Reliability, validity and feasibility of the Project - a component in the Israeli EFL Matriculation test [1]

Tziona Levi
Tel Aviv University

Abstract

In 2003 an assessment component, the *Project*, was added to the English matriculation exam of the Israeli Ministry of Education (henceforth: MoE) to reflect changes incorporated in the national English curriculum introduced in 2001. The quality of the *Project* was investigated by examining its inter-rater reliability, content validity and practical features. Inter-rater reliability was high with respect to the overall grade, but not with respect to the component scores assigned against the official rating scale. A content analysis revealed inconsistencies between the *Project* guidelines and their reflection in the underlying construct of the curriculum. There was a positive correlation between teachers' understanding of the *Project*'s practical aspects and the weight attached to this component by the MoE. A strong positive correlation was found between teachers' positive perception of the *Project* and their willingness to act as external *Project* raters or to promote this component in other schools.

[1] The findings in this paper were presented at the ACROLT symposium "Academic Committee for Research in Language Testing: Assessment, Policy and Education, Ma'ale-Hachmisha, Israel. June, 2005.

Melbourne Papers in Language Testing 2008 Volume 13.2 pp. 1-34. The Language Testing Research Centre, The University of Melbourne.

Introduction

Matriculation examinations are a powerful tool in the Israeli education system (Shohamy, 1991). These national, state-mandated tests are designed by non-school parties who develop, administer, score, report and publish test results (Hamilton, Stecher, & Klein, 2002). They are recognized as strongly influencing the educational system, having control over the curricula and eliciting the introduction of new teaching methods. In addition, the matriculation examinations focus, channel and encourage learning and are assumed to create accountability by successfully measuring and reasonably comparing actual learning against projected learning outcomes (Crooks, 2006; Duvall, 2005). Successful scores in these exams also help determine acceptance into institutes of higher-education.

In 2002, the inspectorate for English, which is responsible for test contents, results, and washback, added a *Project* component to the English as a Foreign Language (EFL) matriculation exam (Steiner, 2002a) This component was added to reflect changes in the national English Curriculum for Israeli Schools (2001). The new curriculum classifies language ability and knowledge into four inter-related language domains, each integrating all four language skills in light of a particular communicative act. In addition to setting standards and benchmarks for each domain and establishing levels for assessing language, the new curriculum also incorporates a set of principles that define effective language learning, teaching, and choice of materials, topics, tasks and assessment. The *Project* component was added to the matriculation exam to reflect the principles of performance assessment and the inclusion of multiple language domains (Steiner, 2002b).

Besides describing the Project component of the high-stakes Israeli Matriculation Exam, this study scrutinizes the psychometric criteria used in its evaluation assessing its validity, reliability and practicality.

The Project

A Project, a carefully planned and designed body of work, stimulates authentic language use. (Mann, Shemesh & Shlayer, 2002). As students are taken out of the classroom to investigate their chosen topic the language required for the development of the *Project* is derived from the nature and content of the *Project* itself. Since projects are student-motivated and student-centered the teacher's role is to facilitate planning and implementation together with the students.

*Project*s are complex tasks, usually executed over lengthy periods of time that require extensive student inquiry and that culminate in student products and performances (Wiggins & McTighe, 1999). Projects genuinely reflect content-based instruction, which language teachers find appealing. Due to the vibrant learning environment, it requires stimulation of higher-order thinking skills and enhancement of student accountability and authentic communication (Stoller, 1997). Thus, students learn a variety of study and language skills that prepare them for the range of academic tasks they will encounter later (Brinton, Snow & Wesche 1989; David, 2008). Furthermore, *Project*s incorporate all four domains of language learning, helping students to attain the standards of each of the four domains of the national Israeli curriculum (Mann et. al., 2002) and are especially appealing as they reflect achievement of the curriculum:

> *"Social Interaction - students develop communication skills, both orally and in writing, enriching their vocabulary and improving their use of accurate language. Access to information - while researching their chosen topic, students are exposed to oral and written texts from a variety of sources: Internet, books, magazines,*

> *experts in the field, etc. Presentation - students are given opportunities for presenting information and ideas on a wide range of topics in an organized manner, in a variety of formats in spoken and written English. The oral presentation complements the written presentation. Appreciation of Literature, Culture and Language - research in-depth and learning about the topic allows students to gain cultural, historical and social insight" (2002: 19).*

Given that the *Project* captures so many valuable educational goals, the chief EFL inspector in Israel enforced its inclusion in the matriculation exam to ensure implementation. The Inspectorate bulletin (Steiner, 2002a) defines the *Project* as a three-part unit (process, product and oral presentation) where the *Project* is written collaboratively in class on a topic of the students' choosing based on guidelines provided by the teacher. It is graded twice; once as 10% of the annual/school grade, which itself is worth 50% of the student's final matriculation grade, and summatively as 50% of the oral exam where students discuss their *Project* (Steiner, 2003b).

Issues of implementation, reasoning and efficiency of *Project*s were raised by teachers on the English Teachers' Network of Israel (ETNI) site between the 19th and 25th of June, 2003, which resulted in a revision of the *Project* guidelines and in explanatory rules being added to the written *Project*.

Examination of the English inspectorate bulletins (Steiner, 2002a; 2002b; 2003a; 2003b; 2003c) and a close reading of the English teachers' mailing list archive (http://www.topica.com/lists/etni/read) call for a study of the *Project* component, especially since its psychometric traits have yet to be established, while its impact on the teaching, learning and assessment of English in Israel is significant.

Research goals and study outline

The purpose of this study is to evaluate the quality of the *Project* and to offer empirical evidence about the process of its implementation as an assessment component in a high-stakes matriculation exam. The uniqueness of the study relates particularly to the context of the *Project*. Since policymakers have explicitly defined the role of the external assessment as highly effective for introducing change, this study can define its psychometric traits in operative terms. Additionally, the study examines the usefulness of this specific tool. To match these goals, the quality of the *Project* component as an assessment tool is examined through establishing the inter-rater reliability of teacher-raters, content analysis (Bachman, 1990; Bachman & Palmer, 1996; McNamara, 2000) by experts and the examination of its practicality and usefulness (Bachman & Palmer, 1996; Moss, 1992; Patton, 1986) as judged by teachers implementing the *Project* component in their teaching. For this purpose, a variety of instruments were applied starting with the rating and justification process, elicitation of expert views and finally distribution of teacher questionnaires.

In order to include as much of the *Project* scope as possible, data about the *Project* was collected from three different sources in a three-stage research:

The first stage examined the *Project*'s inter-rater reliability since *Projects* are assessed only by classroom teachers using a rating scale provided by the MoE.

The second stage scrutinized content validity through expert analysis. This stage aimed at determining whether the *Project* could in fact assess the implementation of the new Israeli curriculum (2001) covering all aspects of learning English as a foreign language. This was done by comparing the MoE requirements and criteria for the *Project*s with the content of the national curriculum.

The third stage verified pragmatic aspects and practicality by collecting information through questionnaires completed by teachers who implemented *Project*s. Pragmatics is highly relevant to examining the quality of the *Project* component since it connects performance-based assessment in general, a high-stakes test and the practicality of a specific testing task.

Research questions

The aims of this study led to the formulation of the following research questions:

What is the quality of the *Project* component as part of the matriculation exam in EFL in terms of its

- inter-rater reliability among the teacher-raters?

- content validity as judged by expert analysis?

- pragmatic features focusing on practicality and feasibility as judged by teachers implementing the *Project* component in their teaching?

Study context

Completing a *Project* performance task is a mandatory requirement for a matriculation grade in English in all high-schools throughout Israel. Therefore, the *Project*s examined in this study were collected randomly to reflect this. These *Project*s were written by groups of high-school students from different backgrounds and levels (all from public schools representing different sectors in Israeli society: Jewish religious, secular and Arab students).

Design of study: instruments, procedures and sample

Data collection was structured in a three-stage research program launched by grading and justification of grades, followed by elicitation of expert views and finalized with a teacher questionnaire.

Grade justification was conducted with teacher-raters (henceforth: raters) to gain rater consistency. Raters graded the *Project* using a MoE-provided rating scale that comprised six elements: 1)content of the *Project* (content); 2) sources used for writing the *Project* (sources); 3) level of language used in the *Project* (language); 4) effort invested in writing the *Project* (effort); 5) individual contribution of each group member to the final *Project* product (individual contribution) and 6) the overall grade (grade) (See Appendix 1). Raters used a numerical score. Then, they justified and rationalized the process by providing the criteria for their assessment in writing (Seliger, Shohamy, 1989). Eight randomly collected *Project*s were examined by 12 EFL teachers of grades 7-12 who teach various populations in Arab and Jewish religious and secular public schools, and who habitually implement *Project* work. The rubric was used twice: once as a grading tool to asses the *Project* holistically, and then to justify the grade. The raters looked at each criterion analytically and justified why they had chosen a particular score. The raters were given the choice to either use the descriptors of the rubric itself or to create their own as open answers.

As a result, categories were created and analyzed qualitatively, adding to the level descriptors provided by the official rating scale (See Appendix 2). Each of the 8 graded *Project*s accumulated 4 sets of ratings (for a total of 96 ratings). Four three-rater groups were created to ensure a systematic overlap of the grade *Project*s and to ensure proper statistical analysis. Ratings and justifications were compared to each other to achieve inter-rater reliability.

Expert views were elicited by seven experts who met four out of the following five criteria:

Experts were very familiar with the Israeli curriculum (2001) either by being part of the document writing committee or by promoting its implementation;

Experts were very familiar with the changes in the matriculation exam format;

Experts were very knowledgeable in traditional and alternative assessment procedures;

Experts were experienced EFL teacher who had prepared students for the EFL matriculation exam previously;

Experts were recognized in the EFL field as knowledgeable in two of the above criteria.

The experts were asked to compare the *Project* component requirements and criteria contained in the MoE guidelines (Steiner, 2003c) with the content of the Israeli curriculum (2001). They were free to inspect the two documents at their will. The comparisons were studied qualitatively to achieve content validity. The tables created for analysis listed the categories of the underlying curriculum principles and the problems of the *Project* specifications in light of those underlying principles.

Finally, Sixty two 7-12-year EFL teachers from different public schools completed the questionnaire. These teachers implemented the *Project* component in the previous year, either as a regular part of their teaching routine or for the first time as a result of the new matriculation demands. The questionnaire inquired about pragmatics, implementation and feasibility of the *Project* component. To ensure examination and coverage of practicality quality, most of the items in the questionnaire reflected Nevo and Shohamy's model (1986). The questionnaire contained two sections: the first examined teachers' views regarding the use of the additional test component and the second inquired about teachers' background and experience as EFL teachers and as *Project* users in their classes.

Data analysis and results

Inter-rater reliability

The first research question of the study targets the quality of the *Project* component in terms of its rater consistency as judged by teacher-raters. The grades for the *Project* given by different raters were compared to each other not only to examine rater reliability, but also to validate raters' scoring showing rating stability. To enable comparability of ratings, the 12 raters were divided into 4 groups of 3 raters each. The grades of the three raters in each group correlated highly on the final grade in 3 out of 4 groups, and in the fourth group 2 out of the 3 raters also correlated highly (Table 1).

Table 1: Rater reliability shown by Pearson correlations between raters for each rating-scale criterion

	Between 1st & 2nd rater	**Between 1st & 3rd rater**	**Between 2nd & 3rd rater**
Content			
Group 1	-0.08	0.24	0.55
Group 2	.81*	.88**	.89**
Group 3	.91***	.75*	.77*
Group 4	0.55	.96***	0.55
Source			
Group 1	0.46	-0.07	.81*
Group 2	.62*	0.16	0.48
Group 3	0.59	0.39	.87**
Group 4	0.61	.62*	0.44
Language			
Group 1	0.35	0.56	0.3
Group 2	.63*	0.43	0
Group 3	0.4	0.58	0.19
Group 4	.90***	.94***	.71*
Effort			
Group 1	-0.35	0.17	0.6
Group 2	0.45	0.2	0.16
Group 3	.79*	.93***	.76*
Group 4	0.02	0.39	0.39
Individual Contribution			
Group 1	0.38	-0.06	0.08
Group 2	.93***	.89**	.77*
Group 3	0.6	.95***	0.48
Group 4	0.05	0.06	.80*
Final grade			
Group 1	0.02	0.43	.65*
Group 2	.84**	.91***	.92***
Group 3	.97**	.86*	.83*
Group 4	.76*	.91***	.92***

Significance level: *p< 0.5 **p<.0.1 ***p< .001

Table 1 displays the correlations for each of the 6 criteria rated. The main findings suggest that for Content, rater groups 2 and 3 agreed on the grade (r > .75, p < 0.5). In the criterion of Individual Contribution, although rater correlations were low, raters indicated scarce group work with chores delegated to individuals. In the criterion of Language, raters disagreed in general, saving a high correlation within group 4 (r > . 90, p < .001). This disagreement could be explained by the rater comments where they dealt with grammar and vocabulary corrections. In the criterion of Effort, raters disagreed commenting on the difficulty of objectively observing the individual effort which is considered dependent on internal classroom behavior, the lack of instructions and guidelines provided in general and specifically when rating work of learning-disabled students. Additionally, they commented, on the various levels of support the classroom teacher offered.

Qualitative analysis of inter-rater reliability and analysis of raters' grades

To determine if the criteria for assessment provided by the Israeli MoE (Steiner, 2003c) best reflects teachers' understanding of *Project* scores, raters explained and justified all grades in an open questionnaire for each of the criteria. Content analysis of the explanations revealed a need to supplement the rating scale by including additional teacher-rater comments.

In the criterion of Content, comments were added in 38.5% of the *Project* ratings, referring to missing elements (such as table of contents, bibliography, etc.) and plagiarism. In the criterion of Individual Contribution, comments were added in 63.5% of the ratings. In the criterion of Sources, commentary was added in 21.8% of the ratings relating to plagiarism or to inappropriate citing of sources. In the criteria of Effort and Individual Contribution, comments were added in 13.5% and 38.5% respectively. Yet other annotations showed clearly that these two criteria are subjective, dependent on reflective evidence within the *Project*, especially when there was little or no evidence of individual or group effort both observed in classroom contexts and not by external raters. Other comments specified insufficient descriptors for external features expected in *Project*s.

To recapitulate, raters used the MoE rating scale and level descriptors, but found them ambiguous and lacking, and thus added verbal clarifications to their grade.

Content validity: results and qualitative analysis of expert views

The second research question relates to the content validity of the *Project* as a high-stakes performance-based test component compatible with the requirements and content of the Israeli curriculum (2001). Seven experts in *Project* work, assessment and the Israeli curriculum examined the correlation and coverage of the *Project*s vis-à-vis the curriculum. Table 2 summarizes Curriculum principles found to be explicitly parallel. Column 3 adds information found compatible with the *Project* guidelines according to the sources listed on the official MoE site. A detailed table (dealing with classroom assessment, assessment role and criteria, choices of tasks, choice of materials and student role) would fall beyond the scope of this paper, therefore only the first section is presented as a sample of the detailed process of content analysis. Thus, this Table provides a snapshot of how the experts systematically matched each relevant item in the Israeli curriculum (2001) to the *Project* officially published guidelines and rubrics.

Table 2: Expert comparison of Israeli curriculum (2001) content and Project guidelines (2003)

Underlying Curriculum Principles	***Principles as defined in official Israeli curriculum***	***Reflection of principles in features of Project as appearing in official guidelines***	*Additional sources referred to in Project requirements (Mann, Shemesh, Shlayer, 2002 & Stoller, 1997)*
Language learning is facilitated when students:	use language as a means of gaining information in other areas	No relation to literacy in L1	"leads to the authentic integration of skills and processing of information from varied sources, mirroring real-life" (2002:18)
	take responsibility for their own learning.	Depends on what teachers’write as feedback.	
	interact, share information, exchange ideas and opinions and work together	Effective use of learning strategies not reflected in Project rating scale.	
	have opportunities to use the target language outside the classroom.	Use of language outside classroom – does that mean at home or in the library?	
	are motivated and willing to invest effort and persistence needed for foreign language learning.	Motivated to find out about people, culture, etc. This depends on the choice of the Project made by the students.	
	develop a positive self-image in the target language.	Developing positive self- image depends on the type of feedback the learners receive from their teachers	
	develop confidence in their ability to use the target language		

Language Teaching is more effective when teachers	*allow students to find out what they know and don't know by themselves.*		*"…are suitable for heterogeneous classes, where weak and strong students work together.*
	create a supportive environment that allows students to take risks, to make errors and experiment with the language.	Supportive environment depends on the environment created by teachers in their classrooms, not related to the projects	"can be designed for all levels of language learners….encourage genuine student-student & student-teacher collaboration. `
	provide opportunities for peer interaction.		
	when teachers create a language-rich environment.	Language rich environment is mainly exposure to written text including Internet texts.	
		Success-oriented feedback depends on the teacher's feedback skills during the work with the Project	
	when teachers provide feedback that is on-going and formative.	Not enough evidence of ongoing feedback.	
	encourage students to use English outside the classroom.	No evidence of students' use of English outside the classroom.	
Choice of Content	is unbiased, unprejudiced, inoffensive	if the teachers allow for freedom of choice	"Students' chosen topics are more motivating than teacher's chosen ones" (2002:20)
	caters for a variety of backgrounds	The reading is mainly non fiction, it is reading of information needed for the – Project	
	stimulates students' interest in extensive reading, in the pleasure of literature	We cannot know if it stimulates their interest in reading, literature. As a result of working on the Project they are obliged to read more, but how can we assess their pleasure?	

A similar table analyzed the benchmark coverage as reflected in the curriculum compared to the *Project* specifications and information compatible with the *Project* guidelines and rating scales in sources officially mentioned in the MoE website. The experts found discrepancies between the *Project* guidelines and the Israeli curriculum (2001) it is meant to assess (38.2% instances). There are more curriculum principles with questionable comparability than principles truly covered by the *Project* guidelines. When investigating the variety of inquiries raised by the experts, four categories were defined and are presented in table 3.

Table 3: Categories matching principles and Project guidelines

Categories	*Occurrences*	
Applying a curriculum principle is class-teacher dependent	7 occurrences	22.5%
Non-existence of principle in assessment rubric	7 occurrences	22.5%
Applying a principle depends on students' choice or motivation, which has to be encouraged before possible implementation	5 occurrences	16.1%
General comparability problems about the possible realization of principle in classroom setting or contradiction between principle and classroom reality	12 occurrences	38.7%

Curriculum content-benchmarks and Project guidelines

Only 38.2% of the curriculum principles were reflected in the MoE *Project* guidelines. From 33 curriculum benchmarks examined in the domains of Access to Information and Presentation (including one benchmark from the domain of Social Interaction), only 11 were identified in the *Project* guidelines and described as matching (33%), whereas 67% of the benchmarks were found to be ‘problematic’ (Expert 3).

A large number of the benchmarks are not reflected in the Project guidelines in general and specifically not in the rubric. This supports the comment of Expert 6: "The handbook is very vague .... referring teachers to the assessment guidelines is not enough". Moreover, Expert 4 asks where the criteria of Effort and Individual Contribution to accessing information and presentation appears explicitly, even though it is mentioned in the principles and the benchmarks.

Expert 2 adds that a number of assessment criteria appearing in the *Project* rubrics do not seem to match any benchmark and provides examples to support her point. This expert ends her analysis with the comment that although theoretically an objective criterion of Length is presented as a guide to teachers; it does not serve as an assessment criterion and does not appear in the rubrics.

These comments and observations raise queries regarding the content validity of the *Project* as an assessment tool.

In addition to the analysis through the comparison of the two documents (*Project* requirements according to handbook, 2003 and the *Project* features appearing in The English Curriculum (2001)), the experts evaluated the process of *Project* production as described in the MoE requirements in an explicit table of features and information (as in Table 2).

Practicality and feasibility of the Project

The third research question examined the practicality of the *Project* through a teacher questionnaire. To assess the value and weight of the questionnaire items, a Principal Components factor analysis was conducted. The factor loading shows three factors of which the eigenvalues is greater than 1. Examination of the findings demonstrates that the common concept in factor I is 'collection of *Project* scores and their meaningful use'. The common concept in factor II is 'suitable conditions for *Project* production' and the guiding concept in factor III is '*Project* as a reflection of values'. The factor loadings of the three coefficients present high averages suggesting that teachers tend to accept the *Project* component as significant. Table 4 provides the questionnaire statements grouped by factors.

Table 4: Questionnaire statements grouped by factors

Factor 1: Collection of *Project* scores and their meaningful use

19. Are *Project*s designed and conducted so that rights and welfare of human subjects are respected and protected?

9. Does *Project* work have a positive impact on the teaching and learning process and on the decision making process of all parties associated with them?

18. Are *Project*s based on known and accepted subject matter?

20. Are *Project* scores and consequences respected within the limits of other related principles such as those dealing with public safety and the right of privacy or others?

7. Are *Project* scores disseminated to all relevant audiences (parents, homeroom teacher, coordinator, etc.) so that they can use the findings?

10. Are *Project*s performed by instruments and procedures providing valid information? (Different assessment tools, ongoing teacher feedback measuring what the *Project* intends to measure)

6. Are *Project* scores (and rubrics) presented in forms readily understood by students?

Factor 2: Suitable conditions for *Project* production

11.Are *Project*s performed by instruments and procedures providing reliable information? (Consistent and accurate data collection procedures of the assessment tools)

15. Are *Projects* conducted with minimum disruption of educational and administrative processes at school and with consideration of existing constraints?

2. Is the audience involved in or affected by the *Project* identified so that their needs can be addressed?

12. Are *Project* conditions adequate when relating to and assessing the achievements of each student?

3. Do you feel both trustworthy and competent to perform *Project* work, so that your findings achieve maximum credibility and acceptance?

16. Are *Project*s planned and conducted with anticipation of the different positions in school?

17. Do *Project*s produce information of sufficient value to justify the resources expended?
8. Is the release of *Project* scores timely so that audience (parents, homeroom teacher, coordinator, etc.) can use them best?
13. Are *Project* data appropriately and systematically analyzed to ensure supportable interpretations of scores?

Factor 3: Project as a reflection of values

14. Are *Project* scores reported objectively without distortion by personal feelings and biases of testers?
5. Are the criteria used to determine *Project* scores justified and clearly described?
1. Describe in short what a *Project* is
4. Is the information collected by the use of the *Project* of such scope suitable to address pertinent questions about student achievements?
22. Are *Project*s scores complete and fair in their presentation of strengths and weaknesses of the individual's work?
21. Is the *Project* conducted in accord with social values and does it not stimulate violation of norms and values accepted at school and in society?

In order to examine teacher opinions regarding the practical aspects of the *Project* component, the questionnaire item averages were calculated across participating teachers.

The range of standard deviations shows a sufficient level of agreement among teachers as to the items on the questionnaire, which are of a relatively narrow range. If the questionnaire item averages are organized from those considered most practical in teachers' eyes to those considered least practical, we can gain more information about teachers' views of the *Project*'s practical aspects. Most of the items carrying higher average ratings appear in the third factor, whereas most of those with the lowest average ratings appear in the second factor. Additionally, the items in factor III – *Project* as a reflection of values – are seemingly more significant than the items in the other two factors and figured relatively more importantly in the production and implementation of *Project*s. The factor loadings of the three coefficients present relatively high averages which show that teachers basically accept the *Project* component as important.

A repeated measures ANOVA showed that a significant difference was found between the three factors F (2, 122) = 11.68, P < .001. By analyzing pairs according to Scheffe, a significant difference was found only between the factor of '*Project* – as a reflection of values' – and the other two factors. That is, the items computed in this factor (14, 5, 1, 4, 22, 21) are seen as the most significant regarding the practical aspect of the *Project* component.

Besides the questions of practicality, teachers were asked about their years as EFL and matriculation teachers, their education, experience and their willingness to be involved in the future with the design and production of the *Project* component in different settings. Pearson correlations were computed to find relationships between these items and practicality factors. A significant correlation was found only in regard to the relationship between education and the factor of suitable conditions for the implementation of the *Project* component (R = -.22, P < .05). This negative correlation may suggest that the more years of education the teachers have, the less they see the practical factor of suitable conditions for *Project* work as important, since they have enough experience to overcome possible obstacles.

Additional questions related to the intent of the research participants to assess *Project*s as external raters and their willingness to guide students and/or colleagues in *Project* implementation. While 54% of the teachers answered that they were not willing to rate *Project*s externally, 41.2% of the teachers expressed willingness to guide students and peer-teachers in the process of the *Project* production.

In a one-way MANOVA analysis comparing the willing and non-willing teachers, a significant difference was found: F (3, 50) = 3.25, P

< .05. This significant difference was found only in regard to the factor of '*Project* – as a reflection of values'.

When examining the means in Table 5, it appears that the teachers who expressed willingness to rate *Project*s externally attach greater importance to the factor of *Project* – as a reflection of values –than the teachers who were not willing to rate externally.

Table 5: Mean and standard deviations of pragmatic factor showing teacher's willigness to become external Project raters

	Willing		Non-willing		
Factors	M	SD	M	SD	F (1,52)
Collection of scores and their meaningful use	3.04	.67	2.81	.82	1.33
Conditions for *Project* production	3.00	.50	2.85	.80	.71
Reflection of values	3.46	.37	3.03	.72	7.33**

**P<.01 n=62

Qualitative examination of the open-ended question in which teachers described what a *Project* is facilitated the formulation of the following categories: Teachers mainly regarded the *Project* as research work (43.9%) and as a piece written around a specific subject (26.8%). Additionally, teachers consider a *Project* as an integration of skills (24.4%). Reflection of curriculum methodology and assessment tools involving process work, a *Project* as a piece of work that is presented in an oral or written form and that should include group work all accrued the same level of importance (22.0%). The three categories of the *Project* as an answer to a research question (14.6%), moving from theory to practice (12.2%) and as an authentic way of learning (12.2%) all yielded medium averages.

Discussion

The first research question focused on the inter-rater reliability of the *Project* assessment component as it appears in the EFL matriculation exam. This psychometric quality is highly relevant to the examination of the *Project*'s quality since, according to the written format of the *Project* assessment guidelines (Steiner, 2003b) one teacher-tester alone assesses the *Project* according to rubrics provided by the MoE. Proving that this grade is reliable will provide meaningful information about specific students' performances converting the *Project* into a reliable source of information and confirming the reliability of the MoE rubrics.

The data from the application of the official rubrics show agreement between raters on the overall *Project* grade, but disagreement within the breakdown of the rubrics' various components. Additionally, discrepancies between groups of raters were detected. Since raters used their professional judgment, differences in opinion were inevitable. This supports the findings of Hamilton et. al. (1993) who call for more research on the validation of rating scales, a central issue for consistency in ratings and inter-rater reliability. Specifically, language performance rating scales are often lacking in clarity due to their large variety of components (Swain, 1993).

More specific conclusions can be drawn about the breakdown within the rubrics. The criterion of Content revealed greater agreement among raters. This is apparently because of the relatively straightforward criterion descriptors which define the number of components that should appear relating to coherent organization according to a table of contents. In the criteria of Individual Contribution, agreement is minimal. Raters could not relate to this criterion as there was no apparent way to measure the individual contribution to the final product. Additionally, raters found no reflection in the products that could help them assess the criterion of group work procedures.

Much of the rater commentary related to the criteria of Source and Language. In fact, the *Project* depends on student's ability to find topic-related sources and process the information into their own language. If students are unable to integrate sources into genuine summaries, the criterion of Language is liable to be affected. Moreover, as the discrepancy in the criterion of Source is also high, it seems that raters agreed that students did not process *Project* materials but downloaded them directly from the Internet or copied from books.

Two possible reasons may account for these findings. One relates to the insufficient descriptors in some of the criteria and the other, to the typical difficulties when rating pieces of work that are based on rater's best judgments. To gain objective results and overcome a possible lack in the competence of rating in a relatively subjective setting, raters must go through extensive training to gain trust and experience. (Bachman, 1990; Gipps, 1994; Linn, 1993). Introducing the rater into the assessment process is both necessary and problematic (McNamara, 2000). The assumption in most rating schemes is that if the rating category labels are clear and explicit, and if the rater is trained carefully to interpret them in accordance with the intentions of the test designers and concentrates while involved in the rating procedure, then the rating process can become objective. Yet rating remains intractably subjected to the rater's judgment.

The second phase of the study focuses on the analysis of the rubric used to evaluate the student *Project*s. In most cases, the reason for a particular grade was linked directly to the criterion and the numerical grade was given as a result of the written descriptors. Although none of the raters added a new criterion (only specifications to the criterion descriptors), the criteria are often lacking and call for more details and specifications. Some criteria (Effort and Individual Contribution) are more subjective than others and are difficult to measure as they depend on qualities which are hard to observe in a final written product. Additionally, the criterion of Individual Contribution also calls for substantial evidence of the assistance the class-teacher provides. These do not appear in the MoE guidelines for *Project* component (Steiner, 2003b). As the *Project* guidelines become less defining, the *Project* outcome becomes more difficult to assess. The rubric is used both as a summative assessment tool and as a formative set of instructions for students to use while working on the *Project* (Gordon et. al., 2002). If the class-teacher does not add specific instructions, the final product may become broad with insufficient information upon which to base the final grade. Three raters upheld their observation saying they couldn't rate the *Project*. "I can't give this a grade! I don't know what the class teacher asked them to do."

Analysis of the data obtained reveals that overall grade results show the highest links between raters. The raters mostly agreed on the final grades, however the disagreement among raters regarding the internal criteria and breakdown of the MoE rating scale was great.

The second research goal was to examine if the *Project* carries content validity as an assessment procedure. An obvious link between the *Project* guidelines and the reflection of their underlying content in the curriculum has not been fully determined and too many items in the Israeli curriculum (2001) are not indicated in the guidelines. According to the experts' findings, the rubric criteria are not defined explicitly enough. The rubric often leaves room for class teacher interpretation and elucidation. Although the *Project* requirements all appear in the Israeli curriculum, its specifications do not all appear in the guidelines. The *Project*, while inclusive, is only one task within the curriculum aiming to "steer the teaching of English as a foreign language in Israel" (2001:1) and pertains to the total setting of the teaching of English. Therefore, it includes more than the *Project* guidelines and as such must be entirely reflected in the curriculum. However, many queries have arisen about the *Project* guidelines and practicalities, showing them to be very complicated. To address this, a full coverage of both documents within each other must be guaranteed. To ensure content validity, the curriculum should be reflected in the *Project* guidelines in its full capacity and vice-versa.

This can be done through a one-to-one matching process, as the experts in this study have displayed.

Additionally, although the *Project* serves as part of the final assessment procedure and despite the emphasis that should have been put on assessment, the actual role of students in the assessment is minor (Expert 1).

The third research question revolves around the practicality and usefulness of the *Project* component. This question is important when considering the constraints teachers face when introducing a new element into their teaching, especially when this element is included in the high-stakes matriculation exam. Teachers define a *Project* as an in-depth research-based task which integrates skills and themes. Furthermore, they realize how it reflects the curriculum methodology. The factor analysis revealed a significant outcome: the participants value the impact of the *Project* for social and educational values more than factors dealing with conditions for *Project* implementation and the means for collecting *Project* scores. This finding is noteworthy, especially when considering the voices of disagreement raised when the EFL inspectorate first introduced the *Project* component.

Another point of interest is the relationship between the number of years of education and the factor of suitable conditions for *Project* implementation. The data reveals that the more educated the teachers are, the less they consider suitable conditions as meaningful in *Project* implementation. Experienced and well-trained teachers are able to conduct *Project* work even if the conditions in school are not conducive to practical and efficient execution.

In addition, the averages show that teachers who relate to the *Project* component as a reflection of values and as an opportunity for teaching educational goals also expressed willingness to rate *Project*s externally. Teachers who accept the importance of the *Project* are also willing to contribute more time and effort to its implementation beyond their job as a class teacher.

The uniqueness of the study relates to the context of the *Project* in which policymakers have explicitly defined the role of the external assessment as highly effective for introducing and implementing change. It aims to define the psychometric traits for assessing the *Project* in operative terms and examines the usefulness and feasibility of this specific assessment task.

Conclusions

In light of the findings, the study leads to diverse conclusions verifying Fulcher's description of rating scales (1996), claiming that some of the descriptors of the rubric are more defining and are sufficient to achieve inter-rater reliability, while others are lacking. Specifications should be added to those criteria where raters found it necessary to add their own, particularly to the subjective criteria of Effort, Language and Individual Contribution, all useful criteria for *Project* production processes. Alternatively, teachers must give students better guidelines how to document these aspects of the *Project* process. Another option is that these particular components should not be assessed by the class teacher alone, who actually observed the work process and heard the oral presentation of the material, but in addition by an external tester. Linn et. al., (1991) suggested several such criteria to judge the quality of complex, performance-based educational assessments: consequences, fairness, cognitive complexity, content quality and content coverage. Thus, we may wish to focus on test tasks that rely on a measure of internal consistency based on carefully defined criteria (Swain, 1993), and are more reliant on *Project* features expressing their complex communicative language behavior.

If reliability is defined as consistency among independent measures and that initial disagreement would provide an impetus for dialog (Moss, 1994), validity devoid of reliability cannot subsist.

Accordingly, the process of rating *Project*s should be unified by providing initial and ongoing training to raters. This recommendation is crucial to a high-stakes exam, where objectivity and unified grades are necessary to gain trustworthiness in the exam results.

Consequently, the findings support a re-thinking process to match the rubric criteria and descriptors explicitly to the curriculum, specifically the principles and benchmarks they are meant to assess, in order to achieve full content validity.

The first step to successful implementation of a new agenda is determining whether the program actually moved from an idea to initial implementation. Hence, evaluators should rely on empirical studies of real-world decision-making to increase their effectiveness and work with stakeholders to consciously reach an "active-reactive-adaptive" process (Patton, 1986).

From a different perspective, modern activity theory offers a productive framework for mapping and transforming the complexities of social practice in wide array of life settings (Coughlan & Duff, 1994). In the context of the *Project*, the term 'activity' brings together cognitive/communicative performance as it relates to and in part produces, its social-institutional context. Embedding effective learning and motivation is possible through the quality of the social framework and the activity carried out within the group. Furthermore, completing *Project*s serves as a framework that determines learning outcomes as beneficial for the learners in promoting their study and language skills.

To summarize, stakeholders want accurate, useful and relevant information, but only convey their real intentions when deciding how seriously to pay attention to an evaluation. The ideal is to find the right combination between truth and practicality. In this paper, although the *Project* is used as an assessment component of the curriculum and was added to the EFL matriculation exam, not all aspects of the assessment procedures were fully measured before starting the implementation. However, an 'active-reactive-adaptive' approach together with the Israeli EFL inspectorate efforts in matching the assessment to the curriculum, can add these missing features.

References

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice*. Oxford: University Press.

Brinton, D., Snow, M., & Wesche, M. (1989). *Content-based second language instruction*. New-York: Newbury House.

Coughlan, P., & Duff, P. (1994). Same task, different activities: Analysis of an SLA task from an activity theory perspective. In J. P. Lantolf & G. Appel (Eds.). *Vygotskian approaches to second language research*. (pp. 173-193). Norwood, NJ: Ablex.

Crooks, T.J. (2006). Excellence in assessment for accountability purposes. Keynote address presented at the Northumbria *EARLI SIG Assessment Conference*. 31 August.

David, J. L. (2008). What research says about project-based learning. *Educational Leadership*. 65 (5), 78-80.

Duvall, E. (2005). Beyond horseshoes and hand grenades: Making the case for applying the principles and practice of dynamic assessment to high-stakes standardized assessments for children with disabilities. Distinguished Paper Award. Presentation at the *Annual Pennsylvania Education Research Association Conference*, Hershey, PA, November.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing* 13(2), 208-38.

Gipps, C. V. (1994). Beyond testing. Towards a theory of educational assessment. London: Falmer Press.

Gordon, C., Kemp, J., Levi, T. & Toperoff, D. (2002). *Assessment Guidelines for the English Curriculum.* Jerusalem, Israel: Ministry of Education.

Hamilton, J., Lopes, M., McNamara, T. & Sheridan, E. (1993). Rating scales and native speaker performance on a communicatively oriented EAP test. Melbourne Papers. In: *Applied Linguistics.* Vol. 2. May, 1-24.

Hamilton, L. S., Stecher, B. M., & Klein, S. P. (Eds.). (2002). *Making sense of test-based accountability in education.* Santa Monica, CA: Rand Corporation.

Linn, R. L. (1993). Educational assessment: expanded expectations and challenges. *Educational Evaluation and Policy Analysis*, 15, 1.

Linn, R.L., Baker, E. & Dunbar, S. (1991). Complex, performance based assessment: expectations and validation criteria. *Educational Researcher*. 20(8), 15-21.

Mann, G., Shemesh, R., & Shlayer, J. (2002). Projects at work. *ETJ. English Teachers' Journal. Israel*. June 54, 18-29.

McNamara, T. (2000). Validity: testing the test. In: *Language Testing*. Oxford: Oxford University Press.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: implications for performance assessment. *Review of Educational Research*, 62, 229-258.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*. 23, 5-12.

Nevo, D., & Shohamy, E. (1986). Evaluation standards for the assessment of alternative testing methods: an application. Studies. In: *Educational Evaluation*. 12, 149-158.

Patton, M. Q. (1986). *Utilization-focused evaluation*. Sage Publications.

Seliger, H. W., & Shohamy, E. (1989). *Second language research methods*. Oxford: Oxford University Press.

Shohamy, E. (1991). International Perspectives of Language Testing and Systems and Policy. ACTFL Annual Review of Foreign Languages. *The American Council on the Teaching of Foreign Languages*. 16, 91-107.

Spolsky B., Ben Meir, D., Inbar, O., Orland, L., Steiner, J., & Vermel, J. (2001). *English Curriculum for all grades: Principles and Standards for Learning English as a Foreign Language*. Jerusalem, Israel: Ministry of Education.

Steiner. J. (2002a). *Bulletin of the Chief Inspector for English Language Education*. Ministry of Education October. Issue 1.

Steiner, J. (2002b). *Bulletin of the Chief Inspector for English Language Education*. Ministry of Education. December. Issue 2.

Steiner, J. (2003a). NBA bulletin 4 - More FAQ's. *Bulletin of the chief inspector for English language education.* Ministry of Education, February Issue 5.

Steiner, J. (2003b). Bulletin of the Chief Inspector for English Language Education. Ministry of Education. April, Issue 6.

Steiner, J. (2003c). *The NBA handbook*. Jerusalem: Ministry of Education.

Stoller, F. L. (1997). Project work: a means to promote language content. *English Language Teaching Forum.* October. Vol. 35, 4.

Swain, M. (1993). Second language testing and second language acquisition: is there a conflict with traditional psychometrics? *Language Testing*. 10, 193-210.

Wiggins, G. & McTighe, J. (1999). *The understanding by design handbook*. Association for Supervision and Curriculum Development, Alexandria: VA.

Appendix 1:
Project Rubrics: Suggested Rubric for Assessing the Written Group and Individual Presentation (From 80 points)-Source: NBA Handbook 2003

Criteria	Number of points				
Content	25	20	15	10	5
	Includes all the required components. Content is clear, well-organized, written in students' own words.		Includes five of the required components. Content is mostly comprehensible, organized and is generally written in students' own words.		Includes three or less of the required components. Content is not clear, lacks organization and is not written in students' own words.
Sources	10	7	5	2	0
	Includes the required number of sources. Sources are relevant to the topic, are integrated into the Project, and are cited correctly.		Includes less than the required number of sources. Some of the sources are relevant to the topic, are integrated into the Project, and are cited correctly on the whole.		Did not mention sources used or did not cite correctly
Language	15	12	9	6	3
	Correct sentence structure. Rich and varied vocabulary. Correct spelling and punctuation.		Correct sentence structure most of the time. Appropriate vocabulary. Correct spelling and punctuation most of the time.		Incorrect sentence structure most of the time. Limited vocabulary. Frequent mistakes in spelling and punctuation.

Effort	10	8	6	4	2
	Clear evidence of investment of time and effort. Work is aesthetic.		Some evidence of investment of time and effort. Work is fairly aesthetic.		Little evidence of investment of time and effort. Work is not aesthetic.
Individual Contribution	20	16	12	8	4
	Actively participated and contributed to the group. Carried out tasks s/he was responsible for on time.		Sometimes participated and contributed to the group. Carried out only some of the tasks s/he was responsible for and/or not always on time.		No participation or contribution to the group. Did not carry out the tasks s/he was responsible for.

Appendix 2:

Commentary and clarifications added by raters to official rating scale, turned into categories.

Criteria	Added commentary and clarification as noted by teachers
Content	Cover page is missing- 3 Lacks organization-2 Downloaded from internet-5 Not in student words-7 No table of contents-2 Lacks content-5 Required components are missing-10 Didn't include all components but still did a nice job-2 No bibliography-2
Sources	Lacks content-1 Downloaded from internet-2 Not cited-3 Positive feedback-3 No written bibliography-8 Inappropriate bibliography-5 No evidence of class-teacher requirements regarding sources-1 Too many sources-1
Language	Downloaded from internet-17 Highly structured. Raises suspicions regarding authenticity-2 No summarizing-1 Some mistakes-1 No evidence of drafts-2 Typed in capital letters-1
Effort	The length is enough-1 Positive feedback-1 Most was copied or downloaded-5 Admission of labor-1 No processing of internet sources-2 No drafts-2 Not typed-1 Stapled in Hebrew direction-1

Individual contribution	No evidence of process work-6 No evidence of groupwork-8 Very hard to see. No teacher requirements or guidelines-2 No reflections-12 Too many students working on one Project-1 Reflections are superficial-2 Some students contributed very little-3 Work seems downloaded-1 Positive feedback-1 Summarizing clearly seen. Related to chekeclist-1

*Number of times specific comment was mentioned by different raters