups or in educational contexts of these groups, lack of control groups, test- and selection-effects, short exposure time, weak assessment tests, and preconceived opinions of researchers or teachers (Egenfeldt-Nielsen 2006; Korteling et al. 2013). For these reasons evidence-based studies in which transfer of training has been properly verified are rare (Peck 2012). On the basis of a qualitative meta-review of about 20 review studies on mainly digital serious educational games, Harteveld (2012) concludes that we need to speak of the rise of a potential powerful tool. Gaming has potential, theoretically and based on some of the hints from literature, but we need to figure out how to utilize and proof that potential.

Since transfer of training with games thus has been occasionally proven, getting empirical data on transfer of training, and on the factors that determine this transfer as well, is important to prove the added value of the educational games and to get insight into the most effective didactical, technical, and organizational gaming design properties of specific types of games (Korteling et al. 2013). Generalization of gaming features is restricted; some design features will be more important for particular types of games than others. In this context, the notion *direct transfer* is relevant (Harteveld and Bekebrede 2011). Direct transfer is the more traditional approach with concrete, predefined, and measurable learning objectives to be obtained at the end of the game. In contrast, the open ended learning approach has more broadly defined, abstract learning objectives and general insights to be achieved. Single-player games are more associated with direct transfer than multiplayer games.

1.3 Instructional Support

One major conclusion of a review of 48 empirical research articles on the effectiveness of educational games (Hays 2005) is that these games seem to be more effective if they are embedded in adequate instructional programs that include debriefing and feedback. In addition, instructional support during play increases the effectiveness of educational games. Teacher guidance and intervention when using games may be carried out by a real teacher or coach facilitating the learning process, in terms of steering the trainee in the right direction and also in providing an effective feedback and debriefing. On the basis of a meta-analytic examination of 65 studies which used a comparison group Sitzmann (2011) draws a similar conclusion concerning the role of instructional support. She deduces that trainees learn more, relative to comparison groups, when so-called simulation games are a supplement to other instructional methods rather than stand-alone instruction. The importance of debriefing while using games is also stressed by others (Alklind Taylor et al. 2012; Crookall 2010; Egenfeldt-Nielsen 2006). A debriefing by human coaches is mostly considered as crucial for learning.

However, when the task to be trained is not very complicated, we suppose that an artificial tutoring system or ITS (Polson and Richardson 1988) may (partly) replace the instructor or coach appropriately. Such automated feedback system may be capable of providing instructions, choosing the right training scenario's, guiding the trainee in the right direction, providing extra information when stuck, and providing feedback and debriefing on task performance. The learning tasks could be built up in complexity in a structured way which helps learning, and also situations that scarcely occur in real-life could be practiced (Van Merrienboer and Kirschner 2007). According to Alklind Taylor (2012), Crookall (2010) and Egenfeldt-Nielsen (2006), human feedback can be partly replaced by automation, but facilitation by training staff which should include a form of debriefing will be indispensable for reflection on what the learners have learned. In line with others (e.g., Farmer et al. 1999; Van Merrienboer and Kirschner 2007) we suppose that the amount of automation in relation to human feedback depends on the type of learning tasks. For instance, in educational games that focus on automating procedures and routines, automated feedback is functional and relatively easy to implement. Automated feedback may be supposed to be less suitable in games intended for learning in complex cognitive tasks (Van Merrienboer and Kirschner 2007). At present debriefing by human coaches is the most effective form of feedback here. Nevertheless, some instruction of human coaches in the form of a debriefing stays important in all types of tasks.

These general insights from the literature are important for all professionals working with educational games or trying to enhance the efficiency of training by saving on instructional personnel. Referring to the previously mentioned conclusions and assumptions, an important issue concerns the beneficial effects of stand-alone educational games provided with intelligent tutoring functions (e.g., Polson and Richardson 1988), which are needed to compensate for the lack of human instructional support and feedback (Egenfeldt-Nielsen 2006). Studying this issue will provide more insight into the potential of stand-alone instructional games and (required) supporting components.

1.4 Goal of The Study

This empirical study aims at getting evidence on transfer of training of particular types of educational serious games. We expect that at least for routine task and procedures to be trained tutoring and feedback can be automated in educational serious games if well-designed. Such a system may sufficiently guide and help students and motivate them to get through the training program. In that case, games can replace (parts of) conventional OJT and this is cost-effective for saving instructional staff at the workplace. Also, the preparation on OJT is efficient since learning tasks could be offered in a well-sequenced way including practicing situations that scarcely occur on the workplace. Following this assumption, we investigated the transfer of training of a stand-alone game for job training with an ITS providing instructions, support cues, and feedback on task performance. We hypothesize that the transfer of training of this type of game could be at least as high as conventional OJT. The game involved in this study is the so-called CASHIER TRAINER for new cashier employees in supermarkets. This game, being representative for educational serious games with an ITS, functions as a case in this study.

2 Cashier Trainer

The CASHIER TRAINER was developed by a leading Dutch training enterprise in the retail sector, Jutten Simulation. This CASHIER TRAINER is being used by various retail organizations in The Netherlands. The game is representative of instructional games with a simulated environment that must be done individually, supported by a virtual tutor. The CASHIER TRAINER is a low-cost virtual training program combining a simulation of a complete task and a simplified 3D representation of the job environment in an attractive and accessible way, see Fig. 1. The game is internet-based and runs on each home computer that is connected to the internet.

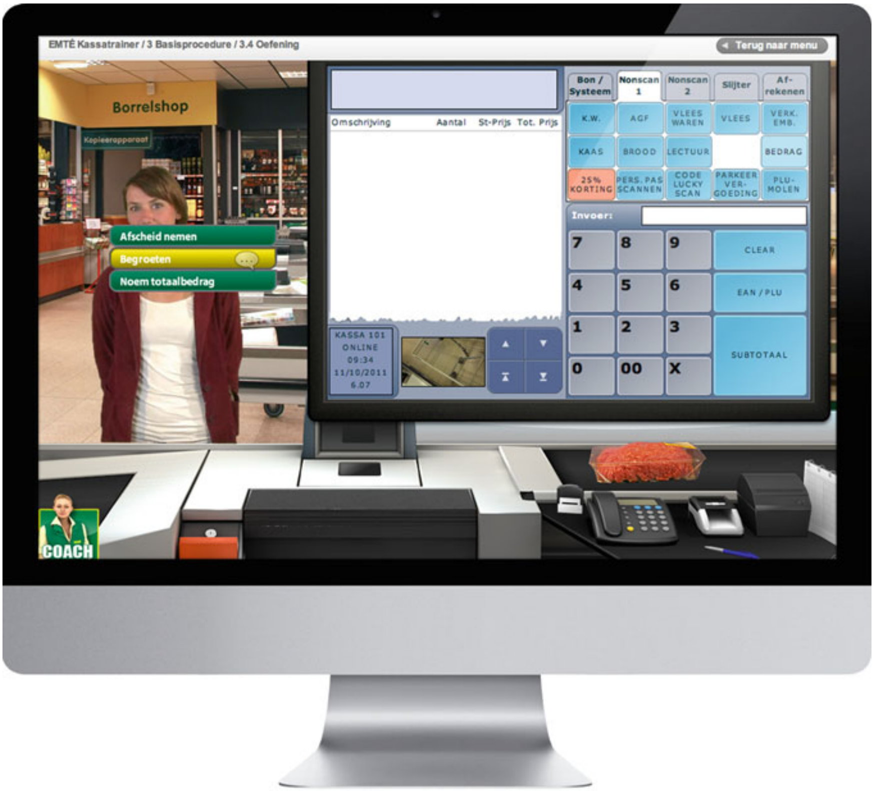

Fig. 1 The CASHIER TRAINER (Jutten simulation)

The CASHIER TRAINER is a single-player game that is done independently at home without involvement of a supervisor, coach, or instructor. The primary goal of the learning program is being maximally prepared by practicing the most important learning tasks before the first working day. In addition, the new employees can practice in a safe environment also for tasks they cannot practice very often because they hardly occur in normal situations on the job. The CASHIER TRAINER program is followed by a brief OJT in the supermarket. This provides a check of the employees' skills before starting the real work and gives them the opportunity to discuss subjects learned in the CASHIER TRAINER as a form of debriefing. They get a short introduction followed by coaching by a supervisor while working at the cash desk.

The CASHIER TRAINER is a typical instructional simulation game with an intelligent tutoring system (Polson and Richardson 1988). The CASHIER TRAINER does not only represent the cash desk, as many other related training systems, but it simulates the whole cashier environment including virtual customers. In this way, the game is intended to simulate the entire cashier task and to immerse the learners in the virtual environment of the supermarket. In the game, new cashier employees follow a strict training program with a well-sequenced order of learning tasks. The CASHIER TRAINER has a sophisticated ITS supporting the learners with explanations, instructions, and feedback on erroneous or incorrect behavior. This feedback is adaptive to the level and type of actions. For instance, the first time that learners execute an incorrect action such as using particular functions of the cash desk, the feedback only consists of 'you are wrong'. After a second mistake, the feedback becomes more concrete and helpful, for instance, 'you uses the wrong button'. A third error follows by very specific feedback, for instance, 'you should use button x'. The errors and feedback are logged and related to a scoring system which is built in as a special game element for motivating and engaging the learners.

The CASHIER TRAINER was developed to train the following learning objectives: registration, operating the cash desk system, scanning of normal and special products, payment procedures (e.g., cash, credit card), safety and control procedures, and communication with clients. These learning objectives are typically procedural so that we assume that feedback could be automated and that the CASHIER TRAINER could replace parts of OJT. The learning tasks are sequenced from part-task to whole task training (Van Merrienboer and Kirschner 2007) and also from only regular tasks to non-regular special tasks. This implies that first only very standard tasks are trained separately, for instance, learners start with scanning regular articles without special assignments such as articles with discount. This is typically part-task training. Next, they combine this with a regular payment method (e.g., cash). Learning tasks become more and more complex, being combined with less regular tasks such as scanning special products (e.g., fruits not weighted) and difficult payment methods (e.g., combination of cash and card). Finally, trainees get mixed training scenarios with all types of regular and non-regular tasks. This can be referred to as whole task training. The level of the trainee determines if he/she has to practice extra on certain tasks or not. Thus, not only the feedback is adaptive but the learning tasks in the game as well.

The quality of this game with regard to simulation fidelity, didactics, game play, and technical aspects were evaluated by gaming and educational experts with the TNO Checklist for Evaluation of Edicational Simulations (TNO-CEES 3.0). This is a new and adapted version of the so-called CONCERT Checklist (Emmerik and Korteling 2002, 2003) that has been developed previously by TNO for the structured evaluation and validation of training simulations. The checklist consists of about 180 items that have been defined on the basis of an extensive literature study on training simulation and Educational gaming (e.g., Korteling et al. 2001). Items include all design, didactical, game play, and fidelity features of the game. Examples of clusters of items are: comprehensiveness of the specification and design process, training program, scenario management, instruction and feedback, intelligent tutoring, game play and mechanics, user interface, models, visual image and content, sound, technical reliance.

The results of this evaluation showed that the CASHIER TRAINER comprised excellent didactical features and human-computer interface features with sufficient physical fidelity and minor technical problems. Major points of possible improvement concerned typical game play features that may be supposed to motivate trainees. The game did not explicitly stimulate or motivate self-directed learning by features that are supposed to enhance entertainment value such as competition, storyline, backstory, or fore shadowing. For the present application of the game for training new cashier employees in retail supermarkets, this extra motivation might not be as important as in other cases, since passing the CASHIER TRAINER program was mandatory for being allowed to start working at the supermarket. In addition, because the cashier task is very procedural as argued, self-directed learning does not seem to be as appropriate here as step-by-step learning with clear corrective feedback provided by the ITS.

In sum, the results of this evaluation were sufficiently convincing that the CASHIER TRAINER could be an appropriate learning device with potentials to replace parts of OJT. In the majority of supermarkets, new cashier employees are only trained by OJT with a supervisor sitting behind the new employee while giving feedback, hints, and instructions. Learning in OJT occurs less structured due to the learning tasks that could not be sequenced and planned as in the simulated environment of the CASHIER TRAINER. This could have consequences for the efficiency of learning in OJT if learners have to wait for non-regular tasks that hardly occur in reality. OJT time could be reduced when the new employees have acquired some basic skills and have experienced non-regular situations already before they enter OJT.

3 Method

3.1 Experimental Design

In order to investigate transfer of training of the CASHIER TRAINER in comparison with conventional OJT, we set up an experiment with two conditions. A total of 45 subjects (7 male, 38 female), who were trained as new cashier employees for a large supermarket chain in The Netherlands, were recruited to participate in the experiment. Subjects were randomly assigned to the control group with conventional OJT ($N = 23$) and the experimental group trained with the CASHIER TRAINER ($N = 22$). Average age was 20 years. An independent t test ($p < 0.05$) did not show significant differences in average age between the two groups.

The OJT group was trained according to the usual training procedures in the retail company. After a short introduction of the general procedures of the cashier task they were trained on the real cash desk under supervision of an experienced cashier employee in the supermarket. The amount of feedback and instruction by the supervisor depends on the progression of the new employees, on the tasks they have to do, for instance, regular or non-regular groceries, and on the instructional skills of the supervisor. Each subject had another supervisor. The experimental group was trained with only the stand-alone CASHIER TRAINER. This group gets access to the CASHIER TRAINER from home circa 3 weeks before their first working day in the supermarket.

Only for the experiment, we tried to fix the training time for both groups at 3 h in total in order to make the control and experimental groups comparable with each other for research purposes. Then it could be compared how much was learned in the CASHIER TRAINER versus OJT within the same time brackets. However, since the study was executed in the real-life practice of supermarket policy, this training time expressed in minutes turned out to be substantially different for all subjects in the OJT group ($M = 213$; $SD = 129$; $SE = 27$) and in the CASHIER TRAINER group ($M = 188$; $SD = 62$, $SE = 13$). We tried to plan the experiment at a specific time and date to fix OJT duration but this was not always possible for the logistics in the supermarket. In addition, accuracy and policy differs per supermarket; we used many supermarkets in the experiment to get a sufficiently high number of candidates. For the experimental group, the subjects were free in spending time in the CASHIER TRAINER in accordance with the purpose of the game and supermarket policy to put responsibility on the new employees themselves. This explains the high variance between subjects in this group. Unfortunately, we did not have the opportunity to put the experiment separately from the practical operations in the supermarkets.

Another methodological restriction of the experiment was that we were not able to do a pretest but only a post-test. Nevertheless, we could argue that the subjects in both groups were comparable since they did not have any experience as cashier employees at all. They all started at a zero competence level. We controlled for other variables such as education, sex, and age. The post-test was administrated by exactly the same performance test for both groups. This test consisted of handling a customer with a shopping cart of preselected groceries that must be paid. The customer was played by a role-player. The candidates must use the cash desk and apply the correct procedures to handle this customer in the real environment of the supermarket at which the employees were working. In this way an experimentally controlled setting on location was created, measuring transfer from the game to the workplace. In order to be able to make a reliable comparison, the task itself was standardized for all subjects in both control and experimental group. This implies that the same groceries have been put into the shopping cart and the same payments must be done by all subjects. The choice of these tasks was based on a task analysis that was carried out in cooperation with the retail company. The assignment for the candidates included registration on the cash desk, regular articles to be scanned, special articles such as price reduction, showing identity card (ID) for alcohol, vouchers, different payment methods, safety and control procedures.

3.2 Measures

The task analysis resulted in a set of training outcomes that were measured in various ways in order to make the measurements more reliable. Many methods for performance measurement exist (e.g. Oprins 2008; Salas et al. 2009). Typically, reliable and valid measurement of transfer of training to performance at the workplace is not trivial (Egenfeldt-Nielsen 2006; Korteling et al. 2013; Salas et al. 2003). In the present case, the learning outcomes could be measured in a relatively easy and reliable way because the cashier task is a highly procedural task. That is, a task consisting of ordered and discrete action sequences are predefined and well-circumscribed for each situation or event that is encountered. These are executed in a rigid or stereotyped way and lead to a clear predetermined experiment. In this experiment we could measure, rate, and evaluate the handling of shopping carts with payments in a supermarket.

We applied a combination of self-assessment and observations. We made a distinction between *performance* during the test: technical errors directly related to the task, and *competence* in a more generic way: the ability of new cashier employees to do their job. They were measured in multiple ways to enhance reliability. First, two well-trained observers noted the type of errors during the test and counted them, called *observed performance*. This checklist contained 15 types of tasks (items) categorized in five predefined rubrics or scales in accordance with the aforementioned task analysis: registration (2 items, e.g., ‘sign up at cash desk’), normal articles (3 items, e.g., ‘scanning three cup-a-soup with different taste’), special articles (4 items, e.g., ‘bananas without registration code’), payments (4 items, e.g., ‘pin card payment’), and control procedures (2 items, e.g., ‘asking for ID with alcohol product’). The observers marked each item on the checklist with 0 (wrong) or 1 (correct). The observers also assessed eight general competences at a five points rating scale, called *observed competence*. The ratings ranged from 1 (*poor*) to 2 (*insufficient*), 3 (*moderate*), 4 (*sufficient*), and 5 (*excellent*). The five anchors were further defined in text to be as precise as possible, for instance, anchor 1 (*poor*) was further defined as ‘many mistakes, not independent, insecure, slowly’. These observers were educated into the requirements, boundary conditions, and performance criteria of the cashier task. The involvement of two observes makes it possible to calculate inter-rater reliability (Oprins et al. 2006). Finally, the observers measured the time of a particular subtask, that is, scanning 15 regular articles, in order to get insight into the speed of working. This subtask was chosen to make this measurement maximally objective.

Second, the subjects did a self-assessment on the performance variables that are directly related to possible errors during the test, called *self-assessed performance*. Because they are not able to use a checklist while they are doing the test like the observers, they rated themselves on comparable tasks as the observers but they were operationalized in a different way. There were five rubrics: cashier desk (3 items; e.g., ‘use of the functionalities of the cash desk’), corrections (3 items; e.g., ‘dealing with problems with the cash desk’), scanning (5 items; e.g., ‘scanning articles with bar code'), payments (6 items; e.g., 'counting the money when giving back to the customer'), and control procedures (3 items; e.g., 'checking the bags of customers on stolen groceries'). They were rated on a rating scale, the same five points rating scale with similar anchors as mentioned above. In addition, they rated the same set of 8 competences as rated by the observers, also on the same anchored five points rating scale, called *self-assessed competence*.

In sum, four outcome measures were used as dependent variables in the experiment: observed performance, observed competence, self-assessed performance, and self-assessed competence. We compared the differences between the experimental condition (CASHIER TRAINER) and the control condition (OJT) for these four outcome measures with a Mann–Whitney *U* test (Mann and Whitney 1947). We used this nonparametric test because the majority of the data were not normally distributed as being checked with the Shapiro–Wilk test. We also correlated the four outcome measures with each other to investigate how they relate to each other (Spearman's rho, two-tailed). For all composite variables the medians of the underlying aforementioned scales have been calculated.

The subjects in the CASHIER TRAINER condition also filled out a questionnaire with items rated on the same five points rating scale concerning the CASHIER TRAINER. This questionnaire includes four rubrics or scales: training technology (3 items, e.g., 'I find the CASHIER TRAINER an appropriate learning device for new cashier employees'), the quality of practice (3 items, e.g., 'I find the amount of exercises sufficiently to do the test'), didactics (5 items, e.g., 'I like the corrective feedback if I do something wrong'), and reality (4 items, e.g., 'I feel immersed in the environment of the supermarket'). Additionally, the subjects could give remarks on positive and negative aspects of the game. We investigated possible influences on the outcome variables by calculating correlations (Spearman's rho, two-tailed).

Finally, structured interviews with retail managers, who had worked with both groups of candidates, were carried out. In these interviews we inventoried subjective observations about task performance and types of errors. We also asked for expected savings in training time with the CASHIER TRAINER, both for new cashier employees and for the supervisors needed in OJT although this was not the main question in this study. These data were qualitatively analyzed and only used to explain the quantitative findings in this study.

4 Results

4.1 Observer Assessments

The observers assessed both task performances during the test and general competence. They also measured the duration of the subtask: scanning 15 regular articles. The average speed hardly varied for the OJT group (34.7 s) and CASHIER TRAINER group (35.4 s). Figure 2 presents the median scores of the two groups for the number and type of errors (averaged per subtask; usually only one

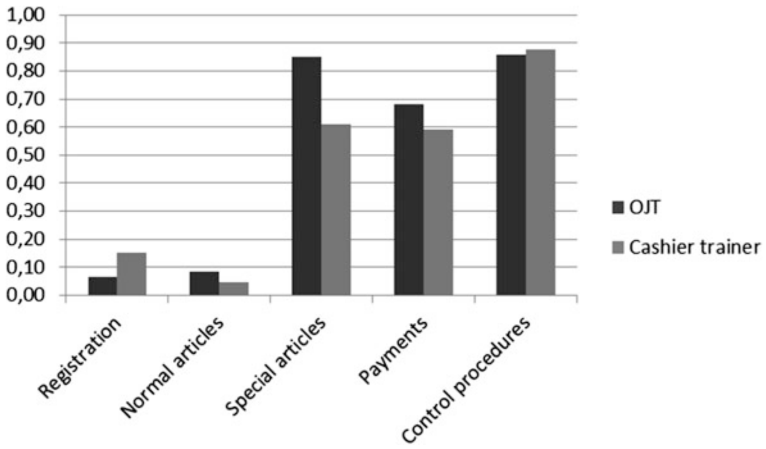

Fig. 2 Median number of errors in observed performance

error could be made per task) made during the performance test, the so-called *observed performance*.

The scores in Fig. 2 are the combined error-scores of the two observers. The reliability was assumed to be sufficiently high: the correlation (Spearman's rho, two-tailed) between the total score of all performance items was .86 significant at $p < 0.001$. The Mann–Whitney U test ($p < 0.05$) did not show significant differences between the two groups. In Fig. 2, we see that the number of errors was lower for regular tasks (registration, normal articles) in comparison with tasks that occur less frequently and are more difficult (special articles, control procedures, payments).

Figure 3 presents the observed median scores on the five-point scale for the two groups on general competence, called *observed competence*.

These are also the combined scores of the two observers. The reliability was assumed to be sufficiently high here as well: the correlation (Spearman's rho, two-tailed) between the total competence scores was 0.69 significant at $p < 0.001$. Figure 3 shows that the scores of candidates in the CASHIER TRAINER condition were slightly superior in 7 out of 8 competences while the scores of the candidates in the OJT conditions were superior only in accuracy. The Mann–Whitney U test ($p < 0.05$) showed that the subtask operating cash desk scored significantly higher for the CASHIER TRAINER than for OJT ($U = 169{,}5$, $p = 0.045$, $r = 0.30$). The rest of them were not significant.

4.2 Self-Assessments

Figure 4 presents the self-assessments of task performance (median scores) of the two groups, OJT and CASHIER TRAINER, called *self-assessed performance*. The variables are not similar to the variables for the observers (see Fig. 2) due to

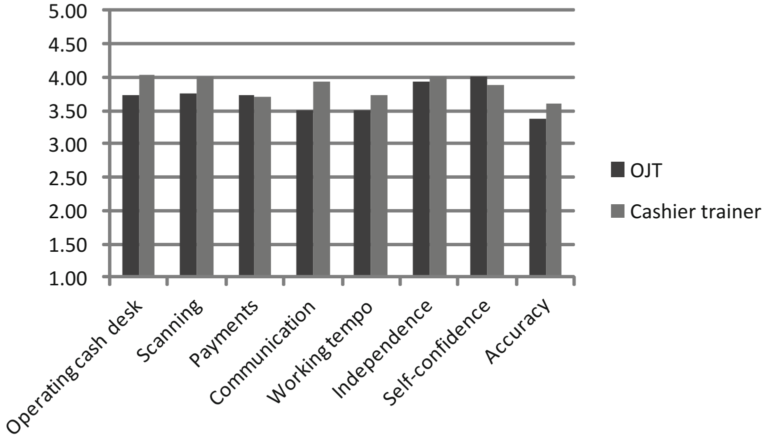

Fig. 3 Median scores for observed competence

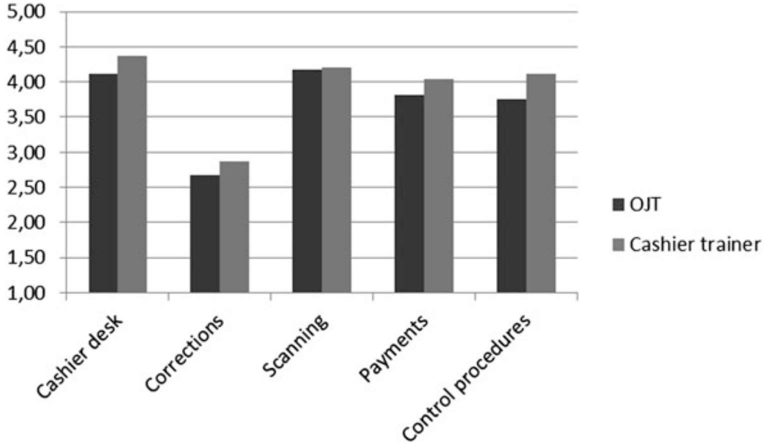

Fig. 4 Median scores for self-assessed performance

the measurement method: the observers counted the errors with value 0 (*wrong*) *or 1* (*correct)* while the subjects rated their task performance on a five points rating scale as stated in the previous section. Figure 4 shows that the candidates in the CASHIER TRAINER condition assessed themselves higher at all five variables than the candidates in the OJT condition. The Mann–Whitney *U* test ($p < 0.05$) showed that two of the five variables were significantly higher rated in the CASHIER TRAINER group than in the OJT group: payments ($U = 164.0$, $p = 0.041$, $r = 0.31$), and control procedures ($U = 167.0$, $p = 0.047$, $r = 0.30$). The values of the other three variables were not significant. Figure 4 also shows that the candidates rated corrections relatively low in comparison with the other variables.

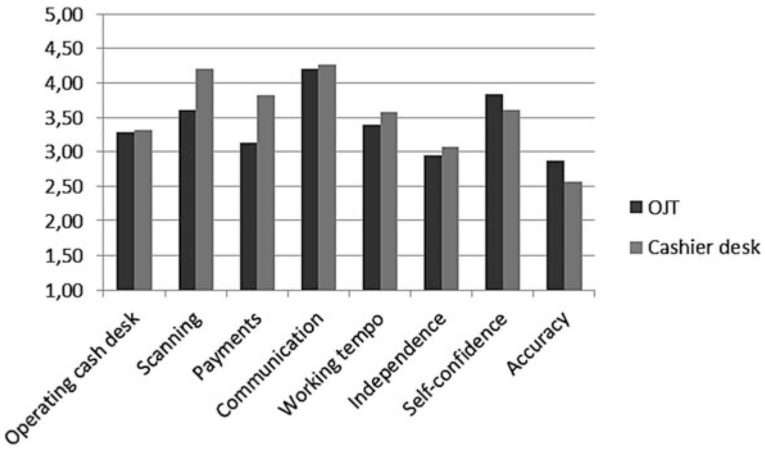

Fig. 5 Median scores for self-assessed competence

Figure 5 presents the self-assessments of general competence (median scores) of both training groups, called *self-assessed competence*. The set of rated competences is similar to the *observed competence* (see Fig. 3) to keep the results comparable. Figure 4 shows that the candidates rated themselves higher at 6 out of 8 competences for the CASHIER TRAINER condition than for the OJT condition. There were two significant differences found with the Mann–Whitney U test ($p < 0.05$): scanning ($U = 152.0$, $p = 0.010$, $r = 0.38$) and payments ($U = 153.5$, $p = 0.018$, $r = 0.35$). It should be noted that the differences between the rated competences are higher for the self-assessments (see Fig. 5) than for the observers (see Fig. 3).

4.3 Total Outcome Measures

Figure 6 presents the median scores of three of the four outcome measures observed competence, self-assessed performance and self-assessed competence. The outcome measures themselves are also based on the medians of the underlying scales. Figure 6 shows that the CASHIER TRAINER group scores slightly higher than the OJT group on the three outcome measures. No significant differences were found with the Mann–Whitney U test ($p < 0.05$). Outcome measure observed performance is not presented in this figure because it is not measured on the same five points rating scale. We did not find any differences between the CASHIER TRAINER group (Mdn = 0.50, SD = 0.048) and the OJT group (Mdn = 0.50, SD = 0.046).

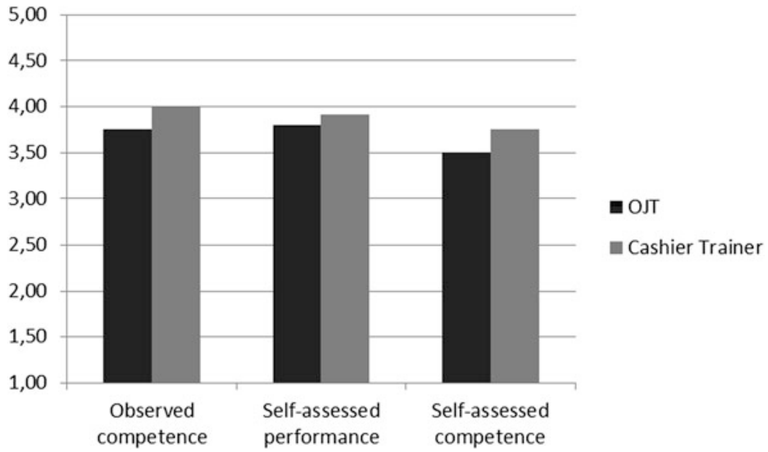

Fig. 6 Median scores for three total outcome measures

4.4 Intercorrelations

Table 1 presents the intercorrelations (Spearman's rho, two-tailed) between the total median scores of the four types of measurements (N = 45):

Table 1 shows that the measurements of performance and competence for the self-assessments are very highly correlated. This implies that performance and competence was strongly related and that the candidates are consistent in their judgments as expected. In the same way, the measurements of performance and competence for the observers are significantly correlated. These correlations are negative because observed performance implies the number of errors which is a reverse scale in comparison to the rating scale from poor to excellent for the other three measures. Observed performance also correlates with the other three measures although not significantly. Interesting is the relatively high and significant correlation between observed and self-assessed competence. This implies that the candidates and the observers have a quite similar picture about their competence. This is less clear for the relationship between observed and self-assessed

Table 1 Correlations (Spearman Rho, 2-tailed) total performance and competence

	1	2	3	4
1. Observed performance	1	−0.68[a]	−15	−0.14
2. Observed competence		1	0.11	0.13
3. Self-assessed performance			1	0.65^{**}
4. Self-assessed competence				1

Note ** $p < 0.01$

Table 2 Correlations (Spearman Rho, 2-tailed) with opinions

	Training technology	Quality of practice	Didactics	Realism
1. Observed performance	−0.03	−0.43*	−0.30	−0.27
2. Observed competence	0.03	0.37	0.11	0.22
3. Self-assessed performance	0.26	0.38	0.20	0.36
4. Self-assessed competence	0.23	0.43*	0.14	0.35

*Note** $p < 0.05$

performance probably due to the different type of measurement, counting errors, or a rating scale.

The candidates' performance was controlled for the possible factors age, education level, gender, and type of job (fulltime, part-time), but no significant correlations were found for any of the four types of measurements, performance or competence. In addition, the amount of time practiced in OJT or in the CASHIER TRAINER was investigated but we did not find significant results for this time factor either.

4.5 Didactical and Motivational Features of the Cashier Trainer

The candidates in the CASHIER TRAINER condition (N = 22) generally had a quite positive opinion on the CASHIER TRAINER. This particularly appears from qualitative data obtained from extra remarks of the subjects. The average ratings of the four scales, rated at the five points scale, varied from 3.7 to 3.9. Table 2 shows the correlations with the various types of measurements. Table 2 shows two significant correlations for quality of practice. In general, the highest correlations although not significant at $p < 0.05$ are found for quality of practice and realism. This suggests that these game features had a positive effect on the outcome of training with the CASHIER TRAINER.

5 Discussion

5.1 Transfer of Training

The objective of the present study was to investigate the transfer of training of the stand-alone CASHIER TRAINER to the workplace in comparison to conventional OJT. The CASHIER TRAINER simulates the entire procedural cashier task, immerses the learners in the virtual environment of a supermarket, and is fitted out with an ITS. The CASHIER TRAINER is therefore representative for single-player educational games with such an automated instruction and feedback functionality. Training outcomes of the experimental group trained with the CASHIER TRAINER were compared to a control group trained conventionally on-the-job. The two groups executed the same performance test at the workplace. Both observed and self-assessed performance as well as observed and self-assessed competence were measured. Transfer of training in this study refers to the (positive) training outcomes of the CASHIER TRAINER group relative to the OJT group. The results show that the outcomes of CASHIER TRAINER candidates are at least as high as the outcomes of the OJT candidates (see Figs. 2, 3, 4, 5 and 6). Although the presented figures suggest that the new employees trained with the CASHIER TRAINER score higher on the most outcome variables than the OJT group, the results of only some of the variables were significant. The overall outcome measures are generally a little bit higher for the CASHIER TRAINER group than for the OJT group but these small differences were not significant either (see Fig. 6).

From these results we may conclude that the new employees in the CASHIER TRAINER condition have learned at least as much as the OJT candidates and even more on certain aspects. We may assume that its automated feedback and instructional system has done its work appropriately; the candidates did not have a human coach when playing the CASHIER TRAINER in contrast with OJT. Thus, transfer of training with the CASHIER TRAINER to the workplace is definitely positive. As argued, we tried to fix the training time in the two conditions to be really able to compare them with each other but we saw that the variance between subjects was very high. Nevertheless, the OJT time was on average somewhat higher than in the CASHIER TRAINER. We did not found any effects of training time on the subjects' individual performance or competence. Therefore, we could derive that the experimental and control group generally have learned as much and that this was probably not caused by a shorter training time.

The practical implication of these findings is that we can assume that the CASHIER TRAINER can replace a certain part of total OJT time. Due to its artificial feedback system, candidates can do this game at home, independently of a human coach, before they enter the supermarket. Therefore, the CASHIER TRAINER can save on teacher or coaching capacity needed in conventional OJT. In the interviews, the supermarket managers estimated these savings on instructional personnel at about 50–70 % but this estimation has not been quantified yet. This does not mean that we can assume that the OJT could be fully replaced by the CASHIER TRAINER; this did not belong to the scope of this study. We did not investigate how much OJT time could be saved exactly.

Considering the cashier trainer as representative educational games with an ITS, this finding about the expected transfer of training of these types of educational games and their possible replacement of OJT time agrees with the expectations found in the literature. Educational games may show high training effectiveness if they are supported with appropriate instructional support (e.g., Egenfeldt-Nielsen 2006; Farmer et al. 1999; Hays 2005; Sitzmann 2011). In this study, we focused on games with an automated feedback system and we did not explicitly investigate the (extra) role of human coaches, for instance, in the form of a debriefing as a follow-up of the game (Alklind Taylor et al. 2012; Crookall 2010). It would be interesting to compare automated and human instruction, feedback, and debriefings with each other in another study to get a deeper insight into the possible surplus value of automation by gaming versus OJT by human supervisors.

The present study was a typical case study with an experimental design for measuring transfer of training with an experimental and control group and executed in real-life working environment (Sitzmann et al. 2011; Stainton et al. 2010). This has caused some methodological restrictions. One of them involves problems with fixing the training time as explained. In addition, due to practical obstacles, we had to leave out a pretest. This would make it possible to compare the delta of learning. From a methodological point of view this is certainly preferred. Moreover, the number of subjects involved was relatively low which makes it even harder to find clear (significant) results. Next, the performance test only consisted of one case with one customer due to practical time restrictions. Two tests would be more reliable. This study should therefore be repeated in a pre and post-test experimental design with fixed training conditions including time, multiple performance tests, and a higher amount of subjects to get more valid evidence.

5.2 *Type of Tasks*

To get more insight into the reasons why training effectiveness appeared to be higher in the CASHIER TRAINER condition, we looked for some trends in the performance and competence data presented in Figs. 2, 3, 4 and 5. The number of errors on relatively easy and regular tasks, such as scanning normal articles and registration at the cash desk, were relatively low (see Fig. 2). The significant differences between experimental and control group that we found, especially for self-assessment (see Figs. 4 and 5) could be referred to as less regular tasks. This is particularly true for payments. In the performance test as well as in the real-life world, these payments are rather complicated for new cashier employees when not standard payment methods are used, for instance, a combination with pin and cash as one of the tasks in the tests. Also back-calculating money is difficult for the most new employees according to the supermarket managers. The CASHIER TRAINER offers the opportunity to practice with a variety of payment methods, from simple to complex, in a safe training environment, including payments that only scarcely occur in reality. We did also find significant results for self-assessment on scanning. This measure comprises also non-regular tasks such as scanning special articles (e.g., price reduction, alcohol). The third significant effect was found for self-assessment on control procedures which again happen less frequently (e.g., asking for ID). Scanning non-regular articles and control procedures can also be systematically trained and practiced in the CASHIER TRAINER. This is much more difficult in the real-life environment of OJT.

These findings provide evidence on earlier statements in the literature that educational games offer the possibility to practice these type of non-regular tasks more often in a well-structured sequence of dedicated scenarios: one of the reasons to use gaming or simulation instead of training in real-life environments (Farmer et al. 1999; Korteling et al. 2012; Oprins 2008; Oprins et al. 2011; Van Merrienboer and Kirschner 2007). The ITS (Polson and Richardson 1988) of the CASHIER TRAINER offers systematic instruction and feedback to learn these non-regular and also relatively difficult tasks systematically and effective in a safe and controlled environment with automated instruction and feedback (Korteling et al. 2012; Oprins 2008).

In our study the learning objectives were predefined and relatively easy to measure, referring to as direct transfer (Harteveld and Bekebrede 2011). In contrast to many other professional tasks, this was possible because the task of cash desk employees is strongly procedural (e.g., Van Merrienboer and Kirschner 2007). This means that the task is based on discrete action sequences, executed in a rigid or stereotyped way and leading to a clear predetermined goal. The CASHIER TRAINER is a stand-alone device developed to train this highly procedural cashier task. On the basis of the positive results so far, we may conclude that educational games—if well-designed and equipped with a well-designed ITS—are effective for this kind of procedural tasks. If other kinds of tasks and learning objectives are involved such as open ended cognitive tasks or complex perceptual-motor tasks (Farmer et al. 1999; Van Merrienboer and Kirschner 2007), it may be more difficult to obtain high positive transfer from these types of educational games with an ITS. It is the question whether human feedback can be replaced by automation with these types of tasks (Alklind Taylor 2012; Egenfeldt-Nielsen 2006).

5.3 Future Learning at the Workplace

Although transfer of training to the workplace is positive, a certain amount of OJT is supposed to be needed after having finished the CASHIER TRAINER program. Especially, the various self-assessed competences were rated differently by the candidates (see Fig. 5). These findings suggest that the candidates are relatively less confident in operating the cash desk in comparison with scanning and communication with customers and theirs working speed, independence, and accuracy had to be further developed according to themselves. This is an expectable outcome since only 3 hours of training is rather low for building routine in any training (e.g., Van Merrienboer and Kirschner 2007). In addition, the retention time must be taken into account here. The time between finishing the CASHIER TRAINER program and the first working day (at which the performance test took place) was about 2 to 3 weeks. This increased the risk of skill loss during this retention period and thus may lead to lower transfer. This risk counts especially for procedural tasks that must be practiced often in order to maintain routine. In general, procedural tasks must be frequently practiced and repeated just before entering OJT in order to achieve maximal transfer (Van Merrienboer and Kirschner 2007). If the time gap between the CASHIER TRAINER program and the first working day had been smaller, for instance, only a few days, the differences in transfer of training with the OJT conditions might have been much higher. For the control group, the performance test was immediately following OJT. This was a rather high contrast with the CASHIER TRAINER which unfortunately we could not control for.

Next to routine in task performance, also some more general competences are supposed to be acquired during prolonged working at the workplace. For instance, the stress component, for instance, rows of waiting customers, is usually missing in instructional games or simulations such as the CASHIER TRAINER. Psychosocial skills such as coping with stress are thus more difficult to train with gaming or simulation because they require face to face or life training (Farmer et al. 1999). In addition, specific communication procedures could be trained in the CASHIER TRAINER but the game did not include natural speech. The virtual customers in the CASHIER TRAINER had a limited arsenal of sentences that could be answered by the candidates in a multiple choice menu. Nevertheless, both the observers and the candidates rated communication as rather high both in the CASHIER TRAINER and OJT condition. Apparently, the virtual customers could replace real-life customers for the acquisition of a specific subset of communication procedures despite of a lack of natural speech. This is particularly relevant in domains where practicing with real customers or clients is difficult to realize, for instance, virtual patients are often used within the medical domain (Akl et al. 2010).

5.4 Observation Versus Self-Assessment

We also investigated the interrelations between the four outcome variables. Within candidates and within observers, the correlations between performance and competence were very high as expected since they refer to their own judgments. Both the performance and competence ratings were based on the same task execution. The relationship between self-assessed competence and observed competence seems stronger than between self-assessed performance and observed performance. The main reason is probably that these two measurement methods were different: counting the formal errors with only two possible scores, correct or wrong, versus more intuitive self-assessing performance more generally rated on a five points scale. It supports the idea that multiple measures should be used to get reliable assessments that can be used for performance measurement in simulation based training (Oprins 2008; Salas et al. 2003, 2009).

It is remarkable that the differences between the various self-assessed competences (see Fig. 5) were higher than for the observed competences (see Fig. 3). This might be explained by the fact that the possibility to get insight into own performance of the trainees by introspection was lacking for the observers. For instance, only the candidates themselves could have a precise feeling about their self-confidence in operating the cash desk independently of the supervisor. The observers could only rely on what they saw. This difference in observation perspective may also have increased the probability of *halo errors*: the tendency to think of a person in general as rather good or inferior and to color the ratings on specific dimensions by this general feeling (Oprins et al. 2006; Oprins 2008; Thorndike 1920). Probably, the observers' ratings were influenced by this well-known rating error.

5.5 Didactical and Motivational Features

Finally, the candidates' opinion on the didactical and motivational features of the CASHIER TRAINER was related to observed and self-assessed performance and competence. In general, the candidates were quite positive on these features. This appears also from the qualitative remarks on the questionnaire as well as from the interviews. The new cashier employees were motivated to do the CASHIER TRAINER, they liked the tutoring and feedback system, and they felt immersed in the simulated environment of the supermarket. These results are consistent with the notion that games are motivating (Csikszentmihali 1990; Malone 1981). They also agree with the idea that fidelity in particular is a very important feature of simulation (Allan et al. 1986; Farmer et al. 1999; Korteling et al. 2012). We would expect that comparable results may be found for other games although this study is restricted to only one case.

However, only the candidates' general opinion on the CASHIER TRAINER was used without really measuring the underlying learning processes such as self-directed learning, way of practicing, repetition of exercises, etcetera. Measuring process variables in combination with outcome variables is crucial to get insight into the factors determining transfer of training: what works and what doesn't work (e.g., Bedwell et al. 2012; Salas et al. 2009) for which groups, tasks, and conditions. Therefore, the relationship between the fidelity, didactical and motivational features and learning processes should be investigated in more detail in the future. In order to get more insight into the most effective design characteristics of instructional simulation games, this kind of research should be based on a larger set of comparable educational games. This conclusion confirms similar kinds of suggestions found in previous reviews (e.g., Ack et al. 2010; Ke 2009; Lee 1999; Randel et al. 1992; Sitzmann 2011; Vogel et al. 2006).

5.6 General conclusion

This is one of the first empirical studies in which real evidence was collected for transfer of training for a stand-alone educational serious game equipped with a well-designed ITS compared to conventional OJT. We found that the amount of learning was at least as high for the candidates who did the CASHIER TRAINER as for candidates who did OJT. Since transfer of training of this representative educational game was sufficiently high; we can conclude that these types of procedural stand-alone games, if well-designed, could reduce the amount of relatively expensive OJT. For particular procedural tasks, an appropriate ITS can replace teachers, coaches or human supervisors, necessary to guide the trainee through the training program and to provide feedback. If this reduces OJT time, this may result in substantial savings on costs of instructional staff. We suppose that this finding may be promising for future development of educational games for similar types of procedural tasks. A simulated environment offers the opportunity to practice both regular as well as non-regular tasks in well-sequenced immersive scenarios. This is not possible in OJT where the learning tasks cannot always easily be planned in advance if non-regular tasks hardly happen in real-life. This serves as an additional advantage of these types of educational serious games.

Acknowledgments This study was a Knowledge Transfer Project of the research program Games for Training and Entertainment (GATE) which was supported by the Dutch ICT-Regie and NWO. We would like to thank Jutten Simulation for using the game CASHIER TRAINER as a case in this study. We also would like to thank the retail company Sligro Foodgroup n.v. in The Netherlands for making it possible to collect data in their EMTÉ supermarkets.

References

Akl EA, Pretorius RW, Sackett K, Scott Erdley W, Bhoopath P, Alfarah Z, Nemann HJ (2010) The effect of educational games on medical students' learning outcomes: a systematic review. BEME guide no 14

Allan JA, Hays JT, Buffardi LC (1986) Maintenance Training Simulator Fidelity and Individual Differences in Transfer of Training. Hum Factors 28:297–509

Alklind Taylor AS, Backlund P, Niklasson L (2012) The Coaching Cycle: A Coaching-by-Gaming Approach in Serious Games. Simul Gaming 20(10):1–25

Baldwin TT, Ford JK (1988) Transfer of training: a review and directions for future research. Pers Psychol 41:63–105

Bandura A (1997) Self-efficacy: the exercise of control. Freeman, New York

Bedwell WL, Pavlas D, Heyne K, Lazzara EH, Salas E (2012) Toward a taxonomy linking game attributes to learning: an empirical study. Simul Gaming 43(6):729–760

Bekebrede G, Warmelink HJG, Mayer IS (2011) Reviewing the need for gaming in education to accommodate the net generation. Comput Educ 57(2):1521–1529. doi:10.1016/j.compedu.2011.02.010

Bell BS, Kanar AM, Kozlowski SWJ (2008) Current issues and future directions in simulation-based training in North America. Int J Hum Res Manag 19:1416–1436

Caro, P.W. (1977). Some factors influencing air force simulator training effectiveness HUMRRO Technical Report tr-77–2. Alexandria, Virginia: Human Resources Research Organisation

Cohn J, Kay S, Milham L, Bell Carroll M, Jones D, Sullivan J, Darken R (2009) Training effectiveness evaluation: from theory to practice. In: Schmorrow D, Cohn J, Nicholson D (eds). The PSI handbook of virtual environments for training and education,pp 157–172

Connolly TM, Boyle EA, MacArthur E, Hainey T, Boyle JM (2012) A systematic literature review of empirical evidence on computer games and serious games. Comput Educ 59(2):661–686. doi:10.1016/j.compedu.2012.03.004

Crookall D (2010) Serious games, debriefing, and simulation/gaming as a discipline. Simul Gaming 41(6):898–920

Csikszentmihali M (1990) Flow: the psychology of optimal experience. Harper and Row, New York

Egenfeldt-Nielsen S (2006) Overview of research on the educational use of video games. Digital Kompetanse 1(3):184–213

Emmerik ML, van, Korteling JE (2002) *Certificering van trainingssimulatoren 2*: de TNO-TM checklist. [Certification of training simulators 2: The TNO-TM checklist] Report TM-02-D010. Soesterberg, The Netherlands: TNO Human Factors research Institute

Emmerik ML, van, Korteling JE (2003) Certificering van trainingssimulatoren 3: Computer-gebaseerd ONdersteuningmiddel voor CERtificering van Trainingssimulatoren. [Certification of training simulators 3: Computer-based Support Tool for Certification of Training Simulators] Report TM-03-D005. Soesterberg, The Netherlands: TNO Human Factors research Institute

Farmer E, van Rooij J, Riemersma J, Jorna P, Moraal J (1999) Handbook of simulator-based training. Ashgate, Aldershot

Fletcher JD, Tobias S (2008) What research has to say (thus far) about designing computer games for learning. Paper presented at the American Educational Research Association, New York

Fowlkes J E, Dwyer DJ, Milham L M, Burns J J, Pierce L G (1999) Team skills assessment: A test and evaluation component for emerging weapon systems. Paper presented at the Interservice/Industry Training, Simulation, and Education Conference, Orlando, FL

Freitas S, de (2006) Learning in immersive worlds: a review of game-based learning (Tech. Rep.). Joint Information Systems Committee Bristol, Bristol

Gee JP (2007) What videogames have to teach us about learning and literacy. Palgrave Macmillan, New York

Gielen EWM (1995) Transfer of training in a corporate setting (doctoral thesis). University Twente, Enschede

Girard C, Ecalle J, Magnan A (2012) Serious games as new educational tools: How effective are they? A meta-analysis of recent studies. J Computer Assist Learn. doi:10.1111/j.1365-2729.2012.00489.x

Harteveld C (2011) Triadic game design: Balancing reality, meaning and play. Springer, London

Harteveld C (2012) Making sense of Virtual Risks: a Quasi-Experimental Investigation into Game-Based Training. IOS Press, Amsterdam

Harteveld C, Bekebrede G (2011) Learning in Single-Versus Multiplayer Games: The More the Merrier? Simul Gaming 42(1):43–63

Hays RT (2005) The effectiveness of instructional games: a literature review and discussion. Technical Report 2005-004. Naval Air Warfare Training Systems Division. Orlando, U.S.A

Jiusto S, DiBiasio D (2006) Experiential learning environments: Do they prepare our students to be self-directed, life-long learners? J Eng Educ 95:195–204

Johnston S, McCormack C (1996) Integrating information technology into university teaching: Identifying the needs and providing the support. Int J Educ Manag 10(5):36–42

Ke F (2009) A qualitative meta-analysis of computer games as learning tools. In Ferdig RE (ed) Handbook of research on effective electronic gaming in education vol I. pp 1–32

Kirkpatrick DI (1998) Evaluating Training Programs: The Four Levels, 2nd edn. Berrett-Koehler, San Francisco

Korteling JE, Helsdingen AS, Theunissen NCM (2012) Serious Games @ Work Learning job-related competences using serious gaming. In: Bakker A, Derks D (eds) The Psychology of Digital Media at Work. London, New York: Psychology Press LTD/Taylor and Francis Group

Korteling JE (2012) Evaluation of a cashier trainer: true evidence for transfer of training of serious gaming. In: Veltkamp R (ed) Growing knowledge for games utrecht NL: Control Magazine 68–69

Korteling JE, Oprins EAPB, Kallen VL (2013) Measurement of Effectiveness for training simulations. In: Wang Z (ed) RTO-MP-SAS-095 Cost-Benefit Analysis of Military Training. Paper presented at the SAS Workshop held in Amsterdam, The Netherlands, 5–6 June 2012. [http://www.cso.nato.int/abstracts.aspx]

Korteling JE, Padmos P, Helsdingen AS, Sluimer RR (2001) *Certificering van trainingssimulatoren 1: kennisinventarisatie* [Certification of training simulators 1: knowledge inventarization] Report TM-01-D003. TNO Human Factors Research Institute, Soesterberg, The Netherlands

Leemkuil H, de Jong T, Ootes S (2000) Review of educational use of games and simulations (IST-1999-13078 Deliverable D1). University of Twente, Enschede

Lee J (1999) Effectiveness of a computer-based instructional simulation: a meta-analysis. Int J Instr Media 26:71–85

Malone TW (1981) Towards a theory of intrinsically motivating instruction. Cognitive Sci 4:333–369

Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 18(1):50–60

van Merrienboer JJG, Kirschner P (2007) Ten steps to complex learning: a systematic approach to four-component instructional design. Lawrence Erlbaum Associates, Mahwah

O'Neil HF, Wainess R, Baker EL (2005) Classification of learning outcomes: evidence from the computer games literature. Curric J 16(4):455–474. doi:10.1080/09585170500384529

Oprins E (2008) Design of a competence-based assessment system for air traffic control training. Maastricht University, Doctoral dissertation

Oprins E, Burggraaff E, Roe R (2011) Analysis of learning curves in the on-the-job training of air traffic controllers. In: D'Oliviera TC (ed) Mechanisms in the chain of safety. Ashgate Publishing Company, Aldershot

Oprins E, Burggraaff E, Van Weerdenburg H (2006) Design of a competence-based assessment system for air traffic control training. Int J Aviat Psychol 16(3):297–320

Oprins E, Burggraaff E, Van Weerdenburg H (2008) Reliability of assessors' ratings in competence-based air traffic control training. Hum Factors Aerosp Saf 6(4):305–322

Peck M (2012) Tools or toys? Training games are popular, but no one knows how well they work. Train simul J dec2011/jan2012

Percival A (1996) Invited reaction: An adult educator responds. Hum Resour Dev Quart 7:131–139

Petraglia J (1998) Reality by design: the rhetoric and technology of authenticity in education. Erlbaum, Mahwah

Phillips DC (1998) How, why, what, when, and where: Perspectives on constructivism in psychology and education. Issues Educ 3:151–194

Polson M, Richardson J (eds) (1988) Foundations of Intelligent Tutoring Systems. Lawrence Erlbaum Associates, Hillsdale

Randal JM, Morris BA, Wetzel CD, Whitehill BV (1992) The effectiveness of games for educational purposes: a review of recent research. Simul Gaming 23:261–276

Roscoe SN, Williges BH (1980) Measurement of transfer of training. In: Roscoe SN (ed) Aviation Psychology. The Iowa State University Press, Iowa

Salas E, Milham LM, Bowers CA (2003) Training evaluation in the military: misconceptions, opportunities, and challenges. Mil Psychol 15:3–16

Salas E, Rosen MA, Held JD, Weissmuller JJ (2009) Performance Measurement in Simulation-Based Training: A Review and Best Practices. Simul Gaming 40(3):328–376

Sitzmann T (2011) A meta-analytic examination of the instructional effectiveness of computer-based simulation games. Pers Psychol 64:489–528

Smith PAC, O'Neil J (2003) A review of action learning literature 1994-2000. Part 1: Bibliography and comments. J Workplace Learn 15:63–69

Stainton AJ, Johnson JE, Borodzicz EP (2010) Educational Validity of Business Gaming Simulation: a research methology framework. Simul Gaming 41(5):705–723

Steffe L, Gale J (eds) (1995) Constructivism in education. Lawrence Erlbaum Associates, Inc., Hillsdale

Stehouwer M, Serné M, Niekel C (2005) A tactical trainer for air defence platoon commanders. In: Proceedings of the interservice/industry, training, simulation, and education conference. Orlando I/ITSEC 2005, Paper no. 206

Shaffer (2006) How computer games help people learn. Palgrave Macmillan, New York

Squire K (2003) Video games in education. Int J Intell Simul Gaming 2(1):49–62

Stubbé HM, Theunissen NCM (2008) Self-directed learning in a ubiquitous learning environment: a meta-review. In: Proceedings of special track on technology support for self-organised learners 2008, pp 5–28

Thorndike EL (1920) A constant error in psychological ratings. J Appl Psychol 4:25–29

Tobias S, Fletcher JD (2007) What research has to say about designing computer games for learning. Educ Technol 47:20–29

Veldhuis GJ, Theunissen NCM (2009) Transfer of Training. Onderzoek naar een maximaal leereffect [Transfer of training. Research on a maximal learning effect]. Opleiding en Train 11:20–22

Vogel J, Vogel DS, Cannon-Bowers J, Bowers CA, Muse K, Wright M (2006) Computer gaming and interactive simulations for learning: a meta-analysis. J Edu Comput Res 34:229–243

Wouters P, van der Spek E, Van Oostendorp H (2009) Current practices in serious game research: a review from a learning outcomes perspective. Games-based learning advancements for multi-sensory human computer interfaces: techniques and effective practices 232–250

Young MF, Slota S, Cutter AB, Yukhymenko M (2012) Our princess is in another castle: A review of trends in serious gaming for education. Rev Educ Res 82(1):61–89