Essays and Related Writing Tasks for Language Testing

Rosanna Tammaro, University of Salerno, Italy
Anna D'Alessio, University of Salerno, Italy

The IAFOR International Conference on Education - Dubai 2015
Official Conference Proceedings

Abstract
For many reasons writing is one of the foundational skills of educated persons. Tests of writing skills are therefore needed. The intent of this paper is to discuss and explore traditional essay techniques and suggests both objective and subjective scoring methods. It is recognized that the problem of quantifying essay tasks is a crucial difficulty in school applications. A method of interpreting learner protocols with a view toward helping learners to overcome language difficulties is focused. Though essay testing may require more work of the teacher and of the students than many other testing procedures, it is considered to be a profitable assessment technique. Certainly it affords a rich yield of diagnostic information concerning the learner's developing expectancy grammar. One of the most often used and perhaps least understood methods of testing language skills is the traditional essay or composition. The freedom allowed by essay tasks may both their greatest strength and weakness- a strength because they require a lot of the examinee, and a weakness because of the judgement required of the examiner. Except for the greater accessibility of the written protocols of learners, the evaluation of writing performances is similar to the evaluation of spoken protocols. The fundamental problem in using essay tasks as tests is the difficulty of quantification - converting performances in scores.

Keywords: language, essay, task, writing skills.

Introduction

Tests in education should be purposefully related to what the schools are trying to accomplish. For this to be so, it is necessary to carry the validation of testing techniques beyond the desk, chalkboard, and classroom to the broader world of experience.

Composition or essay writings are most free writing tasks necessarily qualified as pragmatic tests. Because it is frequently difficult to judge examinees relative to one another when they may have attempted to say entirely different sorts of things, and because it is also difficult to say what constitutes an error in writing, various modified writing tasks have been used. For example, there is the so-called dehydrated sentence, or dehydrated essay. The examinee is given a telegraphic message and is asked to expand it.

Writing tasks may range from the extreme case of allowing examinees to select their own topic and to develop it, to maximally controlled tasks like filling in blanks in a pre-selected (or even contrived) passage prepared by the teacher or examiner. The blanks might require open-ended responses on the order of whole paragraphs, or sentences, or phrases, or words. In the last case, we have arrived back at a rather obvious form of cloze procedure.

Another version of a fairly controlled writing task involves either listening to or reading a passage and then trying to reproduce it from recall. If the original material is auditorily presented, the task becomes a special variety of dictation. This procedure and a variety of others are discussed in greater detail in this article.

For many reasons writing is one of the foundational skills of educated persons. Tests of writing skills are therefore needed. This article explores traditional essay techniques and suggests both objective and subjective scoring methods. It is recognized that the problem of quantifying essay tasks is a crucial difficulty in school applications. A method of interpreting learner protocols with a view toward helping learners to overcome language difficulties is discussed. Though essay testing may require more work of the teacher and of the students than many other testing procedures, it is considered to be a profitable assessment technique. Certainly it affords a rich yield of diagnostic information concerning the learner's developing expectancy grammar.

Essays : technique for evaluating the conformity of a text to normal written discourse

One of the most often used and perhaps least understood methods of testing language skills is the traditional essay or composition. Usually a topic or selection of topics is assigned and students are asked to write an essay of approximately so many words, or possibly to write an essay within a certain time limit. Sometimes, in the interest of controlling the nature of the task, students may be asked to retell a narrative or to expand on or summarize an exposition. For example, the student might be asked to discuss the plot of a novel or story, or to report on the major themes of a non-fictional book or passage of prose.

Essays are probably used so frequently because of the value placed on the ability to organize and express our ideas in written form. This ability is counted among the most important skills imparted by any educational System. It is not known to what extent the acquisition of writing skill carries over into other aspects of language use - e.g., being an articulate speaker, a good listener, or an insightful reader - but there is proof that all of these skills are substantially interrelated. Moreover, there is much evidence to suggest that practice and improvement in one skill area may rather automatically result in a corresponding gain in another skill area. There is even some suggestion in the data from second language learning that practice in speaking and listening (in real life communication) may have as much of an effect on reading and writing as it does on speaking and listening and conversely, pragmatic practice in reading and writing (where communication is the goal) may affect performance in speaking and listening at least as much as reading and writing.

However, in spite of their popularity, there is a major problem with free composition tasks as tests. It is difficult to quantify the results -i.e., to score the protocols produced by examinees. We will see below that reliable methods of scoring can be devised, and that they are only slightly more complex conceptually than the methods suggested for the scoring of dictation protocols. Perhaps we should not stress the difficulty of obtaining reliable scores, but rather the ease and likelihood of obtaining unreliable (and therefore invalid) scores on essay tasks.

Tasks that require the production of sequences of linguistic material that are not highly constrained all present similar scoring problems. For instance, a dictation/composition task, or hear and retell task, or a creative story telling task, and all similar production tasks whether written or oral need persistent scoring problems. The oral tasks require a further step that is not present in the scoring of written protocols - namely transcribing the protocols from tape or scoring them live (an even more difficult undertaking in most cases). Further, the scorer must in many cases try to infer what the examinee intended to say or write instead of merely going by what appears in the protocol.

We are concerned with defining methods of scoring essay tasks in particular, and productive language tasks in general. It is assumed that essay tasks are fundamentally similar to speaking tasks in certain respects - namely in that both types of discourse processing usually presuppose someone's saying (or writing) some-thing for the benefit of someone else. If the writer has nothing to say he is in very much the same boat as the speaker holding forth on no particular topic. If the writer has something to say and no prospective audience, he may be in the position of the child describing its own performances with no audience in mind, or the adult who is said to be thinking out loud.

Of course, these parallels can be drawn too closely. There are important differences between acts of speaking and acts of writing as any literate person knows all too well. The old saw that a person should write as he would speak, or the popular wisdom that unclear writing is the result of unclear thinking, like all forms of proverbial wisdom require intelligent interpretation. Nonetheless, it is taken for granted here that much of what concerning productive oral testing is applicable to writing tasks and need not be repeated here, and conversely that much of what is suggested here concerning the

scoring of written protocols (especially essays) can be applied to transcriptions of oral protocols or to tape recorded oral performances.

Just as it is possible to conceive of non-pragmatic tests of other sorts, it is also possible to invent testing procedures that may require writing, but which are not in any realistic sense pragmatic. Sentence completion tasks, for instance, do not necessarily involve pragmatic mapping onto extra-linguistic context, and they do not generally involve higher order discourse constraints ranging beyond the boundary of a single sentence. A typical school task that requires the utilization of words in sentences (e.g., as a part of a spelling exercise) would not qualify as a pragmatic task. Neither would a written transformation task that requires changing declarative sentences into questions, passives into actives, present tense statements into past tense statements, etc. In general, any manipulative exercise that uses isolated sentences unrelated to a particular pragmatic context of discourse does not constitute a pragmatic writing task.

The key elements that must be present in order for a writing task to qualify are similar to those for speaking tasks. The writer must have something to say ; there must be someone to say it to (either explicitly or implicitly) ; and the task must require the sequential production of elements in the language that are temporally constrained and related via pragmatic mapping to the context of discourse defined by (or for) the writer. Probably such tasks will be maximally successful when the writer is motivated to write about something that has personal value to himself and that he would want to communicate to someone else. Contrived topics and possibly imagined audiences can be expected to be successful only to the extent that they motivate the writer to get personally (albeit vicariously) involved in the production of the text. An unmotivated communicator is a notoriously poor source of information. To the degree that a task fails or succeeds in eliciting a highly motivated performance, it will probably fail or succeed in eliciting valid information about the writing ability of the examinee.

We will consider writing tasks to be pragmatic if they relate to contexts of discourse that are known to the writer and that the writer is attempting to make known to the reader. Protocols that meet these requirements have two important properties that disjointed sentence tasks do not have. First, as Rummelhart (1975) points out, 'Connected discourse differs from an unrelated string of sentences ... in that it is possible to pick out what is important in connected discourse and summarize it without seriously altering the meaning of the discourse' (p. 226). Second, and for the same reasons, it is possible to expand on a discourse text and to interpolate facts and information that are not overtly stated. Neither of these things is possible with disjointed strings of sentences.

Writing about a poignant experience, a narrow escape, recollections of childhood, and the like all constitute bases for pragmatic essay tasks. Of course, topics need not focus on the past, not even on what is likely. They may be entirely fictional predicated on nothing but the writer's creative imagination. Or, the writing may be analytical or expository. The task may be to summarize an argument; retell a narrative; recall an accident; explain a lecture; expand on a summary; fill in the details in an incomplete story; and so on.

There really is no limit to the kinds of writing tasks that are potentially usable as language tests. There is a problem, however, with using such tasks as tests and it has to do with scoring. How can essays or other writing tasks be converted to numbers that will yield meaningful variance between learners? Below we consider two methods of converting essay protocols and the like to numerical values. The first method involves counting errors and determining the ratio of the number of errors to the number of opportunities for errors, and the second technique depends on a more subjective rating System roughly calibrated in the way the FSI rating scales for interview protocols are calibrated.

Difficulty of converting performances to scores: the fundamental problem in using essay tasks as tests.

It is possible in scoring essays to look only at certain so-called grammatical 'functors'. It would, however, be a discrete point scoring method. For instance, the rater might check certain morphemes such as plurals, tense indicators, and the like on 'obligatory occasions'. The subject's score would be the ratio of correct usages to the number of obligatory occasions for the use of such functors. This method, however, does not necessarily have any direct relationship to how effectively the student expresses intended meanings. Therefore, it is considered incomplete and is rejected in favor of a method that focusses on meaning (Evola, Mamer, and Lentz, in press).

To score an essay for its conformity to correct prose, it is first necessary to determine what the essay writer was trying to say. In making such a determination, there is no way to avoid the inferential judgment of someone who knows the language of the essay. Further, it helps a great deal if the reader studies what is said in its full context. Knowledge of the topic, the outline of the material, or any other clues to intended meanings may also be helpful.

Once the reader has developed a notion of what the writer had in mind, it is possible to assess the degree of conformity of what the author said to what a more skilled writer might have said (or to what the text actually said when the task is recall).

A simple method that can be used with essay scoring is to restate any portion of the protocol that does not conform to idiomatic usage or which is clearly in error. The error may be in saying something that does not express the intended.

“ Rewriting the protocol to make it conform to standard usage and also to express the intended meaning of the author may be difficult, but it is not impossible, and it does provide a systematic basis for evaluating the quality of a text. Furthermore, rewriting an essay to express the intended meanings (insofar as they can be determined by the scorer) requires evaluation not just in terms of how well the text conforms to discrete points of morphology and syntax, but how well it expresses the author's intended meanings. There is guesswork and inference in any such rewriting process, but this reflects a normal aspect of the interpretation of language in use whether it is spoken or written. Indeed, the difficulties associated with making the right guesses about intended meanings should reflect the degree of conformity of the essay to normal clear prose. Hence, the guessing involved is not a mere artefact of the task but reflects faithfully the normal use of language in communication” (J.w. Oller, p.386). Once the rewriting has been carefully completed - deleting superfluous or extraneous

material, including obligatory information that may not have been included in the protocol, changing distorted forms or misused items, and so forth - a score may be computed as follows : first , count the number of error-free words in the examinee's protocol; second, subtract from the number of error-free words, the number of errors (allowing a maximum of one error per word of text) ; third, divide the result of step two by the number of words in the rewritten text. These steps can be expressed simply as shown in the following formula:

ESSAY SCORE = [(the number of error-free words in the student's protocol) minus (the number of errors in the student's protocol)] divided by (the number of words in the rewritten text)

In a classroom situation, there are many reasons why the rewriting of each learner's protocol should be done in consultation with the writer of the essay. However, due to practical considerations, this is not always possible. Nonetheless, whether the procedure can be done consultatively or not, rewriting the text to make it express the scorer's best guess as to the intended meaning and to make it do so in idiomatic form is apt to be a much more useful procedure pedagogically than merely marking the errors on the student's paper and handing the paper back. In the latter case, the attention is focused on the surface form of the language used in the essay, in the former, the attention is focused on the intended meaning and the surface form is kept in proper perspective - as a servant to the meaning.

In general there are three types of errors. There are words that must be deleted from the student's protocol; there are words that must be added; and there are words that must be changed.

The point of the scoring method is to provide a technique of converting essay protocols to a quantity that reflects (in the view of at least one proficient reader) the degree of conformity of those protocols to effective idiomatic prose.

In brief, the method probably works as well as it does because it very clearly assesses the overall conformity of the student's writing to idiomatic prose.

Just as in the case of dictations (and cloze tasks) spelling errors are counted only when they distort a word's morphology or pronunciation. In the above protocols, punctuation errors are corrected, but do not contribute to the total score of the examinee. This is not to suggest that punctuation and spelling are not important aspects of writing skill, but for the same reasons that spelling is not scored in dictation these mechanical features are not counted in the essay score either. Because of the results with non-native speakers of English, it is assumed that learning to spell English words and learning to put in appropriate punctuation marks in writing are relatively independent of the more fundamental problem of learning the language. Research is needed to see if this is not also true for native speakers of English(J.w. Oller, p.388).

Many scoring methods besides the one exemplified could be used. For instance, if for whatever reason the examiner wanted to piece a premium on quantity of output, the examinee might be awarded points for errorless words of text. The score might be the number of errorless words of text written in a certain time period.(Brière,1966) even

argued that the mere quantity of output regardless of errors should be considered. In an experimental study he claimed to have shown that learners who were encouraged to write as much as they could as fast as they could within a time limit learned as much as students who received corrective feedback (i.e., whose papers were marked for errors). However, from a testing point of view, it would have to be shown that a mere word count would correlate with other presumed valid measures of writing skill. The best essay is not the longest.

On the whole it would seem best to use a scoring method that awards positive points for error-free words and subtracts points for errors. To keep the focus of the examinee and the scorer on the intended meanings and the clear expression of them, some attention should be paid to the amount of deviation from clear prose in any given attempt at written expression. The scoring method proposed above reflects all of these considerations.

Evaluation content and organization

It has long been supposed that subjective ratings were less accurate than more objective scoring methods. However, as we have seen repeatedly, subjective judgments are indispensable in decisions concerning whether a writer has expressed well his intended meaning and, of course, in determining what that intended meaning is. There is no escape from subjective judgment in the interpretation of normal expression in a natural language. In fact, there are many reasons to suppose that a trained judge may be the most reliable source of information about how well something is said or written. The crucial question in any appeal to such judgments is whether or not they are reliable - that is, do different judges render similar decisions concerning the same samples of writing, and do the same judges render similar decisions concerning the same samples of writing on different occasions. The question of the reliability of ratings is the same whether we are thinking of written or spoken (possibly recorded) samples of language.

More recent work with oral ratings and with the evaluation of written protocols has indicated that even untrained raters tend to render fairly reliable judgments though trained raters do still better. Although Mullen found substantial variability across judges in the calibration of their evaluations, reliability across judges (that is, the correlation of the ratings of the same subjects by different judges) was consistently high.

Essay tasks have often been favored as classroom testing techniques. Educators sometimes appeal to them as a kind of ultimate criterion against which other tests must be judged. However, the greatest virtue of essay tasks may also be their greatest liability. While it is true that such tasks confer the freedom and responsibility of choice on the examinee, they also require thoughtful evaluation on the part of the examiner. The writer may elect to express very simple ideas only in words that he is sure of, or he may venture into more complex or profound meanings that stretch his capacity to put things into words. Whether the writer has charted a conservative or a daring course is left to the judgment of the examiner. For this reason, equal scores may not represent equivalent performances, and unequal scores do not necessarily imply that the higher score represents the better performance.

The main question in interpreting an essay protocol is, 'What was the writer trying to say, and how well did he say it?' There may seem to be two questions here, but for good reasons readers tend to treat them as one. If a writer does not express things fairly well, it will be hard to tell what he is trying to say; similarly, if a writer has little or nothing to say, how can he say it well? It does not take a sage to see the wisdom of saying nothing unless there is something to say. In fact, a person has to go to school for a very long time to become able to write on topics which do not naturally elicit a desire to communicate (though this is not because people lack things to talk and write about).

Once the evaluator is fairly confident that he knows what the writer was trying to express, it is possible to evaluate how well the job was done. Obviously, the evaluator cannot be much better at evaluating than he himself is at writing and this is where the teeth of subjectivity bite hard. The examiner must piece himself in the shoes of the writer and try to figure out precisely what the writer meant and whether he said it well, or how he could have said it better. Thus, the largest part of evaluating essays or other written protocols is inferential, just as the interpretation of speech and writing in other contexts is inferential.

In spite of the criticism that essay tasks allow the writer the freedom to avoid 'difficult structures', such tasks are nonetheless usually quite revealing. A number of problems can be diagnosed by studying a protocol such as number 1 above. Among the glaring errors is the failure to use the definite article where it is required in such expressions as 'The driver of yellow car'. The surface aspect of this error lies in the fact that any noun phrase with a countable head noun (such as 'car' or 'yellow car') requires an article. The deeper aspect of this same error is that without the article, in fact without a definite article, the writer fails to call attention to the fact that the reader knows which yellow car the writer is referring to - namely, the one that ran the red light. If the writer can be made to see this, he can learn to use the article correctly in such cases.

Another noun phrase problem occurs in the phrase 'many damages'. Here, the trouble is that we normally think of damage as a kind of amorphous something which is uncountable. The word 'damage' does not usually appear in the plural form because the referent is not discrete and countable. There can be a little damage or a lot of damage, but not few or many damages. The reader should note that this has no more to do with the syntax of countable and uncountable nouns than it has to do with the concept of what damage is. If we were to conceptualize damage differently, there is nothing in the grammar of the English language that would keep us from counting it or saying 'many damages'. The problem involves meaning and goes far beyond 'pure syntax' even if there were such a thing. If the writer can see the sense or meaning the way the native speaker does, this kind of error can be overcome (J.w. Oller, p.390).

The writer confuses 'give' and 'take'. This distinction has puzzled if not bewildered many a grammarian, yet native speakers have little or no difficulty in using the terms distinctly and appropriately. In this context, 'give' is required because the police were there at the scene. There was no need to take the ticket anywhere. The policeman just handed it to the driver who was presumed to be at fault.

Two other errors involved clause connectors of sorts. The writer says, 'The blue car had a lot of damage that estimated about five hundred dollars'. The problem the writer needs to be made to see is that the way he has it written, the damage is doing the estimating. Another difficulty in joining clauses together appropriately, occurs in the last sentence of the protocol where the learner avows, 'Though I was late getting home that day, but I had an interesting story to tell my parents'. Perhaps the learner does not realize that 'though' and 'but' in this sentence both set up the condition for another statement offering contrary information. In the case of 'but' the contrast has a kind of backward look whereas with 'though' an anticipatory set is cued for information yet to come.

How can the diagnostic information noted above be applied? Surely it is not enough to offer grammatical explanations to the learner. In fact, such explanations may not even be helpful. What can be done? For one thing, factual pattern drills may be used.

The writer knows what the facts are. The deep structure, or the sense, is known. It is how to express the sense in terms of surface structure that the writer needs to discover. The drills should be designed to help the learner see how to express meanings by using certain surface forms.

Much remains to be discovered concerning the nature of writing and the discourse processing skills that underlie it. However, there is little reason to doubt the utility of essay writing as a reasonable testing procedure to find out how well learners can manipulate the written forms of a language. In fact, there are many more arguments in favor of essay writing as a testing technique than there are against it. Its usefulness as a diagnostic tool for informing the organizers of a curriculum cannot be overlooked. Further, such tests can easily (though not effortlessly) be integrated into a sensible curriculum.

Conclusion

The freedom allowed by essay tasks may be both their greatest strength and weakness - a strength because they require a lot of the examinee, and a weakness because of the judgment required of the examiner. Except for the greater accessibility of the written protocols of learners, the evaluation of writing performances is similar to the evaluation of spoken protocols. The fundamental problem in using essay tasks as tests is the difficulty of quantification - converting performances to scores.

A technique for evaluating the conformity of a text to normal written discourse is for a skilled writer (presumably the language teacher) to read the text and restate any portions of it that do not seem to express the author's intended meaning, or which do not conform to idiomatic usage.

For instructional value and also to insure the most accurate possible inferences concerning intended meanings, such inferences are best made in consultation with the author of the text in question. Something to say and the motivation to say it are crucial ingredients to pragmatic writing tasks. A recommended scoring procedure is to count the number of error-free words in the text ; subtract the number of errors ; and divide the result by the number of words in the error-free version of the text. Research has shown that subjective ratings of written protocols are about as reliable as

objective scoring techniques and that the subjective methods generate about as much valid test variance as the objective techniques do.

Research has also shown that attempts to direct attention toward presumed components or aspects of writing ability (e.g., vocabulary, grammar, organization, content, and the like) have generally failed. Apparently, whatever is involved in the skill of writing is a relatively unitary factor that does not lend itself easily to componential analysis. It must be noted that the reliability of ratings of essays by different judges is only marginally related to the calibration of the ratings - it is principally a matter of whether different judges rank subjects similarly. Judges that differ widely in the specific ratings they assign to a set of protocols may agree almost entirely in what they rank as high and what they rank as low in the set.

A given learner's protocol may serve as a basis for an in depth analysis of that learner's developing grammatical System. Further, it may provide the basis for factual pattern drills designed to help the learner acquire difficult structures and usages.

Indeed, factual pattern drills, derived directly from the facts in the contexts can serve as a basis for preparing materials for an entire class or a whole curriculum to be used in many classes. Essays are reasonable testing devices that have the advantage of being easily incorporated into a language curriculum.

References

Brière E. (1966). *Quantity before quality in second language composition*. Language Learning, Volume 16, Issue 3-4, pages 141–151

Evola, J., Mamer, E., & Lentz, B. (1980). Discrete point versus global scoring for cohesive devices. In J. W. Oller & K. Perkins (Eds.), *Research in language testing* (pp. 177-181). Rowley, MA: Newbury House.

Oller, J.W Jr. (1979). *Language Tests at School*. New York: Longman
Rummelhart, D.E. (1975). *Notes on a schema for stories*. New York:Academic Press

Willliam E. Coffman (1972). The validity of Essay Tests in Glenn H. Bracht, Kenneth D. Hopkins, and Julian C. Stanley (eds.), *Perspectives in Educational and Psychological Measurement Englewood Cliffs*. New Jersey:Prentice-Hall

Celeste M. Kaczmarek, (in press). Scoring and Rating Essay Tasks in J. W. Oller and Kyle Perkins (eds.) *Research in Language Testing. Rowley*. Mass.: Newbury House

Nancy Martin, Pat D'Arcy, Bryan Newton and Robert Parker (1976). *Writing and Learning across the Curriculum 11-16*. Leicester, England: Woolaston Parker

Mullen, K.A., (1980). Evaluating writing proficiency in ESL. In: Oller Jr., J.W., Perkins, K. (Eds.), *Research in Language Testing*. Newbury House Publishers Incorporated, MA, pp. 160–170.

Contact email: rtammaro@unisa.it; dalessioanna1@libero.it.