Technical Report #21

Assessment Tools for Teaching and Learning

Standard Setting for asTTle Reading:

A comparison of methods

Submitted by the Assessment Tools for Teaching and Learning team,

Auckland UniServices Ltd

University of Auckland

August 2003

Standard Setting for asTTle Reading:
A comparison of methods

This report provides an explanation of the standard setting for asTTle reading. asTTle is funded by the Ministry of Education to Auckland Uniservices Ltd. at the University of Auckland to research and develop an assessment application for Reading, Writing, Mathematics, Pānui, Pāngarau, and Tuhituhi for Years 5-7 (Levels 2-4) for New Zealand schools. We acknowledge this funding, and thank the Ministry of Education for their continued assistance in the development of this project.

Four different standard setting methods were used in setting curriculum level cut scores for asTTle assessments developed to provide performance measures of close reading in the English curriculum. The report contrasts the four methods (examinee-centred, item-centred, test-centred, performance threshold-centred) and describes how the four methods were reconciled. It concludes with recommendations for future standard setting and demonstrates the robustness of the asTTle curriculum level scores.

I would like to acknowledge the participation of various personnel in these studies. Professor Jim Tognolini, UNSW, ran and analysed the test-centred method. Dr Gavin Brown ran and analysed the item-centred method in conjunction with Ms Trish Meagher-Lundberg. Dr Brown was responsible for pulling together the curriculum level definitions of reading from the various workshops listed in Appendix 3. Mr Peter Keegan and Dr Brown assisted with the running of the test-centred method workshop.

John Hattie
Project Director, asTTle
August 2003

The bibliographic citation for this report is:

Hattie, J.A., & Brown, G. T. L. (2003, August). *Standard setting for asTTle reading: A comparison of methods*. asTTle Technical Report #21, University of Auckland/Ministry of Education.

Table of Contents

Standard Setting for asTTle Reading: A comparison of methods

Setting performance standards is a process of eliciting reasoned judgments from experts who are (a) knowledgeable about the demands of the test or assessment for which a standard is to be set, (b) understand the meaning of scores at various levels on the scales used to summarize examinees' performances, and (c) fully comprehend the definitions of achievement associated with the performance standards that they have been asked to establish. It is important that the standard setting method reflects the nature of the decision process, be replicable (Brennan, 1995), and that there is evidence to support the intended interpretations and to refute competing interpretations (Kane, 1992; Shepard, 1993).

It is the case that many test developers concentrate more on the qualities of the items and processes to ensure comparability. When developing diagnostic tests with an emphasis on formative interpretation it is as critical to also deal with the process of how standards are set, that then underlie any comparisons and particularly to elucidate the meaning of these standards. In the development of tests of reading and mathematics in English and Maori (hence four sets of tests) related to the New Zealand curricula it is critical to use defensible standard setting methods. It is not defensible to set up "committees" to debate issues, decide on standards and then get some buy-in from other groups. Such a method has no psychometric rigour, and is often swayed by beliefs from a very small number of persons in the committee. Instead, standard setting has become a major focus of many research studies, the basis of many court decisions, and there is a large body of literature on how to set standards (Jaeger, 1989; Cizek, 2000) and in recent years comparisons between the standards set by different methods (Plake, 1995). The focus of this study is to outline how the standards were set for the NZ based asTTle (Assessment Tools for Teaching and Learning) English reading tests, and the consequential interpretations of these standards.

The context for standard setting

Setting performance standards is an example of a larger class of problems relating to judgment or decision-making tasks (Pitz & Sachs, 1984). A judgment or decision-making task is characterised by uncertainty of information or outcome, or by an outcome

dependent on personal preferences, or both. If judges are to provide reasoned recommendations on an appropriate performance standard, the process through which they consider alternatives and come to a judgment must help to resolve their uncertainties. Thus setting standards is a social judgment process that involves decision-making under uncertainty. Jaeger (1994) has argued that resolution of uncertainty occurs through an iterative process that incorporates initial judgment, information on the consequences of initial judgments, the opportunity to learn about the rationales underlying fellow judges' judgments, and opportunities to reconsider initial judgments in light of the information gained from these sources.

Typically in standard setting methods, one or more panels of judges are assembled for the purpose of recommending what examinees should know and be able to do to achieve some valued end and to specify an associated score on a test that is regarded as an indicator of requisite knowledge and ability. This associated score is commonly called a cut-score, but it is important to distinguish between the notion of cut score (the score on the asTTle assessment chosen to select or classify examinees with respect to the performance standard), and a performance standard ("the minimally adequate level of performance for some purpose," Kane, 1994, p. 425). Thus, cut-scores are points on the asTTle reading scale, for example, that form boundaries between levels of performance, and performance standards are the specifications of the knowledge, skills, and abilities needed to accomplish various levels of performance (see Appendix 3 for asTTle reading level standards).

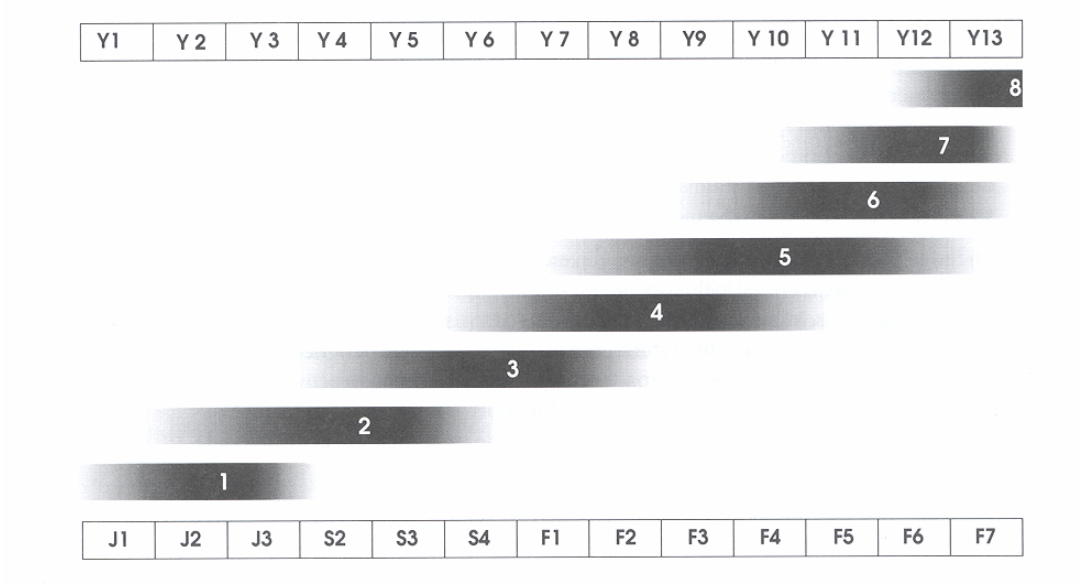
The research on standard setting indicates that different methods for setting standards may lead to different results (Jaeger, 1989). Berk (1986), over 15 years ago documented 38 methods for standard setting, and a recent review (Zieky, 2001) outlined many more methods. Standard setting has been called the "Achilles Heel" of educational testing (Hambleton & Plake, 1998) primarily because there is no clear consensus on the best choices among numerous methods and because the results of applying any method cannot easily be validated (Kane, 1994). Thus choosing a standard setting method may lead to different cut scores, and it seems defensible to use a variety of methods to ensure that any cut score is not a consequence of choosing a particular method.

Standard setting within the asTTle project.

asTTle is a computer-based application that allows teachers to create tests based on what they intend to or have been teaching to cover a teacher-specified distribution of ability. The application then chooses a best-set of items, using a linear programming method (Fletcher, 2000) optimising the test information function (Hambleton, Swaminathan & Rogers, 1991) while attending to about 15 constraints including content, difficulty, time (40 minutes maximum), minimising previously used items. The teacher can then administer the test, input the scores, and examine interactive reports. The application is particularly focused on helping the teacher or school leader interpret the meaning of the scores relative to New Zealand norms, assessing the curriculum objectives each student (and for the group) can do relative to their ability (objectives achieved, strengths, gaps, and to be achieved), suggests teaching resources relative to enhancing the achievement of the class, and relative to the curriculum levels. Thus, asTTle is designed to generate tests and report performance relative to curriculum levels. Hence, it is critical to defensibly set standards.

For the past 15 years the NZ curriculum has been based on levels of achievement. Figure 1 indicates the relationship between the Years and levels of the curricula. It can be seen that a level typically covers two years of instruction, that there may be some overlap between the levels, and that levels 2 to 4 (the focus of this version of asTTle) relates to most children in Years 5 and 7, who are approximately aged 10-13. Classifying students within a two-year level is relatively easy from a psychometric viewpoint, but of limited information to teachers, and because there inevitably is tremendous variability within a two year period as to what students within a level are expected to know and be able to do (a child beginning level 3 is much different in skills from a child at the end of level 3). Further, the overlap as indicated in the Figure 1 may mean that students classified as level 2 could be understood as doing tasks at the same levels of proficiency as a child at level 3, and hence classified at two levels simultaneously. There has been wide and prolonged debate as to common understandings to the levels (see Brown, 1998 for a brief discussion) and we experienced remarkable strengths of convictions on the meaning of a particular level as we met teachers in the development of asTTle, and these convictions were remarkably variable.

Figure 1.
The relation between Years of schooling and Levels
(from the NZ English Curriculum, 1994, p. 20)



With the release of various curriculum-referenced assessment tools (i.e., Assessment Resource Bank, asTTle, National Exemplars in 1996, 2003, 2002 respectively) this problem of non-shared understandings is being addressed on a national level. The common complaint is that one method (e.g., exemplars) is providing a different standard to asTTle, and this again is different from the Assessment Resource Bank. In terms of the reading curriculum, there were no other sets of nationally moderated descriptions of levels-related performance available. It is considered a major success of asTTle if debate is created about meaning of "levels", as only then can we address the more important question of whether the standards developed by asTTle are sufficiently rigorous and appropriately set at the desired levels.

Levels within levels.

To remedy the problems of the overlapping levels and the wide two-year spread, two solutions were adopted: one dealing with the across levels issue, and the other dealing with the within level issue. First, it was proposed that there be no overlap in the levels. This leads to a tighter explication of the proficiencies underlying each level and

thus aims to assist teachers in setting more challenging tasks. This adaptation does not mean, however, that all students progress in the same way through the levels. Second, because of the wide variation within levels the levels were further demarcated into three within-level groups. These three within-level groups are thirds and it is critical to note that the major objective is first to classify a student correctly into the level, and only then to further refine this to one of the thirds.

There was little debate about separating the students within a level into thirds, but a major concern related to the naming of these "thirds". The names have been the source of major debate in other education systems, and the major issues when deciding upon labels is that they convey normative meaning to teachers, that they are seen as sequential, that they do not have wide variation in interpretation, and that they are not already being used to denote other aspects of teacher meaning.

We began by reviewing terms commonly used in New Zealand Schools to indicate achievement. In a study of 159 School Reports (Peddie & Hattie, 1999) we identified 6 major methods: Expectations (19%), Frequency (10%), Competence (18%), Confidence (1%), Excellence (6%), and relative/normative labels (25%), and open-ended methods (21%). In no school where developmental labels used, although our advice in resolving the labelling has most often centered on such labels.

We found that labels with "developmental" connotations are most liked by teachers although they have the widest variation in meaning. For example, labels such as "developing", and "emerging" are most variably interpreted. Relative and normative labels such as "i", "ii", and "iii" (which are provisionally adopted by the MoE Exemplar project) imply a normative allocation into thirds rather than a substantive meaning of what a student knows and can do, and will exacerbate the major variability in meaning. Using other normative labels, such as letter grades "above and below average", "low and high have similar problems, particularly as they are most often interpreted relative to the class group, which may be demonstrably different from a national sample (see Robinson & Timperley, 1999). Expectations are recursive, in that they imply a set of standards without necessarily specify information about that standard. For example, above expectation begs the notion of what is "expectation". Frequency label (always, consistently, sometimes, usually, most of the time) refers appropriately to the number of

times a student can perform a task and implies little about the expected proficiency (e.g., I can sometimes put in commas, but this says nothing about the desired level of "comma" performance at the level). Confidence labels (confident, practices, introductory) are more about self-efficacy of the task than about the standards of the task. Excellence notions (excellent, very good, fair, poor) imply a distribution and come closer to the notion of standards, but are interpreted most variably, usually contingent on the situation in which the task is performed, and thus have less generality. Competence terms include "Distinction, doing well, outstanding, skilled, competent, capable" are closest to elaborating the notion of a standard.

After extensive debate on this issue, the National Assessment Governing Board (NAGB), who oversee the very extensive National Assessment of Education Progress (NAEP) in the USA adopted the terms "Basic", "Proficient", and "Advanced" to denote the three categories within levels (across all subjects). Following from NAGB (1996, 2000) usage, we adopted these terms. Basic refers to items that require partial mastery of knowledge and skills that are fundamental for proficient work at the level. Proficient refers to items that are simple applications of the knowledge and skills of the given level, and advance refers to items that are difficult applications of the knowledge and skills at this level. Many other education authorities have now adopted these terms, and thus they are becoming more commonly used – primarily because they are often introduced with little prior use and thus their meanings can be stipulative, the variability in meaning is low, and the interpretative power of these concepts high.

There was some debate as to whether a fourth within level should be introduced. Michael Scriven (consultation meeting with asTTle team, March 7, 2001) argued for a fourth label, Below Basic – given that Basic was not sufficiently descriptive of those "barely" acceptable within the level. Various education authorities have included Below Basic (e.g., Wisconsin uses "minimal"). We did not include this notion as there seemed insufficient discrimination between "Below basic" and "Basic", and strictly according to the NZ Curriculum Below Basic means a level Below – that is, Below Basic level 3, is strictly somewhere in Level 2.

It is also important to note that the standards of basic, proficient and advanced, and of the Levels 2, 3, 4 can be applied in two ways. First, they can be used to provide an

overall, best judgement, statement about the Level of performance of the student (she is at Level 3 in Reading). Second, they can be used to provide more specific information about particular dimensions of reading (she can perform at level 3 in punctuation and level 4 at inference, and so on). Further, it is expected that there will be measurement error in classifying students and a major aim of asTTle is identify and reduce the various sources of measurement error (both in estimating a proficiency, and in classifying into a level).

Hence for the asTTle standard setting process, there are eleven (between and within) levels of interest in each curriculum subject: Below level 2, level 2 Basic, Proficient, and Advanced, level 3 Basic, Proficient, and Advanced, level 4 Basic, Proficient, and Advanced, and Above level 4.

Methods for standard setting

There are three major methods of setting standards: test-centered (judgements about items), examinee-centred (judgements about examinees), and processing-centred methods (judgements about processing items). Given the difficulties of finding a single best method (see Jaeger, 1989), it seemed a worthwhile research task as well as a defensible process, to use one of each of the three major methods of standard setting, assess the comparability in cut scores, and attempt to understand the dynamics of the three methods on the cut score/standard. We used one of each of the three methods in the setting of the Reading Standards, and a different method in the setting of the Mathematics Standards (reported in the asTTle Manual—Hattie, Brown, & Keegan, 2002). Although not the focus of this study, for Writing and Tuhituhi we reversed the process and set the Standards first, and then made scoring rules relative to these standards (see Glasswell, Parr, & Aikman, 2001 for a description of standard setting in writing).

A. Test-centered methods.

The “test-centered” procedure focuses on judgments of the properties of tests or items (Jaeger , 1989). The most well-known and well-researched is the modified Angoff method. This involves judges evaluating the proficiency of the "minimally acceptable candidate" at each cut score. Judges are asked to imagine a student whose relevant

knowledge, skills, and abilities are just at the level of the performance standard (say, just at Level 3), and then to estimate judgmentally the probability that such an examinee would answer each test item correctly (for reviews see Berk, 1986; Jaeger, 1989; Kane, 1994). The Angoff method is the most widely used standard-setting procedure in large-scale educational assessment.

There have been criticisms of the Angoff method (Cizek, 1993; Kane, 1993; Pellegrino, Jones & Mitchell, 1999; Shepard, Glaser, Linn, & Bohrnstedt, 1993; U. S. General Accounting Office, 1993), and these criticisms point to the inconsistent results across different types of test items, and that the method imposes a judgment task that is beyond the cognitive capacity of most judges (Chang, 1996; DeMauro, 1995; Impara & Plake, 1996; Quereshi & Fisher, 1977; Taube & Newman, 1996; Thorndike, 1980; Wheeler, 1991). The current implementation of the modified Angoff method made adaptations aimed to deal with many of these criticisms.

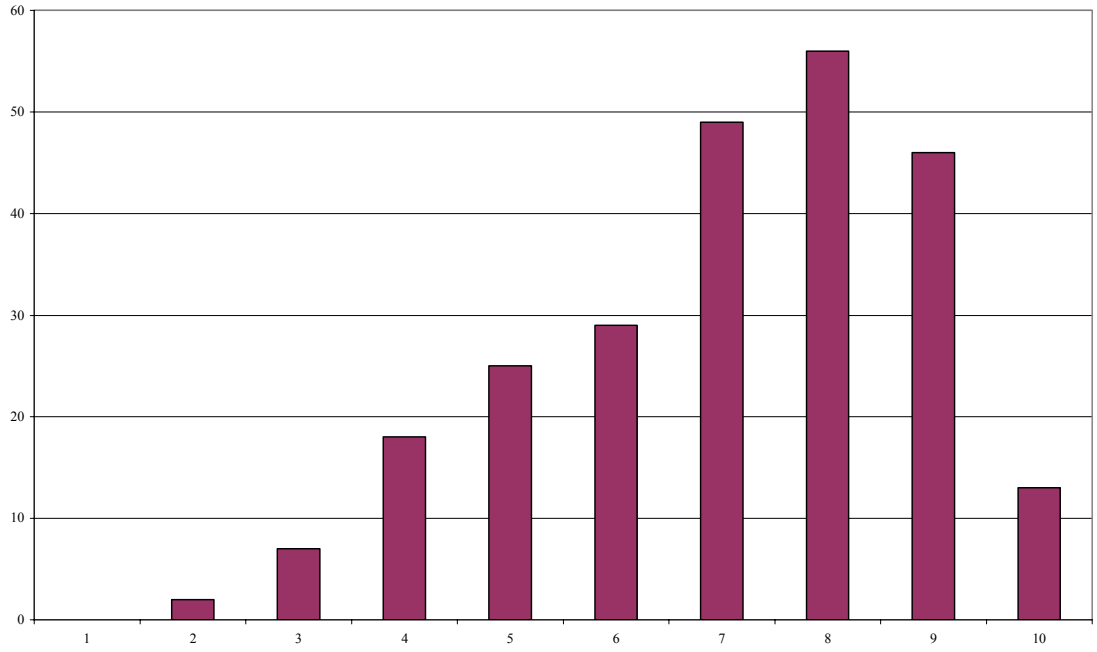
The data for this task evolved from a cut-score setting exercise undertaken by 14 judges using 245 items taken from the asTTle Reading assessment tasks. Jaeger and Mills (2001) demonstrated that panel sizes of about 15 are about optimal in standard setting workshops. Fourteen panellists were recruited for a two-day workshop from a large number of practising primary school teachers nominated by principals as expert teachers in literacy instruction and curriculum assessment. The teachers were selected from across a range socio-economic schools, most had taught 10 or more years, and held senior positions with responsibility for either English or assessment within their schools.

1. Setting the cut score.

The first task involved calculating the mean probability and standard deviation for each item. The teachers in this cut-score setting exercise were required to consider the test items, one item at a time, and judge the probability of a minimally competent student (i.e., a student who just meets the standard described by the levels) at any level correctly answering that item. These judgements were made by requesting a judgement on a 1 to 10 scale (10 certain to attain). The 14 judges' estimates of the probability that a benchmark student at each level (2, 3, and 4) will succeed on the items were averaged for each Reading item. The standard deviation of each item is used as an indicator of agreement.

The number of items at each of the ten percent probability bands for the minimally competent student at Level 3 is depicted in Figure 2. The average was .69 (*SD* = 1.45) indicating a relatively easy set of items and a high degree of variance.

Figure 2.
Number of items at each 10 percent probability band for minimally competent Level 3 student



When the judge's estimate for an item is summed across all the items, the result produces an estimate of the raw cut score (i.e. the score above which a student is considered to be above the benchmark standard) for that judge. The mean of the judges' estimates then provides the estimate of the raw cut score for the "average" judge. This mean would be the same as that obtained from summing the average of the judge's estimates across all 245 items and then dividing by the number of judges.

2. The performance of the teachers/judges.

One of the concerns is whether the teachers/judges were excellent in their judgments about performance on the items by the "minimally competent student", and whether there are items for which there is more than expected variability. The Rasch Unidimensional Models for Measurement (RUMM) Program was used to estimate the location of the judges and the items using the ratings of 14 judges on the 245 items.

These estimates were used to determine the fit of the items to the model and the internal consistency of the judges.

The judges were then provided with information regarding how well their estimates accorded with the estimates of the average judge, taking account of the relative "harshness" of the judges. Judges then had the opportunity to reconsider their estimates. The new estimates became the data set for the remainder of the analyses. The RUMM Program was then used to reanalyse the modified data set. Table 1 shows a sample of similar information for the judges, and this provides an indication as to how internally consistent each judge has been in his/her estimates. Clearly, judges 001 to 004 showed acceptable fit characteristics (Chi square) while judge 005 exhibited low fit. The non-significant probability statistics for judges 001 to 004 indicate that their ratings were relatively similar while the statistically significant probability value for judge 005 indicates high internal inconsistency in ratings relative to the mean.

Table 1.
Judge fit statistics based on RUMM.

Judge	Fit(z)	Fit(ChiSq)	Probability
J001	-0.402	3.028	0.541
J002	-2.309	2.771	0.586
J003	0.541	2.279	0.676
J004	-0.135	4.757	0.295
J005	1.191	15.620	0.000

3. Translating the logit cut score onto pre-developed asTTle Literacy scales.

It is noted that only 245 items from the total item pool of 825 items were used in the modified Angoff standard setting. Clearly, if these items are relatively easy then a modification needs to be undertaken to place the cut score on the underlying asTTle ability scale (similarly if these sampled items were too difficult then there needs to be a transformation to bring the sampled items onto the distribution as the total item scale). To make this correction, a translation constant was derived to transfer the logit cut score onto the asTTle scale. This constant is the difference between the average logit of the 245 items and the average logit of all 825 items, and was .173, indicating that the sample

items were easier than the total by a small margin. The final cut scores for the level 2, 3, and 4 were then calculated, and the Basic, Proficient, and Advanced cuts within each level were estimated at one-third within each level. The cut score between level 2 and 3 was 604, and between level 3 and 4 was 710.

4. Concluding comments

There was much debate among the teachers about the meaning of the "minimally competent student" at each level, and it was clear that this was not a concept that they were comfortable in using. The standard error of the cut scores was rather high and this reduced the confidence in the use of the cut-scores. Most critically, the time and effort (two days with 14 judges) was exorbitant to merely set two cut scores, and there was no evidence that these teachers could have dependably set cut scores for the levels within levels using this method. It was too difficult, too prone to error, and the confidence of the teachers in their work was not very high.

B. Item Centred Method

The item-centred approach involves professional judges (teachers) assigning assessment items to the various levels on the basis of their interpretation of the New Zealand Curriculum documents. These documents, to varying degrees of specificity, detail what students are expected to know and be able to do as they progress through the levels. Many of these specifications, in reading, are listed as relatively global learning objectives that require further detailed specification (Brown, 1998). A curriculum mapping exercise (Limbrick, Keene, & Girven, 2001) provided initial guidance as to level nature of achievement objectives. An item-signature process (Burstein, Koretz, Linn, Sugrue, Novak, Baker, & Lewis, 1995/1996) had previously been used to identify the curriculum characteristics of reading texts and items (Meagher-Lundberg & Brown, 2001a & 2001b). Items were considered to have a signature where there was agreement of two-thirds or more and where agreement was not reached discussions ensued to ensure consensus and to further refine understanding of judgement rules. Thus, significant information about what each item required students to do in terms of curriculum objectives, curriculum processes, cognitive processes, and reading comprehension processes was available to judges in assigning a curriculum level to each assessment item. Furthermore, the

appropriate curriculum level for each reading passage had been assigned in the item-signature workshops.

A two-day item-centred standard setting workshop was run to rate reading items to assign curriculum level to each item. The workshop participants were currently practising teachers, and they had previously participated in item signature methods before commencing the classification work. Fourteen teachers were purposefully selected in light of their previous involvement with the project, and their experience in the area of literacy. These teachers were highly experienced with 13 having more than ten years teaching service, currently taught either senior school (Years 4 to 6) or intermediate (Years 7 to 8), and they had wide ranging previous teaching experience from new entrant to junior high school level. They also had responsibility for curriculum development in their schools across a range of curriculum areas including English, Literacy leadership, assessment, literacy projects and contracts, had qualifications higher than the necessary minimum and four had University degrees with three currently undertaking Masters degrees.

1. Item Classification

There were 295 reading items based on 44 reading texts that had been previously classified according to a number of signatures but which still required difficulty/levelling. The items and texts were classified as level 2, 3, 4, or greater than level 4. Within each level a further delineation was made for Basic, Proficient, and Advanced. Basic items were those simple enough for students whose reading skills and knowledge required significant scaffolding or support at this curriculum level. Proficient items were those considered appropriate for students whose reading skills and knowledge were competent enough to use the text in instruction at this curriculum level. Advanced items were those considered appropriate for students whose reading skills and knowledge were independent at this curriculum level.

2. Level Cut Scores

An additional task of this workshop was to create preliminary descriptions of the performance standards or rules for assigning items to curriculum levels and sub-levels. Thus, a significant proportion of time was spent in whole group discussion in response to reading items and the creation of a common set of rules. Thus, less time was spent rating

item level difficulty individually. In the end 108 of the 295 items were rated individually based on the commonly developed preliminary reading level performance standards (Appendix 1). The inter-rater agreement index was used to assess the extent to which scores assigned by raters agreed with each other (> .6 is considered acceptable). Inter-rater agreement for item difficulties was excellent (standardised α = .985).

For each teacher the cut-score of the logits of the items was estimated, and then averaged across the teachers. Table 2 provides the distribution of levels as assigned by the teachers. The correlation between the logits of these items and the final assigned curricula levels was .67.

Table 2.
Cut scores based on the item-centred method for the asTTle Reading items

Levels	Number of items	Ave Logit	Cut score
Above level 4 Advanced	0		
Level 4 Advanced	7	2.536	754
Level 4 Proficient	14	1.585	659
Level 4 Basic	16	1.418	642
Level 3 Advanced	81	.520	552
Level 3 Proficient	41	.893	589
Level 3 Basic	60	.009	501
Level 2 Advanced	10	-.019	498
Level 2 Proficient	43	-.871	413
Level 2 Basic	23	-1.969	303
Below Level 2 Basic	0		

3. Conclusion

This method successfully generated performance descriptions of standards and successfully guaranteed agreement among the judges operating independently the performance standards. Nevertheless, this group of judges had required significant training before they were able to accurately agree on the curriculum level difficulty of items. Earlier work in the item signature workshops had shown relatively low levels of agreement on item difficulty and so the success of this process can be attributed to prolonged exposure, extensive discussion, and high levels of expertise in literacy.

C. Examinee-centred method.

The emphasis in this method is on a holistic "on-balance judgement" of the total performance of a single student hence the notion of examinee-centered methods. Judges are requested to categorise examinees based on some criterion assessment of the

examinee's level of performance relative to performance standards. The cut scores are then set by identifying points on the score scale that would be most consistent with these categorisation decisions. There are many examinee-centered methods, including the borderline-groups method (Livingston & Zicky, 1982), the contrasting groups method (Livingstone & Zicky, 1982), and the method used in this study, the generalised examinee-centered method. Briefly, judges are asked to evaluate a representative sample of student asTTle scripts using a rating scale that is defined in terms of the performance standards (e.g., level 2, 3, and 4). These ratings are then linked to the student's test scores to generate a relation between scores and the ratings, which is then used to assign a cut score to each performance level.

The method used in this study is outlined in detail in Cohen, Kane, and Crooks (1999). They begin by noting the advantages of the examinee-centered methods, including that the judges are asked to rate the performance of students rather than perform the less familiar task of rating test items, the emphasis on-balance or global judgements rather than an item-by-item judgement, and teachers are typically more familiar with judging overall student performance making many decisions about how to professionally weight the differential performances of students across many items. The generalised examinee-centred method involves using all ratings to set each cut-score in a simultaneous manner, while building a profile of the performances desired at each performance level. There are four major steps in the model.

1. Draw a sample of examinees from the population of test takers

The selection of judges follows that of most standard setting methods. We asked 17 currently practising teachers, nominated by their principals as having much experience and expertise in teaching literacy to the students represented in the sample (Years 5 to 7) to be part of a two-day standard setting workshop. These teachers were highly experienced with 16 having more than ten years teaching service, currently taught either senior primary school (Years 4 to 6) or intermediate (Years 7 to 8), and who had wide ranging previous teaching experience from new entrant to intermediate school level. They also had responsibility for curriculum development in their schools across a range of curriculum areas including English, Literacy leadership, assessment, literacy projects

and contracts; had qualifications higher than the necessary minimum and three had University degrees with three currently undertaking Masters degrees.

2. Rating performance using performance standards.

The ratings were based on 50 students' responses to the asTTle reading tasks. Each student booklet consisted of about 40 closed reading items (e.g., multiple choice, short constructed responses, matching, editing). Thirteen papers were chosen for training, and five sets of 10 papers each were selected for the actual rating. As there are 16 forms of the asTTle booklets used in standardisation (each with link items), a random sample of 6 forms was chosen in this study. Papers were chosen as representative of the distribution of performances in the population of examinees, by randomly sampling one paper from each decile in the distribution of total test scores. As noted by Cohen, Kane and Crooks (1999), this range is intended to provide judges with information about both the range and the distribution of performance in the population of students tested. The judges were told that the papers in each set were roughly representative of the distribution and were asked to read over all papers before rating them. There was no information provided about scoring multiple-choice items (thus requiring judges to emulate the cognitive processing of the students as they tackled each item), the scores of the open-ended responses were provided, and judges were implored to not merely sum up scores but to depend on their "on-balance judgements" of the student's performances.

The judges were introduced to preliminary descriptions of the performance standards (developed in the item-centred workshop) associated with each of the performance categories. Appendix 1 provides an outline of the descriptions, which was subjected to much discussion, modification, and agreement over the training period. These levels lead to the performance rating scales intended to help judges rate each student's performance relative to the performance standards. We placed particularly attention on first assigning the students' performances to level 2, 3 or 4, and only then assigning to Basic, Proficient, and Advanced. We asked judges to make three piles representing the levels, and then work within each to make the finer judgements. The performance rating scale was 0 for Below level 2, 1 for 2 Basic, and so on.

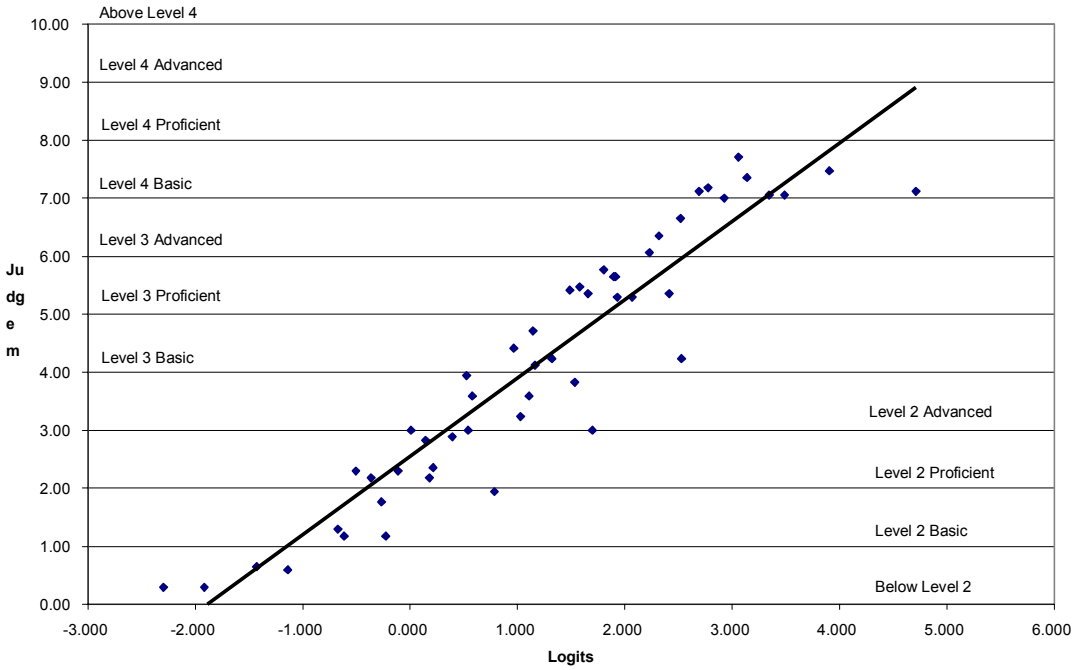
Judges were trained following the recommendations in Cohen, Kane and Crooks (1999) and most standard setting methods (Raymond & Reid, 2001). They were

introduced to the preliminary performance standards, the rating scale, and then practised using them with one student script. After much discussion about the assigned performance rating, the descriptions were improved, and this continued for three3 more scripts, handled one at a time. Batches of three, and then four further scripts, were rated, and a final description of performance standards was agreed. Emphasis was placed on rating the scripts relative to the performance standards, and not to use a norm-referenced or scoring system.

3. Setting the cut scores

Across the 50 scripts, the 17 judges made 850 ratings, leading to an overall mean on the performance scale of 3.85 (sd = 2.24), or when translated to the asTTle scale, the mean was 567 (sem = 16). Note this value is close to Level 3 Basic. Figure 3 presents the scatterplot of ratings on test scores using the sunflower method (which gives an indication of density). The relation appears linear with greater spread or variance as levels increased. The overall correlation between ratings and test scores was .93, which is remarkably high and provides confidence in the cut-score judgements of these teachers.

Figure 3.
Relation between Logits and Examinee-Centred Method Judgements



As recommended by Cohen, Kane and Crooks (1999), the equating relation zs=zr was used to specify the relation between ratings and scores. Table 3 presents the cut score estimates for the asTTle reading items.

Table 3.
Cut scores on the Reading asTTle scale Examinee-Centred Method

		Cut scores	sem©	Total rating error
10	Above level 4 Advanced	738	3.62	8.72
9	Level 4 Advanced	698	3.67	8.77
8	Level 4 Proficient	659	3.74	8.84
7	Level 4 Basic	620	3.82	8.92
6	Level 3 Advanced	581	3.93	9.03
5	Level 3 Proficient	542	4.04	9.14
4	Level 3 Basic	503	4.17	9.27
3	Level 2 Advanced	463	4.31	9.41
2	Level 2 Proficient	424	4.46	9.56
1	Level 2 Basic	385	4.63	9.73
0	Below Level 2 Basic	346	4.80	9.90

4. Estimating the precision of the cut scores

Cohen, Kane and Crooks (1999) recommend estimating the magnitude of errors based on two potential sources. The first is the error inherent in the estimation of the equating relations, and the second is the sampling error associated with the selection of raters, in particular the variability in the general levels at which the raters tended to set the standards. For the errors due to variability in equating, recall that the judges were provided with five sets of 10 scripts, each set roughly representative of the distribution of students across the total score scale. Further, judges were provided with breaks between each set, and each set was from a different test form, thus enhancing the sense that these were (quasi-) independent replications. The relations between the scores and performances were developed separately for each of the five sets of 10 test performances and the resulting five equating functions were used to generate five values of each of the cut scores. Treating the cut scores based on the full set of ratings of the 50 test performances as equivalent to the average of these five separate estimates, the standard error in the overall estimate is given by the standard deviation of the estimates, c, divided by the square root of the number of estimates (in this case 5): These errors are presented in Table 3 above, and they average 4 score points on the final asTTle score scale (Mn = 500, sd = 100); a relatively trivial value. There was considerably more variability at level 2 than at level 4 (r between ratings and sem = .45). Thus, the judges were more consistent in their rating of the upper compared to the lower levels.

The second source of error is due to raters, and this reflects whether raters tended to be more or less severe in their ratings of individual student papers. The mean judges' rating was estimated, and the standard error of this mean (sd / Sqrt n, where n = 17) was estimated. This was then translated to the test score using $SEM_R(S) = m\ SEM(R)$, where m is the slope of the linear equation relating ratings to scores. This standard error due to raters was 5.10; again a relatively trivial value.

The total cut-score error is the addition of these two above sources of error (see Table 3). The magnitude of these total errors is very small. This evaluation is based on Cohen, Kane and Crooks' (1999) recommendations that the standard errors of the cut scores should be considerable smaller than the standard errors of the test scores (which is 16.48). On average the total cut score error is about half the test score error, and thus

there is little added to the overall error of measurement. The misclassification rate into levels, or levels within levels would be quite small indeed. As Cohen, Kane and Crooks noted, a standard error in the cut score that is one half the standard error of measurement adds relatively little to the overall error and, therefore, will have relatively little impact on the misclassification rates.

5. Concluding comments

The judges were confident of their standard setting judgements by the end of the session, especially for Levels 2 and 3. Confidence on agreement about Level 4 was lower. Nevertheless, the standard errors of measurement were very small, while the cut scores were linear and consistent across forms. The teachers found the discussion about the nature of reading at each curriculum level helpful to their own teaching and assessment practice.

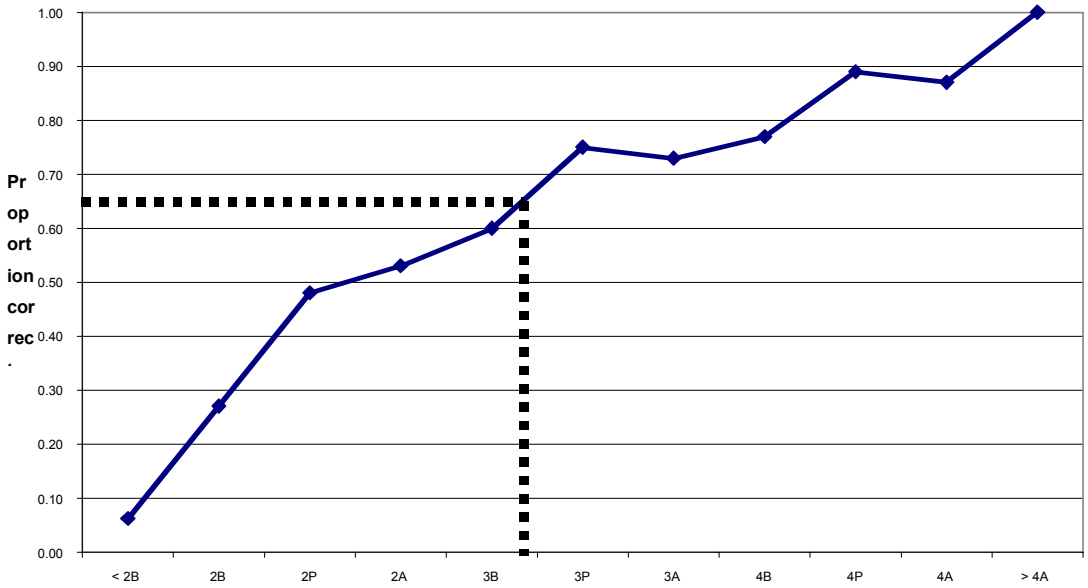
D. Performance Threshold cut score method

The predominant argument in standard setting is that it is not possible to use the statistical information from the students to defensibly set the cut-scores. While we agree with this claim, we did include a fourth performance threshold method to complement the above methods to assist in setting the standards. A response probability or performance threshold can be used to indicate the average percentage of students getting an item correct for an achievement objective to be understood as being something that students at a particular stage of development 'can do' (Zwick, Senturk, Wang, & Loomis, 2001). Thus, an item belongs to say, Level 3 Basic if around an appropriate percentage of students considered 3 Basic get the item correct, while fewer at 2 Advanced score it correct, and more at 3 Proficient score correct on the same item. Commonly 50% response probability is considered appropriate to assign an item as something that students can do, provided there is no guessing involved. Given the effect of guessing (estimated for asTTle reading at around p = .18, Leeson & Fletcher (2003, December)—it is considered appropriate to use a response probability around .68 for tests involving 100% multiple-choice items as the cut score for a score region. asTTle reading adopted .65 as an appropriate performance threshold to identify items as belonging to a given curriculum level. We determined the best "cut scores" from the above three methods and

then assigned the students into 11 score bands based on the asTTle total reading scores. We then identified the performance threshold at which 65% of the students scored the item correct.

For example, Figure 4 presents the proportion correct for item 10 from Form G in reading. Sixty-five percent of the students who scored within the range of 3 basic answered this item correctly. The graph shows the proportion of students who scored the item correct within each level. As with all items there was a monotonically increasing slope across the levels. Any item that did not show this increase had already been discarded as a result of an earlier item analysis. It is noted, however, that the slope varied across items and the optimal slope is a 45 degree angle such as exhibited for item 10G in Figure 4.

Figure 4
Performance on Reading item QG10



Each item was than classified as belonging to the level where the 65% proportion was located. The average item logit of all those classified within a level was calculated and scaled into the asTTle reading scale. Table 4 presents the number of items in each level and the corresponding cut-score for that level. It is noted that there are many more Level 2 than 4 items, and this was subsequently rectified by adding in more level 4 items in later calibrations of asTTle reading assessments.

Table 4.
Number of items using 65% Performance-Centred Method

Levels	Number	Cut-score
Above level 4 Advanced		
Level 4 Advanced	16	785
Level 4 Proficient	9	718
Level 4 Basic	20	725
Level 3 Advanced	43	650
Level 3 Proficient	57	622
Level 3 Basic	64	570
Level 2 Advanced	106	509
Level 2 Proficient	108	422
Level 2 Basic	67	309
Below Level 2 Basic	145	224

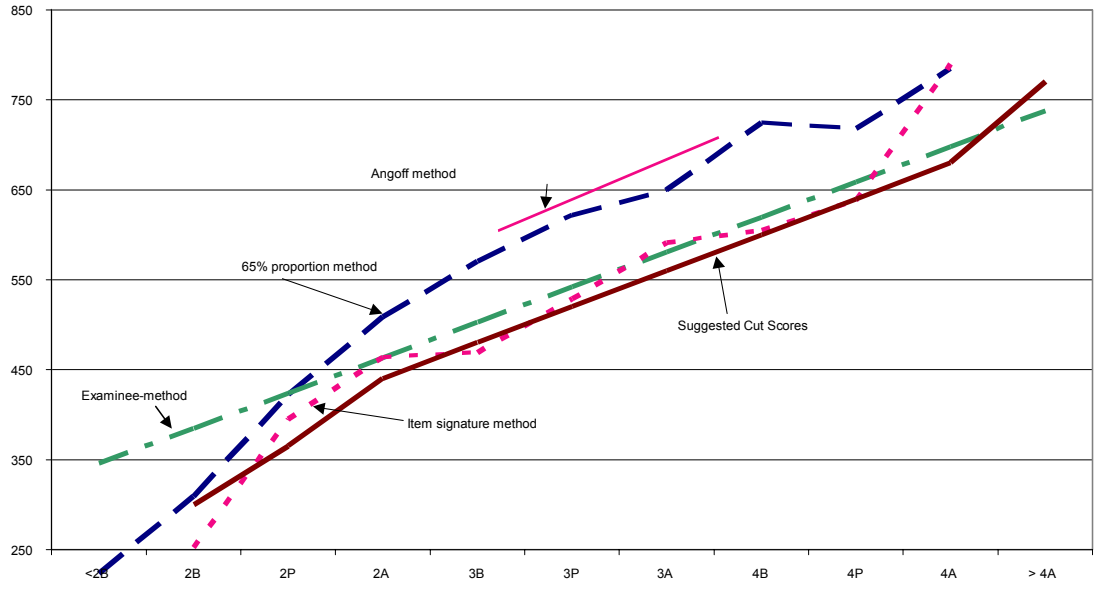
Concluding comments

The scores set using this method, a norm-reference procedure, show similarity to those set with the modified Angoff, examinee-based method. The 65% performance threshold method of setting cut scores is relatively efficient but depends on a defensible process by which to assign students to curriculum levels in the first place. Thus it functions best as a confirmatory method for identifying the kinds of skills students already assigned to curriculum levels can perform.

Comparison of the four methods

Figure 5 presents the cut scores from the four methods. There were similarities between the Angoff method and the proportion method, but both these methods would have led to cut-scores that were 1 standard deviation greater than the other two methods. There was more agreement, however, at the lower levels were there many more items. In this case the examinee, signature, and proportion method were more consistent in the choice of cut-scores.

Figure 5.
Standard Setting via four methods



As this was the first time that a curriculum level indexed assessment tool at the level of item performance would be available to NZ teachers in the elementary years, we were cautioned to use more conservative cut-scores. There was one other reason for choosing more conservative levels and that related to the low number of students in the expected age range (Years 7-8) who performed at level 4. If we had used the 65% proportion threshold or Angoff methods there would be almost no students in the sample of 18,000 classifiable as Level 4. Many New Zealand teachers would most likely not see this as defensible.

We did note a major reason for this high standard was that teachers in the various workshops had little difficulties identifying a level 2 or 3 piece of work on the basis of what a child could do. When they initially assigned a piece of work as Level 4, the debate quickly shifted to what the child could not do (e.g., could not punctuate, could not find information from Level 3 text) and thus level of overall performance for the student was soon shifted down to level 2 or 3. It is suggested that teachers grade work at the higher levels of elementary school more on the basis of what they do **not** see in the student's work relative to their conception of level 4; whereas for level 2 and level 3 they

look for the 'best-fit' overall judgement of the level of proficiency exhibited by the student. This conjecture needs further investigation.

We thus set the standards using the information more from the examinee and item signature methods and these are included in the asTTle application. The standard error using these two methods is less than half the gap between any two sub-levels, hence the scores and levels can be used with much confidence. The last phase was to use these cut-scores and then assess the competencies of the items within these cut scores. Appendix 1 lists many of these competencies and it is suggested that NZ teachers and those interested in standards begin to discuss whether they are satisfied with these assignments of competencies to levels. If not, there may need to be changes in the competencies, which would then lead to changes in the asTTle cut-scores (and certainly not the other way around whereby we arbitrarily change the asTTle cut-scores to ensure "more" students at a level).

Concluding comment

The overall message of this paper is that there are wide variations in the decisions made depending on the standard setting method being used. It certainly suggests that by asking judges to consider the "minimally competent student" at each level (as in the Angoff method) the results are more in line with the actual performances of students (as crudely assessed by our 65% method). Judges are possibly using real norm-referenced performance to make decisions as to where a curriculum level begins and ends. The other two methods are more related to the competencies and standards and less to the actual performances of students, and thus are more likely to reflect the "desired" or tacit standards intended by the curriculum documents as to where each curriculum level begins or ends.

Most important, the use of four methods provides more confidence that the cut-scores within the asTTle application for reading are defensible and not merely based on a committee's opinions after whatever length of debate. The methods capture the processes and beliefs leading to best judgements about standards, and led to a clearer articulation of the proficiencies within each level. When teachers interpret the levels and sub-levels within asTTle they need to reference the meanings to these competencies – rather than to some other non-curriculum scale such as reading ages and not subjected to rigorous

standard setting procedures. The more important debate is not such a comparison to other tests, but whether the competencies listed in the Appendix 3 (as determined by defensible standard setting methods) are deemed adequate for our NZ students. This is a debate about desired standards relative to the curriculum and to the aspirations of those who are invested in desired rates of progress. Project asTTle (version 1 and 2) reading assessments can be used confidently to monitor children's curriculum level progress because of the processes described in this report.

References

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56*, 137-172.

Brennan, R. L. (1995). *Standard setting from the perspective of generalizability theory.* Paper presented at the Joint Conference on Standard Setting for Large-Scale Assessments, Washington, DC.

Brown, G. T. L. (1998). The New Zealand English Curriculum. *English in Aotearoa, 35*, 64-70.

Burstein, L., Koretz, D., Linn, R., Sugrue, B., Novak, J., Baker, E. L., & Harris, E. L. (1995/1996). Describing performance standards: Validity of the 1992 National Assessment of Educational Progress achievement level descriptors as characterizations of mathematics performance. *Educational Assessment, 3*(1), 9-51.

Chang, L. (1996). Does a standard reflect minimal competency of examinees or judge competency? *Applied Measurement in Education, 9*, 161-173.

Cizek, G.J. (1993). On the disappearance of standards. *Education Week, 13*, (19), 32,24.

Cizek, G. J. (Ed.). (2001). *Setting Performance Standards. Concepts, methods and perspectives*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A generalizable examinee-centred method for standard setting on achievement tests. *Applied Measurement in Education, 12*(4), 367-381.

DeMauro, G. E. (1995, March). *Construct validation of minimum competence in standard setting.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Fletcher, R. B. (2000). *A Review of Linear Programming and its Application to the Assessment Tools for Teaching and Learning (asTTle) Projects* (Technical 5). Auckland: University of Auckland.

Glasswell, K., Parr, J., & Aikman, M. (2001). *Development of the asTTle Writing Assessment Rubrics for Scoring Extended Writing Tasks* (Unpublished technical report). Auckland: University of Auckland, asTTle project.

Hambleton, R. K., & Plake, B. S. (1998). *Categorical assignments of student work: An analytical standard-setting method designed for complex performance assessments with multiple performance categories.* Paper presented at the Annual meeting of the American Educational Research Association and the National Council on Measurement in Education, San Diego, CA.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamental of item response theory*. Newbury Park, CA: Sage.

Hattie, J., & Peddie, R. (in review). School reports: " Praising with faint damns".

Hattie, J. A. C., Brown, G. T. L., & Keegan, P. J. (2003). *Assessment Tools for Teaching and Learning Manual: Version 2, 2003*. Auckland: University of Auckland.

Impara, J. C., & Plake, B. S. (1996, April). *Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

Jaeger, R. M. (1989, April). *Selection of judges for standard setting: What kinds? How many?* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

Jaeger, R. M. (1994). *Setting performance standards through two-stage judgemental policy capturing.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Jaeger, R. M., & Mills, C. N. (2001). An integrated judgement procedure for setting standards on complex, large-scale assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates Ltd.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Kane, M. (1993). *Comments on the NAE evaluation of NAGB achievement levels*. Washington, DC: National Assessment Governing Board.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*, 425-461.

Leeson, H., & Fletcher, R. (2003, December). *An investigation of fit: Comparison of 1-, 2-, 3- parameter IRT models to project asTTle data.* Paper presented at the Joint NZARE/AARE Conference, Auckland.

Limbrick, E., Keenan, J., & Girven, A. (2000). *Mapping the English Curriculum* (Technical report 4). Auckland: University of Auckland, asTTle Project.

Livingston, S. A., & Zicky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Meagher-Lundberg, P., & Brown, G. T. L. (2001a). *Item signature study 2: Report on the characteristics of reading texts and items from calibration 2.* (Technical Report 16). Auckland, New Zealand: University of Auckland, Project asTTle.

Meagher-Lundberg, P., & Brown, G. T. L. (2001b). *Item signature study: Report on the characteristics of reading texts and items from calibration 1.* (Technical Report 12). Auckland, New Zealand: University of Auckland, Project asTTle.

Ministry of Education. (1993) *The New Zealand Curriculum Framework.* Wellington, Learning Media.

Ministry of Education. (1994) *English in the New Zealand Curriculum*. Wellington, Learning Media.

Ministry of Education. (2003).*National Exemplars Trial.* Retrieved August, 15, 2003, from http://www.tki.org.nz/r/assessment/exemplars/index_e.php

National Assessment Governing Board (NAGB). (1996,2000). *Mathematics Framework for 1996 and 2000 National Assessment of Educational Progress.* Washington, DC: NAGB

New Zealand Council for Educational Research (NZCER) (2003) *Assessment Resource Banks.* Retrieved August, 15, 2003 from http://arb.nzcer.org.na/nzcer3/nzcer.htm

Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (1999). *Grading the nation's report card. Evaluating NAEP and transforming the Assessment of Educational Progress*. Washington DC: National Academy Press.

Pitz, G. F., & Sachs, N. J. (1984). Judgment and decision: Theory and application. *Annual Review of Psychology, 35*, 139-163.

Plake, B.S. (1995). The Performance Domain and the Structure of the Decision Space. *Applied Measurement in Education, 8,* 1, 3-14.

Quereshi, M. Y., & Fisher, T. L. (1977). Logical versus empirical estimates of item difficulty. *Educational and Psychological Measurement, 37*, 91-100.

Raymond, M. R., & Reid, J. B. (2001). Who Made Thee a Judge? Selecting and Training Participants for Standard Setting. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, methods and perspectives* (pp. 119-158). Mahwah, NJ: Lawrence Erlbaum Associates.

Robinson, V., & Timperley, H. (2000). The Link between Accountability and Improvement: The Case of Reporting to Parents. *Peabody Journal of Education, 75*(4), 66-89.

Shepard, L.A. (1993). Evaluation test validity. In L Darling-Hammond (Ed.) *Review of Research in Education, 19,* (pp. 405-450). Washington, DC: American Educational Research Association.

Shepard, L. A., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement*. Stanford, CA, Stanford University: National Academy of Education.

Taube, K. T., & Newman, L. S. (1996, April). *The accuracy and use of item difficulty calibrations estimated from judges' ratings of item difficulty.* Paper presented at the Annual meeting of the American Educational Research Association, New York.

Thorndike, R. L. (1980). Item and score conversion by pooled judgement, *Proceedings of the Educational Testing Service conference on testing equation*. Princeton, NJ: Educational Testing Service.

US General Accounting Office. (1993). *Educational achievement standards: NAGB's approach yields misleading interpretations*. Washington DC: [GAO/PEMD-93-12].

Wheeler, P. (1991). *The relationship between modified Angoff knowledge estimation judgements and item difficulty values for seven NTE specialty area tests.* Paper presented at the Annual Meeting of the California Educational Research Association, San Diego, CA.

Zieky, M. J. (2001). So much has changed: How the setting of cut-scores has evolved since the 1980s. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 19-52). Mahwah, NJ: Lawrence Erlbaum Associates Ltd.

Zwick R., Senturk D., Wang J., & Loomis S.C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice, 20*, 2, 15-25

Appendix 1 Curriculum Level Definitions: Close Reading Level 2

According to the curriculum, students should:

Respond (i.e., perform answering or corresponding action; show sensitiveness to by behaviour or change; say or do something in reply or as a reaction) **to**: language, meanings, and ideas in different texts.

Relate (i.e., bring into relation, establish relation between; establish a causal connection between) **to**: personal experiences.

Students make a general link to their own world when responding to the text.

Comparisons are simple, literal, and list-like.

Level 2 Students can

Respond to the text

Retrieve facts and locale

Recall and retell

Ask questions

Use cues in text to help understand

Synthesis and sequence ideas

Understands word beginnings, endings, and medial sound s (diagraphs, diphthongs, blends, rhymes)

Infer at a simple level/one step thinking

Use visual language to interpret the title, illustrations, formats

React to messages and detect author's message

Can predict what happens next

Can distinguish between fact and fiction

Uses own experience to answer questions

Major headings

Text Structure

Imagery

Understanding text

Sophistication of ideas

Ability to link text to deeper levels of thinking

BASIC

Task Difficulty: This item is answered easily, quickly, and without support. Basic readers will attempt and may answer correctly but only with considerable effort or support.

Experience: Students refer literally to their own experience, or make a seemingly obscure or tangential comment, the interpretation of which depends on knowing the personal experience of the student. Making a link between the text and their own experience

Vocabulary: Task requires students to verbatim recall from or locate in the text. They can retrieve information from the text

Language: Language of task question and text being read should be easy for children at this level. Focused more on the mechanics of reading

Cognitive Processing: Only one step is required to answer the task. Can gain meaning from the text usually in a literal way (e.g., repeating the text words)

Scaffolding: Significant support is offered either in the text or the task making the completion of the item relatively easy.

Comprehension: Literal comprehension influenced by own personal experience.

Surface Features: Some level of difficulty spelling frequent common words

PROFICIENT

Task Difficulty: most readers within the level answer this item easily, quickly, and without much effort or support.

Experience: Reference to own experience is more complex, relevant and explicitly related to text or task.

Vocabulary: Task requires students to locate information in almost identical words or very close to the prompts in the task. No inference is required to understand terms.

Language: Language of task question and text being read provides some challenge for children at this level.

Cognitive Processing: Two steps are required to answer the task. Consistent location of verbatim material. Able to handle simple sequencing tasks.
Comprehension: Some inferring tasks are handled orally not in writing. May give one part to a two-part question. Retelling of text is still highly dependent on text.
Texts: Tasks relate to simple format texts, especially prose narrative.
Surface Features:
Punctuation: Sentence beginnings and endings handled accurately (i.e., Capitals for first word and proper nouns and final mark (i.e., full stop or question). Other marks are infrequent.
Grammar: reasonably accurate tense usage but not consistent
Spelling: Accurate spelling of all of the 300 most frequent common words.

ADVANCED
Task Difficulty: only the best readers within the level answer this item easily, quickly, and without much effort or support. At this level proficient readers will attempt and may answer correctly but only with considerable effort or support. Basic readers may not attempt this item, or, despite considerable effort or support be unable to answer it correctly, or, may respond only through guessing.
Purpose, Intention, Audience: Able to recognise text purpose, intention, or audience of basic text forms, such as a letter, invitation, or narrative.
Experience: Able to reflect beyond the text. Empathises with character through own experiences. Can relate to own experience is explicitly related to text being read and may invoke thematic sense of experience rather than literal relationship. Empathises with characters through own experience.
Vocabulary: Task requires students to use substitution or synonymy to answer the task.
Language: Language of task question and text being read is challenging for children at this level.
Cognitive Processing: Three or more steps are required to answer the task.
Comprehension: Involvement with one piece of information demonstrated. Able to make simple reorganisation or restructuring of information such as beginning, middle, and end.

LEVEL 3

According to the curriculum, students should:
Discuss (i.e., Examine by argument, debate; talk about so as to reach a decision; to talk or write about a topic in detail): language, meanings, and ideas in a range of texts.
Relate (i.e., Bring into relation, establish relation between; establish a causal connection between) **understandings** to: personal experiences; and other texts
Discussion involves comparison, classification, and contrast. No longer do students just relate the text to their personal experience, and when those links are made they are much more specific. Students are much more able to interpret and explore text than they were in Level 2.

Level 3 Students can
Respond to the text
Retrieve and reorganise facts and locale
Recall and retell in own words, and use retelling to predict
Ask questions
Use cues in text and own understanding to help understand
Synthesis and sequence ideas
Understands word beginnings, endings, and medial sound s (diagraphs, diphthongs, blends, rhymes)
Can skin scan, and select key ideas
Can summarise and analyse the message in text, using imagery and syntax
Can use and understand synonyms and antonyms
Infer and interpret the underlying meaning of a text
Use opinion based on the text or own understanding
Use visual language to interpret the title, illustrations, formats
React to messages and detect author's message
Answers questions logically
Can evaluate and compare

Be able to read & understand all the questions
Texts are at a level 3 difficulty by noun frequency (reading age 10-12)
Imagery treated literally not explored deeper levels
Verbatim copying of text
Low-level understanding of text
Imagery higher level 2 or more steps
High-level lateral thinking
Factual questions skimmed over, but higher order questions inferred and answered
Deeper links to text, not understanding themes
Accurate retrieval of information

BASIC

Task Difficulty: This item is answered easily, quickly.
Experience: Reference to own experience is subsumed to that of the characters, or authorial intent revealed in text.
Vocabulary: High frequency affixation of vocabulary (e.g., negatives- 'un', 'in', 'dis') is recognised and understood. More synonyms are known.
Language: Language of task question and text being read is easy for children at this level.
Cognitive Processing: Only one step is required to answer the task. Use more strategies to comprehend
Scaffolding: Significant support is offered either in the text or the task making the completion of the item relatively easy.
Comprehension: Able to suggest alternative endings. Simple, straightforward inferences are made. Task requires students to verbatim recall from, or locate in, the text. Links between ideas and meanings in text are beginning to be made. Starting to use text cues to make inferences. Able to locate information in graphs & tables.
Surface Features:
Punctuation: Able to independently recognise and use speech marks.
Grammar: Complex sentence structures (i.e., one subordinate clause per sentence) used and recognised. Consistent tense use.

Spelling: Has spelling conscience & awareness.

PROFICIENT

Task Difficulty: most readers within the level answer this item easily, quickly, and without much effort or support. Basic readers will attempt and may answer correctly but only with considerable effort or support.

Vocabulary: Task requires students to use synonymy or find words with the same meaning.

Language: Language of task question and text being read is medium difficulty for children at this level.

Cognitive Processing: Two steps are required to answer the task. The student needs to look in more than one place in the text. Reorganisation of information in text.

Level 3

Comprehension: Use text for inference and meaning, not just vocabulary knowledge. Make links within texts. Students make explicit references to text, for example for detail or support.

Texts: Increasing diversity of text layout or format.

ADVANCED

Vocabulary: Task requires students to use substitution or synonymy to answer the task.

Language: Language of task question and text being read is difficult for children at this level.

Cognitive Processing: Three or more steps are required to answer the task.

Comprehension: Students should make a statement of reason. Able to transform, reorganise, and restructure information. Meaning restated using different words.

Task Difficulty: only the best readers within the level answer this item easily, quickly, and without much effort or support. At this level proficient readers will attempt and may answer correctly but only with considerable effort or support. Basic readers may not attempt this item, or, despite considerable effort or support be unable to answer it correctly, or, may respond only through guessing.

Purpose, Intention, Audience: Able to answer questions about text purpose, intention, or audience.

LEVEL 4

According to the curriculum, students should:
Discuss (i.e., Examine by argument, debate; talk about so as to reach a decision; to talk or write about a topic in detail): language, meanings, and ideas in a range of texts. The task requires more reasoning about decisions than Level 3 in the form of explicit statements.
Relate (i.e., Bring into relation, establish relation between; establish a causal connection between) understandings to experiences, purposes, audiences, and other texts.
At Level 4, the author's intention and the purpose of the text are important. Students should be able to determine the audience, intent, or purpose from a relatively long (more than 250 words) prose or poetic text. At this level, students have to infer and construct in their own words what they understand the text's audience, purpose, or intent to be. In addition, students are expected to handle the nature of the language used in the text, that is, the language features, characteristics of text, and the effect of those characteristics. In other words, what are the language features that contribute to a text's meaning?
At level 4 texts are longer and more difficult, while the task is subtler and requires inference. Low frequency vocabulary occurs e.g. "inquisitive" and there is no or little support in text. Students are expected to evaluate the qualities of a text.

Level 4 students can:
Rich vocabulary including phrases
Greater inference
Greater understanding brought to text by reader
Broader adult understanding of the readings
Link parts of text to understand text
Identify writer's voice
Show 3 or more steps of processing
Ability to paraphrase
Attention to the writer's voice, technique, effect
Relate parts of text to each other to answer questions (cross referencing)

BASIC

Task Difficulty: This item is answered easily, quickly, and without much effort or support by all but the newest readers within the level.

Experience: Able to predict kinds of texts that well known-people might enjoy.

Vocabulary: Recognises low frequency vocabulary and able to generate understanding from word parts and connections in text. Less dependent on context for word knowledge.

Language: Language of task question and text being read is easy for children at this level.

Cognitive Processing: Only one step is required to answer the task.

Comprehension: Able to retell intention of a text in own words relying on key words and exploiting the thematic issues of a text rather than just literal verbatim repetition. Begins to analyse character, plot, and theme of text. Recognise message of text. Close attention to detail paid. Use of text to support ideas. Able to clarify, summarise, and justify opinions.

Scaffolding: Significant support is offered either in the text or the task making the completion of the item relatively easy.

PROFICIENT

Task Difficulty: most readers within the level answer this item easily, quickly, and without much effort or support. Basic readers will attempt and may answer correctly but only with considerable effort or support.

Experience: Reference to own experience is much more relevant and explicitly related to text being read or task being required.

Vocabulary: Low frequency vocabulary or connotations are understood. Task requires students to locate information in almost identical words or very close to the prompts in the task.

Language: Language of task question and text being read is medium difficulty for children at this level.

Cognitive Processing: Two steps are required to answer the task.

Purpose, Intention, Audience: Students have to infer and construct in their own words what they understand the text's audience, purpose, or intent to be.
Texts: Typography of texts is relatively dense (e.g., smaller font, use of multiple columns). Specialised, longer, and inconsiderate (e.g., muddled sequence) texts are understood within time limits. These texts will contain embedded texts of a different format or purpose.
Comprehension: Abstract generalised inferences are made. Evaluations and decisions are made about the text. Alternatives are developed. Able to compare texts written by an author with those written at other times by the same author or those written by different authors; and speculate on reasons for differences (e.g., different audience).

ADVANCED
Task Difficulty: only the best readers within the level answer this item easily, quickly, and without much effort or support. At this level proficient readers will attempt and may answer correctly but only with considerable effort or support. Basic readers may not attempt this item, or, despite considerable effort or support be unable to answer it correctly, or, may respond only through guessing.
Purpose, Intention, Audience: Students are expected to handle the nature of the language used in the text, that is, the language features, characteristics of text, and the effect of those characteristics. In other words, what are the language features that contribute to a text's meaning?
Experience: Reference to own experience is explicitly related to text being read and may invoke thematic sense of experience rather than literal relationship.
Vocabulary: Task requires students to use substitution or synonymy to answer the task.
Language: Language of task question and text being read is difficult for children at this level.
Cognitive Processing: Three or more steps are required to answer the task.

Appendix 2 Standard setting evaluations by teachers

Examinee-Centred method

Statement	Strongly Disagree	Usually Disagree	Slightly Agree	Moderately Agree	Usually Agree	Strongly Agree
1. My work was explained and directed well.			1		6	6
2. The training for this workshop gave me a clearer understanding about reading Curriculum Levels.			1	2	4	6
3. Scoring instructions were clear.			1	1	3	8
4. Curriculum Level standards for items based on this work will help me plan student teaching better.				1	2	10
5. The workload of the workshop was just right.				1	2	10
6. I will encourage my colleagues to participate in Project asTTle activities this year.					2	11
7. The definitions used to rate items were consistent with New Zealand Curriculum Levels 2 to 4 and bands within those Levels.				1	2	10
8. The location was conducive to our work.					4	9
9. The food and beverages were satisfactory.					3	10
10. Curriculum Level standards for items, based on this work, will help me assess students better.				1	1	11
Total Ratings	0	0	3	7	29	91

Item-Centred Method

Statement	Strongly Disagree	Mostly Disagree	Slightly Agree	Moderately Agree	Mostly Agree	Strongly Agree
1. My work was explained and directed well.					8	8
2. The food and beverages were satisfactory.				1	8	7
3. The training for this workshop gave me a clearer understanding about reading Curriculum Levels.				3	8	5
4. Scoring instructions were clear.				3	8	5
5. Curriculum Level standards based on this work will help me to plan my teaching better.				4	5	7
6. The workload of the workshop was just right.				1	9	5
7. I will encourage my colleagues to participate in Project asTTle activities this year.					5	11
8. The definitions used to rate items were consistent with New Zealand Curriculum Levels 2 to 4 and bands within those levels.				2	11	3
9. The location was conducive to our work.				1	9	6
10. Curriculum Level standards based on this work will help me assess students better.					6	10
Total Ratings				15	77	67

Test-Centred Method

Statement	Strongly Disagree	Usually Disagree	Slightly Agree	Moderately Agree	Usually Agree	Strongly Agree
1. My work was explained and directed well.					5	9
2. The training for this workshop gave me a clearer understanding about reading Curriculum Levels.					9	5
3. Scoring instructions were clear.				3	6	5
4. Curriculum Level standards for items based on this work will help me plan student teaching better.				1	2	11
5. The workload of the workshop was just right.			2	1	5	6
6. I will encourage my colleagues to participate in Project asTTle activities this year.				1	3	10
7. The definitions used to rate items were consistent with New Zealand Curriculum Levels 2 to 4 and bands within those Levels.				1	9	4
8. The location was conducive to our work.					3	11
9. The food and beverages were satisfactory					3	11
10. Curriculum Level standards for items, based on this work, will help me assess students better.					4	10
Total Ratings	0	0	2	7	49	82

Appendix 3 asTTle reading curriculum levels descriptors

LEVEL 2

According to the curriculum, students should:

- **Respond (**i.e., perform answering or corresponding action; show sensitiveness to by behaviour or change; say or do something in reply or as a reaction) **to**: language, meanings, and ideas in different texts.
- **Relate** (i.e., bring into relation, establish relation between; establish a causal connection between) **to**: personal experiences.

Students make a general link to their own world when responding to the text. Comparisons are simple, literal, and list-like.

Level 2 Students can

- *Respond or react to messages or text using own experience*
- May use cues in text to help understand but tends to rely on personal experience
- Retrieve and locate facts
- Recall and retell
- Ask questions
- Understand word beginnings, endings, and medial sounds (e.g., digraphs, diphthongs, blends, rhymes)
- Put facts or events into simple beginning, middle, & end sequence
- Infer at a simple level/ one step thinking
- Use connection between visual and text to get basic understand of title page, cover, or illustrations
- Can predict what happens next
- Can distinguish between fact and fiction

Characteristic	2 Basic	2 Proficient	2 Advanced
TEXT	This item is answered easily and quickly because the text is written at a simple 8-10 reading age level. Language of task question and text being read should be easy for children at this level.	The text, such as a letter, invitation, or narrative, requires reading at the 8-10 reading age level. Language of task question and text being read provides some challenge for children at this level.	Language of task question and text being read is challenging for children at the 8-10 reading age. Basic text forms, such as a letter, invitation, or narrative, are read with ease.
STUDENT RESPONSE	Students refer literally to their own experience. Their responses may appear to be tangential or irrelevant. However, they will make sense to the student. Focused more on the mechanics of reading	Reference to own experience is explicitly related to the text or task.	Able to reflect beyond the text. Empathises with character through own experiences. Student can respond to an overarching theme in text rather than just express a literal relationship to own life.
TASK: DEEPER FEATURES	**FI** Students retrieve verbatim information from the text. Only one step is required to answer the task. **K** **U** Meaning is gained from the text in a literal way (i.e., repeating the text words). Highly reliant on personal experience. **C** Connections made to personal experience only. **I** Inference and interpretation defined by personal experience	**FI** Task requires students to locate information in almost identical words or very close to the prompts in the task. Consistent location of verbatim material. **K** No inference is required to understand words. **U** Retelling of text is still highly dependent on text. Two steps can be handled to answer the task. May give one part to a two-part question. Distinguish fact and fiction. **C** Makes connection between visual and text for basic understanding of title page, cover, or illustrations **I** Able to handle simple sequencing tasks. Some inferring tasks are handled orally not in writing.	**FI** **K** Task requires students to use substitution or synonymy to answer the task. **U** Three or more steps are required to answer the task. Involvement with one piece of information demonstrated. Able to make simple reorganisation or restructuring of information such as beginning, middle, and end. **C** **I** Straightforward inference may be necessary. Able to recognise text purpose, intention, or audience of basic text forms, such as a letter, invitation, or narrative
TASK: SURFACE FEATURES	Some level of difficulty with spelling frequent common words	Sentence beginnings and endings handled accurately (i.e., Capitals for first word and proper nouns and final mark (i.e., full stop or question). Other marks are infrequent.	Reasonably accurate tense usage but not consistent Accurate spelling of all of the 300 most frequent common words.
External Reference Texts	*Are You My Mother?* Dr Seuss **Note:** *Hop on Pop* Dr Seuss is Level 1	*The Cat in the Hat* Dr Seuss	

According to the curriculum, students should:

- **Discuss (i.e.,** Examine by argument, debate; talk about so as to reach a decision; to talk or write about a topic in detail): language, meanings, and ideas in a range of texts.
- **Relate** (i.e., Bring into relation, establish relation between; establish a causal connection between) **understandings** to: personal experiences; and other texts

Discussion involves comparison, classification, and contrast. No longer do students just relate the text to their personal experience, and when those links are made they are much more specific. Students are much more able to interpret and explore text than they were in Level 2.

Level 3 Students can

- Retrieve and reorganise facts and locale
- Recall and retell in own words, and use retelling to predict
- Ask questions
- Use cues in text and own understanding to help understand
- Synthesis and sequence ideas
- Understands word beginnings, endings, and medial sounds (digraphs, diphthongs, blends, rhymes)
- Can skin scan, and select key ideas
- Can summarise and analyse the message in text, using imagery and syntax
- Can use and understand synonyms and antonyms
- Infer and interpret the underlying meaning of a text
- Use opinion based on the text or own understanding
- Use visual language to interpret the title, illustrations, formats
- React to messages and detect author's message
- Answers questions logically
- Compare and contrast messages, ideas, facts
- Imagery treated literally not explored deeper levels
- High level lateral thinking
- Factual questions skimmed over, but higher order questions inferred and answered

Characteristic	3 Basic	3 Proficient	3 Advanced
TEXT	Language of task question and text being read is easy for children at this reading age level 10-12	The text requires reading at the 10-12 reading age. Language of task question and text being read provides some challenge for children at this level. Increasing diversity of text layout or format including tables, graphic display, and embedded text forms.	Only the best readers within the level answer this item easily, or quickly. Language of task question and text being read is challenging for children at the 10-12 reading age.
STUDENT RESPONSE	Reference to own experience is subsumed to that of the text. Low-level understanding of text is shown.	Inconsistent success with answers still evident. Deeper links to text are made.	Consistency in answering questions correctly evident Clear understanding of text themes is shown. Retrieve information accurately and consistently.
TASK: DEEPER FEATURES	**FI** Location and verbatim copying of text is required. Able to locate information in graphs & tables **K** High frequency affixation of vocabulary (e.g., negatives- 'un', 'in', 'dis') is recognised and understood. More synonyms are known. **U** Only one step is required to answer the task. **C** Links between ideas and meanings in text are beginning to be made. **I** Able to suggest alternative endings. Simple, straightforward inferences are made. Starting to use text cues to make inferences.	**FI** The student needs to look in more than one place in the text. **K** Task requires students to use synonymy or find words with the same meaning. Able to use antonyms. **U** Two steps may be required to answer the task. Reorganisation of information in text is required. Opinions based on text. **C** Make links within texts. Students make explicit references to text, for example for detail or support. **I** Use text for inference and meaning, not just rely on prior vocabulary knowledge. Imagery treated literally.	**FI** **K** Task requires students to use substitution or synonymy or to restate meaning using different words. **U** Two to three steps may be required to answer the task. A statement of reason may be required. Information may have to be transformed, reorganised, or restructured. **C** Compare and contrast messages, ideas, facts **I** Able to identify the text's purpose, intention, or audience, though not required to offer an explanation or justification.
TASK: SURFACE FEATURES	Able to independently recognise and use speech marks. Complex sentence structures (i.e., one subordinate clause per sentence) used and recognised. Consistent tense use. Has spelling conscience & awareness.		
External Reference Texts		*Sneetches* Dr Seuss	

LEVEL 4

According to the curriculum, students should:

- **Discuss** (i.e., Examine by argument, debate; talk about so as to reach a decision; to talk or write about a topic in detail): language, meanings, and ideas in a range of texts. The task requires more reasoning about decisions than Level 3 in the form of explicit statements.
- **Relate** (i.e., Bring into relation, establish relation between; establish a causal connection between) understandings to experiences, purposes, audiences, and other texts.

At Level 4, the author's intention and the purpose of the text are important. Students should be able to determine the audience, intent, or purpose from a relatively long (more than 250 words) prose or poetic text. At this level, students have to infer and construct in their own words what they understand the text's audience, purpose, or intent to be. In addition, students are expected to handle the nature of the language used in the text, that is, the language features, characteristics of text, and the effect of those characteristics. In other words, what are the language features that contribute to a text's meaning?
At level 4 texts are longer and more difficult, while the task is subtler and requires inference. Low frequency vocabulary occurs e.g. "inquisitive" and there is no or little support in text. Students are expected to evaluate the qualities of a text.

Level 4 students can:

- Rich vocabulary including phrases
- More adult-like understanding of vocabulary, and general life brought to text by reader
- Ability to paraphrase
- Evaluate and compare and contrast messages, ideas, facts
- Interpret and infer underlying meanings
- Evaluate the merit or worth (providing one reason) of author's language
- Identify language techniques and their effects, providing at least one reason for opinion
- Attention to the writer's voice, technique, effect
- Link parts of text to understand text
- Relate parts of text to each other to answer questions (cross referencing)

Characteristic	4 Basic	4 Proficient	4 Advanced
TEXT	Language of task question and text being read is easy for children at this reading age level 12-14.	The text requires reading at the 12-14 reading age. Language of task question and text being read provides some challenge for children at this level. Typography of texts is relatively dense (e.g., smaller font, use of multiple columns). These texts will contain embedded texts of a different format or purpose. Relatively long (>250 words) read.	Only the best readers within the reading age 12-14 answer this item easily, or quickly. Language of task question and text being read is challenging for children at the 12-14 reading age. Specialised, longer, and inconsiderate (e.g., muddled sequence) texts are understood within time limits.
STUDENT RESPONSE	Able to predict kinds of texts that well-known-people might enjoy.		
TASK: DEEPER FEATURES	**FI** Able to retell intention of a text in own words relying on key words and exploiting the thematic issues of a text rather than just literal verbatim repetition. **K** Recognises low frequency vocabulary and able to generate understanding from word parts and connections in text. Less dependent on context for word knowledge. **U** Begins to analyse character, plot, and theme of text. Recognise message of text. Close attention to detail paid. Use of text to support ideas. Able to clarify, summarise, and justify opinions. **C** **I**	**FI** **K** Low frequency vocabulary or connotations are understood. **U** Able to paraphrase. Able to identify language techniques and their effects providing at least one reason for opinion. **C** Compare and contrast messages, ideas, or facts. Able to compare texts written by an author with those written at other times by the same author or those written by different authors. Able to cross-reference or link parts of text to each other. **I** Abstract generalised inferences are made. Evaluations and decisions are made about the text. Alternatives are developed. Able to speculate on reasons for differences between texts (e.g., different audience). Students have to infer and construct in their own words what they understand the text's audience, purpose, or intent to be	**FI** **K** Task requires students to use substitution or synonymy to answer the task. **U** **C** **I** Students are expected to handle the nature of the language used in the text, that is, the language features, characteristics of text, and the effect of those characteristics. In other words, what are the language features that contribute to a text's meaning?
TASK: SURFACE FEATURES		.	
External Reference Texts			