1993

Developing a measure of student literacy competencies at a tertiary level using Rasch measurement

Barry Sheridan

Les Puhl

EDITH COWAN UNIVERSITY

PERTH WESTERN AUSTRALIA

DEVELOPING A MEASURE OF STUDENT LITERACY COMPETENCIES AT A TERTIARY LEVEL USING RASCH MEASUREMENT

Barry Sheridan and Les Puhl

RESEARCH REPORT No: 3

June 1993

MEASUREMENT ASSESSMENT and EVALUATION LABORATORY

DEVELOPING A MEASURE OF STUDENT LITERACY COMPETENCIES AT A TERTIARY LEVEL USING RASCH MEASUREMENT

Barry Sheridan and Les Puhl

RESEARCH REPORT No: 3

June 1993

Measurement Assessment and Evaluation Laboratory
Edith Cowan University

Paper presented at the National Council for Measurement in Education Meeting
Atlanta, April 1993

ACKNOWLEDGEMENTS

The authors are grateful for the assistance offered by
Professor David Andrich, Murdoch University
during the conduct of the research featured,
and for a critical review of the draft revision of,
this report

Developing a Measure of Student Literacy Competencies at a Tertiary Level using Rasch Measurement

Abstract

This paper reports an Australian investigation into concerns about student writing at University level and the construction of an objective measure of literacy. The English Skills Assessment (ESA) test, involving multiple choice items, and an essay, marked according to specified criteria, was administered to newly enrolled students (N = 495) in a University's Education programme. Analyses reveal inconsistencies between subtests of the ESA test, but the written test shows more promise. While some association is observed between the two measures, comparable subtests (such as spelling, punctuation, sentence structure) do not appear to be measuring the same thing.

Key words: literacy, measurement, Rasch, tertiary

Developing a Measure of Student Literacy Competencies at a Tertiary Level using Rasch Measurement

Introduction

This study investigates the psychometric properties of a direct measure of literacy used by the Faculty of Education at the Edith Cowan University in Western Australia and assesses its validity in association with the English Skills Assessment (ESA) test. The purpose of the study is to produce outcomes which will contribute to the development of an ecologically valid measure of the literary competencies of students at a tertiary level. An ecologically valid measure, according to the researchers, would be both theoretically sound from the perspective of literacy and measurement theory, and also meet the exigencies of the context in which the measure is administered. In short, the study aims to contribute to the production of an effective measure of literacy which is cost effective and not too time consuming to administer and score on the one hand, and reliable and valid on the other. An important feature of this study is that it involves the use of a measurement model which draws on recent developments in psychometric theory. The employment of this model makes it possible to construct variables which assist in the development of both an understanding of the nature of literacy as a variable and the creation of an objective measure of literacy. The special ability of this measurement model to handle multiple scoring categories associated with both the different subtests of the ESA test and the different aspects identified as comprising the literacy variable is a significant feature of this study.

The Problem

Anecdotal evidence indicates that there is a perception, among Faculty of Education staff at Edith Cowan University, that the standards of written literacy of many of its students are inadequate. The teaching of English has received considerable criticism in the media because of the apparent inability of secondary schools to produce students who understand basic concepts in English such as spelling, grammar, and punctuation (Brandeth, 1988; Back to Basics, 1988; Brock, 1990). The criticism levelled at secondary schools has in turn been blamed on the low–level literacy competencies of teachers (Australian Education Council, 1990). A literacy problem may be manifest more in the case of students enrolling in teacher education courses than in other courses at tertiary institutions because teacher education students have amongst the lowest scoring–profiles in tertiary entrance examinations conducted throughout Australia (National Board of Employment, Education and Training, 1990; Prichard, 1990). Prospective students in teacher education programmes in the United States appear to reflect similar characteristics. The problems are of such a magnitude that many of the faculties of education in Universities in the United States either run across–the–board literacy programmes for all incoming students or run mandatory remedial writing programmes for the students diagnosed by the faculties as being at risk (Duke, 1985; Carpenter & Johnson 1990).

The Faculty of Education's response at Edith Cowan University to the perceived inadequacy of the written literacy standards of education students is to assess the literacy performances of the entire student cohort using testing procedures which they believe provide adequate measures of literacy. These procedures have included administering the English Skills Assessment (ESA) test (ACER, 1982b), and analysing written products such as essay assignments or essays written specifically for this purpose under test conditions. Because of the time factor involved in assessing every student enrolled, the Faculty has until recently favoured the use of the ESA test because of the ease of scoring its multiple choice format. However, the limited scope of a multiple choice test to assess many of the underlying knowledge structures contributing to literacy is seen by many to be a severe shortcoming of the ESA test. Staff are concerned that there are important structural and organisational inadequacies in students' writing which are not identified by this type of test and, therefore, are not attended to by students. As a consequence, the ESA test is not perceived to be a valid measure of literacy.

The staff's concern about the validity of the ESA test as a measure of literacy is supported by the literature. This literature indicates that the task of writing is far more complex than is suggested by the writing related skills measured by indirect measures such as the ESA test. A synthesis of recent literature dealing with cognitive process models for writing (Berieter & Scardamalia, 1983, 1985; Flower, 1989a, 1989b; Flower & Hayes, 1981; Hayes & Flower, 1980, 1983; Stein, 1985; Stotsky 1990) indicates a need to go far beyond a conceptualisation which views writing as simply a product which can be assessed by examining the surface features such as spelling, punctuation, and grammar. Nightingale (1988), Parry (1989), and Taylor and Nightingale (1990) identified the lack of understanding of the underlying content, and the structures and organisational formats needed to express that understanding, as being the problem rather than the mechanics of tertiary students' writing. Lack of purpose and poor organisation result in incoherent writing (McCulley, 1985). In contrast technical errors, such as spelling and punctuation, rarely cause entire essays to be incoherent.

An Australian study comparing the performances of 226 tertiary–level students in the ESA test with their performances in the writing of a narrative was carried out by Holbrook and Bourke (1989). This study showed that the types of errors made by students in the categories measured by the ESA test did not necessarily equate to the kinds of errors made by students in the same categories when writing a narrative, for example, students who performed poorly in the spelling component of the ESA test did not necessarily manifest poor spelling in their narrative writing.

The ESA test appears, therefore, to be flawed from the perspective of writing theory, and the research evidence also shows that it may be inadequate for the tasks for which it is used. In addition, the alternative direct measures trialed by the Faculty of Education, whilst appearing to reflect more theoretically adequate conceptualisations of literacy as described in the literature, have not been assessed satisfactorily in terms of their internal validity and consistency. To address this specific issue as well as the broader aspects of what might constitute an objective measure of literacy, the present study was undertaken within the University.

Methods and Techniques

The Design

The present study follows the general approach reported in Australia by Holbrook and Bourke (1989) but extends considerably their conceptualisation and measurement techniques by adopting a measurement model capable of assessing objective measurement in a way not covered by the earlier study.

An immediate cause for concern with the ESA test is the use of the multiple choice format as an end in itself. This is a method of test design that does not enjoy a lot of support today due to the limited range of scores available. In addition, there is a growing concern regarding lack of attention to conditional, or local, independence between dichotomous test items (see, for example, Andrich, 1985b; Rosenbaum, 1988; Wilson & Adams, 1992). In its present form, the issue of conditional independence is not addressed by the ESA test Manual of Procedures (ACER, 1982a), despite the fact that the conceptual framework for this test identifies different subgroups within each of the subtests comprising the instrument. By examining the individual dichotomous item responses collectively as subsets, or item bundles, conditional independence can be addressed and accounted for with the measurement model employed. Use of the ESA test in its original direct dichotomous format is, therefore, not the best way to investigate the capabilities of this test as a measure of literacy because this approach is unable to assess the conceptual framework as presented in the Manual of Procedures. Andrich (1985b) provides a detailed discussion of the strategy employed for this analysis.

By combining the dichotomous responses of the ESA test into the respective subgroups as specified in the Manual, it is possible to address the relevance of the individual subgroups as aspects of the global variable identified here as literacy. This technique also applies to the structure of the essay assignment with its emphasis on extended writing, and referred to in this paper as the Assignment–Essay Marking Key (AEMK) test. As the formats for both tests now feature the same basic structure, individual analysis outcomes can be compared in a logical manner. At the same time, subgroups from both tests can be examined in combination and any association between presumed areas of similarity assessed accordingly. The latter technique involving the degree of association between both tests provides a powerful means of establishing the construct validity of the tests as measures of literacy.

An advantage of using the above technique is that it allows for a more parsimonious approach for understanding the meaning of the variable measured by the tests. Of special interest to this study is the opportunity to assess the contribution of both *surface features* and *deep structures* to an understanding of what the ESA test is measuring. The former refers to aspects such as spelling, punctuation, and vocabulary, whilst the latter is more concerned with paragraph and sentence structures and the development of logical thought processes. It is this feature of the ESA test that has not received the attention it deserves in previous investigations into its psychometric properties and provides one of the major focuses of the investigation for this study.

The degree to which the different subtests of both the ESA test and the AEMK test fit the measurement model would then provide evidence for validity of the measures involved as well as insights into the nature of the variable of literacy as conceptualised. Any relationships present could then be examined both within and across the two forms of assessment to provide a measure of association through the employment of a non–linear logistic model. By offering person free measures, which is critical for the development of objective measurement, this type of model provides a means of assessing the nature of the variable literacy in a way not possible with normal linear modelling.

Another aspect of the design that could influence the precision of the measure of literacy is different levels of marker severity. This problem is not present with the ESA test but is a reality for the AEMK test. With an appropriate data design, the severity of markers can be built directly into the measurement process and any differences accounted for by the measurement model. Because of the restricted nature of the initial investigation, the large number of students involved and time limitations imposed on the marking schedule, it was not possible to adopt such a design as this would have required each paper to be marked at least twice by different markers. However, an alternative strategy was developed to provide knowledge of marker severity and marker inconsistency. As this was an exploratory study, and given the limitations imposed by the markers, this strategy was regarded as more than useful to the tasks at hand.

Measurement Model

The measurement model employed in this study is the extended model of Rasch (Andrich, 1985a, 1985b, 1988). Rasch (1960/80) models provide for "separable person and item parameters and hence sufficient statistics ... which makes possible 'specifically objective' comparisons of persons and items and thus fundamental measurement" (Masters & Wright, 1984, p.529). This model is especially suitable for the present study because of its facility to handle polychotomous categorical data in a meaningful way and to address the behaviour of the thresholds located between the different item categories. A set of thresholds are conceptualised as boundaries between the response categories of an item and specify the change in probability of a response occurring in one or the other of two categories separated by the threshold. If the threshold estimates for a particular item do not appear in a sequential, ordered, manner then this is evidence of misfit to the construction of the model (Andrich, 1985a; Sheridan, 1993). Threshold disorder can often provide valuable insights into the nature of the variable under review.

Another test–of–fit of data to the model involves a person–item interaction statistic in which the behaviour of individual items and individual persons can be assessed (Andrich & Sheridan, 1980). Items can also be evaluated both on an individual basis and collectively across the whole test using an item–trait interaction test–of–fit which assesses the stability of items across the range of person abilities or attitudes. A further test–of–fit to the model involves the person separation index (Andrich, 1982a) which is the Rasch model equivalent of the Cronbach Alpha. This index provides the degree to which a test can separate persons in a meaningful way along the latent trait continuum and thus provides a measure of the power of the other tests–of–fit employed.

The Instruments

This study employed two instruments: the English Skills Assessment (ESA) test and an assignment essay to be referred to in this paper as the Assignment–Essay Marking Key (AEMK) test.

The English Skills Assessment (ESA) test: The ESA test is a two part battery of standardised tests intended for use with students in Years 11 and 12 of secondary school and the first year of post–secondary education. It was adapted for Australian conditions from two American tests: the Sequential Tests of Education Progress Series I, for grades 10 to 12, and the Descriptive Tests of Language Skills for College Freshman (ACER, 1982b). For ease of specification, each of the eight tests comprising the two parts of the ESA will be referred to as a subtest of the one ESA test.

Part I of the ESA test consists of three timed subtests: Spelling; Punctuation and Capitalization; and Comprehension I while Part II consists of five timed subtests: Comprehension II; Usage; Vocabulary; Sentence Structure; and Logical Relationships. Comprehension I contains three extended text passages each of 400 to 500 words in length and Comprehension II contains five separate single paragraph texts of 60 to 70 words each. The conceptual framework adopted by the authors to guide the development of the ESA test is based on 26 subgroups that contribute collectively to different aspects of literacy. These subgroups are subdivisions within the different subtests, with from three to five subgroups identified per subtest, except for Vocabulary which has a default subgroup of one only. Item 10 in the Spelling subtest was not considered as it did not contribute to the multiple category concept involved. Thus, the Spelling subtest is defined by four subgroups relating to type of spelling error: (a) initial syllable or sound, (b) medial syllable or sound, (c) final syllable or sound, and (d) consonants; the Punctuation and Capitalization subtest by four subgroups defined as capitalisation, apostrophe, comma, and miscellaneous punctuation, with a fifth type containing no errors present, and so on. Each subtest contains a series of multiple choice statements generated in accordance with the subgroup specifications for the subtest, and each statement scored as a dichotomous item. The number of items per subgroup and, in turn, per subtest varies making a total of 188 items in all, with 95 in Part I and 93 in Part II. The complete list of subgroups and the number of items per subgroup is provided in the Manual of Procedures available for the ESA test (ACER, 1982b) and is reproduced as Appendix A.

Of special interest to this study is the conceptual framework guiding the ESA test. On closer inspection, this test appears to cover a very broad range of aspects generally associated with the notion of literacy. From reports in the literature on research related to the ESA, little or no attention has been directed at the fundamental significance of this conceptualisation as a means of addressing the measurement capabilities of the instrument. As described earlier in the Design section, this test appears to comprise two major aspects of literacy, *surface features* and *deep structures* and it is this broad categorisation that provides the chief focus for investigating the psychometric characteristics of the ESA.

The Assignment–Essay Marking Key (AEMK). One of the purposes of the investigation carried out by this study is to compare student performance on the ESA test with that in real, academic writing situations. An assignment requiring an essay response was chosen because most Faculty of Education staff prefer students to use an essay format for assignments requiring extended writing. An assignment common to all first year primary and secondary teacher education students was identified in the first semester core education unit. The task was considered suitable because: (a) it involved various attributes of academic writing such as research and extended writing, (b) the task was common to all students in the sample investigated, and (c) students had a real purpose for carrying out the task since it constituted an assessable part of their course work.

The marking key for the essay assignment was broken down into a number of subgroups and component segments which included aspects of structure, organisation, cohesion, grammar, writing conventions, and content. As well as marking the content of the assignments, lecturers were asked to rate as excellent, satisfactory, borderline, or unsatisfactory the competencies of the students that were operating in each of the writing subgroups. Lengthy discussions were held with the marking team about the nature of writing in general and academic writing in particular. Specific definitions and meanings were established for the various items on the marking key and a marking guide developed to help ensure consistency among markers. The marking guide included: (a) a general description of the ratings, and (b) a description of each item on the marking key together with an explanation of how the ratings would be applied to those items.

The marking key featured twelve subgroups of literacy. These twelve subgroups are derived from two aspects which are believed by the researchers to constitute the major aspects in writing. The first major aspect consists of specific knowledge about the superficial aspects of language (hence *surface features*). By superficial aspects of language, we refer to the accepted, because of convention, apparent features of language as opposed to the semantic content underlying those features. These aspects include grammar and the mechanics of writing, that is, spelling, punctuation, elements of sentence structure, and word usage.

The second major aspect refers to the semantic content or *deep structures* underlying the surface features. This aspect includes the writer's global organising conceptual framework for a text. McCulley (1985) and others (Landis, 1990; McKenna, 1988; Stein, 1985) argue that coherence in writing arises from writers' understanding of the content and their systematic communication of that content to an audience using organisational formats which meet their purposes for writing and the needs of their audience. Thus, effective writers construct a network of main ideas and sub–ideas that convey a conceptual hierarchy for the reader. This conceptual hierarchy is reflected in the overall structure of the discourse, in the way in which ideas are sequenced and chunked into paragraph clusters and paragraphs, and the way in which ideas are expressed in sentences.

If the point of writing is to communicate, a text can still be an effective piece of communication even though it contains some problems with respect to its surface features. For example, it could contain a number of spelling errors, punctuation errors, and sentence structure errors such as comma splices. However, given there are no other problems, it would require a great number of surface feature errors for a text to be entirely incomprehensible. As a consequence, in any measure of literacy the surface features are, in relative terms, of less importance that the deep structures.

To communicate effectively, the surface features of language must reflect an already existing underlying coherent content base. Put simply, irrespective of a writer's understanding of the surface features of language it is not possible for the writer to communicate effectively content which the writer does not fully comprehend. From the perspective of writing and literacy theory, the second variable is not only more important than the first, but also represents a higher order level of cognitive skills. These skills include analysis, synthesis, and the ability to coordinate into a single meaningful text a complex range of ideas.

The first three subgroups in the marking key: Essay Structure, Paragraph Sequence, and Paragraph Structure were seen to convey the underlying understanding of the content of an essay (that is, deep structures). The four subgroups of Sentence Structure, Usage, Spelling, and Punctuation, were included in the marking key because they correspond to subtests in the ESA test. The logic of this decision was firstly to establish if the ESA test could specify the same students as those identified as having problems in authentic writing situations and, secondly, to provide an opportunity to establish the validity of both measuring instruments. The remaining five subgroups, Consistency of Person, Consistency in Tense, Agreement, Referencing, and Essay Length, were included because the lecturers/markers thought that they represented important technical aspects of academic writing in which students often display deficiencies. A listing and description of the twelve subgroups comprising the AEMK appears as Appendix B.

The Sample and Data Collection

The ESA test was administered during the first semester to all newly enrolled students in the Education degree programme offered at the Edith Cowan University which is the largest provider of teacher trainees in Western Australia. Responses were collected from 838 students using special optical scoring sheets. During the same semester, a written assignment was completed by students as part of the requirements for a core education unit of study and marked according to the set of criteria presented earlier in this section. Because of time and organisation limitations with the markers, only 495 of the total number of students were assessed for the AEMK test, but this number was sufficient for the analyses to be undertaken.

Results

The results of the investigation undertaken by this study revolved around two main analyses. Firstly, an examination of the properties of the English Skills Assessment (ESA) test as a measure of literacy was undertaken with particular reference to its conceptual framework as defined in the Manual of Procedures. The second main analysis involved the Assignment–Essay Marking Key (AEMK) test as a measure of literacy assessed directly from students' written work. In addition, the association between these two tests provided a focus for investigating the construct validity of the measures derived. The over–riding purpose of this investigation was to arrive at a meaningful understanding of what is literacy as measured by these instruments and how this understanding can assist in determining the level of literacy in students enrolled at the tertiary level. All item analyses reported in this paper were undertaken using the computer program ASCORE (Andrich, Lyne & Sheridan, 1991) which incorporates the extended model of Rasch.

The ESA as a Multiple Choice Test

The initial analysis of the ESA test involved all 188 multiple choice statements scored as individual dichotomous items. The analysis showed that the majority of items fit the model. However, there was no clear pattern evident for the approximate 10 percent of items misfitting the model because these were distributed across all eight subtests. A person separation index of 0.91 is favourable but with dichotomous tests this value can be inflated due to the inevitable presence of dependencies between items of this type. An examination of the distribution of person ability estimates relative to the item estimates reveals that the test is on the easy side for tertiary students and thus not as well targeted as it could be. These distributions are presented in Figure 1. This mismatch will be addressed in more detail at a later stage when assessing the threshold order for the subgroup items. While the prognosis for a test of this size is quite favourable in terms of the normal *empirical* specifications associated with an item analysis of this type, the fundamental meaning of the variable being measured is far from clear. The problem here is trying to make sense of the location of items along the continuum in accordance with the original conceptual framework posited and how this provides meaning as a measure of literacy in accordance with literacy theory.

Individual analyses of each of the eight subtests is also possible. Motivation for attempting these analyses comes from the recommendations presented in the Manual of Procedures which encourage the use of total scores for both the individual subtests and the ESA test as a whole. As total scores provide the sufficient statistics for Rasch model analyses, it is clear that the procedure adopted in the Manual would be addressed in accordance with the requirements of objective measurement. Generally, however, the analyses do not reveal anything conclusive with at least 25 percent item misfit and indices of person separation below 0.7. Whilst these individual subtests may have some use for specific diagnostic purposes as suggested in the Manual, the whole exercise merely fragments attempts to provide an understanding of the variable of literacy. At best, these analyses provide only superficial insights into the measurement of literacy which can be addressed in a more meaningful way by appealing to the original conceptual framework of the ESA test and employing the special features of the extended model of Rasch to investigate the measurement implications of this conceptualisation in terms of multiple category scoring per subgroup.

Subgroup approach to the ESA test

As specified in the Manual of Procedures, the ESA test is based on a conceptual framework comprising 26 subgroups distributed in approximately equal numbers across seven of the eight subtests, with the Vocabulary subtest conceived as a single entity. For ease of interpretation and precision and in accordance with the scientific goal of observing the law of parsimony, it is an advantage with polychotomous analyses to have the number of categories per subgroup, or item, approximately equal. Apart from two subgroups in the ESA test, this goal is within reasonable limits. Accordingly, ten items were selected from the 18 comprising the "Final syllable or sound" subgroup in the Spelling subtest and

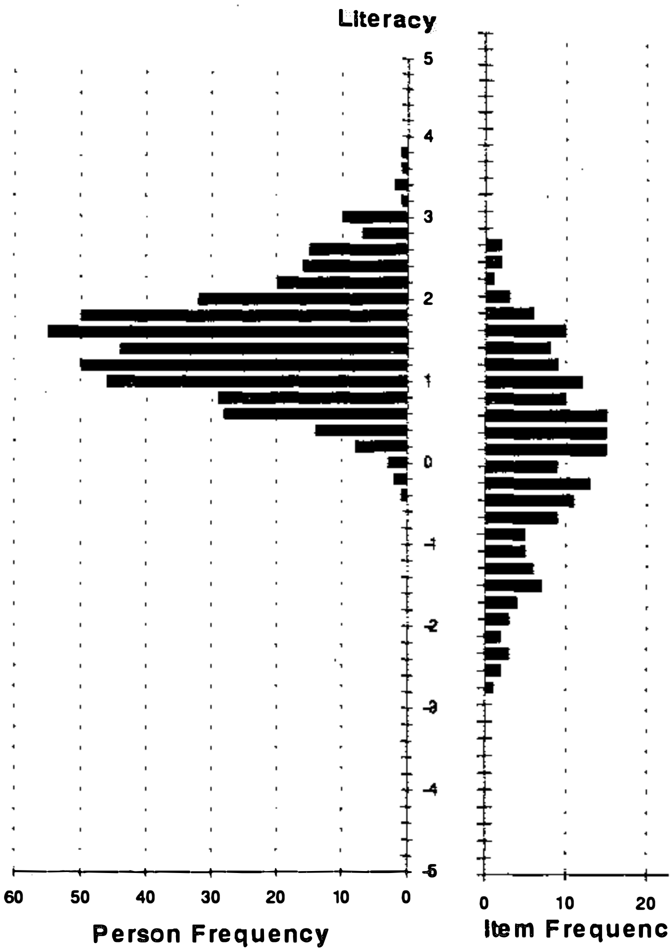

FIGURE 1: Distribution of Tertiary Student Literacy and Item Location Estimates for the 188 multiple choice ESA test

ten from the 20 item Vocabulary subtest/subgroup; this was achieved by selecting alternate items. The individual dichotomous scores were then summed across all items in each subgroup to produce an amended data file comprising 26 subgroup scores in place of the original 188 item scores. An item analysis involving subgroups in place of individual dichotomous items (Andrich, 1985b) was then undertaken on the ESA data.

The first analysis involved all 26 subgroups and revealed threshold disorder for several subgroups. Significantly, most of this disorder was localised amongst those subgroups identified as tapping the surface features of literacy: Spelling, Punctuation and Capitalization, and Vocabulary. A further subgroup, Comprehension II, also exhibited threshold disorder. Of interest here is that the latter subtest comprised very short text paragraphs compared to much longer sets of text passages for subtest Comprehension I. In this situation, it is tempting to conclude that the substance of Comprehension II leans more towards surface features than is the case for Comprehension I. Because disordered thresholds represent a fundamental problem for the construction of a variable, no further examination of the details for this run of the analysis need be considered. As a consequence of the clear dichotomy expressed in the literature regarding surface features and deep structures, all subgroups comprising the three subtests identified as tapping surface features were not considered for analysis with a second run of ASCORE.

The elimination of these three subtests did not change the situation for Comprehension II in that the threshold disorder remained unresolved. Because of the rationalisation offered earlier regarding the nature of this subtest, Comprehension II was also eliminated from further analysis and another run of ASCORE undertaken. Threshold disorder was still evident but not as global as with the earlier subtests. Suspect targeting between the persons and items alluded to in the previous section would now appear to be a contributor to the problem of threshold disorder. To explore this further, data would be required from students in the secondary school which would almost certainly minimise the target mismatch. This procedure would also address the problem of under–representation in the extreme categories whilst at the same time remaining faithful to the original specifications and conceptualisation of the ESA test.

To obtain a feel for the meaning of the variable which this test is measuring, an examination of the 13 subgroup statements relative to the location of the subgroups on a literacy continuum would be instructive. Figure 2 displays the subgroups in difficulty order from most to least difficult. If the conceptualisation of literacy as comprising a lower order aspect (called *surface features*) and a higher order aspect (called *deep structure*) were supported by the data, then the most difficult subgroups would relate to deep structure and the least difficult subgroups would relate to surface features. According to the analysis, the most difficult subgroups are *translation and inference*, and *understanding main ideas*; and the least difficult subgroups are *analysis*, and *drawing analogies*. However, a problem occurs here in that both the most difficult and least difficult subgroups relate to deep structures. To complete successfully those subgroups involving *translation and inference*, *understanding main ideas*, and *analysis* would require students to go beyond the surface features of a text and to understand the ideas and concepts underlying the text. Similarly, a subgroup involving the "drawing of analogies" also require that students go beyond the surface and detect underlying conceptual relationships. Moreover, in many tests, including IQ tests, the drawing of analogies constitute one of the more difficult item types in these tests. While some subgroups which purportedly tapped deep structures appear as relatively easy, other subgroups involving surface features, such as *diction and idiom,* appear more difficult.

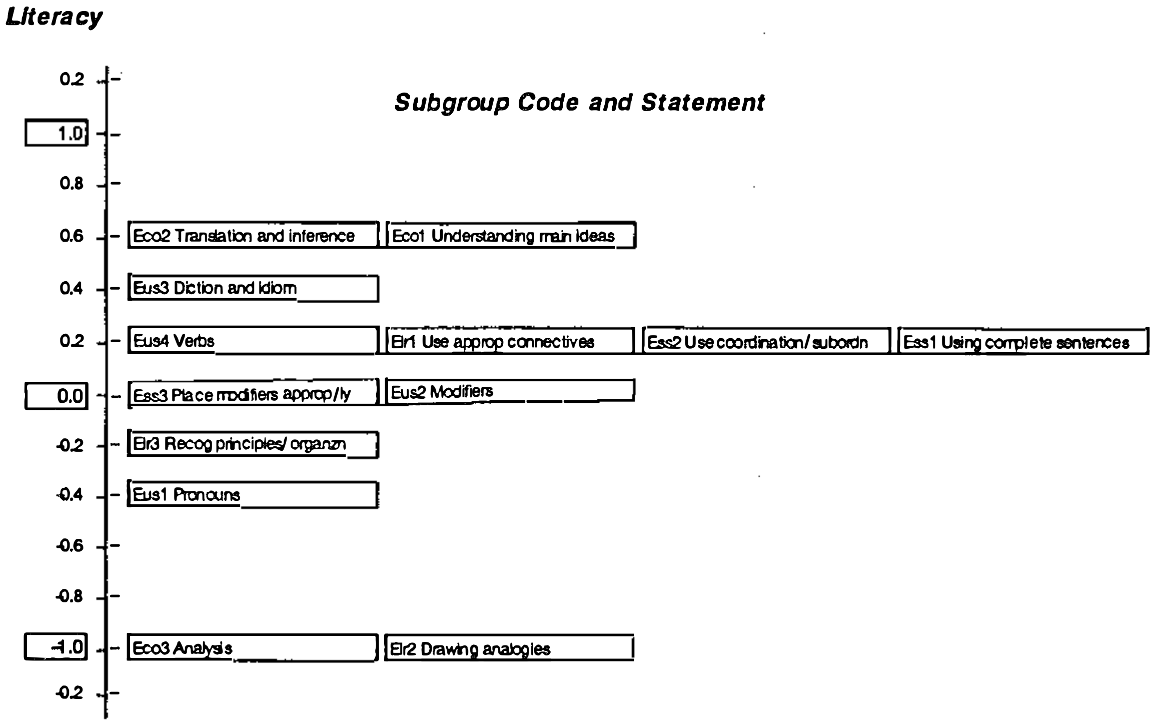

FIGURE 2: Distribution of Subgroups along the ESA Literacy continuum displaying code and statement content

In an attempt to try and understand the logic behind the ordering of the various subgroups, the individual multiple choice items comprising these subgroups were examined. This examination revealed some possible reasons for the subgroup ordering obtained in the analysis. One possibility is that in the ESA test, surface features such as *diction and idiom* might appear as difficult items in the analysis because of their idiosyncratic nature. Within a particular population, the knowledge tested by some particular items may be obscure. This obscurity might arise from changes in teaching fashion where, for example, in the 1970's and 1980's it was not fashionable to teach formal grammar in Australian schools. As a consequence, students who attended schools during this period may have gaps in their knowledge about the surface features of language. Another reason for this obscurity is that some of the language used in textbooks and tests may predate the language in current use. Similarly, other items which tap deep structure such as *drawing analogies* may appear as easier items because of the idiosyncratic nature of the population being tested. If there is a mismatch between the language of the population being tested and the language used by the test, then a shift in the frame of reference for the variable to be measured has occurred. This would then affect the requirements for specific objectivity associated with the measurement model.

When the individual test items were examined other potential problems became apparent. In particular, the uneven number of items within each subgroup may be contributing to the unexpected results in the ordering of the subgroups. In the ESA subtest Comprehension I, for example, there were 15 test items. Seven of these were ***translation and inference*** items, and six were ***understanding main ideas*** items. However, there were only two items in the ***analysis*** subgroup. Given the small number of items in the ***analysis*** subgroup, there is a high probability that the rating obtained was a result of chance alone. This also has implications for the threshold estimates which may have been influenced by this discrepancy in the number of categories across different item

Overall, then, it would appear that whilst part of the ESA test demonstrates a coherence consistent with the conceptualisation of literacy as outlined by the researchers, there are also some inconsistencies. These inconsistencies may be the result of problems with some of the items and in the uneven nature of the construction of the test. In particular, there appears to be a problem with targeting as alluded to earlier. This problem indicates that further research is required using a more appropriately targeted calibration sample drawn from students at the secondary school level before any further conclusions about the ESA test could be consolidated.

The AEMK test

The second phase of the analysis for this study involves the development of a measure of literacy based on twelve subgroups associated with both surface features and deep structures. For this purpose, measures from the Assignment–Essay Marking Key (AEMK) test derive from scoring associated with extended written material.

An initial analysis with ASCORE involving all twelve subgroups was undertaken using a total calibrating sample of 389 students which comprised all students with a complete record for both the ESA and AEMK tests. This analysis revealed threshold disorder for subgroups 9 to 12 and a cursory glance at the test–of–fit information indicated extreme misfit to the model. Subgroup 9 ("Spelling") and subgroup 10 ("Punctuation") are clearly surface features whilst subgroup 11 ("Referencing") and subgroup 12 ("Essay Length") would not command a high priority for inclusion as components of literacy as conceptualised for the study. Considering this outcome, and because subgroups 11 and 12 were included only at the request from the lecturers/markers, it was decided that these two subgroups be withdrawn from future analyses involving the AEMK test, but that the two surface feature subgroups 9 and 10 should remain for the present. However, a new analysis involving subgroups 1 to 10 did not reveal much improvement, with subgroups 9 and 10 still revealing threshold disorder and that only a small improvement in the test–of–fit situation resulted.

At this stage it was decided to examine the implication of both marker severity and marker consistency on the analysis. As explained earlier, the design restrictions forced on the study did not allow for incorporating marker severity directly into the analysis, but an examination of the item–person interaction standardised residual test–of–fit statistics would provide a basis for examining marker consistency. The strategy involves dividing the sample into the 14 marker–groups and calculating the mean value for the standardised residual statistics for the students in each group. This statistic is an indicator of the degree of concordance across the response vector for each student and is an indicator of the level of attainment of the Guttman response pattern. An increasing negative value for this statistic indicates increasing agreement between the actual and expected patterns whilst increasing positive values implies increasing discordance between the two response patterns. If no significant disparity is present between the markers then the average values across the 14 groups should not differ from each other by any significant degree.

As Table 1 reveals, there is a wide range in the mean test–of–fit values present across the different marker groups. For this exercise, the original calibrating sample (N = 389) and an edited subset of this sample were used to provide the two sets of mean values noted. The reduced sample involved an edited data base whereby extreme scores to specific subgroups were deleted from individual student response records, as well as the total elimination of students who had extreme response patterns. The information used to create this edited file was based on the initial analyses referred to earlier for the first 10 subgroups of the AEMK test. In both cases, three markers displayed clear discrepancies relative to the others. A second set of values, called "leniency", also appears in this table. These values are the mean ability estimates for the students in their respective marker groups and provide an indication of the severity, or leniency, of the marker, based on the assumption that no obvious, known, bias was present to influence the original allocation of students to markers.

From the evidence presented for marker consistency, it was decided to amend the original calibrating sample by deleting all students assessed by markers 4, 7 and 8 and rerun ASCORE for subgroups 1 to 10. While no change was observed for the disorder of thresholds for subgroups 9 and 10, a dramatic change occurred in the tests–of–fit associated with these subgroups. Because of the clear implication that subgroups 9 and 10 represent surface features of literacy as defined, these two subgroups were now omitted from further analyses of this test. This time, as Table 2 reveals, all thresholds were

TABLE 1

Indicators of Marker Consistency and Marker Leniency based on Standardised Residual Fit Statistics

Marker Consistency (Fit Stats) and Leniency (Ability Est.) - Sorted by Fit Statistics

Calibrating Sample (extreme responses and extreme respondents omitted)

Marker:	6	1	12	9	14	13	2	5	3	10	11	8*	4*	7*
N:	21	15	10	13	24	16	29	39	15	30	16	17	32	43
Consistency:														
Mean:	-1.30	-1.13	-0.91	-0.89	-0.86	-0.78	-0.64	-0.49	-0.33	-0.32	-0.23	0.25	0.42	0.85
SD:	1.18	1.62	0.90	0.92	1.13	0.84	1.10	0.83	0.59	0.98	0.97	1.21	0.75	0.89
Leniency:														
Mean:	0.82	1.40	0.53	-0.05	0.60	1.95	0.16	1.44	1.71	2.43	0.17	3.53	0.44	2.41
SD:	0.86	0.88	0.48	0.76	1.25	1.13	0.86	0.72	1.67	1.37	0.84	1.60	1.63	1.21

Original Calibrating Sample

Marker:	1	12	6	9	2	14	11	5	13	3	10	4*	8*	7*
N:	32	16	36	18	34	31	21	44	17	19	36	32	30	50
Consistency:														
Mean:	-2.96	-2.46	-2.18	-1.81	-1.35	-1.32	-0.83	-0.78	-0.51	-0.43	-0.09	0.77	0.89	1.36
SD:	1.99	1.69	1.81	1.51	1.76	1.65	1.22	1.30	0.95	1.54	1.58	0.93	2.06	1.15
Leniency:														
Mean:	0.86	0.40	0.47	-0.02	0.05	0.45	-0.03	0.84	1.43	0.94	1.71	0.11	2.41	1.31
SD:	0.70	0.40	0.52	0.56	0.59	1.11	0.67	0.59	1.17	1.02	1.13	0.95	1.22	0.95

Rationale: The more negative (–ve) the fit statistic, the more consistent a marker's ratings.
The more positive (+ve) the fit statistic, the less consistent a marker's ratings.

The higher the ability estimate, the easier the marker – that is, more lenient.
The lower the ability estimate, the harder the marker – that is, less lenient.

*Indicates least consistent markers – also includes two of the four easiest markers.

correctly ordered and all subgroups fitted the model in accordance with the theoretical specifications of literacy theory. One point needs emphasis, however, and this refers to the disproportion of negative standardised residual fit statistics. This trend implies a tendency towards overfitting which is associated with high discrimination and is possibly a manifestation of an easy test overall, as was the case for the ESA test.

TABLE 2

Threshold Estimates and Subgroup fit to the Model for the AEMK test under the hypothesis of no misfit for 8 subgroups

	Item-trait Interaction		Standardised Residual	Threshold Estimates		
Subgroup*	Chi Sq**	Prob	(189 df)	1	2	3
WES1	7.29	0.03	-2.83	-4.55	-1.09	5.63
WPS2	0.61	0.89	-2.60	-4.66	-0.89	5.55
WPA3	1.63	0.64	-1.85	-4.24	-1.27	5.51
WSS4	1.76	0.61	-1.55	-3.54	-2.46	6.00
WWC5	4.83	0.16	-1.09	-3.80	-2.39	6.19
WCP6	0.36	0.95	-0.57	-5.93	-1.58	7.51
WCT7	2.48	0.46	-2.02	-4.51	-2.80	7.32
WAG8	0.86	0.83	-1.17	-6.46	-0.67	7.13

* Key for subgroup label code appears in Appendix B.

**The overall χ^2 is 19.82 which has a probability of 0.53 on 21 degrees of freedom. Individual χ^2 probabilities are based on 3 degrees of freedom.

It would be instructive at this point to examine the content descriptions for the eight subgroups of the AEMK test relative to their location on the literacy continuum. Figure 3 displays these subgroups in difficulty order from most to least difficult. Although the location of some of the subgroups gives support for the two aspect conceptualisation of literacy as outlined earlier, *surface features* and *deep structures*, the results need to be viewed with some caution. While the location of the subgroups: *sentence structure*, *paragraph structure*, and *paragraph sequence* appear to support the two aspect conceptualisation of literacy, there are some misgivings regarding the quality of both the marking key and the markers employed. Three causes for concern became apparent following a re–examination of the student assessments carried out after these analyses were undertaken. Of these concerns, two related to the markers and one to the marking key.

With respect to the markers, it was apparent that they had a better understanding of how to score the surface features of language than they did for scoring the deep structures. For example, it was noted that markers were able to detect problems at the sentence level but failed to detect the fact that many of these sentence problems were a manifestation of problems at the paragraph and overall discourse levels. A review, by the researchers, of a sample of papers showed that in many instances students appeared to have no obvious communicative purpose for their writing. As students did not have a clear idea of why they were writing, they also did not seem to know what to write. Ideas appeared on paper as they appeared in the writer's head; as a stream of consciousness, without any systematic

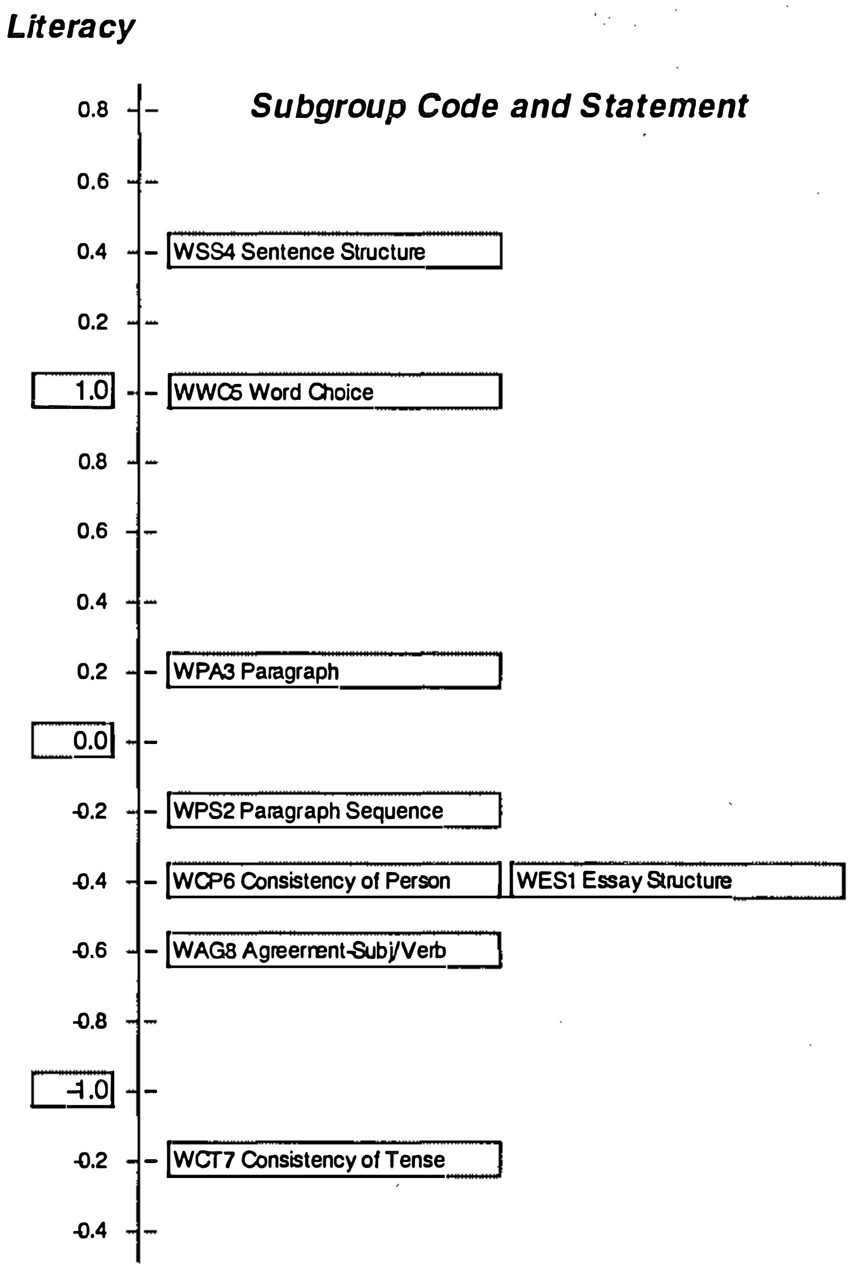

FIGURE 3: Distribution of Subgroups along the AEMK test Literacy Continuum displaying code and statement content

form of organisation which links the ideas to a central theme or thesis. Lack of purpose and poor organisation resulted in incoherent writing at all levels: overall discourse, paragraphs, sentences. That is, in a sample of papers viewed it seems that many of the sentence problems arose, at least in part, as a consequence of the lack of underlying coherence in the content of the discourse. As a consequence; the accuracy of scores assigned by markers to the categories of essay structure, paragraph sequence, and paragraph structure must be viewed with caution. This is a more fundamental issue than that relating to marker consistency referred to earlier. Markers also appeared to have a problem scoring one of the surface feature subgroups: *tense*. Therefore, it is possible that *tense* may have been a more difficult subgroup than indicated by its location on the measurement continuum and as displayed in Figure 3. Regarding the marking key, its construction would also appear to be suspect. Some of the elements of the marking key which were treated as separate subgroups such as *agreement* and *tense* are, upon reflection, really subsets of the subgroups: *usage* and *sentence structure*.

Construct validity of the AEMK test

In spite of these misgivings, it appears from the evidence available that the first eight subgroups of the AEMK test do provide a reasonable basis for measuring literacy directly from students written work. This means that the student literacy estimates obtained are defined in terms of a synthesis of essay, paragraph, and sentence structures, consistency of person and tense, agreement between subject and verb, and appropriate word choice. The fact that all of these subgroups are in accord with the requirements of the measurement model and were generated from a conceptual framework itself derived from literacy theory, is evidence for the construct validity of the instrument. Further demonstration of the construct validity could be sought in comparing the AEMK test with the ESA test, as both instruments appear to derive from similar conceptualisations but using different mechanisms for recording student responses.

The application of non–linear models to assess the degree of association between measuring instruments does not appear to be wide. One significant advantage the Rasch model enjoys over the familiar linear models in current use is the availability of "person–free" measures. Thus, if two tests are seen as measuring the same, or similar, constructs, and if the different items associated with each test are included in the one item analysis, and if these items all fit the model, then this is clear evidence of the construct validity of both tests. Attention is now directed to an assessment of the construct validity of the ESA and AEMK tests employing this strategy.

The amended data file, with the three inconsistent markers removed, was employed for this analysis. Table 3 displays the threshold estimates and test–of–fit values when subgroups 1 to 8 of the AEMK test are incorporated with the 13 subgroups from the Comprehension I, Usage, Sentence Structure, and Logical Relationships subtests of the ESA test. Following an analysis using ASCORE, only *consistency of tense* (WCT7 from the AEMK test) and the *drawing analogies* (Elr2) subgroup from the Logical Relationships subtest of the ESA test exhibit threshold disorder. Also, some subgroups

TABLE 3

Threshold Estimates and Subgroup Fit for AEMK test and ESA test combined under the hypothesis of no misfit for 21 subgroups

	Item-trait Interaction		Standardised Residual	Threshold Estimates								
Subgroup	Chi Sq**	Prob	(240 df)	1	2	3	4	5	6	7	8	9
WES1*	8.35	0.06	-0.49	-2.05	-0.76	2.81						
WPS2	10.07	0.01	-0.16	-2.98	-0.31	3.28						
WPA3	7.83	0.07	-0.18	-2.74	-0.47	3.21						
WSS4	11.35	0.00	-0.73	-2.39	-1.23	3.62						
WWC5	7.38	0.09	-0.50	-2.41	-1.09	3.50						
WCP6	11.30	0.00	-0.94	-3.16	-1.12	4.28						
WCT7	13.36	0.00	-1.70	-1.52	-3.30	4.82						
WAG8	18.66	0.00	-1.11	-3.25	-1.21	4.46						
Eco1	3.99	0.39	0.24	-1.70	-1.25	-0.63	0.16	1.13	2.28			
Eco2	5.12	0.26	1.35	-2.76	-1.64	-0.71	0.11	0.87	1.64	2.49		
Eco3	8.96	0.04	-1.04	-1.17	1.17							
Eus1	4.92	0.28	0.00	-1.53	0.39	1.14						
Eus2	3.84	0.41	-0.13	-1.56	-1.11	-0.24	0.87	2.04				
Eus3	3.14	0.52	-1.51	-1.83	-0.84	-0.20	0.30	0.87	1.71			
Eus4	4.14	0.37	-0.38	-1.97	-1.35	-0.61	0.25	1.25	2.42			
Ess1	10.48	0.01	-0.16	-1.43	0.12	1.31						
Ess2	7.36	0.09	1.03	-1.10	-0.88	-0.72	-0.57	-0.37	-0.06	0.42	1.14	2.15
Ess3	12.29	0.00	-0.07	-1.85	-0.92	-0.36	0.14	0.86	2.12			
Elr1	5.69	0.20	0.64	-1.43	-1.18	-0.87	-0.49	-0.02	0.56	1.28	2.15	
Elr2	1.04	0.90	-0.75	-1.28	-1.42	-0.33	1.09	1.94				
Elr3	2.44	0.65	0.68	-1.83	-0.94	-0.36	0.06	0.44	0.94	1.69		

*Key for subgroup label code appears in Appendix A (for ESA test) and Appendix B (for AEMK test)

**The overall χ^2 is 161.69 which has a probability of 0.00 on 80 degrees of freedom. Individual χ^2 probabilities are based on 4 degrees of freedom.

show misfit for the item–trait interaction test–of–fit but all appear in order with the standardised residual fit statistic. As suggested earlier in the reporting on the ESA test, the minimising of the skewed targeting by including secondary school students in the calibrating sample, could possibly address the problem of threshold disorder, even though only 2 of the 21 subgroups exhibit this feature in the present situation. Also no consistent misfit to the model is evident from these data though several subgroups exhibit some misfit for the item–trait interaction statistic. In addition, a person separation index of 0.88 indicates that the tests of fit have sufficient power.

Before leaving the combined ESA and AEMK test, two further displays should be consulted. The distribution of student literacy estimates relative to the subgroup estimates is displayed in Figure 4. This distribution continues to demonstrate the presence of skewed targeting for tertiary students. However, as this distribution involves the AEMK test as well, the scoring key for the in depth written assessment needs to be re–examined as was suggested in the discussion earlier. The display of statement contents for all 21 subgroups as provided in Figure 5 reveals a similar distribution to that evident from the two tests when considered individually and presented in Figure 2 and Figure 3 respectively.

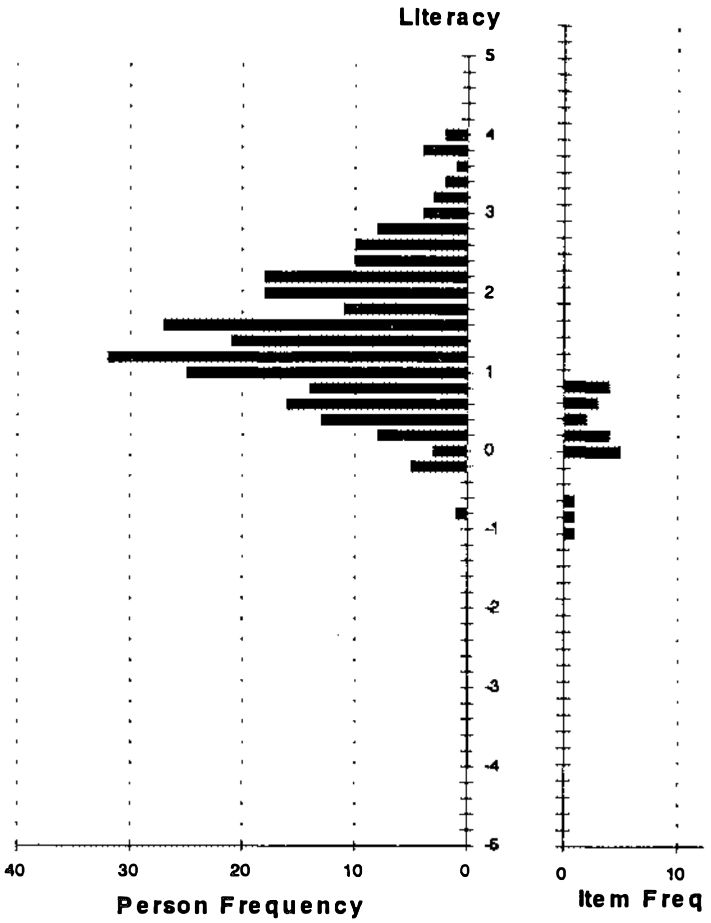

FIGURE 4 Distribution of Tertiary Student Literacy and Item Location Estimates for 13 ESA subgroups and 8 AEMK subgroups combined

Figure 5 shows the distribution of the subgroups for the ESA and AEMK tests along the measurement continuum. The seven most difficult subgroups are: *understanding main ideas and direct statements* (Eco1), *translation and inference* (Eco2), *diction and idiom* (Eus3), and *sentence structure* (WSS4), which are all located at 0.6 logits on the continuum; and *modifiers* (Eus4), *sentence structure* (Ess1), and *word choice* (WWC5), which are located at 0.4 logits on the continuum. As indicated earlier, the subgroup locations generated when individual analyses of both tests were undertaken give partial support to a conceptualisation which views literacy as consisting of two aspects. The combined analysis gives further support to this conceptualisation because three of the four subgroups shown as being the most difficult in the distribution include items tapping deep structures.

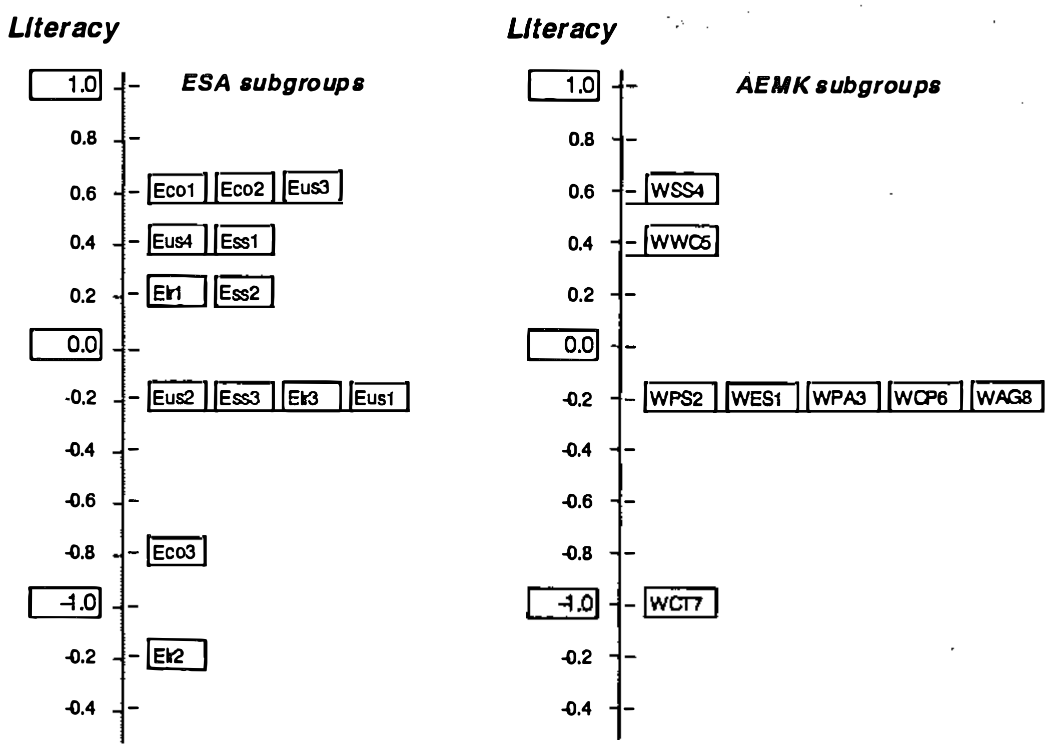

FIGURE 5: Distribution of subgroups for ESA and AEMK tests along the Literacy continuum displaying code and statement content

It is interesting to note that in both the ESA and the AEMK tests, subgroups relating to sentence structure are among the more difficult subgroups and, in the case of the AEMK test, the most difficult subgroup. In an earlier discussion, it was indicated that the location of some of the subgroups in the AEMK test, including *sentence structure*, may have been influenced by the level of understanding of markers which, in turn, affected their ability to score adequately some of the test's subgroups. However, another factor contributing to the level of difficulty of this subgroup may relate to the nature of the subgroup itself. Though some subgroups have been described as tapping surface features and others as tapping deep structures, the subgroups are not, in reality, exclusively one or the other. Whilst the subgroup *structure of the discourse* may be categorised as pertaining to deep structures, because it reflects the global organisation of the underlying content, it also has surface features. That is, it reflects conventions of written language which a writer may or may not know. For example, an essay consists of an introduction, a main body, and a conclusion. A student's inability to write an adequate essay may be a reflection of (a) a lack of understanding of the content, and/or (b) a lack of understanding of some of the elements of the essay form of writing required in a particular context. Other subgroups containing items such as the use of apostrophes relate more clearly to conventions and, thus, to surface features. However, sentence structure contains strong elements of both surface and deep structures. For a sentence to make sense the underlying content must be clear, but sentences also are governed by strong conventions such as word order, etc.

The way in which the surface features and deep structures interrelate have important implications for the design of a measure of student literacy competencies. These implications are discussed in more detail at the conclusion of this paper.

Overall, then, there is evidence to support the notion that both the AEMK test (subgroups 1 to 8) and the ESA test (after omitting the surface features subtests of Spelling, Punctuation and Capitalization, and Vocabulary, and the Comprehension II subtest) exhibit construct validity and produce a measure of literacy whose meaning approximates with the conceptualisation for literacy outlined.

Educational Importance of Study

It appears from the literature that the fundamental prerequisite for a coherent text is that writers have an adequate conceptual or content base for their writing. This study has shown that the relationship between surface features and deep structures in texts is complex. This complex relationship will need to be taken into account in the development of a direct measure of literacy. Some of the problems which have arisen in the analyses carried out in this study derive from the fact that, in the design of the AEMK, an attempt was made to construct a single measure for the assessment of students' literacy competencies. The internal consistency of this measure was then evaluated. However, given the different nature of surface features and deep structures, the way in which this test was evaluated may not have been appropriate. The data available seems to indicate that surface features and deep structures are related but different properties of a text, in much the same way as height and weight are related but different characteristics of people. As these properties are related but different, it is quite possible for a person of high ability, as shown by their performance in the subgroups assessing deep structure, to perform relatively poorly in specific subgroups assessing surface features. For example, a good writer may be a poor speller.

The apparent unsuitability of analysing all the subgroups comprising the AEMK as components of a single variable, indicates that any assessment of literacy competencies may need to include two separate measures: a measure of competency with respect to deep structures and a measure of competency with respect to surface features. This suggestion would be in accord with the literacy requirements for student teachers as there is not only a requirement that these students be able to extract, organise, and communicate information in a logical and coherent fashion, but there is also a requirement that they understand and adopt the conventions and forms required of them in the context within which they operate. A literacy competency requirement of students attending a tertiary institution is technical accuracy. Technical accuracy is also a necessary requirement of the profession education students have chosen to enter.

Having considered that a test of literacy competencies could contain two measures, the placement of the subgroups into either one or the other measure is still problematic. Earlier in this paper, it was indicated that none of the subgroups could be categorised as either exclusively pertaining to deep structures or exclusively pertaining to surface features, although, it is possible to classify most subgroups by the extent to which they belonged to one of these components of literacy. However, the *sentence structure* subgroup is difficult to classify because it contains strong elements of each of these two components. A possible solution to this problem would be to include sentence structure in both measures. Thus, one measure would contain a sentence structure subgroup which assessed sentences only in relation to their semantic aspects, and the second measure would contain a sentence structure subgroup which assessed sentences only in relation to their technical correctness.

The aim of this study was to produce a measure of tertiary students' literacy competencies which was cost effective and not too time consuming. The ESA test was used by the Faculty of Education because it fulfilled both these criteria. However, staff perceived that this test was inadequate as a measure of literacy. The evidence provided by this study regarding the use of the ESA test as a measure of literacy indicates that this test should not be employed in the form prescribed in the Manual of Procedures. At best, these procedures are superficial and it is difficult to see how the reporting of raw scores obtained from summing the respective dichotomous item responses for the eight subtest relates in a meaningful way to the 26 subgroup descriptions. Unless the variable constructed for an instrument is shown to relate in a meaningful way to some prescribed theoretical base that underpins the test, then it is unrealistic to claim that the test is a measure of any construct whose meaning can be established.

The analyses reported for this study reveal that the ESA test can be assigned some meaning as a measure of literacy provided the separate 188 statements are not used as dichotomous items on an individual basis. Rather, the selection of the 13 non surface–feature–subgroups identified by the study would provide a reasonable measure of literacy as a variable defined in accordance with the statements as listed in the Manual of

Procedures, though some clarification of the labels assigned the subgroups would be required. These findings suggest that the ESA test, as used to date by the Edith Cowan University, and other institutions, is not an adequate measure of literacy as required for the screening of students entering a teacher education degree programme.

The ESA test is claimed to be suitable for students in their last two years at secondary school as well as for students in their first year at tertiary level. By selecting students towards the upper end of this range it would appear that the test is not properly targeted as it is too easy for many students. While the presence of such skewness between student ability and item difficulty need not necessarily invalidate the test as a measure of a proscribed variable, it will reduce the precision of the estimates obtained. In addition, this shift between item difficulty and person ability produces a disproportionate number of responses in the extreme categories of many of the items, or subgroups involved, which can influence the threshold order for such items. As a consequence, it is not possible to investigate fully the threshold structure of the ESA test without recourse to responses from students located at the upper secondary school level. This means that the conclusions reached regarding the ESA test as a measure of literacy must be viewed with some caution. Thus the results of this study tend to support the staff's perceptions about the lack of suitability of the ESA test as a measure of tertiary literacy.

This study also assessed the possibility of using a direct measure as an alternative from the ESA test for evaluating students' literacy competencies. This direct measure involved the assessment of students' literacy competencies in an assignment that was already carried out as a part of the students' course work. This method was considered as potentially suitable because it imposed no additional burden on the students and only a small additional burden on markers. However, inconsistencies arising from a lack of expertise on the part of the markers indicate that it is not feasible to carry out a direct assessment of students' literacy competencies within their existing course work. Instead, an alternative form of direct assessment which employs expert markers will have to be devised. As noted earlier, the subgroups comprising the AEMK will also have to be revised so that item groups which are really subsets of other subgroups do not appear as separate subgroups in their own right.

To increase the speed of marking and thus reduce the additional costs involved in administering a separate assessment may require different approaches to marking than used in this study. For example, markers may be asked to count errors instead of classifying students into one of four ratings. An error count should be faster and reduce the potential for markers to make classification errors. Using different marking systems may also necessitate the employment of a measurement model capable of handling error counts. A possible approach would involve a consideration of the Poisson distribution. Outcomes from studies undertaken by Andrich, (1973), Hake (1986), and Rasch (1960/1980) should give guidance regarding error counts for the development of an appropriate measurement model.

References

ACER (1982a). *English Skills Assessment Interim Manual.* Hawthorn, Melbourne: ACER.

ACER (1982b). *English Skills Assessment Parts I and II.* Hawthorn, Melbourne: ACER.

Andrich, D. (1973). *Latent trait psychometric theory in the measurement and evaluation of essay writing ability.* Unpublished doctoral dissertation, The University of Chicago.

Andrich, D. (1982a). An Index of Person Separation in latent trait theory, the traditional KR.20 Index, and the Guttman Scale response pattern. *Educational research and perspectives, 9* (1), 95–104.

Andrich, D. (1982b). Using Latent Trait measurement to analyse attitudinal data: A synthesis of viewpoints. In D. Spearitt (Ed.). *The improvement of measurement in Education and Psychology* (pp.89–126). Melbourne: ACER Ltd.

Andrich, D. (1985a). An elaboration of Guttman scaling with Rasch models for measurement. In N. Brandon-Tuma (Ed.). *Sociological Methodology,* (pp. 33–80). San Fransisco: Jossey-Bass.

Andrich, D. (1985b). A latent-trait model for items with response dependencies: Implications for test construction and analysis. In S. E. Embretson (Ed.). *Test design: Developments in psychology and psychometrics,* (pp. 245–275). Orlando: Academic Press.

Andrich, D. (1988). A general form of Rasch's extended logistic model for partial credit scoring. *Applied Measurement in Education, 1* (4), 363–378.

Andrich, D., Lyne, A., & Sheridan, B. (1991). *ASCORE: A FORTRAN IV program for the general analysis of dichotomous and graded response data to achievement and attitude instruments.* Perth: School of Education, Murdoch University.

Andrich, D., & Sheridan, B. (1980). *RATE: A Fortran IV program for analysing rated data according to a Rasch model* (Research Report. No. 5). Perth: University of Western Australia, Department of Education, Measurement and Statistics Laboratory.

Australian Education Council (1990). *Teacher education in Australia.* Canberra: Australian Government Publishing Service.

Back to basics (1989, May 4). *The West Australian,* p. 10.

Bereiter, C., & Scardamalia, M. (1983). Levels of inquiry in writing research. In P. Mosenthal, L. Tamor, & S. Walmsley (Eds), *Research on writing: Principles and Methods.* New York: Longman.

Bereiter, C., & Scardamalia, M. (1985). Levels of inquiry into the nature of expertise in writing. *Review of Research in Education, 13,* 259–282.

Brandeth, S. (1988, May 3). Teaching hopefuls failing on literacy. *The West Australian*, p. 4.

Brock, P. (1990). A review of some of the literary, political, and mythological contexts of reform and regression in Literacy Education. In J. Howell, A. McNamara, & M. Clough (Eds.), *Social Context of Literacy* (pp. 15–33). Carlton: Australian Reading Association.

Carpenter, K., & Johnson, L. (1991). Program Organisation. In R. Flippo & D. Caverly (Eds), *College reading and study strategy programs* (pp. 28– 69). Newark: International Reading Association.

Duke, C. (1985). Developing a writing assessment of candidates for admission to teacher education. *Journal of Teacher Education*, *36*(2), 7–11.

Flower, L. (1989a). Cognition, context, and theory building. *College Composition and Communication*, *40*, 282–311.

Flower, L. (1989b). *Studying cognition in context: introduction to the study*. (Eric Document Reproduction Service No. ED 306593)

Flower, L., & Hayes, J. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*, 365–387.

Glaser, R. (1984). Education and thinking: the role of knowledge. *American Psychologist*, *39*(1), 93–104.

Guttman,L. (1950). The problem of attitude and opinion measurement. In S. A. Stouffer and others (Eds), *Measurement and Prediction.* New York: Wiley.

Guttman,L. (1954). The principal components of scalable attitudes. In P. F. Lazarsfeld (Ed.), *Mathematical Thinking in the Social Sciences.* New York: Free Press.

Hake, R. (1986). How do we judge what they write? In K. L. Greenburg, H. S. Wiener, & R.A. Donovan (Eds.), *Writing assessments: Issues and strategies* (pp.153–167). New York: Longman.

Hayes, J., & Flower, L. (1980). Identifying the organisation of the writing process. In L. Gregg, & E. Steinberg (Eds), *Cognitive processes in writing*. Hillsdale: Erlbaum.

Hayes, J., & Flower, L. (1983). *A cognitive model of the writing process*. (ERIC Document Reproduction Service No. ED 240608)

Holbrook, A., & Bourke, S. (1989). Assessment of the English skills of tertiary students. *Higher Education Research and Development, 8* (2), 161–179.

Landis, K. (1990). The knowledge of composition. Paper presented at the 41st Annual Meeting of the Conference of College Composition and Communication, Chicago.

Masters, G.N., & Wright, B.D. (1984). The essential process in a family of measurement models. *Psychometrika*, *49*(4), 529–544.

McCulley, G. (1985). Writing quality, coherence and cohesion. *Research in the Teaching of English*, *19*(3), 269–282.

McKenna, M. (1988). The development and validation of a model for text coherency. Paper presented at the 38th Annual Meeting of the National Reading Conference, Tucson.

National Board of Employment, Education and Training (1990, September). *Teacher Education in Australia* (Commissioned Report No. 6). Canberra: Australian Government Publishing Service.

Nightingale, P. (1988). Understanding processes and problems in student writing. *Studies in Higher Education*, *13*(3), 263–283.

Parry, S. (1989). Achieving academic literacy: Disciplined discourse. *Higher Education Research and Development*, *8*(2), 147-158.

Prichard, M. (1990, November 16). Low-mark pupils go teaching. *The West Australian*, p. 7.

Rasch, G. (1960/80). *Probabilistic Models for Some Intelligence and Attainment Tests.* (Expanded ed.) Chicago: University of Chicago Press.

Rosenbaum, P.R. (1988). Item bundles. *Psychometrika, 53,* 349-359

Sheridan, B. (1993). *Threshold location and Likert-style questionnaires.* (Research Report No. 1). Perth, Western Australia: Edith Cowan University, Measurement Assessment and Evaluation Laboratory.

Stein, N. (1985). Knowledge and process in the acquisition of writing skills. *Review of Research in Education, 13* (pp. 225-257).

Stotsky, S. (1990). On planning and writing plans - or beware of borrowed theories. *College Composition and Communication, 41,* 37-57.

Taylor, G., & Nightingale, P. (1990). Not mechanics but meaning: Error in tertiary students' writing. *Higher Education Research and Development*, *9*(2), 161-175.

Wilson, M., & Adams, R. J. (1992). Rasch models for item bundles. Paper presented at the annual meeting of the American Education Research Association, San Fransisco, April 20 - 24.

APPENDIX A

Conceptual Framework for the ESA test

The conceptual framework for this test comprises 26 subgroups as follows:

No.	Code	Subtest	Subgroup	Item s
1	Esp1	Spelling*	Initial syllable or sound	3
2	Esp2	- do -	Medial syllable or sound	8
3	Esp3	- do -	Final syllable or sound	18
4	Esp4	- do -	Consonants	10
5	Epc1	Punctuation & Capitallization*	Capitallization	10
6	Epc2	- do -	Apostrophe	8
7	Epc3	- do -	Comma	10
8	Epc4	- do -	Miscellaneous punctuation	5
9	Epc5	- do -	No error present	7
10	Eco1	Comprehension I	Understanding main ideas	6
11	Eco2	- do -	Translation and inference	7
12	Eco3	- do -	Analysis	2
13	Eco4	Comprehension II*	Understanding main ideas	5
14	Eco5	- do -	Understanding direct statements	6
15	Eco6	- do -	Drawing inferences	4
16	Eus1	Usage	Pronouns	3
17	Eus2	- do -	Modifiers	5
18	Eus3	- do -	Diction and idiom	6
19	Eus4		Verbs	6
20	Evo1	Vocabulary*	Synonyms	20
21	Ess1	Sentence Structure	Using complete sentences	3
22	Ess2	- do -	Using coordination & subordin.	9
23	Ess3	- do -	Placing modifiers appropriately	6
24	Elr1	Logical Relationships	Using appropriate connectives	8
25	Elr2	- do -	Drawing analogies	5
26	Elr3	- do -	Recognizing principles of organization	7

* omit from recommended ammended version of test.

APPENDIX B

Conceptual Framework for the AEMK test

The conceptual framework for this test comprises twelve subgroups as follows:

No.	Code	Subgroup	Description
1	WES1	ESSAY STRUCTURE	Ideas are introduced and logically developed.
2	WPS2	PARAGRAPH SEQUENCE	Relationship between paragraphs clearly signalled
3	WPA3	PARAGRAPH	Main points clearly stated and amplified within paragraphs.
4	WSS4	SENTENCE STRUCTURE	Sentences clear and unambiguous. No syntax errors.
5	WWC5	WORD CHOICE	Appropriate word usage.
6	WCP6	CONSISTENCY OF PERSON	Appropriate person is maintained throughout the essay.
7	WCT7	CONSISTENCY OF TENSE	Appropriate tenses are maintained throughout the essay.
8	WAG8	AGREEMENT	There is subject/verb and noun/pronoun agreement.
9	WSP9*	SPELLING	Accurate and consistent.
10	WU10*	PUNCTUATION	Accurate and consistent.
11	WR11*	REFERENCING	Appropriate conventions observed, in and out of text.
12	WL12*	ESSAY LENGTH	As appropriate.

* omitted from final version of test.