ELSEVIER

Learning and Instruction 18 (2008) 172–186

Learning and Instruction

www.elsevier.com/locate/learninstruc

The effect of practical experience on perceptions of assessment authenticity, study approach, and learning outcomes

Judith T.M. Gulikers [a,*], Liesbeth Kester [a], Paul A. Kirschner [a,b], Theo J. Bastiaens [a,c]

[a] *Educational Technology Expertise Center, Open University of the Netherlands, The Netherlands*
[b] *Research Centre Learning in Interaction of the Utrecht University, The Netherlands*
[c] *Educational and Media Institute, Fern University of Hagen, Germany*

Received 23 May 2006; revised 4 December 2006; accepted 23 February 2007

Abstract

Does authentic assessment or the perception of it affect how students study and learn? Does practical experience affect how assessment authenticity is perceived? And does practical experience influence how an authentic assessment affects student learning? Mixed methods design yielded insight into the answers to these questions. This article presents the results of a study on the relationships between authenticity perceptions of different cohorts of students, who differed in the amount of practical experience, their study approach and their perceived degree of professional skill development. The results showed some salient differences in how freshman- and senior-student groups perceive the same authentic assessment and how this assessment influences their learning. These results suggest possible guidelines for developing and using authentic assessments during a curriculum in which learning and working are intertwined.

Keywords: Authentic assessment; Work-based learning; Student perceptions; Study approach; Vocational education

1. Introduction

One of the main characteristics of new modes of assessment that focus on higher-order skills or competencies that are relevant for successful job performance, is that they are *authentic* (Boud, 1990, 1995; Dierick & Dochy, 2001; Gielen, Dochy, & Dierick, 2003; Messick, 1994; Segers, 2004; Tillema, Kessels, & Meijers, 2000). Such authentic assessment aims at linking learning and working by creating a correspondence between what is assessed in the school and what students need to do in the workplace during an internship or after finishing their education (Boud, 1995; Gulikers, Bastiaens, & Kirschner, 2004; Messick, 1994). By creating this correspondence, authentic assessments

* Corresponding author. Present address: Education and Competence Chair group of the Wageningen University, P.O. Box 8130, 6700 EW Wageningen, The Netherlands. Tel.: +31 317 484332; fax: +31 317 484573.
E-mail address: judith.gulikers@wur.nl (J.T.M. Gulikers).

0959-4752/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.learninstruc.2007.02.012

are expected to positively influence student learning and better prepare students for their future careers. Authentic assessments are expected to (a) stimulate students to learn more deeply (Birenbaum, 1996; Dochy & McDowell, 1998; Frederiksen, 1984; McDowell, 1995); (b) stimulate students to develop professionally relevant skills and thinking processes used by professionals (Gielen et al., 2003; Savery & Duffy, 1995); and (c) motivate students to learn by showing the immediate relevance of that what is learnt for professional practice (Herrington & Herrington, 1998; Lizzio & Wilson, 2004a; Martens, Gulikers, & Bastiaens, 2004; McDowell, 1995).

1.1. Perception of assessment authenticity and student learning

Unfortunately, the influence of authentic assessment on student learning is not this straightforward. It is complicated by two important issues. First, authenticity is not an objective construct (Honebein, Duffy, & Fishman, 1993; Petraglia, 1998). This means that people can differ in their perception of the authenticity of the same assessment. The problem in this case is that *student's perception* of assessment characteristics seems to have more influence on student learning than the ''objective'' characteristics of assessment themselves (Entwistle, 1991; Van Rossum & Schenk, 1984). If assessment authenticity is defined by the degree of correspondence between the assessment and the professional practice situation it is purported to reflect (Gulikers et al., 2004), then the influence of an authentic assessment on student learning depends on how a student perceives the resemblance between this assessment and professional practice.

The second complicating issue is that a student's ideas of professional practice might change as a result of the experiences in professional practice (Handal & Hofgaard Lycke, 2005; Honebein et al., 1993; Lizzio & Wilson, 2004a). Many types of education, especially vocational, are trying to integrate learning and working in order to smoothen the transition from school to the workplace (Biemans, Nieuwenhuis, Poell, Mulder, & Wesselink, 2004; Boshuizen, Bromme, & Gruber, 2004). This approach entails increasing the student's opportunity to gain experience in professional practice during schooling. The evidence for the effect of this experience on professional practice, however, is unequivocal. Following the previous line of reasoning, changed ideas about professional practice might change the perception of authenticity of an assessment, which, in turn, might influence how an authentic assessment affects the learning of students with different experiences in professional practice. Whether this is true is a highly important question in current educational practice, because it affects how authentic assessments should be used and operationalised during educational trajectories.

Some studies suggested that students do not change their perceptions of professional practice and of assessment during their years of studying, even if they gain more experience in professional practice through internships (Handal & Hofgaard Lycke, 2005; Winning, Elaine, & Townsend, 2005). Handal and Hofgaard Lycke (2005) showed that students' perception of professional practice as well as their approach to studying did not change during the years of study, but that they did change drastically after 1 year of work. It was argued that as long as students stay in school they are mostly guided by school requirements rather than by their possibly changed ideas of professional practice or assessment. Based on the results of these studies, it can be assumed that the relationship between perception of assessment authenticity and student learning and professional skill development is stable over the years. This would imply that there is no need to change the form of authentic assessment during students' educational trajectory. Important to note, however, is that both of the above mentioned studies were conducted at the university level. The internship regimes are likely to be different from the internship regime in vocational education in the Netherlands – the context of this study – in which students start doing internships from the start of their studies and where learning and working are alternated on a regular basis. This might be a crucial difference.

In support of the influence of practical experience, Honebein et al. (1993) and Messick (1994) suggested that students with different levels of practical expertise might learn better with different kinds of assessments. As an example, they argued that when students have had enough opportunity to get a good picture of professional practice, the physical context of an assessment might become self-imposed. In other words, experienced students would be able to create a realistic physical context for themselves and do not need assessments to be situated in a high fidelity context to stimulate their learning. Inexperienced students might not yet be able to frame an assessment task in a realistic context, because they have had too little practical experience for this framing. This implies that inexperienced students benefit more from a contextualised assessment than more experienced students. On the other hand, it has also been suggested that as students get closer to their graduation and thus closer to actually working, the need for a very authentic physical context (preferably the real workplace) is increasing, instead of decreasing, and is needed to positively influence learning

(Klarus, 2000). All of these studies imply that the same authentic assessment differentially influences the learning of students with varying degree of practical experience. Specifically, the same authentic assessment might be effective for students with little experience in professional practice, while being less effective for students who have more professional experience or vice versa. In practice, this would favour the use of different kinds of authentic assessments for students with different amounts of experience in professional practice.

The present study aimed at examining perceptions of assessment authenticity and their relationship with student learning. A growing body of literature and research on new modes of assessment stresses that the effects of assessment on student learning should always be examined in light of the whole learning environment along with students' perception of the learning environment (Biggs, 1996; Birenbaum, 1996; Segers, Dierick, & Dochy, 2001; Struyven, 2005). This approach stresses the need for alignment between instruction and assessment to positively influence student learning. This means that when authentic assessment is directed towards stimulating deep study activities and the development of professional skills, then instruction should also be perceived to require these same skills (Gulikers et al., 2004). When students perceive a mismatch between the kind of learning stimulated by the instruction and the kind of learning that is needed for the assessment, the expected positive effect on learning does not occur (Segers et al., 2001; Struyven, 2005). Therefore, this study not only considered student perceptions of assessment authenticity but also their perceptions of alignment between the authentic assessment and instruction.

1.2. The hypothesised model

Fig. 1 shows the hypothesised model that describes the expected relationships in this study (Gulikers et al., 2004; Lizzio, Wilson, & Simons, 2002).

The independent variables are *perception of assessment authenticity* and *perception of alignment* between instruction and assessment. They are depicted in the left column of Fig. 1. "Assessment authenticity" is defined as a multidimensional construct with five assessment characteristics (i.e., dimensions; Gulikers et al., 2004), namely the assessment task, the physical context of the assessment, the social context of the assessment, the assessment form,

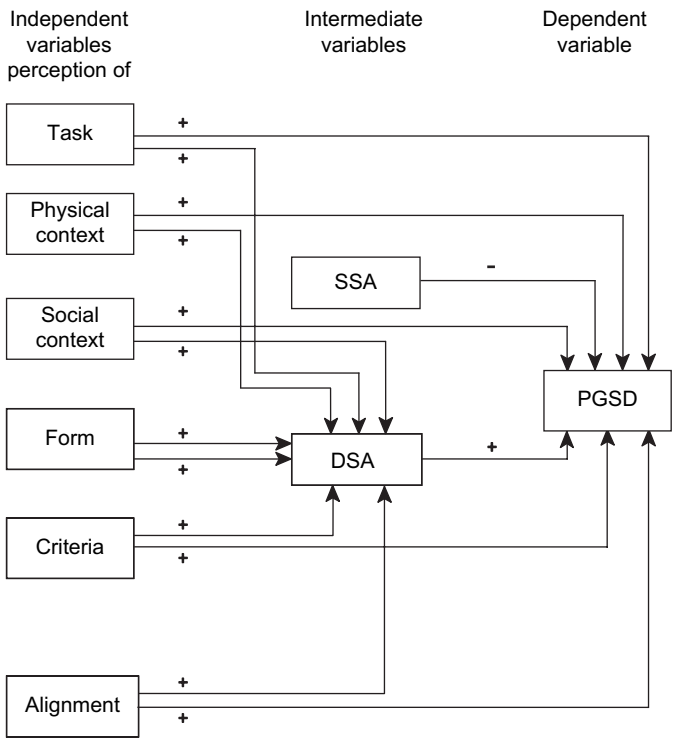

Fig. 1. The hypothesised model (DSA = deep study approach; SSA = surface study approach; PGSD = perception of generic skill development).

and the assessment criteria. The perception of assessment authenticity should be also rated along these five dimensions and, therefore, they are depicted as five separate variables in the model.

The dependent variable is the *perception of generic skill development* (PGSD), which indicates to what extent students feel that an assessment stimulates the development of six generic professional skills, namely problem-solving, planning, collaborating, communicating, dealing with unknown situations, and thinking analytically (Lizzio et al., 2002; Wilson, Lizzio, & Ramsden, 1997).

The intermediate variable, namely *study approach*, is split up in deep study approach (DSA) and surface study approach (SSA) (Biggs, Kember, & Leung, 2001), both depicted in the middle column and, thus, situated between the independent variables and the dependent variable PGSD. In line with the goals of authentic assessment, this model hypothesised that an increase in the perception of authenticity of any of the five assessment dimensions and/or of alignment results in increased development of generic skills either directly or indirectly through encouraging a deep study approach (Lizzio et al., 2002).

2. The present study: the effect of practical experience

The present study examined how two student groups that differ in the degree of practical experience perceive the authenticity of the same kind of authentic assessment and what kind of study approach and degree of PGSD they report in response to the assessment. It also examined if the hypothesised model adequately describes the *relations* between perceptions of assessment authenticity, study approach, and PGSD for both student groups and whether these relations are differentiated as a result of more practical experience. The following predictions were formulated:

1. The two student groups that differ in the degree of practical experience should differ in their perception of assessment authenticity and alignment, in their reported study approach, and in their degree of PGSD in response to the same assessment (Hypothesis 1).
2. In line with the relationships described in the hypothesised model, an increase in the perception of assessment authenticity in the various assessment dimensions should result in deeper studying and higher PGSD for both student groups (Hypothesis 2).
3. The relationships described in the hypothesised model might be affected by students' differential practical experience meaning that the same authentic assessment differently influences the relationship between study approach and PGSD for students with differing degrees of practical experience (Hypothesis 3).

3. Method

3.1. Participants

Two groups of students from a Vocational Education and Training (VET) College for Social Work in the Netherlands participated in the study. They were 81 freshman students (mean age = 17.82, SD = 1.57, 17 males, 64 females) and 118 senior students (mean age = 19.16, SD = 1.14, 48 males, 33 females), respectively. These two groups were selected because they reflected the moderator variable ''amount of practical experience''. Freshman students were at the beginning of their studies and had little experience in professional practice. They had experience with only one institute where they had been doing an internship 1 day a week for half a year. Senior students were almost at the end of their studies and had a lot more and different kinds of practical working experience. They did internships in several institutes and of varying durations – from 1 day a week through 6 months full-time.

3.2. Materials

3.2.1. The assessment

This study made use of two regular summative assessments that were designed to be authentic assessments by the participating VET College for Social Work. Both student groups performed the same kind of assessment (a 10-minute

role-play) situated in a classroom and based on a social-work-related case description that students could prepare beforehand. A teacher played the role of client and students were handed a list of 10 assessment criteria 1 week before the assessment. Students individually had to show their competence in dealing with the problem situation described in the case description. During their performance, students were observed and assessed by two assessors. After every student assessment, the assessors shortly discussed their observations and decided whether the student succeeded or not. The assessments of the two assessors were aligned to a competency-based instructional period of 9 weeks in which students gained knowledge, skills and attitudes via problem-based learning, self study, and skills training during which students practiced with performing role-plays.

3.2.2. Perception of assessment authenticity

A 24-item questionnaire based on the five-dimensional framework for assessment authenticity (Gulikers et al., 2004) was used to examine to what degree students perceived the five assessment characteristics (the task, the physical context, the social context, the form, and the criteria) to resemble professional practice. The respective five subscales of the questionnaire were based on a factor analysis of VET students' scores (Gulikers, Bastiaens, & Kirschner, 2006). The items were scored on a five-point Likert scale ranging from 1 (totally disagree) to 5 (totally agree). All subscales, except for the social context one, had a reasonable internal consistency, shown in Cronbach's alpha ranging from .63 to .83. The social context scale was excluded from further analysis due to its low reliability ($\alpha = .48$).

3.2.3. Perception of alignment

The perception of alignment was measured by a five-item questionnaire developed for this study. It examined whether students perceived the instruction to convey the same message as the assessment with regard to what kind of learning is valued (e.g., ''During the instructional phase I had to use my knowledge in the same way as during the assessment'' or ''Based on the instruction, I expected a different kind of assessment''). Confirmatory factor analysis confirmed unidimensionality – the five items loaded on one factor with factor loadings ranging from .48 to .85. Cronbach's alpha for this scale was .75.

3.2.4. Study approach

Study approach was measured with the Revised-Study Process Questionnaire-2 Factors (R-SPQ-2F; Biggs et al., 2001), a revision of the Study Process Questionnaire (Biggs, 1987). The R-SPQ-2F is a 20-item questionnaire that is used to distinguish between two study approaches, namely a deep study approach (DSA; 10 items) and a surface study approach (SSA; 10 items). Several studies indicated reliable coefficients for the two subscales (Biggs et al., 2001); moreover, the items were short, all positively stated and without difficult wording. These were important considerations, since the research population involved students in VET instead of higher education. In addition, the questionnaire (e.g., Scouller, 1997) was successfully used in previous research to examine relations between study approaches and learning outcomes. The original questionnaire was translated into Dutch and contextualised to the authentic assessment that was the object of this study. Confirmatory factor analysis largely confirmed the two scales. All deep study items loaded on factor one and 8 out of the 10 surface items loaded on factor two with factor loadings larger than .4 (Field, 2000). Results indicated that the two subscales of the translated version had a reasonable internal consistency in the VET context ($\alpha = .81$ for DSA, and $\alpha = .66$ for SSA). Item analysis showed that deleting the two surface items resulted in an internal consistency for the SSA subscale of $\alpha = .68$. Due to this very small impact of deleting these two items, it was decided to keep all 10 items of the original and validated subscale in the analysis (Biggs et al., 2001).

3.2.5. Perceived qualitative learning outcome

The perceived qualitative learning outcome was measured with a Dutch translation of the Generic Skill Development subscale of the Course Experience Questionnaire (CEQ; Wilson et al., 1997). This subscale, consisting of six items, measured the extent to which students felt that a certain learning activity (in this case, studying for the authentic assessment) contributed to the development of six respective transferable generic skills (i.e., problem-solving, analytic skills, teamwork, confidence in tackling unfamiliar situation, ability to plan work, and written communication skills). Lizzio et al. (2002) showed that this subscale could be used as a measure for qualitative learning outcome. The

unidimensionality of this translated version was corroborated by a confirmatory factor analysis. The six items loaded on one factor with factor loadings ranging from .57 to .80 with a good internal consistency in the VET context ($\alpha = .74$).

3.2.6. Focus groups

To gain a deeper insight into the perceptions of authenticity of the assessment, the way these perceptions influenced study approaches and perceived learning outcome, and the effect of practical experience on these relationships, the quantitative data were complemented with qualitative data obtained from semi-structured focus-group interviews with freshman and senior students. A random selection of students participated in five interviews; two freshman groups (total $n = 18$) and three senior groups (total $n = 27$). In these group interviews, participants were encouraged to freely express their perceptions and experiences and to respond to each other, based on initial stimuli provided by the interviewer. The interview schedule used in this study was based on a combination of the five-dimensional framework for authenticity assessment (Gulikers et al., 2004), and the interview schedules of Sambell, McDowell, and Brown (1997) and Lizzio et al. (2002) focusing on perceptions of assessment characteristics and on the consequential validity of the assessment. The interviews focused on (1) how students perceived the authenticity of the five characteristics of the assessment; (2) how they prepared for the assessment and if this depended on the authenticity of the assessment characteristics; and (3) what kind of learning they believed was being assessed in this way.

3.3. Procedure

Almost immediately after performing the assessment, every student was informed about the fail/succeed decision of the assessors. All students filled in the questionnaires in their next class meeting after the assessment, 2 days later. Every class meeting was supervised by a teacher. The focus-group interviews were conducted 2 weeks after performing the assessment.

3.4. Analyses

3.4.1. Quantitative analysis

A MANOVA was used to compare the mean levels of perceived authenticity of the assessment characteristic and alignment, study approaches, and PGSD between the two student groups.

To examine if more perception of authenticity and alignment resulted in deeper learning and PGSD, structural equation modelling (SEM) with AMOS (Byrne, 2001) was used. This analysis was chosen because – as opposed to regression analysis – it is appropriate for examining direct as well as indirect relationships between *continuous variables* and detecting small differences *within* one group (Joreskog, 1993). To examine if the relationships in the hypothesised model differ between freshman and senior students, multi-group structural equation modelling was used. This method first examined if the hypothesised model (Fig. 1) adequately described the relationships *within* each group separately and then compared the found relationships *between* the groups. The goal was to make inferences about group (freshman vs. senior) differences in the relationships. This is done in a three-step manner (Byrne, 2001): (1) assess the tenability of the hypothesised model for each group separately; (2) assess the tenability of the hypothesised model simultaneously across both groups (baseline model); (3) assess group differences on individual parameters linking the various variables by (a) constraining all theoretically interesting parameters to be equal across groups and comparing this to the baseline model and (b) by sequentially releasing constraints if the Modification index indicates a significant improvement in data-model fit. Parameters whose constraints are released are inferred to differ across groups, while those whose constraints are not released are inferred to be stable across groups.

3.4.2. Qualitative analysis

The focus-group data were used to complement the quantitative data, but they were especially used to explain differences between groups or to explain unexpected findings. The interview data for analysis were first parsed in fragments and coded based on the themes of the interview schedule used. To minimise the influence of personal interpretation and increase the reliability of the conclusions, the interpretations of the codes were first discussed

and a selection of the interviews was coded, by two researchers independently to find a common ground for data coding. This resulted in an interrater agreement of 74%. After that, the interpretation of all interviews rested upon careful reflection and discussion between two researchers, one of whom was not involved in the conducting of the interviews.

4. Results

4.1. Group differences

The 2 (group) × 8 (perceptions of authenticity, study approach, and PGSD) MANOVA was significant, Pillais' trace = .14, $F(8, 190) = 4.00$ $p = .00$, Cohen's $f = .41$.[1] The univariate tests showed that freshman students perceived the assessment task and physical context as more authentic than senior students, $F(1, 197) = 4.94$, $p = .03$, $f = .16$; $F(1, 197) = 12.35$, $p = .001$, Cohen's $f = .25$, respectively, and they differed in PGSD, $F(1, 197) = 10.13$, $p = .002$, Cohen's $f = .23$, in favour of the freshman students, while the two groups did not differ in their study approaches and in their perception of assessment form or criteria authenticity and their perception of alignment. Table 1 displays the mean scores of both student groups. To avoid Type I error due to the large number of univariate tests, Bonferroni correction on the alpha level, $\alpha = .05/8 = .006$, was applied. Following this stringent alpha level, the two groups of students did not differ in their perception of the assessment task. This was also indicated by the low effect size of the respective univariate test.

4.2. Relationships between the variables

Table 2 displays the data of the steps taken in the multi-group SEM analyses.

First, the hypothesised model (Model 1) was tested as a whole and after that, all parameters were compared individually. The hypothesised model turned out to be tenable for both groups separately, $\chi^2(6, n = 118) = 7.71$, $p = .26$, CFI = .99, and RMSEA = .05 for senior students and $\chi^2(6, n = 81) = 1.52$, $p = .96$, CFI = .99, and RMSEA = .00 for freshman students. In addition, the model was tenable for the two groups together, $\chi^2(12, N = 199) = 9.22$, $p = .68$, CFI = 1.0, and RMSEA = .00. This means that, within both groups, students who perceived the assessment as more authentic, employed deeper learning and reported higher PGSD than students who perceived the assessment as less authentic. As shown in Fig. 2, not all hypothesised relationships were significant. Nevertheless, authenticity perceptions of all assessment dimensions had either an indirect or a direct significant effect on the PGSD, supporting the expectation that the different dimensions of assessment authenticity influence deep studying and/or PGSD. Authenticity perceptions of the task, physical context and criteria influenced PGSD indirectly through DSA. The influence of the authenticity perception of the assessment form on PGSD was only direct, without influencing DSA. The perception of alignment had no significant effect on DSA nor on PGSD.

To test the equality of the individual parameters linking the independent, intermediate, and dependent variables in the hypothesised model for the two student groups, the values of these parameters were constrained to be equal across groups. The χ^2 of this constrained model (Model 2) was then compared to the hypothesised model (Model 1). The χ^2 difference was significant, $\Delta\chi^2(12) = 33.50$, $p < .05$, which means that not all parameter values were equal across the two student groups. To locate the non-equivalent parameter values in the model, the Modification Indexes (MI) were studied. The MI showed that the link between criteria authenticity and DSA did not have an equal value for the two student groups. Releasing the equality constraint of this parameter, thereby allowing the link between criteria authenticity and DSA to differ between groups, resulted in a significant improvement of Model 2, $\Delta\chi^2(1) = 21.87$, $p < .05$. This means that the value of the path of perception of criteria authenticity to DSA, and indirectly on PGSD, differed between freshman and senior students. For freshman students the value was $\beta = .29$, while for senior students it was $\beta = -.41$. This finding shows that, in line with the prediction, an increase in perception of criteria authenticity stimulated freshman students to deeper studying and higher PGSD. Senior students, on the other hand, reported less deep studying and lower PGSD when they perceived the assessment criteria as more

[1] Cohen's f is provided as a measure of effect size, with $f = .10$, $f = .25$, and $f = .40$ corresponding to a small, medium, and large effect, respectively (Cohen, 1988).

Table 1
Mean (and SD) perceptions of authenticity and alignment, study approach and PGSD of senior and freshman students

	Senior students ($n = 118$)	Freshman students ($n = 81$)
Task	3.10 (.77)	3.33 (.61)
Physical context	2.53 (.92)	2.99 (.86)
Form	3.31 (.74)	3.41 (.67)
Criteria	3.20 (.62)	3.23 (.46)
Alignment	3.41 (.74)	3.28 (.63)
Surface study approach	2.64 (.51)	2.77 (.50)
Deep study approach	2.85 (.64)	2.93 (.57)
PGSD	2.81 (.61)	3.10 (.59)

PGSD, perception of generic skill development.

authentic. There were no other parameters that differed between groups (no other modification indexes were significant).

In short, the two groups differed in the perception of the physical context of the assessment and their PGSD. In addition, the quantitative results hinted a small difference in perception of the authenticity of the task. On the other hand, the hypothesised model regarding the relationships between perceptions of assessment authenticity, study approaches, and PGSD was in the main confirmed for both student groups, suggesting that the effect of the perception of authenticity on study approach and perceived learning outcome is fairly stable. Students in both groups, who perceived the task and the physical context as more authentic, reported deeper studying and, through it, more PGSD than students who perceived these assessment characteristics as less authentic. The effect of form on PGSD, however, was direct and not through DSA. Finally, the influence of perception of criteria authenticity on deep studying was very strong for both groups, but in opposite directions. For freshman students, an increased perception of criterion authenticity resulted in an increase in DSA and PGSD, while, contrary to the expectations, this relationship was reversed for senior students. Finally, perception of alignment had neither direct nor indirect effect on PGSD.

4.3. Qualitative results

The focus-group data were used to explain the differences found between the two student groups. Illustrative quotations from both freshman and senior students are given to support the agreement or dissimilarity between the student groups. (Fragment numbers combined with a ''f'' for freshman and ''s'' for senior focus group and the interview number are given in parentheses.)

4.3.1. Perceptions of task and physical context authenticity

The authenticity of the assessment task referred to the degree to which the content of the assessment resembled activities of professional practice. All student groups agreed that the assessment task referred to activities and problems that students encounter in their internships:

Table 2
Goodness-of-fit statistics for tests of invariance across freshman- and senior-student groups

Model description	Comparative model	χ^2	*df*	$\Delta\chi^2$	Δdf	*p*
Theoretical model for 118 senior students		7.71	6			
Theoretical model for 81 freshman students		1.52	6			
Combined baseline model 118 + 81 (Model 1)		9.22	12			
All relations constrained equal	Model 1	51.04	33	41.82	21	<.05
Factor loadings constrained equal (Model 2)	Model 1	42.50	24	33.50	12	<.05
Factor loading constrained equal except Criteria → DSA	Model 2	20.63	23	21.81	1	<.05

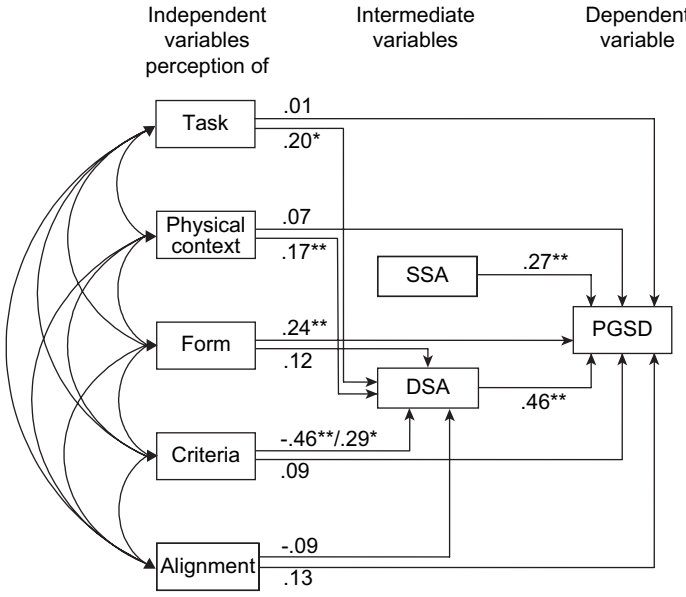

Fig. 2. The structural model with unstandardised path coefficients (DSA = deep study approach; SSA = surface study approach; PGSD = perception of generic skill development). *$p < .05$, **$p < .01$.

> The cases are realistic; they are real. Look, we are social workers and these cases are really directed at the activities of social workers (73, s4).
>
> *Or*
>
> [are the cases realistic?] Yes definitely, you can encounter them anytime at your workplace (56, f2).

A difference that was found was that the senior groups favoured assessment tasks that they could personalise to fit with their own interests or working context, or assessment tasks that dealt with extreme, more specialised cases instead of the general social work activities reflected in the current assessment. Seniors said that they would learn more from this kind of assessment tasks and would be more motivated to learn from them:

> I think it would be better if you can choose your own tasks (…) I think you learn more from situations that you really have problems with (59, s5)
>
> *Or*
>
> [in the case of dealing with unwanted behaviour] if the case would deal with real, hard aggressive situations, then I would learn more from it or be more willing to study for it. For example, that you have to deal with a really aggressive client who is throwing chairs, instead of someone who is writing in his agenda while you are speaking, which is the kind of unwanted behaviour you get in the assessment (118, s3).

With respect to the physical context, both freshman and senior students agreed that they would respond differently in real professional practice than in a role-play. However, freshman felt that role-play assessments were appropriate for assessing job performance:

> It assesses how you would respond in a certain situation, when it would take place in practice (8, f1).
>
> *Or*
>
> Instead of writing things down on a piece of paper, a role-play resembles a real situation and asks you to bring it [theory] in practice (61, f1).

Seniors, on the other hand, felt that assessing with role-plays in school is redundant:

> I think that they should not do such an assessment; they just need to come to our workplace (88, s4).
>
> *Or*
>
> We already have a lot of work and life experience that you think: 'do we really have to do this in school?' For us it often feels as if we are making up things [role-plays] just to stay busy (108, s3).

4.3.2. Authentic criteria and student learning

In both groups, the criteria had the greatest influence on student learning (see Fig. 2), but the relationships were opposite. When asked ''What determines your study activities for the assessment and your behaviour during the assessment?'' All student groups agreed that this were the criteria:

> You just have to perform all the steps that are described in the criteria. In this case, regarding a Jehovah's Witness, it meant that you did not have to know actually anything about what a Jehovah's Witness is (51, f2).
>
> *Or*
>
> You get that list with criteria and you are completely focused on performing these points, otherwise you will fail (15, s3).

Also, both groups agreed that the criteria were realistic, in essence and on a broad level:

> I think they are good points. At my workplace they also pay attention to these points (41, f1).

However, the concretisation of the criteria was too specific, theoretical, or based on schoolbooks or rules:

> Theoretically speaking the criteria are really good but they do not always work like that in practice (16, s5).
>
> In the assessment, I have to meet all kinds of school rules like explicitly appointing behaviour-feelings consequences, while in practice I will not always explicitly mention how I feel or what the consequences are (61, f2).
>
> *Or*
>
> You use them [the criteria] in practice all the time, but still it is different from how you've learnt it here in school. You use them, but you just translate them in their [your clients] language (23, s5).

Differences were found in how the realistic but specific criteria influenced freshman or senior student learning. In short, it seemed as if freshman felt more comfortable with specific criteria, because this gave guidance and structure to their learning, whereas seniors felt that the criteria were appropriate on a general level, but the specification inhibited them to respond naturally. Freshman students said:

> I think this description of the criteria is logical, otherwise you don't know, I mean, they are just ''helping points'' like: there are three causes mentioned in the case, so you have to appoint three causes in your role-play, I think that is a good thing (40, f1).
>
> *Or*
>
> I think that the criteria are realistic because the conversation has to be structured of course. You have to receive guidelines like ''you have to say something about this and about that'' (92–93, f1).

Seniors, on the other hand, argued:

> The big difference between the assessment and practice is that in the assessment you have to show all the criteria in 10 minutes, so you are focussing on ticking of all the criteria, while in practice, these criteria will come naturally during a conversation of half an hour (113, s3).
>
> *Or*
>
> In practice you will also get to the point where everything is ok again and the client is satisfied, but you just get there in another way then you have to do here [at school] (9, s4).

Moreover, seniors explicitly referred to the fact that having more experience in professional practice changed how students dealt with assessments:

> The more time you've been doing internships, in practice, the more experience you get and the better you possess all the skills and that is good for the assessments (177, s5).
>
> *Or*
>
> The assessments deal with skills that we already possess, but I think that younger students cannot automatically carry out the assessments to a successful conclusion, that is a difference with us (128, s3).

5. Discussion

This study compared two student groups from a VET College for Social Work, who differed in their practical experience. It studied the effect of this practical experience on students' perception of assessment authenticity and alignment, and on the relations between perceptions of assessment authenticity, study approaches, and PGSD.

Overall, the results showed that when students, both freshmen and seniors, perceive an assessment as more resembling their future professional practice (i.e., as authentic), they are stimulated to more deep studying and more generic skill development. This means that Hypothesis 2 was confirmed. In terms of practical implications, this supports the use of authentic assessments during a VET curriculum and also suggests that it is unnecessary to develop completely different kinds of authentic assessments for freshman and senior students.

However, the MANOVA and SEM analysis showed some salient differences between freshman and senior students. Hypothesis 1 was partly confirmed, namely freshman and senior students differed in the perception of the physical context and in their degree of PGSD. Combined with the qualitative results, Hypothesis 1 also tended to be confirmed for perception of the assessment task. Hypothesis 3 was only confirmed for the relationship between perception of criterion authenticity and a deep study approach, which was the only relationship that differed in the two groups. Focus-group interviews were conducted to find out if these differences could be explained by the fact that the two student groups differed in their degree of practical experience. First, senior students referred explicitly to the fact that having more experience in practice made them more skilled to naturally deal with assessments of job performance. Specific assessment criteria (even though perceived as being authentic realistic steps taken in practice) inhibited seniors from performing the assessment naturally. It appears that seniors no longer need these analytic steps to successfully perform the assessment task. This was supported by Govaerts, van der Vleuten, Schuwirth, and Muijtjens (2005) who found that senior students became demotivated by analytic criteria, while freshmen needed analytic criteria to guide their learning. The ''expertise reversal effect'' (Kalyuga, Ayres, Chandler, & Sweller, 2003) argued that once students have gained expertise, they need less instructional guidance, because they have internalised the information. As a result of more experience with performing the assessment tasks in practice, seniors might perceive performing in practice differently (i.e., in a more integrated instead of a step-by-step way) than freshman students. Instructional guidance, in this case analytic performance criteria, becomes redundant for more experienced students and no longer contributes to their learning or even hinders it (Kalyuga et al., 2003).

In addition, the qualitative results illustrated that seniors experienced the tasks as focussing too highly on general social-work activities. Seniors preferred more specialised instances, because they would learn more from them. It appears that the general tasks were only 'more of the same' for the more experienced senior students. Furthermore, seniors perceived the physical context, situated in an in-school role-play with a teacher, as redundant because of their experience on the work floor. They already performed these tasks in professional practice, which made them perceive performing these tasks in a role-play in school as less authentic. These results concerning the authenticity of the assessment task and physical context suggest that senior students already had experience with performing the tasks used in the assessments in professional practice. This experience possibly explains their feeling that they can perform the tasks in the assessments naturally, based on their experience. A combination of senior students' perceptions of the low authenticity of the assessment task, physical context and criteria might explain why they reported developing fewer professional skills than freshmen in response to the assessment; seniors simply felt that there was less to learn because of their previous experiences in professional practice.

In short, when students (i.e., freshmen as well as seniors) perceive an assessment as more authentic, they report studying more deeply and developing more professional skills. What students perceive as authentic depends on

how they perceive professional practice and performance in professional practice (Lizzio & Wilson, 2004a; Messick, 1994). In addition, this perception of what professional practice and performing in practice is can change when students gain more practical experience. When students do not perceive the assessment as appropriately reflecting professional practice, even if the assessment was developed to be authentic, it might not support, or even hamper, their learning. Taken together, this might mean that students with varying degrees of professional practice benefit more from different kinds of authentic assessment.

Finally, the introduction argued for the importance of perceived alignment between instruction and assessment for an assessment to be effective. In this study, students were, in general, positive about the alignment (see Table 1), but the results showed that perception of alignment in itself did not have an effect on study approach or PGSD. A possible explanation could be that perceived alignment is a precondition for letting the other positive effects of the assessment to become visible. This explanation is corroborated by previous studies (e.g., Segers et al., 2001) that only found negative effects of the assessment on study approach when students perceived instruction and assessment to be *not* in alignment.

5.1. Practical implications

For educational practice, at least for vocational types of education in which learning and working are intertwined, our findings imply that using authentic assessment is useful and effective during a competency-based curriculum, but some critical issues need to be considered in its operationalisation for students with differing practical experience. Based on this study, students with more practical experience might learn more from assessments when (a) they have holistic criteria that reflect how these students perceive performing in professional practice; (b) they use assessment tasks that reflect more specialised (out of the ordinary) professional activities instead of general professional activities practice or assessment tasks that allow students to tailor the tasks to their personal interests or working context; and (c) they are situated in real professional practice instead of in a role-play in school. Students with little practical experience, on the other hand, prefer (a) more analytic criteria because performing in practice is also still a stepwise process for them and specific steps help them learn; (b) they are satisfied with assessment tasks that reflect more general professional activities, and (c) for them assessing in the workplace is not absolutely necessary, since assessing in a simulated or role-play setting can appropriate reflect performing in practice.

5.2. Limitations

The present study and its implications are of relevance for schools where the goal of the assessments is to stimulate professional skill development and where learning and working are strongly integrated. With respect to the generalisability of the results of this study, three aspects need to be taken into account. First, VET in various countries is not necessarily using internship regimes as in the Netherlands. This might make it difficult to generalise outside of the VET context in the Netherlands. Moreover, this study examined the difference between students with little and much experience in professional practice and not between freshman and senior students in general. The results of this study cannot be transferred to freshman and senior college students or higher education students, where learning and working are much less integrated. Second, this study was done using one specific type of authentic assessment, namely role-play. The results might be transferable to, for example, patient simulations in nursing or medicine, but not to completely other types of (authentic) assessments. Therefore, this kind of study should be replicated in other domains, with other student groups and with other kinds of authentic assessments.

In interpreting the results of this study it should be taken into account that because senior students have more experience in professional practice, they also have more experience with assessments at the work floor, while freshman students only have experienced authentic assessments in school. Not only previous experiences with professional practice, but also previous assessment experience might influence how students perceive an authentic assessment and what kind of authentic assessment they need to stimulate their learning.

The learning outcome measure used in this study was students' perception of their generic skill development. It would be valuable to add a more objective measurement such as the achievement score based on the assessment criteria and rated by assessors. This was not possible in this study, because the achievement score was only reported as a succeed/fail decision. This is a dichotomous variable, with little variance, that cannot be used in these kinds of quantitative studies.

5.3. Future research

Future research should study authenticity perceptions and their effect on study approach and professional skill development in educational contexts other than VET, to get a grip on the most effective operationalisations of authentic assessments in different contexts. Interesting contexts are types of education where learning and working are not as integrated or where the future work field is much broader and therefore less clear. Professional development and assessment literature (Boshuizen et al., 2004; Kasworm & Marienau, 1997; Segers, Dochy, & Cascallar, 2003) advocate the use of authentic learning tasks or assessments early in the educational trajectory. However, previous studies suggested that students with no practical experience might have unrealistic perceptions of professional practice (Lizzio & Wilson, 2004b; Pena, 1997). The role of authentic assessment in this phase of an educational career might well be to help students create a more realistic idea of professional practice. The question then would be what kind of authentic assessment can give inexperienced students a realistic preview of professional practice. Future research should examine how the authenticity should be operationalised to stimulate the learning of the inexperienced students.

By comparing student perceptions about the authenticity of the same kind of authentic assessment, both within one group and between groups, this study gave indications about the important elements of assessment authenticity and how these should be operationalised to be effective for different student groups. The next step should be to compare assessments that do or do not take these elements into account and examine their impact on student learning and professional skill development. How do students perceive the authenticity of these assessments? What kind of study activities do they employ in response to the different assessments? The problem with this kind of studies, and especially with conducting them in an ecologically valid setting, is that this often requires using and comparing different student groups, most likely from different schools, because one school probably does not have various versions of assessment. Since the impact of assessments is dependent on the whole learning environment (e.g., Biggs, 1996; Struyven, 2005) it might be difficult to distil the impact of the assessment. Thoroughly describing the research/school context and using qualitative data to examine the influence of a complex mix of factors is imperative in these cases (Birenbaum, 2003).

The present study showed that perceived authenticity is an important element of new modes of assessment aiming at developing competencies or assessing job performance. Even though authenticity is not the only criterion for valid assessment (Baartman, Bastiaens, Kirschner, & van der Vleuten, 2006; Dierick & Dochy, 2001), we argue that an assessment that is perceived as authentic by students is an important step in the direction of bridging the gap between learning and working. Additionally, this study helps build this bridge throughout a curriculum in which learning and working are intertwined, by giving insight and guidelines for using authentic assessments that are perceived as authentic by students with different degrees of experience in professional practice.

Acknowledgments

We thank Frans Bleumer, Lisan van Beurden, Marja van de Broek and the students from the Baronie College who made it possible for us to conduct this study.

References

Baartman, L. K. J., Bastiaens, Th. J., Kirschner, P. A., & van der Vleuten, C. P. M. (2006). The wheel of competency assessment: presenting quality criteria for competency assessment programmes. *Studies in Educational Evaluation, 32*, 153–170.

Biemans, H., Nieuwenhuis, L., Poell, R., Mulder, M., & Wesselink, R. (2004). Competence-based VET in the Netherlands: background and pitfalls. *Journal of Vocational Education and Training, 56*, 523–538.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*, 347–364.

Biggs, J. B. (1987). *The Study Process Questionnaire Manual*. Melbourne, Australia: Australian Council for Educational Research.

Biggs, J., Kember, D., & Leung, D. Y. P. (2001). The revised Two-Factor Study Process Questionnaire: R-SPQ-2F. *British Journal of Educational Psychology, 71*, 133–149.

Birenbaum, M. (1996). Assessment 2000: Towards a pluralistic approach to assessment. In M. Birenbaum, & F. J. R. C. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. 3–29). Boston: Kluwer.

Birenbaum, M. (2003). New insights into learning and teaching and the implications for asessment. In M. Segers, F. J. R. C. Dochy, & E. Cascallar (Eds.), *Optimising new modes of assessment: In search of qualities and standard* (pp. 13–36). Dordrecht, The Netherlands: Kluwer.

Boshuizen, H. P. A., Bromme, R., & Gruber, H. (2004). *Professional learning: Gaps and transitions on the way from novice to expert*. Dordrecht, The Netherlands: Kluwer.

Boud, D. (1990). Assessment and the promotion of academic values. *Studies in Higher Education, 15*(1), 101–111.

Boud, D. (1995). Assessment and learning: contradictory or complementary? In P. Knight (Ed.), *Assessment for learning in higher education* (pp. 35–48). London, England: Kogan Page.

Byrne, B. M. (2001). *Structural equation modeling with AMOS. Basic concepts, applications and programming*. Mahwah, NJ: Erlbaum.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Dierick, S., & Dochy, F. (2001). New lines in edumetrics: new forms of assessment lead to new assessment criteria. *Studies in Educational Evaluation, 27*(4), 307–329.

Dochy, F. J. R. C., & McDowell, L. (1998). Assessment as a tool for learning. *Studies in Educational Evaluation, 23*(4), 279–298.

Entwistle, N. J. (1991). Approaches to learning and perceptions of the learning environment. Introduction to the special issue. *Higher Education, 22*, 201–204.

Field, A. P. (2000). *Discovering statistics using SPSS for Windows: Advanced techniques for the beginner*. London: Sage.

Frederiksen, N. (1984). The real test bias, influences of testing on teaching and learning. *American Psychologist, 39*(3), 193–202.

Gielen, S., Dochy, F., & Dierick, S. (2003). Evaluating the consequential validity of new modes of assessment: the influence of assessment on learning, including the pre-, post-, and true assessment effects. In M. Segers, F. Dochy, & E. Cascallar (Eds.), *Optimising new modes of assessment: In search of quality and standards* (pp. 37–54). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Govaerts, M. J. B., van der Vleuten, C. P. M., Schuwirth, L. W. T., & Muijtjens, A. M. M. (2005). The use of observational diaries in in-training evaluation: student perceptions. *Advances in Health Sciences Education, 10*(3), 171–188.

Gulikers, J., Bastiaens, Th., & Kirschner, P. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development, 52*(3), 67–85.

Gulikers, J. T. M., Bastiaens, Th. J., & Kirschner, P. A. (2006). Authentic assessment, student and teacher perceptions: the practical value of the five-dimensional framework. *Journal of Vocational Education and Training, 58*, 337–357.

Handal, G., & Hofgaard Lycke, K. (2005, August). From higher education to professional practice: Learning among students and novice professionals. Paper presented at the 11th biannual meeting of the European Association for Research on Learning and Instruction, Nicosia, Cyprus.

Herrington, J., & Herrington, A. (1998). Authentic assessment and multimedia: how university students respond to a model of authentic assessment. *Higher Educational Research & Development, 17*(3), 305–322.

Honebein, P. C., Duffy, T. M., & Fishman, B. J. (1993). Constructivism and the design of learning environments: context and authentic activities for learning. In T. M. Duffy, J. Lowyck, & D. H. Jonassen (Eds.), *Designing environments for constructive learning* (pp. 88–108). Berlin: Springer.

Joreskog, K. G. (1993). Testing structural equation modeling. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 294–316). Newbury Park, CA: Sage.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*, 23–32.

Kasworm, C. E., & Marienau, C. A. (1997). Principles for assessment of adult learning. *New Directions of Adult and Continuing Education, 75*, 5–16.

Klarus, R. (2000). Beoordeling en toetsing in het nieuwe onderwijsconcept. [Evaluation and assessment in the new educational philosophy]. In J. Onstenk (Ed.), *Op zoek naar een krachtige beroepsgerichte leeromgeving. Fundamenten voor een onderwijsconcept voor de bve-sector* (pp. 12–27). 's Hertogenbosch, The Netherlands: Cinop.

Lizzio, A., & Wilson, K. (2004a). First-year students' perceptions of capability. *Studies in Higher Education, 29*(1), 109–128.

Lizzio, A., & Wilson, K. (2004b). Action learning in higher education: an investigation of its potential to develop professional capability. *Studies in Higher Education, 29*, 469–488.

Lizzio, A., Wilson, K., & Simons, R. (2002). University students' perceptions of the learning environment and academic outcomes: implications for theory and practice. *Studies in Higher Education, 27*, 27–51.

Martens, R., Gulikers, J., & Bastiaens, Th. (2004). The impact of intrinsic motivation in e-learning with authentic computer tasks. *Journal of Computer Assisted Learning, 20*, 368–376.

McDowell, L. (1995). The impact of innovative assessment on student learning. *Innovations in Education and Training International, 32*(4), 302–313.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Pena, E. (1997). Great expectations: the reality of the workplace. *Australian Journal of Career Development, 6*, 25–32.

Petraglia, J. (1998). *Reality by design: The rhetoric and technology of authenticity in education*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Sambell, K., McDowell, L., & Brown, S. (1997). But is it fair? An exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation, 23*(4), 349–371.

Savery, J. R., & Duffy, T. M. (1995). Problem based learning: An instructional model and its constructivist framework. In B. G. Wilson (Ed.), *Constructivist learning environments: Case studies in instructional design* (pp. 135–150). Englewood Cliffs, NJ: Educational Technology Publications.

Scouller, K. M. (1997). Students' perceptions of three assessment methods: assignment essay, multiple choice question examination, short answer examination. *Research and Development in Higher Education, 20*, 646–653.

Segers, M., Dierick, S., & Dochy, F. (2001). Quality standards for new modes of assessment. An exploratory study of the consequential validity of the OverAll test. *European Journal of Psychology of Education, 16*(4), 569–586.

Segers, M., Dochy, F., & Cascallar, E. (Eds.). (2003). *Optimising new modes of assessment: In search of qualities and standards*. Dordrecht, The Netherlands: Kluwer.

Segers, M. S. R. (2004). Assessment en leren als een twee-eenheid: Onderzoek naar de impact van assessment op leren [Assessment and learning as twofoldness: research on the impact of assessment on learning]. *Tijdschrift voor Hoger Onderwijs, 22*(4), 188–220.

Struyven, K. (2005). The effects of student-activated teaching/learning environments of students' perceptions, student performance and pre-service teachers' teaching. Unpublished doctoral dissertation, University of Leuven, Belgium.

Tillema, H. H., Kessels, J. W. M., & Meijers, F. (2000). Competencies as building blocks for integrating assessment with instruction in vocational education: a case from the Netherlands. *Assessment and Evaluation in Higher Education, 25*(3), 265–278.

Van Rossum, E. J., & Schenk, S. M. (1984). The relationship between learning conceptions, study strategies and learning outcome. *British Journal of Educational Psychology, 54*(1), 73–83.

Wilson, K. L., Lizzio, A., & Ramsden, P. (1997). The development, validation and application of the course experience questionnaire. *Studies in Higher Education, 22*(1), 33–53.

Winning, T., Elaine, L., & Townsend, G. (2005). Student experiences of assessment in two problem-based dental curricula: Adelaide and Dublin. *Assessment and Evaluation in Higher Education, 30*(5), 489–505.