Alternative assessment: Self-assessment beyond the mainstream

Anne-Mieke Janssen- van Dieten
University of Nijmegen

Abstract

The majority of language testing research concentrates on target groups with high levels of education. In this article the focus of attention is on adult migrants with a low educational level. It describes an experiment into the effect of training upon the quality of self-assessments, which is thought to enhance the efficiency of self-directed learning. The function of self-assessment in education is discussed. The article concludes that self-assessment is a very worthwhile assessment procedure, provided it is undertaken to enhance learning and is solidly embedded in an educational approach which is consistent with the underlying principles of self-direction and a reflective way of learning.

1. Introduction

One of the characteristics of alternative assessment procedures is that they '....set expectations that are appropriate within the cognitive, social and academic development of the learner' , and '.... allow for a more valid interpretation of information than that obtained from more traditional standardized tests', which make them .... 'particularly valuable for second language learners who come from culturally diverse backgrounds and who may have atypical educational experiences' (Hamayan, 1995: 215). Self-assessment is often mentioned as one of those procedures, along with portfolios or observation (e.g. Shohamy, 1995; Norris, Brown, Hudson & Yoshioka, 1998). Self-assessment, however, is not more valid simply because it is an alternative procedure. It seems worthwhile to further explore the function and applicability of self-assessment in educational settings, in particular in language courses for adult migrants with a low level of previous education. This will be done on the basis of findings of a research project on self-assessment that was carried out

Melbourne Papers in Language Testing 2000 Volume 9.2 pp. 18-29. The Language Testing Research Centre, The University of Melbourne.

between 1987 and 1992 with adult migrant learners of Dutch. One study in that project involved the effect of training.

What is self-assessment?

Self-assessment is providing judgement on your own language behavior or learning behavior based upon reflection upon that behavior. The judgement, however, can be based on direct observation of either simulated or real-life performance, or on introspection. It makes a difference whether the stimulus concerns performance in a real life situation, in a simulated situation or in an imaginary one, although all of them may reflect a 'real' performance context. Therefore we have to be specific about what we mean by self-assessment.

Why self-assessment?

Arguments for the application of self-assessment can be divided into two main categories:

- practical considerations;
- considerations in which the learner and the learning process play a dominant role.

Practical considerations

The practical advantages of self-assessment are that it is less expensive and less time consuming than performance testing, that it is possible to get information about language activities that cannot easily be tested, and that it is easier to adjust to the specific needs of different learner groups. In this category of arguments, self-assessment is viewed as a practical substitute for testing. That is why most of the research in this field has concurrent criterion validity as its main object. Most of the research in this domain is based on introspection elicited via questionnaires. In a nutshell, the results can be summarized by saying that the internal consistency coefficient, if reported at all, is high. If reliability is high, this tells us that respondents have a high degree of consensus in their perception of the order of difficulty of tasks and that their self-assessments were seriously carried out. In order to establish concurrent validity, self-assessment scores have to be compared with an external criterion. This is generally done by correlation or accuracy measures, which have yielded a variety of outcomes, ranging from very low to very high correlation coefficients or proportions of agreement. The majority can be characterized as moderately high (Oskarsson, 1984; Blanche & Merino, 1989; Oscarson, 1997) and both overestimation and underestimation occur, but the former tends to be more common (Janssen-van Dieten, 1989, 1992). It could be argued, of course, that the agreement between self-assessment and external criteria is not the purpose at all of using self-assessment as a source of information about language proficiency. This is a legitimate and attractive point of view, which fully reflects Holec's (1991) opinion that the learner is autonomous in using his/her own criteria and which fits in an autonomous educational setting, where the responsibility for learning and learning outcomes lies with the learner. In a setting in which results have to be reported to external bodies, however, the question of how to interpret the outcomes remains to be answered.

Learning process oriented considerations

The second category of arguments deals with the learner and the learning process. Among them we find the development of a reflective attitude, gaining insight into evaluation criteria and stimulation of goal orientation, task analysis, diagnosis and remedial follow-up. In other words, self-assessment is regarded as a tool to enhance the learning process and its only, but very important, function in education is a formative one. In this respect, it has been argued that the quality of the self-assessments, which means the extent of agreement with the judgements of experts, will benefit the efficiency of the learning process (Painchaud & LeBlanc, 1984; Janssen-van Dieten, 1992). The question of whether it is possible to improve the quality of self-assessment was the central issue in the field experiment that is described in the following section.

2. The experiment

Background

In 1985 the European Commission commissioned the Nijmegen Department of Applied Linguistics and CITO, the Dutch Institute for Test Development to develop a test battery to be used in adult migrant language education, for a target group preparing for entrance to lower level vocational training. Inspired by the successful Swedish application of self-assessment (von Elek, 1985), we developed both a test and a self-assessment version of the test. We will not present the results in any detail (for a more detailed discussion see: Janssen-van Dieten 1989; 1992) but some of the findings are interesting in this context. In both test administrations we found low to moderate correlations and accuracy indices. We did not find any relationships between the accuracy of self-assessment and background variables such as gender, age, length of residence, western or non-western country of origin and previous education. What was rather alarming, in my opinion, was that there was no relationship at all between instruction in L2 Dutch and accuracy of self-assessment. What was even more alarming was that the self-assessment version, which was intended to stimulate a more process-oriented approach, appeared not to be used at all. When teachers have a choice between self-assessment or tests, they seem to prefer tests because they believe that self-assessment is a very 'unreliable' substitute for a test. The first finding was alarming because language instruction did not appear to improve insight in the learner's language proficiency and it did not lead to a better insight into evaluation criteria. The second finding points to a clear misinterpretation of the intended and repeatedly emphasized formative function of the self-assessment version. That is why it was decided to carry out an experiment with self-assessment as the core of a training program. Furthermore, the experiment was carried out deliberately in schools for Basic Adult Education (Basiseducatie). In the Netherlands this term refers to education for people who have enjoyed a maximum of eight years of previous education. The reasons for doing this were manifold. One of them was that current descriptors of language levels, on which tests are based, do not seem to do justice to people with a low level of education. Quality descriptors go hand in hand with task descriptors; the more cognitively complex the task the better the performance. It is very difficult for people with primary school as the highest level of education to show progress in second language mastery on tasks that are designed for college level students (Janssen-van Dieten, 1997). The most important reason for carrying out the experiment in basic education, however, was the belief, frequently expressed by teachers, that self-assessment requires a level of reflection that is beyond the reach of learners with a low level of education.

The study

In the study it was hypothesized that:

1. training learners to reflect on language use and language learning would have a positive effect on the quality of the subjects' self-assessment of their proficiency;
2. this training would have a positive effect on their proficiency.

The experiment was carried out in three different schools. In each school two parallel classes were involved, one as the control group and one as the experimental group. We started out with 96 subjects. Unfortunately the statistical mortality rate was very high, due to the fact that people were forced to take any job that came their way. The result was that we ended up with 38 subjects. This did not affect the internal validity, but it did reduce the power of the experiment. The design was a pretest-posttest design with four months of training in between, which was limited to writing instruction for practical reasons. During training, self-assessment was the starting point for different kinds of learning activities, all of them considered to be conducive to a reflective and more self-directed learning process (Janssen-van Dieten, 1992; 1993). The intended self-assessment activities were twofold: classroom activities and real life activities. The teachers in the experimental group got a two-day training, in which the focus was on the principles of 'learning how to learn' and in which they were asked to, at least, carry out the following activities. In the classroom, along with each writing assignment, learners should hand in a self-assessment paper in which they gave a global judgement about the quality of the product in general and an analytical one, focussing mainly on linguistic features. The teacher should provide positive and informative feedback on the self-assessments. The written assignments should then be corrected by individual learners or in pairs. Learners also should indicate which aspects they wanted to work on, select exercises and comment on what they had learned from them.

The addition of self-assessment of real life activities was aimed at raising awareness of potential learning opportunities outside the classroom. Every week students handed in what we called 'learning charts', in which they commented on their real life Dutch language activities. They described the activities, assessed them, reported what they did when things went wrong and reflected on the usefulness of the strategies adopted. These charts were regularly discussed in class. Analyzing these charts was revealing and exciting. They provided insight into the frequency and nature of language contacts. With some exceptions these contacts appeared to be limited to more or less formal, obligatory contacts with doctors, schools, lawyers, hairdressers etc. Some students reported that the mere fact that they had to hand in a chart made them talk to neighbors or watch Dutch television. The exciting part was to see how well they managed to notice aspects that enhanced or impeded communication and how some of them progressed in using and reporting strategies (Janssen-van Dieten, 1992).

The self-assessment procedure used in the training program was direct observation of either simulated tasks in classroom or real life performance outside school. For practical reasons, the procedure used in the pre- and posttest was introspection. Before they did the test, which consisted of 40 short tasks (level 2 and 3 of the test described in Janssen-van Dieten, 1989), the learners indicated whether they believed themselves to be able to complete them in an acceptable way, by answering: 'Yes, I can', No, I can not' or '?', meaning 'I am not quite sure'.

During the training period, both the control and experimental groups were observed and the results led us to hypothesize that one experimental sub-group would outperform the other two, with regard to progress in the quality of self-assessment. The hypothesis with regard to progress in proficiency was left out because the group in question had only one third of the instruction time of the majority of the control group with which they had to be compared. Expecting more progress in language proficiency from a group with 3 hours instruction time than from a group with an average of 9 hours instruction time would have been too optimistic a prediction.

Results

The results did not confirm the original hypothesis, concerning the superiority of the experimental group as a whole, but they did confirm the readjusted expectations after observation of what happened in class. There was no significant effect of training for the experimental group as a whole (Table 1).

group	*n*	mean diff.	*sd*	*t*-value	*df*	*p* one-tailed
experimental	18	2.89	5.04			
				1.00	33	.16
control	17	1.41	3.48			

Table 1. Mean gain scores in number of accurate assessments in experimental group and control group and statistical test of difference

Observations revealed that the differences in both the quality and the quantity of the training in the experimental sub-groups were so great that we expected one group, subgroup 3, to outperform subgroups 1 and 2. This expectation was confirmed. Contrasting the experimental sub-groups individually with the total control group showed that group 3 was the only group whose gain scores were significantly higher at a 5 percent level, with an average positive gain score of 5.43. We may conclude, therefore, that training can have a positive effect on the quality of self-assessment, provided it is conducted in a proper way.

Further qualitative analyses showed that shifts in the self-assessment behavior of the experimental group were more positive than those in the control group. The most striking shift was the decrease of uncertainty, particularly in group 3. Moreover, it appeared that the improvements in group 3 were very evenly spread over four separate sub-skills of writing, and that this group made significantly more progress in the assessment of performance at the textual level.

As far as progress in writing proficiency is concerned, the expected superiority of the experimental group could likewise not be confirmed (Table 2). Here, too, gain scores came up to expectations, but the difference with the control group was not significant.

group	*n*	mean diff.	*sd*	*t*-value	*df**	*p* one -tailed
experimental	18	4.61	5.30	1.25	24	.11
control	17	2.88	2.47			

* number of degrees of freedom readjusted for uneven range

Table 2. Mean gain scores writing of the experimental group and the control group and test of difference

It is remarkable, however, that every single experimental sub-group made more progress than any of the control sub-groups, in spite of the great differences in instruction time between sub-groups.

Discussion

As the expected results were not achieved, at least not in all respects, we considered several factors that may have affected the findings. Since we did obtain positive results in subgroup 3, we explored the factors that could have played a part in the success of their training in more detail. We did this in cooperation with the teachers, who evaluated their own experiences. We hypothesized that the main factors were teacher-dependent, the most important one being a teacher's strong faith in her learners' abilities to direct their own learning process. Such conviction promotes teachers' willingness to change their teaching style and to act in accordance with the underlying principles. For example, teachers 1 and 2 asked learners what their needs were, but there was no follow-up. They asked learners to assess themselves but feedback consisted mainly of error correction. Learners in sub-groups 1 and 2 had no choice. It was the teacher who decided on home work and class activities. In other words, there was reflection, but without any consequences. Learners did not get the opportunity to develop a feeling of personal causation. Teachers 1 and 2 could be characterized as acting from a technical cognitive interest, in which activities were central, whereas teacher 3 seemed to be moved more by a practical interest, which means that the underlying idea of self-direction was the starting point for her actions (Grundy, 1987). Another important difference between group 3 on the one hand and groups 1 and 2 on the other was that in the school to which sub-group 3 belonged, there was a disciplined school climate in which students had mutual responsibility for learning. According to DeCharms, Plimpton, and Koenigs (1976) this seems to be a condition for the development of responsibility for one's own learning process.

From these findings it may be concluded that self-assessment is not simply an alternative testing procedure. A fruitful application is heavily dependent on the presence of a learning environment in which self-assessment is embedded.

3. Conclusion

The answer to the question of whether self-assessment is a useful assessment procedure is , in my opinion, a firm 'yes', accompanied by a strong 'but'. The crucial question is: 'Assessment for what purpose?'

In large-scale research projects or in other cases in which comparative information about language proficiency is needed and in which testing is impossible, self-assessment can be a practical substitute for testing (Oscarson, 1997). That, however, is not the issue of this discussion. We are not interested in some global indication as to whether group or person X seems to perform better than Y.

Since we are discussing alternative assessment, an umbrella term for different kinds of performance assessment, what we want to know is how students perform on tasks (either simulated or authentic) that reflect real life performance and that are representative of (future) needs and/or educational goals. There are two main reasons for assessing students: accountability assessment and educational assessment (Gipps, 1994).

Self-assessment and accountability

Let us assume that self-assessment in an authentic setting is used as one of the assessment procedures to account to external bodies. The idea, as such, of involving in evaluation the very person that experiences how it is to carry out a specific task, in order to counterbalance external judgment, seems to be a sound one. But is it as fair as it is believed to be? Apart from the ethics of asking someone to be honest when stakes are high, we have to deal with public acceptability. What are the consequences for the person in question if the results of other assessment procedures are not convergent with his or her self-assessment? If the divergence is a positive one on the part of self-assessment, I really fear that stakeholders will not only tend to ignore the self-assessment outcomes but will draw conclusions about the personality of that person, which were not intended to be reported on. A negative divergence may also lead to second thoughts about a person's eligibility. And what about the reporting of the judgment of external observers along with self-assessment on the same task? One student, from another study, who received job training in a shop, reported not to have any problems in dealing with clients. Later on she explained that she had problems every now and then, especially with names of specific items. But then she told the client she did not understand and asked a Dutch colleague in the presence of the client what the item meant and where it was, so that she would know next time. This was an effective strategy, so she did not see any problem at all. Would her boss, asked to judge her functioning, be of the same opinion? And whose opinion has most weight eventually, if the addition about strategy use is not included in the report? Like assessment, self-assessment requires reflection on and discussion about perspectives and criteria underlying the judgement.

Another problem is the comparability of tasks and the use and interpretation of criteria. The problem of generalizability across tasks in different contexts which are presumed to measure the same construct is a problem in all performance assessment, as is that of the interpretation of criteria by different, even thoroughly trained, raters. This is even more problematic for individuals who assess single cases and have no objective standards, apart from their own experience, for comparison. An example from my data about communication outside school may illustrate this. Two parents with children in the same class reported a conversation with the teacher about their children at a parents' evening. Both conversations involved the same teacher and both parents were apparently faced with a comparable task. One of them reported that she had no problem: she understood everything and did not experience any difficulties in communication apart from the fact that she did not know a few words, which however she was able to paraphrase. Her daughter was doing very well at school. The other parent reported difficulties in understanding the teacher and said that she was not able to say what she wanted to say. Her daughter had a lot of problems at school. Even if this background information is reported, which can be very threatening for a person who does not want those private things to be revealed, how should both assessments be interpreted? Finally, there is a practical problem. As the learning charts in our research revealed, learners' contacts with native speakers are scarce, at least for this particular population, and often limited to formal institutional contexts. In those contexts survival is so important that circumstances for optimal functioning are not favorable.

I think that it is for the benefit of the learner to be very careful with reporting self-assessment data.

Self-assessment for learning

On the other hand, when the aim is not accountability but measurement for learning, self-assessment can be a very useful procedure. In that case, however, it should not just be one of a set of alternative assessment procedures that are applied to provide a combination of data from different angles. It should be solidly embedded in an educational approach which is consistent with the underlying principles of self-direction and a reflective way of learning. Only then does it offer detailed feedback both to the teacher and the learner and is not threatening, because both parties profit from it. It provides teachers with a better insight into the learning process and into specific problems or needs of learners. Teachers in the project, for instance, were very surprised to discover that learners could be uncertain about utterances that were completely adequate and accurate. Learners, on the other hand, seldom receive positive feedback on such aspects they are uncertain about but which remain unnoticed if they are not far beyond what was expected. Self-assessment helps them build a realistic picture, which in turn helps them direct their learning more efficiently. As was indicated before, it was not easy for some teachers to really act in accordance with the underlying principles, and this was reflected in the attitudes of their class. The observed behavioral changes, especially in sub-group 3, the fact that, formerly very teacher-dependent, students started telling the teacher what they did and did not want to learn, set themselves tasks and told me, the observer from outside, to change the focus of attention in the learning charts, are more convincing than the figures reported.

With regard to the general belief among most teachers that the level of reflection required for self-assessment is beyond the reach of learners with a low level of education, which was one of the reasons for carrying out the experiment in basic education, we can cautiously say that that belief is not justified. The sub-group that did best, was the sub-group with the lowest average level of previous education.

In brief, self-assessment is a worthwhile assessment procedure when assessment is undertaken to enhance learning, but we should think twice before applying it for other purposes in which individuals are involved, and not groups.

References

Blanche, P. and Merino, B. 1989. Self-assessment of foreign-language skills: Implications for teachers and researchers. *Language Learning 39*, 3, 313–40.

DeCharms, R., Plimpton, F. and Koenigs, S. 1976. The origin classroom is different. In: R. DeCharms (ed.) *Enhancing motivation: Change in the classroom.* New York: Irvington Publisher Inc., 161–76.

Elek, T. von, 1985. A test of Swedish as a second language: An experiment in self-assessment. In Y. Lee, A. Fok, R. Lord and G. Low (eds.) *New directions in language testing*. Oxford: Pergamon Press, 47-57.

Gipps, C. 1994. *Beyond Testing: Towards a theory of educational assessment.* London: The Falmer Press.

Grundy, S. 1987. *Curriculum: Product or praxis.* Deakin Studies in Education Series, 1, London: The Falmer Press.

Hamayan, E. 1995. Approaches to alternative assessment. *Annual Review of Applied Linguistics 15*, 212–26.

Holec, H. 1991. Apprendre à l'apprenant à s' évaluer: Quelques pistes à suivres. *Études de Linguistique Appliquée 79*, 39–47.

Janssen-van Dieten, A. 1989. The development of a test of Dutch as a second language: The validity of self-assessment by inexperienced subjects. *Language Testing 6*, 1, 30–46.

Janssen-van Dieten, A. 1992. *Zelfbeoordeling en tweede-taalleren: Een empirisch onderzoek naar zelfbeoordeling bij volwassen leerders van het Nederlands*, doctoral thesis, KU Nijmegen.

Janssen-van Dieten, A. 1993. Self-assessment and adult second language learning. *Toegepaste Taalwetenschap in Artikelen 46/47,* 2/3, 131–139.

Janssen-van Dieten, A. 1997. Adult second-language policy in the Netherlands: Some considerations. In Th. Bongaerts and K. De Bot (eds.) *Perspectives on foreign-language policy, Studies in honour of Theo van Els.* Amsterdam: John Benjamins Publishing Company, 201–17.

Norris, J., Brown, J., Hudson, T. and Yoshioka, J. 1998. *Designing second language performance assessment.* Technical Report 18, Second Language Teaching & Curriculum Center, Manoa: University of Hawai'i.

Oscarson, M. 1997. Self-assessment of foreign and second language proficiency. In C. Clapham and D. Corson (eds.) *Language testing and assessment.* Encyclopedia of language and education, Volume 7, Dordrecht: Kluwer Academic Publishers, 175–87.

Oskarsson, M. 1984. *Self-assessment of foreign language skills: A survey of research and development work.* Strasbourg: Council of Europe.

Painchaud, G. and LeBlanc, R. 1984. L'auto-évaluation en contexte scolaire. *Études de Linguistique Appliquée 56,* 88–98.

Shohamy, E. 1995. Performance assessment in language testing. *Annual Review of Applied Linguistics 15,* 188–211.