Self-assessment in second language testing: a meta-analysis and analysis of experiential factors

Steven Ross *Kwansei Gakuin University*

Self-assessment has been used widely in language testing research, but has produced variable results. In many quarters self-assessment is considered a viable alternative to formal second language assessment for placement and criterion-referenced interpretations, although variation in self-assessment validity coefficients suggests potential difficulty in accurate interpretation. This article first summarizes the research literature with the use of a formal meta-analysis conducted on 60 correlations reported in the second language testing literature. These are the basis for estimates of median effect sizes for second language speaking, listening, reading and writing tests. The second phase of the study is an empirical analysis of the validity of a self-assessment instrument. 236 'just-instructed' English as a foreign language learners completed self-assessments of functional English skills derived from instructional materials and from general proficiency criteria. The learners' teachers also provided assessments of each of the 236 learners. The criterion variable was an achievement test written to assess mastery of the just-completed course materials. Contrastive multiple regression analyses revealed differential validities for self-assessment compared to teacher assessment depending on the extent of learners' experience with the language skill self-assessed.

It is with a cyclical regularity that the issue of self-assessment, also known as self-evaluation, finds its way into journals dealing with educational measurement and applied linguistics. While the issues related to self-assessment are those traditionally dealt with in measurement theory, that is, reliability and construct validity, there has been surprisingly little discussion of the value of self-assessment as an alternative to more expensive and logistically viable approaches to proficiency and achievement assessment, particularly in the area of second and foreign language testing. The scarce empirical work that has been done on self-assessment seems to be the byproduct of more complex studies of construct validity – the well-known multitrait-multimethod approaches used by Bachman and Palmer (1981; 1982), and more recently, by Buck (1992). This is perhaps not surprising since self-assessment has traditionally been viewed as antithetical to

Address for correspondence: Steven Ross, School of Policy Studies, Kwansei Gakuin University, 2–1 Gakuen, Sanda, Japan 669–13; e-mail: z95001@ksc.kwansei.ac.jp

Language Testing 1998 15 (1) 1–20 0265-5322(98)LT129OA

the concerns of educational measurement, and because, while there are some discernible advantages of using it, e.g., increasing student and teacher motivation, self-assessment often introduces more unwanted measurement facets than can be dealt with.

Of the few literature reviews dealing with self-assessment in language testing, the most comprehensive is that of Blanche and Merino (1989), which summarizes the sampling, methodology and criterion variables of several second and foreign language proficiency studies utilizing self-assessment. The Blanche and Merino review does an adequate job of introducing the issues related to self-assessment and a summary of the major findings, but suffers from the literature review approach the authors adopt. The 'big picture' of self-assessment is confined to the realm of subjective evaluations of tendencies:

> The emerging pattern is one of consistent overall agreement between self-assessments and ratings based on a variety of external criteria. The accuracy of most students' self-estimates often varies depending on the linguistic skills and materials involved in the evaluations, but these estimates are generally *good* or *very good* (Blanche and Merino, 1989: 315).

Their review provides the point of departure for this analytical summary of self-assessment in the current language testing literature. In contrast to the prose-based approach to summarizing findings, this summary will utilize meta-analysis (Johnson, 1989; Hunter and Schmidt, 1990). This approach provides a coherent empirical methodology for estimating average effect sizes and testing the homogeneity of findings (Rosenthal, 1984; Wolf, 1986), and in doing so, gives a cumulative picture of the state of knowledge about the research area of interest.

I Self-assessment in language learning

Starting with the most recent summary of studies involving self-assessment in language testing (Blanche and Merino, 1989), previous literature was accessed via bibliographic tracing and compact disk searches on the topic. For the purposes of this analysis, only studies that dealt with self-assessment/evaluation in second and foreign language testing were selected for the summary and meta-analysis. A larger and more diverse literature on the subject exists in the general field of educational measurement (e.g., Falchikov and Boud, 1989), but a coverage of the topic in general would not allow for the kind of summary needed for evaluating the status of self-assessment in second language testing. The criterion used for selection of articles for the analysis was that there was an empirical basis for evaluating the relationship between self-assessment and a second or foreign language criterion variable. By far the most common metric used in the

literature was the product-moment correlation. The major analytical goal of the meta-analysis therefore was the estimation of the average effect size for correlation coefficients, and the homogeneity of the findings across the literature sampled.

The most common approach to self-assessment in the second and foreign language literature reviewed involved correlating self-assessment scales with an outcome measure according to specific skill areas. The proclivity for authors to organize self-assessment studies according to different language skills is fortuitous, for it allows the meta-analyst to specify the relative strengths and weaknesses of the self-assessment approach across skill areas. Our preliminary analysis of the self-assessment literature, however, will be based on all correlational studies across four language skill areas. Subsequent analyses will focus on specific skills.

The data entered into the general self-assessment meta-analysis database came from ten studies of second language proficiency. The studies covered a wide range of second and foreign language contexts. The studies and their major characteristics are listed in Table 1.

Table 1 The studies and their major characteristics

Authors	Subjects	Languages	*n* size
Bachman and Palmer, 1981	Mixed	ESL	75
Bachman and Palmer, 1982	Chinese	ESL	116
Buck, 1992	Japanese	ESL	220
Janseen-van Dieten, 1989	Mixed	Dutch	33–116
Wongsotorn, 1981	Thai	ESL	22–27
Blanche, 1990	USA	French	32
LeBlanc and Painchaud, 1985	English	French	878
LeBlanc and Painchaud, 1985	French	English	861
Milleret, Stansfield and Mann-Kenyon, 1991	USA	Portuguese	11–12
Ferguson, 1978	Swiss	ESL	89

II Preliminary results

The combined meta-analysis summarizes the relationship between 60 correlations and various measures of second language proficiency. Although in some of the studies other aspects of second language proficiency were correlated with the self-assessment scales, e.g., pragmatic and sociolinguistic competence, for this analysis we will limit our scope to four second language skill areas – reading, speaking, listening and writing. Table 2 lists the summary of the meta-analysis of the 60 correlations. The effect size *g* is derived from standardized

Table 2 Effect sizes for 60 heterogeneous self-assessments

Study	*g*	95% CI	*r*	*p*	Deviation	Homogeneity
Overall	+1.6384	+1.60/+1.67	+0.6337	0.0001	0.639	–25.602

correlation coefficients. The average correlation for the 60 self-assessments is a robust 0.63, with less than one chance in one-hundred thousand that the observed effect would emerge serendipitously. The homogeneity statistic, however, indicates that there is considerable variation across the correlations sampled in the study. This comes as no surprise, since Table 2 encompasses all four skills in both foreign and second language contexts.

At the heart of the meta-analysis is the effect size coefficient. For coefficients of correlation, the most common metric used in establishing the validity of self-assessment with criterion variables in second/foreign language testing, the effect size in each study is estimated with coefficient *g*:

$$g = \frac{2r}{\sqrt{1 - r^2}} \tag{1}$$

The effect size *g* provides an index for the comparison of validity coefficients across studies, and thus provides an indication of the extent of the null hypothesis for the population (Cohen, 1988).

Following Rosenthal (1984) we will represent the distributions of the correlations as untransformed *r*s in a stem and leaf plot in order to examine their range and magnitude. The stem and leaf representation of the 60 correlation coefficients is as follows:

```
Minimum is              0.090
Lower hinge is          0.390
Median is               0.489
Upper hinge is          0.648
Maximum is              0.800

      0
      1
      1                 56
      2
      2                 55689
      3                 2
      3     H           5555999
      4                 02223
      4     M           5566778889
```

5		001123
5		66688
6	H	034
6		5566
7		01134
7		58889
8		0

The range of the self-assessment correlations suggests that there is considerable variation in the ability learners show in accurately estimating their own second language skills. There are many reasons why this may be so. The crafting of self-assessment scales requires considerable finesse, and they have to involve language skills that learners have had enough instruction or language contact to develop in order for them to give adequate self-evaluations. The across-skill summary above provides the boundaries of self-assessment validity, and in general it concurs with Blanche's (1990) summary.

In order to investigate the differential validity of self-assessment further we will have to consider the ease with which learners can provide self-assessments in specific skill areas. We will then investigate through manipulation of the direct experience factor some possible mediating influences impinging on self-assessment.

III Self-assessment of reading skills

The literature review and meta-analysis covered the most 'important' second language skills, primarily for English for second/foreign language learners. For the purposes of the meta-analysis, and the limited information available in published literature on the topic, we must assume that all the criterion variables had equal reliability. As Hunter and Schmidt (1990) point out, however, the impact of the error of measurement may not be symmetric across studies, which may lead to larger homogeneity estimates.

The largest number of correlations were between second language reading criterion variables and self-assessments in reading (Table 3). Not unlike the picture observed in the across-skills effect sizes, the average correlation and effect size for reading appear robust. Considering the fact that the sampling of subjects includes both second language and foreign language contexts, it is perhaps not surprising

Table 3 Meta-analysis of 23 self-assessment reading *r*s

Reading/SA	*g*	95% CI	*r*	*p*	Deviation	Homogeneity
Overall	+1.5555	+1.50/+1.61	+0.6139	0.0000	0.464	–11.492

that self-assessment with reading correlations are strong. Reading tends to be the skill that is first taught in the foreign language context, and given the fact that most of the subjects were recruited from universities, subjects were most likely very experienced in using their reading skills.

The homogeneity of self-assessment and reading skill correlation is considerably greater than self-assessment correlations in general:

Minimum is		0.320
Lower hinge is		0.460
Median is		0.490
Upper hinge is		0.572
Maximum is		0.700
3		2
3		9
4	H	034
4	M	5778899
5		1344
5	H	68
6		
6		888
7		00

It appears that self-assessment of this skill is relatively more valid than that of lesser developed skills. A plausible reason for this slight advantage for reading may relate to the extent of experience learners have with second language reading. In many foreign language contexts, exposure to the written word predates extensive opportunities for listening and speaking practice, and thus may influence to some degree the relative accuracy of self-assessment. This experience factor is explored in detail below.

IV Self-assessment of listening skills

The self-rating of listening skills provides another strong average correlation. There is, however, a wide range of variation in subjects' accuracy in the self-assessment of this skill. This variation may be due to the possibility that subjects' experience with listening to the second/foreign language may be less extensive than their experience with reading. There is also a possibility that listeners in the EFL context would evaluate their skills in relative terms rather than in absolute terms. For instance, learners who progress to the fourth semester of an EFL listening curriculum may conceptualize their skill in terms of their abilities in relation to less advanced peers. When their self-assessments are correlated with criterion tests, the correlations may

indeed be low. Another possibility is that alluded to by Blanche and Merino (1989) – low proficiency listeners may overestimate their skills, and high proficiency listeners, especially those in the foreign language context, may underestimate their skills (Table 4).

Table 4 Effect sizes for 18 self-assessment listening *r*s

Study	*g*	95% CI	*r*	*p*	Deviation	Homogeneity
Overall	+1.7128	+1.65/+1.77	+0.6505	0.0000	0.856	–51.007

The range of correlations, although based on a limited sampling of 18 correlations, is more variable than those observed in the reading data set, possibly owing to the relatively infrequent experience learners may have with second language listening compared with second language reading:

Minimum is 0.250
Lower hinge is 0.350
Median is 0.473
Upper hinge is 0.690
Maximum is 0.810

2 558
3 H 559
4 M 2568
5 011
6 H 9
7 889
8 1

The magnitude of self-assessment of listening nevertheless appears to be in general close to what we have seen in reading. The average correlation is strong – again concurring with the Blanche and Merino evaluation. It is interesting to note that reading and listening are receptive skills that do not require the speaker to preplan and execute specific production strategies. If indeed metacognitive and metalinguistic awareness are instrumental in successful interlanguage communication, we might expect learners to be relatively more aware of their own proficiency in the productive skills of speaking and writing.

V Self-assessment of speaking

The survey of the research literature revealed a surprising number of correlations of self-assessment with speaking skills, suggesting that

there is an expectation that experience should be related to self-assessment skill. The summary of the meta-analysis for speaking skill self-assessment suggests that in contrast to this expectation, learners are actually less adept at estimating their own speaking skills. Furthermore, there is more homogeneity in the correlations of the self-assessments with the speaking criteria, usually in the form of teacher ratings of speaking skill or oral proficiency interviews (Table 5). While the average correlation and effect sizes are nontrivial, they are considerably smaller than those observed for reading and listening. It should be noted that most of the very low correlations observed came from the Dutch as a second language context (Janssen-van Dieten, 1989) in which three different levels of learner proficiency were sampled. Factorial designs invite an interaction of proficiency and test difficulty and subsequent nonlinearity. The stem and leaf distribution for the correlations between self-assessment and speaking skill criteria is as follows:

Minimum is		0.090
Lower hinge is		0.390
Median is		0.530
Upper hinge is		0.662
Maximum is		0.780
0		9
1		
1		59
2		
2		69
3		3
3	H	59
4		2
4		669
5	M	013
5		668
6		04
6	H	56678
7		0
7		568

The large range of correlations with speaking criterion measures suggests that the self-assessment of speaking skill is quite susceptible to

Table 5 Effect sizes for 29 self-assessment speaking *r*s

Study	*g*	95% CI	*r*	*p*	Deviation	Homogeneity
Overall	+1.3313	+1.26/+1.40	+0.5541	0.0000	0.512	–8.391

extraneous factors in the self-assessment process. It is also important to consider that the criterion measures of speaking skill are likewise open to variation. Speaking skill is often assessed *post hoc* and holistically, by structural interviews that are biased towards formal control of grammar, or by noninterval rating scale criteria, which could lead to a truncated correlation. Second language speakers may assess their own abilities in the light of their communicative intentions rather than the actual effect of their efforts to convey messages to an interlocutor. Interestingly, the other product-dependent measure, that of writing skill, also reveals a relatively lower average correlation between self-assessment and the criterion.

Table 6 Effect sizes for 15 self-assessment writing *r*s

Study	*g*	95% CI	*r*	*p*	Deviation	Homogeneity
Overall	+1.2334	+1.15/+1.31	+0.5249	0.0000	0.416	–7.567

As in the example provided by the speaking correlations, the methods of assessing writing skill may not result in interval scaling. Many assessments instead utilize nominal or categorical scales that do not readily lend themselves to correlational analysis. We might therefore suppose that the correlation between self-assessment and the two product-orientated skills of speaking and writing would be higher than the overall average correlation observed in this meta-analysis.

The range of observed correlations is fairly homogeneous with a gap between the lowest observed *r* and the lower hinge of the stem and leaf plot:

```
Minimum is          0.160
Lower hinge is      0.420
Median is           0.560
Upper hinge is      0.636
Maximum is          0.680

1       6
2
2
3
3       55
4   H   13
4       5
5       0
5   M   68
6   H   034
6       568
```

Unlike reading, listening and speaking, the writing skill appears to have an outlier ($r = 0.160$), which influences the homogeneity more than the average correlation. Hunter and Schmidt (1990: 263) warn that outliers may be artifacts of printing errors or poor research design.

VI Some meta-analysis issues

The general picture of the concurrent validity of self-assessment with criterion skills suggests that there is ample evidence of robust correlations. There are, however, a few inferential issues that must be checked before the correlations can be interpreted directly. One deals with the adequacy of the sampling, and addresses the 'file-drawer problem' (Rosenthal, 1984: 107). That is, there is a well-known bias towards the public dissemination of 'significant' results in the research literature. The file drawer problem involves an estimation of the number of studies (correlations) that would be needed to provide counterevidence to the observed results. The obtained coefficient (Orwin, 1983) provides an index of the 'fail safe' number of correlations or effect sizes needed to reverse the pattern the meta-analysis indicates:

$$N_{fs} = \frac{N(\bar{d} - \bar{d}_c)}{\bar{d}_c} \qquad (2)$$

Where N is the total number of studies in the meta-analysis; $\bar{d}$ is the average effect size (g, above); and $\bar{d}_c$ is the criterion effect size selected. The criterion effect size is often based on one of those suggested by Cohen (1988): $d = 0.2$ (small effect); $d = 0.5$ (medium effect); $d = 0.8$ (large effect). Since all the meta-analyses results above involve large effect sizes, i.e., all the observed gs are larger than 0.80, we will proceed with the assumption of a large effect size in evaluating the results of the meta-analysis of the four skill areas (Table 7). The fail-safe threshold is related to the average effect size ($\bar{d}$). The higher the $\bar{d}$, the larger the number of studies with no effect a researcher would need to find in order to provide counterevidence to

Table 7 Fail-safe N estimates

	$\bar{d}$	N	N_{fs}
Reading	1.555	23	22
Listening	1.712	18	20
Speaking	1.331	29	19
Writing	1.233	15	8

the putative effects indicated by the literature review that forms the basis of the meta-analysis. For this analysis, we would need only eight correlations showing no relationship between self-assessment and writing skill in order to 'overturn' our meta-analytical finding of a robust effect size. For listening, which had the largest average effect size, we would need a larger number of studies with null findings than those sampled in order to conclude that the product of this analysis is a matter of sampling bias.

The second inferential problem with meta-analysis findings centres on the fact that more than one effect size may be obtained from each of the studies sampled. Since self-assessment studies typically involve more than one criterion measure, a limited number of groups of learners provide data for correlations between self-estimates of different skills with criterion measures of those skills, and in doing so violate the assumption that each correlation represents a unique self-assessment. Persons inclined to under- or overestimate their own proficiency in one skill are likely to do the same in another skill. The pattern of correlations in a matrix therefore may potentially contain some degree of autocorrelation since the sampling characteristics of the group are not independent. Although the literature on meta-analysis is unclear about the severity of this problem (Wolf, 1986: 16), one potential remedy is to convert the effect sizes from each study as a single statistic. In the case of correlation coefficients, this involves first transforming the reported correlations into Fisher's *Z* and then averaging them before converting the average *r* to an effect size.

In the foregoing analyses a number of the studies sampled involve multiple dependent variables and therefore more than one correlation between the self-assessment and the criterion. We will therefore rerun the meta-analysis using one average effect size per study so as to examine the difference in magnitude observed thus far in the effect sizes with the effect sizes after the multiple criterion variable bias has been reduced to an average. For the sake of brevity, and for our secondary goal of investigating the experiential basis for self-assessment in listening, we will limit the reanalysis to the relationship between self-assessment and listening comprehension.

Table 8 summarizes the outcome of averaging correlations within studies. In contrast to the 18 correlations summarized in Table 4,

Table 8 Effect sizes for four self-assessment listening *r*s (averaged)

Study	*g*	95% CI	*r*	*p*	Deviation	Homogeneity
Overall	+2.286	+2.18/+2.39	+0.752	0.0000	0.864	–69.93

whose average *r* was +0.65, the combined analysis results in even larger effect sizes and an average correlation of +0.75. Both approaches to the analysis of the relationship between self-assessment and second/foreign language listening strongly support the meta-analysis finding that the relationship is robust.

A number of second/foreign language testing studies have employed self-assessment for the purpose of accessing a trait with a unique method. In these multitrait-multimethod studies (Bachman and Palmer, 1981; 1982; Buck, 1992), self-assessment has usually demonstrated high correlations with various criterion variables, but has also usually been correlated with the other self-assessment measures across traits. The implication of these studies is that self-assessment introduces a systematic method facet even if it has some partial concurrent validity with important criterion variables. Depending on measurement needs and logistical constraints, self-assessment may be viewed as providing too cloudy a picture of proficiency for some testing decisions, e.g., candidate selection, or may be viewed as sufficiently accurate for other 'low stakes' decisions, e.g., placement within programmes or rough-and-ready needs analysis instruments.

VII Factors affecting self-assessment

The correlations between self-assessment and the criterion variables in Table 4 indicate considerable variation. In this section we will consider factors that influence the magnitude of correlations between self-assessment and criterion measures. As was mentioned in the discussion of the speaking and writing skills, the likelihood that researchers use ordinal scales for these skills may lead to a depression in the correlation coefficient. Assuming, however, that both the self-assessment scale and the criterion scale are ordinal, we can consider other subject-specific factors that affect concurrent validity. One is the possibility that the descriptions used on the self-assessment are situational: 'I can understand the dialogue in French films.' When such statements are used for self-ratings, each subject may understand the ordered categories in differing ways. Some, for instance, may interpret the ordinal scale literally and may assume partial comprehension is the 'middle' of the scale. Others may consider themselves in relative terms – relative to peers, or may view the ordinally scaled item relatively, that is, understanding French in films is easier than understanding French radio broadcasts. These factors contribute to the method artifact often uncovered in construct validation studies.

One other factor influencing self-assessment is more subtle. It involves the matching of the self-assessment items to second language skills in contexts that subjects can be expected actually to have

experienced. Since many self-assessment scales are constructed so as to capitalize on contextualized, ordinally scaled definitions of proficiency, which may not be directly related to the second language learners' actual experience with the language, the experience factor is potentially an important source of variation in self-assessment. Fortunately, the influence of experience can be examined empirically through the manipulation of a self-assessment survey and language learning course content.

VIII An analysis of experiential factors in self-assessment

A self-assessment validation study was conducted for a language training programme at a large Japanese electronics company. The overall design of the study was to have beginning and elementary-level subjects ($n = 254$) complete skill-focused self-assessment batteries (20 items) in Japanese prior to completing a 60-item achievement test (KR20 = 0.838) devised to cover the skills featured in a year-long English as a foreign language course taught in-house. The company employees were mostly male, college educated and ranged in proficiency from 250 to 550 on the Test of English for International Communication (TOEIC).

The achievement test covered only the content of the coursebook[1] used during the 90-hour instructional period. The self-assessment was designed to match the content of the coursebook thematically and linguistically. Each subject's teacher ($n = 8$) was also provided with an English translation of the self-assessment. Teachers were asked to review their own class records and complete the assessment form so as to provide a third perspective on each subject's performance and mastery of the coursebook content. By having the teachers' assessment of the students' performance, the basis of a triangulated view of self-assessment was established. Of main interest was the degree to which student and teacher evaluations of each subject's achievement would correlate with the criterion achievement test. At the elementary level the achievement test consisted of ten sections, most of which were taken verbatim out of the coursebook. A few of the test sections, however, were modified so as to test the functional content of that particular text section, but in a format different from that used during classroom exercises. This manipulation of the test format made for potential variation in the subjects' recognition of the content

[1] *BBC beginner's English stage one* (Garton-Sprenger, J. and Greenall, S., 1986. London: BBC English).

as being directly traceable to specific language lessons. This manipulation leads directly to a hypothesis about the recoverability of classroom experiences and their applicability to the self-assessment process. Specifically, we expect that self-assessment ratings will show a larger multiple correlation with an achievement test criterion variable that is identical to test sections the subjects had experienced directly than with test sections using different formats from those appearing in the coursebook. We can also anticipate that teacher assessments of student performance follow the same principle – that assessment is most accurate when based on experiences observing individual students learning specific coursebook content. Stated another way, these expectations anticipate the relative accuracy of self-assessment when the criterion is achievement (experienced) as opposed to self-assessment on proficiency-based (abstract) criteria.

1 Materials

Three sections of the elementary-level achievement test will be featured in this analysis. The listening skill will be considered here, primarily because it is the skill area most amenable to instructional influences in this foreign language context, and secondly because the majority of the text exercises were related to this skill.

For the purposes of this study, one of the listening sections was rewritten into a listening cloze format. The subjects were presented with the text of the passage with specific lexical items deleted in a 'rational cloze' manner. This criterion represents the manipulated variable in that the subjects had originally heard the passage *History of London* (see the Appendix) in a listening-for-gist task with no accompanying text. While the theme is potentially recoverable, the actual task on the test was different from that experienced by the students in the context of the classroom lessons.

The other two test sections represented the content of the coursebook as closely as possible. One involved listening to a weather broadcast and marking a map of Europe (see the Appendix). The other was based on a conversation between two of the text characters. The subjects' task was to fill in a chart with information from the dialogue (see the Appendix). The format and content of these two sections most closely resembled the experience learners had in their language lessons, and are therefore akin to assessments of achievement, or mastery of lesson content. Initially each of the ten test sections was entered into contrastive standard regression analyses as a

separate dependent variable ($n = 205$).[2] One set of predictors was the 20 self-assessment ratings in Likert scale format (1–7) provided by the students in their native language, Japanese, and an identical set provided by each student's teacher in English. Figure 1 provides a graphic representation of the contrastive multiple regressions. As might be expected, teacher evaluations of their students' performances in general show a higher correlation with the test section dependent variables.

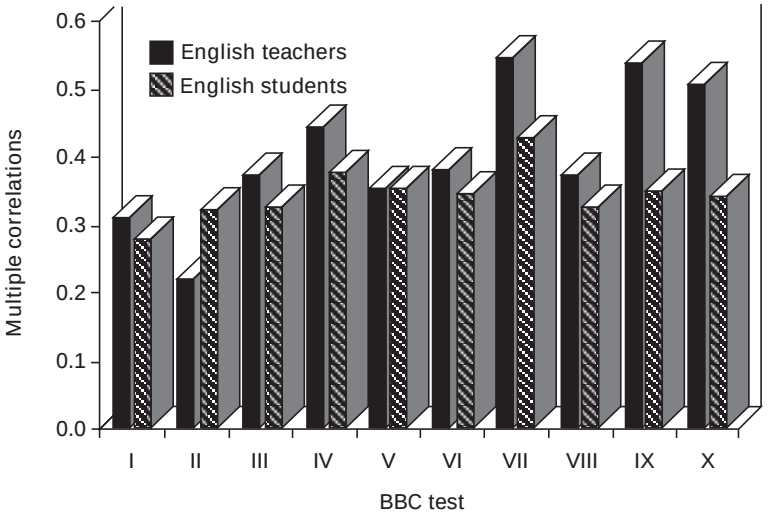

Figure 1 Multiple regression results

2 Results

For the testing of the main object of interest we will focus on the regressions of the students' self-assessment with the modified format test section (III) relative to the multiple regression results based on the students' assessment of their own functional abilities for the exact-match test sections (the sum of VIII and IX). Sections VIII and IX were summed in order to create an approximately equal number of total points for the modified and unmodified sections of the test. Table 9 summarizes the results.

[2]The multiple regression analyses were based on correlation matrices calculated with pairwise deletions of missing self-assessment responses. The *n* of 205 was the smallest in a range of *n*s from 205 to 254. Ninety per cent of the cases contributed to the correlations between the relevant BBC test sections and the self-assessments.

Table 9 Multiple regression results for Sections III, VIII and IX

III	0.391	<0.05
VIII and IX	0.501	<0.0001

Although the self-assessments result in a multiple correlation with the modified format section of the achievement test (III) not dramatically different in magnitude from those produced by the teachers' ratings of the students (Figure 1, III), there is a considerably larger multiple correlation with the exact-match section of the test. This provides evidence that learners will be more accurate in the self-assessment process if the criterion variable is one that exemplifies achievement of functional ('can do') skills on the self-assessment battery. When the battery contains items of a more abstract nature, which may assess language proficiency, learners can be expected to have had less direct experience in practising those language skills, and the resulting self-assessment may be less accurate. This finding suggests that episodic memory of using particular skills in the classroom experience would enhance the accuracy of self-assessment. Teachers' assessments, in contrast, may be considered more generalizable, and based on cumulative experience in observing student performance in the classroom context. Teacher assessments are perhaps best used when the criterion variables are of general proficiency, as opposed to mastery of specific course objectives. Learner self-assessment of course objectives, mediated by trial, error, feedback and revision in the learning process, may better assess the learners' confidence in the degree content mastery.

IX Discussion

The results of the meta-analysis concur with the Blanche and Merino (1989) conclusion that self-assessment typically provides robust concurrent validity with criterion variables. The close-up examination of the process of self-assessment, mediated by variation in direct experience in language learning tasks, suggests that the degree of experience learners bring to the self-assessment context influences the accuracy of the product. We might assume that when the criterion is one that does not invoke episodic memory, the self-assessor may have to rely on a recollection of his or her own general proficiency to make the assessment. It is perhaps at this point that the methodological artifacts of self-assessment are most likely to interfere with the process. Sub-

jects may also resort to relativity (Moritz, 1995), or be influenced by self-flattery.

The main limitations of this study are concerned with the methodology of meta-analysis. The literature base for self-assessment in second language learning is not extensive, but fortunately is growing at a steady rate. The magnitudes of the correlations observed are not homogeneous, suggesting that the studies sampled may have been constrained in a number of ways. Differing *n* sizes, lack of statistical power, low internal consistency, truncated scales, and the like, serve to cloud the relationship we seek to examine. This is an unfortunate reality for meta-analysis as a quantitative method, which makes it only slightly more advantageous than a hermeneutically inspired literature review, unless pains are taken to remove biases stemming from these factors (Hunter and Schmidt, 1990).

While these correlational results provide some confirmatory evidence for the role of direct experience as a mediating factor in self-assessment, it should be noted that the self-assessment battery comprised 20 individual functional ('can do') statements for student and teacher ratings. These were also correlated with the total achievement test variance in a standard regression format so as to ascertain their cumulative correlation. This approach revealed that some of the self-assessment items may contribute very little to achievement test variance, and that there may be, in some instances, only weak evidence of construct validity when there is a mismatch between the content of the self-assessment items and criterion skills. This finding underscores the need to design self-assessment of language learning achievement according to specific curricular content. Provided that the content validity requirement is met, the overall picture indicates that there is clear potential for predictive accuracy of criterion skills based on self-assessment measures. Under what circumstances self-assessment procedures will be sanctioned in language teaching and testing programmes remains, however, an open question.

X References

Bachman, L. and **Palmer, A.** 1981: The construct validity of the FSI Oral Proficiency Interview. *Language Learning* 31, 67–86.

—— 1982: The construct validation of some components of communicative proficiency. *TESOL Quarterly* 16, 449–65.

Blanche, P. 1990: Using standardized achievement and oral proficiency tests for self-assessment purposes: the DLIFLC study. *Language Testing* 7, 202–29.

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

Buck, G. 1992: Listening comprehension: construct validity and trait characteristics. *Language Learning* 42, 313–57.

Cohen, J. 1988: *Statistical power analysis for the social sciences* (2nd edn). Hillsdale, NJ: Lawrence Erlbaum Associates.

Falchikov, N. and **Boud, D.J.** 1989: Student self-assessment in higher education: a meta-analysis. *Review of Educational Research* 59, 395–430.

Ferguson, N. 1978: Self-assessment of listening comprehension. *International Review of Applied Linguistics* 16, 149–56.

Hunter, J.E. and **Schmidt, F.L.** 1990: *Methods of meta-analysis*. Newbury Park, CA: Sage.

Janssen-van Dieten, A.-M. 1989: The development of a test of Dutch as a second language: the validity of self-assessment by inexperienced subjects. *Language Testing* 6, 1–13.

Johnson, B.T. 1989: *DSTAT: software for the meta-analytic review of research literatures*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Le Blanc, R. and **Painchaud, G.** 1985: Self-assessment as a second language placement instrument. *TESOL Quarterly* 19, 673–87.

Milleret, M., Stansfield, C. and **Mann-Kenyon, D.** 1991: The validity of the Portuguese speaking test for use in a summer study abroad program. *Hispania* 74, 778–87.

Moritz, C. 1995: Self-assessment of foreign language proficiency: a critical analysis of issues and a study of cognitive orientations of French learners. Unpublished PhD dissertation, Cornell University.

Orwin, R. 1983: A fail-safe *N* for effect size. *Journal of Educational Statistics* 8, 157–59.

Oskarsson, M. 1980: *Approaches to self-assessment in foreign language learning*. Oxford: Pergamon Press.

Rosenthal, R. 1984: *Meta-analytic procedures for social research*. Beverly Hills: Sage.

Wolf, F. 1986: *Meta-analysis: quantitative methods for research synthesis. Sage University Paper Series on Quantitative Applications in the Social Sciences*. 07-059. Newbury Park, CA: Sage.

Wongsotorn, A. 1981: Self-assessment in English skills by undergraduate and graduate students in Thai universities. In Read, J., editor, *Directions in language testing*, Singapore: Regional Language Centre.

Appendix

III Listen to the story about London. Fill in the blank spaces with the words you hear.

In 1860 London was a large city of two and a half 13) _____ people. By 1910 London was the biggest city in the world, with a 14) _____

of five million. During that fifty years there were many important 15) _____ in the city. In the first half of the Nineteenth 16) _____, transport was a big 17) _____ in London. One important means of transport was the horse bus. But it was too 18) _____ for most people, and it was very slow. The other 19) _____ means of transport was the River Thames. Many people travelled across London by 20) _____ on the river. But again, it wasn't very cheap, and if the weather was bad, it wasn't very safe. So thousands of people 21) _____ several miles to work every day. The world's first 22) _____ railway opened in 1862. It was cheaper, safer and faster than the horsebus and the steamboat. And millions of people used it every year. Then Karl Benz invented the first 23) _____ in 1885. And in 1907 two hundred motor buses came to London. Many people started travelling by bus, and people who were very rich, bought their own cars. Another important change came with 24) _____, and in the first part of the Nineteenth Century, London had 25) _____ The streets were dark, dirty, and often dangerous. Things became much brighter when electric street lighting came to London in 1878. In the same year, people started using the telephone, and by 1891 the people of London could make phone calls to Paris. Five years later, in 1896, the first 26) _____ opened in the center of London. People started going to the movies. At the turn of the century, a large number of theaters, hotels, restaurants, and department stores opened, and by 1910, London was a modern city very different from London in 1860.

VIII Listen to the weather forecast and mark the map of Europe.

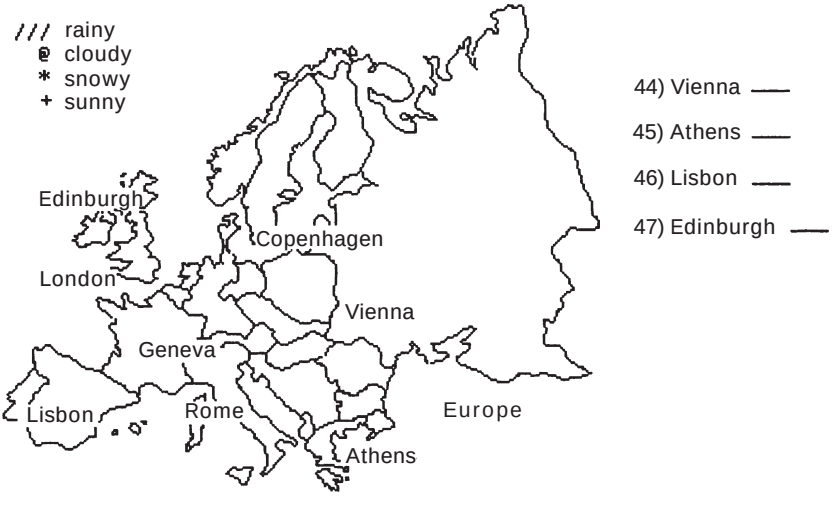

Figure 2

IX Listen to Koji and Diana talking. Mark the chart to show who does what.

	Koji	Diana
48) Confirm flight		
49) Pack suitcases		
50) Buy presents		
51) Pay hotel bill		
52) Type report		
53) Book restaurant		
54) Phone wife		
55) Book taxi		