Technical Report # 22

Validation Evidence of asTTle Reading Assessment Results: Norms and Criteria

Abstract: This report presents an analysis of arguments about assessment that contribute to the validity of asTTle reading comprehension scale and curriculum level scores. These principles relate to the importance of curriculum-linked interpretations, detailed as opposed to general analysis of achievement, the weakness of age or year related normative score interpretation, and methods for establishing legitimate expectations of performance. Furthermore, analysis of a small set of school data shows that the asTTle Surface Cognitive Processing and Finding Information close reading scores provide similar results to generalised measures of reading comprehension and teacher judgements of achievement. However, the richness of asTTle achievement results provide meaningful insight into teachable objectives that students still need to master, notwithstanding overall high norm-referenced ranking on generalised ability measures. It is argued that the current data provide a robust evidence for the validation of the asTTle reading comprehension scores.

Submitted by the Assessment Tools for Teaching and Learning team,

Auckland UniServices Ltd

University of Auckland

November 2003

Validation Evidence of asTTle Reading Assessment Results: Norms and Criteria

asTTle is funded by the Ministry of Education to Auckland UniServices at the University of Auckland to research and develop an assessment application for Reading, Writing, Mathematics, Pānui, Pängarau, and Tuhituhi for Years 5-11 (Levels 2-6) for New Zealand schools. We acknowledge this funding, and thank the Ministry of Education for their continued assistance in the development of this project.

This report presents an analysis of arguments about assessment that contribute to the validity of asTTle reading comprehension scale and curriculum level scores. These principles relate to the importance of curriculum-linked interpretations, detailed as opposed to general analysis of achievement, the weakness of age or year related normative score interpretation, and methods for establishing legitimate expectations of performance. Furthermore, analysis of a small set of school data shows that the asTTle Surface Cognitive Processing and Finding Information close reading scores provide similar results to generalised measures of reading comprehension and teacher judgements of achievement. However, the richness of asTTle achievement results provide meaningful insight into teachable objectives that students still need to master, notwithstanding overall high norm-referenced ranking on generalised ability measures. It is argued that the current data provide a robust evidence for the validation of the asTTle reading comprehension scores.

I would like to thank the tens of thousands of students and hundreds of schools and teachers across the country who gave their time to completing the tasks and whose performance is reported here. Their contribution was invaluable. I would also like to thank the many researchers and content experts who provided guidance, feedback, and ideas for the creation of the asTTle assessment tasks and frameworks; their contribution can be seen in the many asTTle Technical Reports.

This report is the product of dialogue among the asTTle team members over many months of data entry, data verification, statistical analyses, and interpretation. I would like to thank Dr Gavin Brown, Dr Peter Keegan, Earl Irving, and Andrea MacKay for assisting with those processes. They were assisted by Damian Mooyman, Tim Sutherland, and Paulmi Patel who did much of the detailed data checking.

John Hattie
Project Director, asTTle
November, 2004

The bibliographic citation for this report is:
Hattie, J.A.C., Brown, G.T.L., Keegan, P., Irving, S.E., MacKay, A.J., Sutherland, T., Mooyman, D., & Patel, P. (2003, November). *Validation Evidence of asTTle Reading Assessment Results: Norms and Criteria.* asTTle Tech. Rep. 22, University of Auckland/Ministry of Education.

Table of Contents

Validity refers to the degree to which there is evidence that supports the proposed interpretations or uses of an assessment (AERA/APA/NCME, 1999). A common form of evidence for the validity of an assessment is its relationship to other variables such as measures of the same underlying construct. In New Zealand, teachers have been using the levels of the curriculum framework as a basis for describing student performance since the early 1990s. Teachers have been using norm-referenced scores (e.g., percentiles and stanines) derived from general ability measures of literacy and numeracy published since the late 1960s by the New Zealand Council for Educational Research to report or monitor student achievement.

The asTTle resource produces norm-referenced scale scores (i.e., 500 represents the mean of Year 6 with a standard deviation of 100) and curriculum level standards-referenced scores (Level 2 Basic to Level 6 Advanced) for literacy and numeracy. Over the years since asTTle's release, teachers have found surprising differences between their own judgements or impressions of student attainment relative to the NZ Curriculum Levels and what asTTle produces (e.g., "*I know my students can read Level 4 texts but asTTle says they are at Level 3*"). Additionally, teachers have questioned the difference between student performance on standardised

general ability assessments and their scores as reported in asTTle (e.g., *how can my students be in Stanine 9 on PAT, but asTTle says they are only in Level 3?*"). Because asTTle has already provided validation evidence based on analysis of test content (see the many Technical Reports on curriculum mapping and item signatures and teacher feedback), it is important that these score differences be understood and resolved, so that users do not treat the results from asTTle as irrelevant.

Thus, this report has two major objectives. First, it reviews issues around the interpretation of test scores, including norm- and criterion-referenced scores, general-ability and detailed outcomes-referenced tests, the nature of curriculum level progress, and the importance of standards. Secondly, it examines student reading data taken from asTTle, teacher judgements, and PATs and examines the nature of concurrence and discrepancy. The report concludes with the assertion that the apparent discrepancy is small and is largely a function of differing interpretive approaches. It is claimed that teachers and principals can use asTTle validly to understand, report, and respond to student skills, knowledge, and ability in reading comprehension.

Interpreting Assessments

A number of important issues and assumptions about the nature of responding to assessments must be addressed before examining in detail the validation evidence for asTTle. These include:

- Is the test valid or is it our interpretations?
- Is one assessment enough?
- Is comparing performance to that of others enough?
- How does general ability in a subject relate to specific abilities?
- What level of performance is expected according to the curriculum?
- How are expectations about levels of performance created?

Valid Tests vs. Valid Interpretations

A common interpretive approach to assessment is to suggest that some assessments are by definition either summative or formative (for example, Carr, 2001; Torrance & Pryor, 1998; Philipp, Flores, Sowder, & Schappelle, 1994) with the implication that summative is bad and formative is good assessment. Brown (2004) has demonstrated that this simplistic dualism is insufficient to describe how teachers actually conceive of assessment. Further, it is clear that a test is neither summative nor formative– only the use or interpretations of tests can be considered formative or summative (Linn & Gronlund, 2000; Messick, 1989). For example, consider the "piece of paper" on which these words lie. The paper could also be used to start a fire, make a dart, or to cover a window. The paper is still paper, but the *use* of the paper differs. Similarly for tests. An assessment tool can be used for many purposes – it is the *nature of the interpretation* that is diagnostic, formative, or summative, or all of these. The *test itself* can never be so categorised. A test in itself is neither reliable nor valid, it is the interpretation or use that we make that is reliable or valid.

So any claims that asTTle is diagnostic and therefore cannot be used for summative or formative purposes, or that it is summative and therefore cannot be used for diagnostic interpretations is misleading. It is not correct to claim that asTTle was devised as diagnostic. It was designed so that valid and reliable interpretations could be made – for diagnostic, summative, and/or formative purposes – about teaching and learning decisions. These purposes are in the mind of the user, not the developer of the test; hence the reliability and validity relate to the decisions and interpretations derived from the mind of the user. The quality of assessment is in the quality of interpretations made and assessors must make sure those decisions are worthwhile, valid, significant, and address the educational needs of students.

Multiple Assessments

It is a fundamental axiom in test theory that no decision should be made on the basis of one data point. The purpose of tools like asTTle is to provide feedback to teachers, students, principals, or whomever – feedback that is sometimes surprising, sometimes revealing, and probably of little value if merely confirmatory (however much such confirmation may be welcomed). When any test reveals a new way of looking at teaching or learning we must immediately focus our attention on triangulating or cross-validating this belief. Thus, reference to multiple assessments, whether of different method or of multiple frequency, when making critical decisions and interpretations is essential.

General Ability vs. Specific Ability Interpretations

Given the outcomes oriented curriculum framework that New Zealand has adopted since 1993, assessments must provide information relative to the content or curricula that teachers are required to deliver. Too often in New Zealand test history, we have had tests that have provided limited, if any, feedback to teachers about their fundamental role in improving the quality of student learning. Feedback from assessment ought to help teachers interpret how well they, as teachers, are enhancing their students' attainment of curricula objectives, and what they can do to move students onto more challenging tasks (Popham, 2000). Part of the reason for this lack of feedback is that any test that can be administered fairly to students in school time cannot evaluate all the various detailed facets of learning a subject; there are simply too many outcomes for a single test.

Thus, the tradition in New Zealand has been for general tests that cover a little bit of everything. These assessments result in interpretations that can only comment on a learner's general ability in that subject. Furthermore, generalised tests of school

achievement tend to have very high correlations with general intellectual ability or prior achievement (Kline, 2000). We know from many studies that reading vocabulary is one of the best measures of "intelligence", and if such measures are used to assess the power of teachers and schools, it is unlikely that we will see any effects from "teachers and schools". For example, an analysis of children's performances on the *School Entry Assessment* (given about 6 weeks into the first year of schooling) and their *PAT Reading Vocabulary* scores 5-7 years later (N = 2000+ students), indicates that there is a remarkably high correlation r = .7 or R^2 = .48 between the two measures. In other words, almost half the variance of reading vocabulary in Years 6-9 can be accounted for by the students' prior performance in the sixth week of Year 1.

A good assessment scheme focuses on content in which students' performance can be improved through instruction. Without the ability to analyse student performance in terms of the "rich ideas" of close reading (e.g., in asTTle these are finding information, knowledge, understanding, connections, inference, and surface features), teachers are left only with generalised ability measures. These may not be able to indicate with sufficient accuracy what students need to learn nor show an instructional program has had. Detailed analyses may require more testing opportunities, larger banks of items, and the ability to aggregate results from multiple detailed tests; all of which are possible through the asTTle resource.

Norm-Referenced Interpretation

Generally, educational assessments support norm-referenced interpretations about performance of individuals relative to a nationally representative population. The ease by which such interpretations can be reported (percentiles and stanines) can interfere with some of the critical decisions teachers need to make about learning. If

we know anything in the business of schooling, it is that there is a high correlation between past ability and future ability – so tracking normative information is too often self-fulfilling. In other words, students who are at Stanine 9 this year were likely to have been at Stanine 9 last year and are also likely to be in the top 4 or 5 percent next year. However, such students may have learnt nothing from the New Zealand curricula in the intervening year, especially if the whole year cohort made little progress. The norm-referenced score interpretation simply indicates that such students have just kept their rank order. This dependence on rank order interpretations can obscure lack of real learning.

For example, we might expect about 40% of Year 7 students to have mastered the objectives as specified for Level 4 of the Reading curriculum. On the basis of 8000+ Year 7 students (from the asTTle sampling of New Zealand students), the percentage attaining four selected reading objectives ranges at Level 4 between 33 and 51 per cent (Table 1).

Table 1

Percent Year 7 Students Attaining Selected Level 4 Reading Objectives

Objective	Average Attainment
Identify and understand main ideas	33%
Makes links between aspects of texts	34%
Consistently read for meaning	40%
Find, select, and retrieve information	51%

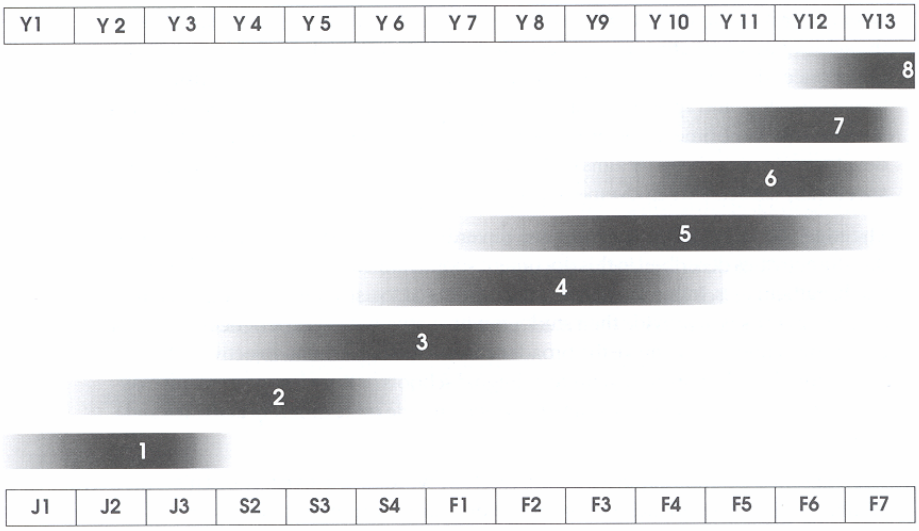
Clearly, an overall average obscures the detail possible with a disaggregated measure. Further, this evidence suggests that students could maintain their rank order position, without making significant progress in the more complex objectives of understanding main ideas and making connections. Thus, overdependence on a total, general ability, rank order score interpretation may conspire to prevent student learning.

Curriculum Level Interpretations

The New Zealand Curriculum Framework is predicated on identification of major themes or strands of learning objectives and on the sequencing of such learning into 8 major levels of progress. The asTTle Project has done extensive work in mapping these rich ideas in reading, writing, mathematics, pānui, tuhituhi, and pāngarau against the expected levels of attainment (see Technical Reports 4, 6, 11, 13, 23, 25, 34, 36, 37, 38, 39). The fundamental figure in all Ministry of Education curriculum documents that spells out the "expected" relation between levels and ages (the latter through the proxy of years of schooling) is Figure 1 taken from the *English in the New Zealand Curriculum* document page 20.

It is noted that the Levels diagram (Figure 1) could be improved because it permits multiple interpretations. It is likely that the diagram indicates that across any one year level the range of ability among students will vary across levels (e.g., students in Year 4 are expected to mostly in Level 2, with some in Level 3, and a small possibility that a few will still be in Level 1). Another possible interpretation is that a student could be simultaneously at multiple levels. While this could be the case across objectives (e.g., Tom can be at Level 3 for "Consistently reading for meaning" and Level 4 for "Identify and understand main idea"), it cannot be the case that Tom can be classified as being simultaneously at both Level 3 and Level 4 in "Reading". A non-overlapping step function drawing would be more defensible, and still allow for classes of students to be in multiple levels.

Figure 1.
Expected Relationship of School Years to Curriculum Levels



From this diagram a set of expectations about the relationship of levels to school years can be ascertained (Table 2). The relationship of year to level was determined by looking down from the beginning, mid, and end points of each Level bar to the year. Thus, it is can be determined that the typical student is expected to take about two years to advance though each of Levels 2 to 5.

Table 2
Curriculum Level Start and End Years

Level	Commence Level	Possible Range of Expected End of Level	
1	Beginning of Y1	end of Y2	Beginning of Y4
2	End of Y2	end of Y4	Beginning of Y6
3	End of Y4	end of Y6	Beginning of Y8
4	End of Y6	end of Y8	Beginning of Y10
5	End of Y8	end of Y10	Beginning of Y12
6	End of Y9	end of Y11	Beginning of Y13

A key question is the age of students at each Year. It is assumed that most New Zealand students enter Year 1 at age 5, though it must also be assumed that, with entry distributed more or less equally throughout the year, the median age of students in Year 1 is more likely to be 5.5 years. Furthermore, it is assumed that a curriculum level is best associated to the age of students representing the mid-range of relevant years. Thus, students in the 25th to 75th percentiles of a level would be expected to fall

in the year and age ranges outlined in Table 3. Accordingly, it is possible to interpret the expected performance of Level 4 as that typically associated with 12.5 to 13.5 year olds.

Table 3
Middle 50 Percent Distribution of Levels by Age Range

Level	Mid-Percentile Range	Age Range
1	Students in Y1 to Y3	5.5—7.5
2	Students in Y2 to Y4	6.5—8.5
3	Students in Y4 to Y8	8.5—12.5
4	Students in Y5 to Y9	9.5—13.4
5	Students in Y8 to Y12	12.5—16.5
6	Students in Y9 to mid Y13	13.5—17.5

Hence, according to this Ministry chart and the assumptions outlined above, it could be estimated that between the start and end of the school year about 40% of Year 7 students should have attained curricula objectives at Level 4. In addition to determining an expected rate of progress through the curriculum, it is useful to examine the assumptions implicit in a levels based curriculum about the nature of learning tasks. In reading for example, students' tasks, given an appropriate difficulty of reading material, at level 4 must be:

- more challenging than those listed in Levels 3 or 2 or 1 of the close reading strand,
- taught before students advance to Level 5 objectives,
- appropriate for students in Years 7 (although adults could be at this Level, and some 6 year olds might be at this Level),
- what secondary teachers could legitimately "expect" the majority of students to know and be able to do when they enter secondary school (at Year 9), and
- the kinds of literacy skill or ability that about 40% of Year 7 students should have attained.

From this discussion of how the curriculum levels seem to support age or year related normative interpretations, the user of an assessment may have some justification for assuming that what the average child can do in any year defines the level of performance needed for a curriculum level. However, this norm-referenced interpretation is not consistent with the intent of the curriculum framework which suggests that objectives or outcomes are not linked inevitably to learner ages; the outcomes represent stages of learning, not an inherent and universal developmental sequence (Brown, 1998). If an assessment scheme simply accepts the age/year basis for standards, then norm-referenced interpretations are inevitable. However, the approach utilised in asTTle is to provide both year related norms and criterion-based descriptions of the skills, knowledge, and ability associated with each Level of learning regardless of the age of the learner. The procedure for developing such descriptions is standard setting.

Standards-Based Interpretations

Although it is possible, and possibly even justifiable given the previous discussion, to use year as the key to level, this is not the fundamental connection underlying the asTTle resource. The key question is "*What are the expected levels of performance or proficiency at each of the Levels 2 to 6?*" With these definitions it is possible to then distinguish degrees of progress within each level. In asTTle, performance within each level is demarked into thirds (i.e., Basic, Proficient, and Advanced). A defensible process for setting the standards for reading is outlined in asTTle Technical Report #21 which commences:

> 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).

There are many different methods of determining the cut scores for the various levels – and three methods, each set by different panels of teachers, were used for setting the standards for the asTTle reading comprehension materials in Levels 2 to 4 (details are provided in Technical Report #21). What can be asserted with confidence is that the interpretation of asTTle performance into curriculum level scores has been done in a robust and defensible manner and that the curriculum levels assigned to students are a fair and valid representation of their underlying ability. The asTTle Curriculum Level scores do not represent judgements based only on the ability of the most elite students; rather they represent consensus opinion about the underlying abilities represented by those levels in the context of primary school.

Interpreting asTTle Data

asTTle has data obtained from its standardisation of hundreds of curriculum-based tasks from tens of thousands of primary and secondary school students. The data have been analysed to create the underlying comparison norms that teachers can make use of when interpreting test performance through the asTTle reports. Those data have provided a robust description of what students can do at each year level and, more importantly, of what parts of each curriculum statement are easy or difficult for students at each year level. From the data, validation evidence can be obtained in order to compare asTTle's measurements of student ability against those of curriculum expectations, teacher judgements, and nationally standardised general ability assessments.

Curriculum Expectations

In order to make explicit the relationship of the curriculum expectations relative to the asTTle data, four reading objectives that are supposed to be attained by Level 4 students given appropriately difficult texts were identified for detailed examination. These were:

- Consistently read for meaning
- Makes links between aspects of texts
- Find, select, and retrieve information
- Identify and understand main ideas

See asTTle Technical Report 4 and Limbrick, Keenan, and Girven (2001) for discussions of how these objectives are derived from the curriculum statement for English. From the asTTle reading and mathematics data, the percentage of students in Years 5 to 7 succeeding at Level 4 items was calculated for Years 5 to 7 (Table 4). Note this is preliminary data and will be further verified with the extension of asTTle data into Year 8 to 12 and Levels 5 and 6.

Table 4
Actual Proportion of Students in Level 4 or better by Year and Subject

Subject	Year 5	Year 6	Year 7	Expected in Level 4 at Year 7
Reading	13%	19%	21%	40%
Mathematics	3%	8%	18%	40%

From these data, it is possible to identify major interpretations. Very few students in Years 5 to 7 are working in or above Level 4. Only 13% of students in reading and 3% in mathematics are at Level 4 or better in Year 5, 19% in Year 6 are achieving at Level 4 or better in reading and only 8% in mathematics, and by the end of Year 7 just 21% are achieving at Level 4 or better in reading and 18% in mathematics. Given that the top 11% constitute stanines 8 and 9, while the top 23% represent stanines 7 to 9 in a normal distribution, students in Level 4 or above in these two subjects in Years 5 to

7 are in the top quarter of the population (i.e., stanines 7 to 9). This result is quite different to the age based interpretation of Figure 1 which would indicate that on average 40% of students should attain Level 4 by the end of Year 7. Note also that in Years 5 and 6, a much larger proportion of students are reading in Level 4 than are in mathematics at Level 4, and that the gap between subjects closes significantly in Year 7.

Note that analytic charts of attainment, using the asTTle data, could be done at the whole curricula level (e.g., Reading), at the rich idea level (e.g., Understanding, Finding information), or at the objective level (e.g., consistently read for meaning). Further, it is critical that such an analysis be undertaken for all appropriate groups – Maori, Pasifika, boys, students with languages other than English at home, etc. No interpretation of assessment data should hide the problems or successes of subgroups (including years) in "averages" or "overall performance". Note that schools can carry out such analyses for themselves using the asTTle Console and NZ Comparisons reporting tools.

Having established, the empirical realities of student performance against standards judged by New Zealand teachers and against the expectations derived from the curriculum, it is possible and appropriate to ask whether the percentage of students reaching each level of performance is acceptable. This is a major task for not only each school community but also for the nation. Nevertheless, given the performance of over 92,000 students in Years 4 to 12 embedded in asTTle, educators have the ability to examine what effort might be needed to achieve a desired level of performance.

Teacher Judgement and General Ability

The most commonly used norm-referenced, standardised, general ability assessments used in New Zealand are the Progressive Achievement Test (PAT) series from NZCER. Table 5 provides reading achievement data from 17 students supplied by an anonymous teacher and school, including Progressive Achievement Tests (PAT) in reading comprehension and vocabulary, asTTle test of close reading, and from the teacher's own classroom based judgement of reading proficiency. The PAT scores are Class Stanines, while the teacher and asTTle scores are curriculum levels using the Basic, Proficient, and Advanced sub-categories within each level.

Table 5
Reading Achievement Data for 17 Anonymous Year 7 students

Student	PAT Class Stanine		Teacher's Judgement Curriculum Level	asTTle Curriculum Level
	Reading Comprehension	Reading Vocabulary		
Marla	9	7	4A	3A
May	9	7	>4A	4B
Kyle	8	8	4P/A	3A
Bobbi	8	7	4P/A	3A
Tom	**8**	**8**	**4A**	**3A**
Joel	8	8	4P/A	3A
Dick	**8**	**8**	**4P**	**3P**
Mary	7	7	4P	3B
Harry	**7**	**8**	**4P**	**3P**
James	6	6	4B	3P
Judy	9	9	>4A	4P
Billy	9	7	4A	4B
Jan	8	6	4A	4B
Dilbert	9	9	4A	4B
Joan	9	6	4A	4B
Harvey	9	9	4A	4B
Isobel	8	7	4P	3A
Average	*8.2*	*7.5*	*4A*	*3A*

The average stanine for PAT results shows that this class scores in the top quarter of the distribution for Year 7 students in New Zealand for reading comprehension and vocabulary (i.e., all but James are in stanine 7 to 9). The average teacher judged curriculum level for this class is just under 4 Advanced, while the

average asTTle curriculum level is just over 3 Advanced. Thus, there appears to be a major discrepancy, as was indicated at the beginning of this report, between the seemingly high scores from the PAT and teacher judgements and the much lower scores given by asTTle curriculum levels. To illustrate what is happening in this example, consider Tom who has a PAT reading comprehension and vocabulary stanine of 8 (i.e., he is performing in the top 11% of students in Year 7), the teacher considers him Level 4 Advanced, but asTTle places Tom at Level 3 Advanced. Note that Level 3 Advanced for Year 7 reading in the asTTle norming sample is equivalent to the 71st to 79th percentiles which fall in the range of 6th-7th stanines. Thus, it is possible that the difference may be a function of random error rather than some other systematic differences.

Nevertheless, assuming there is a systematic, non-chance difference between asTTle reading scores and the two other measures, what alternative interpretations could account for this difference? According to the asTTle Individual Learning Pathways Report, Tom was not able to perform the following objectives, with the number of items related to each objective shown in brackets, despite his generalised reading vocabulary and comprehension scores placing him among the top 11% of students:

- Consistently read for meaning (on 6 items)
- Make links between aspects of texts (3 items)
- Find, select, and retrieve information (5 items)
- Identify and understand main ideas (3 items)

Another piece of information that helps explain the disjuncture is, according to the asTTle reports, that Tom is 4 Basic at Surface cognitive processing and 3 Proficient at Deep cognitive processing as classified by the SOLO taxonomy (see Technical

Reports 12, 16, & 28 for descriptions of how items were identified according to SOLO and Technical Report 43 for an in-depth explanation of the taxonomy). Thus, Tom has some significant weaknesses in the deep processing objectives of close reading related to making links, reading for main ideas and meaning, and even processing information within text.

Consider the percentage of Year 7 students who could attain the same objectives at Level 4 (Table 6). Clearly on average about 40% of Year 7 students can do tasks related to these same outcomes or objective. This is a value close to the expected 40% value mentioned above—but clearly Tom cannot do these objectives. There is some confidence to the claim Tom is probably not reading at Level 4—indeed he may be best placed at just a little lower; at Level 3 Advanced—regardless of his rank relative to other Year 7 students on a more generalised reading ability test. Tom is still overall "brighter" than his peers, but the purpose of school-based assessment is to ascertain what Tom knows and is able to do – not just establish his distribution score compared to an age cohort.

Table 6
Mean Performance of Year 7 Students on Selected Level 4 Reading Objectives

Objective	Mean Correct
Consistently read for meaning	40%
Find, select, and retrieve information	51%
Identify and understand main ideas	33%
Makes links between aspects of texts	34%
Average	*40%*

As a second example, consider the case of Dick, who scored stanine 8 for PAT reading comprehension and vocabulary, received a 4 Proficient from his reading teacher, and only a Level 3 Proficient on an asTTle reading test (note this curriculum level score falls in stanine 6 or 59th to 71st percentile of asTTle reading scores). According to the asTTle report, Dick is 3 Basic in Understanding, 3 Basic in Connections, and 2 Advanced in Inference. It is clear that Dick has poor skills in

Inference, and although his PAT scores indicate he is at Stanine 8 in vocabulary and comprehension, he has skills in Inference well below those of other Year 7 students.

And now look at Harry, who, according to the same asTTle reading test, is 4 Proficient at Surface, but 3 Proficient at Deep cognitive processing. Harry's overall score is in Level 3 Proficient which, as mentioned above, falls in stanine 6 or 59th to 71st percentile of asTTle reading scores, while his surface score falls in stanine 8 or the 89th to 93rd percentile of asTTle reading scores. He scored in the 7th and 8th stanines for the PAT and Level 4 Proficient according to his teacher's judgement. It is likely that a generalised reading vocabulary test and the teacher are more influenced by Harry's surface knowledge rather than his deep understanding. Indeed, the evidence seems to support such an interpretation across all three students.

The message in these examples is that normative distributions are not necessarily related to what students can and cannot do, or to what students know or do not know. They merely indicate that students can be ranked via a normal distribution. This ranking is based on tests that are generalised estimates of intelligence (such as reading vocabulary), rather than being measures of what the New Zealand curricula stipulate that students should learn. Interestingly, however, there is a great deal of commonality between the asTTle scores, especially those for surface processing, and the normal distribution scores derived from other measures. It just so happens that the asTTle curriculum level scores associated with high scoring students are lower than the levels teachers or even the curriculum seem to expect. Fundamentally, the asTTle Surface Cognitive Processing and Finding Information scores appear to give the most similar results to both the PAT and teacher judgement. Furthermore, use of asTTle makes possible significantly richer interpretations about what and how well students

have learnt; much more useful educationally than simply a rank-order score on a relatively stable and immutable attribute like intelligence.

Final comments

The asTTle resource provides feedback information about not only relative standing but also patterns of strength and weakness which can be used to identify where to begin instruction and how to monitor effectiveness. This applies to students at both ends of the distribution as well as those in the middle. We argue that it is the job of educators to lower the correlation between performance in week 5 of Year 1 and performance in Year 7. Finding that the same students were in stanine 1 in Year 1 and stanine 1 again in Year 9 tells us little other than someone had to be there (that is how normal distributions, whether reported with percentiles or stanines, work).

It remains to be seen whether having rich criterion, standards, and norms-referenced scores available through asTTle will help teachers' move from an over-reliance on norm-referenced interpretations. The data examined and the arguments made in this report should give teachers confidence in the validity of the asTTle close reading assessment materials. We claim that the assertion that asTTle scores are invalid just because they are different to teacher expectations or PAT scores is incorrect; rather we claim that the evidence presented here provides support for the validity of asTTle's measurement of the construct 'reading comprehension'.

References

AERA/APA/NCME (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: AERA.

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.

Brown, G.T.L. (2002). *Item signature study: Report on the characteristics of reading texts and items from calibration 3.* (Tech. Rep. 28). Auckland, NZ: University of Auckland, Project asTTle.

Brown, G.T.L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Policy, Principles and Practice, 11*(3), 305-322.

Carr, M. (2001). *Assessment in early childhood settings: Learning stories*. London: Paul Chapman.

Coogan, P., Hoben, N., & Parr, J. M. (2003). *Written language curriculum framework and map: Levels 5-6* (asTTle Tech. Rep. 37). Auckland, NZ: University of Auckland/Ministry of Education.

Christensen, I., Trinick, T., & Keegan, P. J. (2003). *Pāngarau curriculum framework and map: Levels 2-6* (asTTle Tech. Rep. 38). Auckland, NZ: University of Auckland/Ministry of Education.

Ell, F. (2001). *Mathematics in the New Zealand Curriculum - A concept map of the curriculum document.* (asTTle Tech. Rep. 11). Auckland, NZ: University of Auckland, Project asTTle.

Fairhall, U., & Keegan, P. J. (2001). *Pāngarau curriculum framework and map: Levels 2-4.* (asTTle Tech. Rep. 13). Auckland, NZ: University of Auckland/Ministry of Education.

Glasswell, K., Parr, J., & Aikman, M. (2001). *Development of the asTTle writing assessment rubrics for scoring extended writing tasks.* (asTTle Tech. Rep. 6). Auckland, NZ: University of Auckland, Project asTTle.

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

Hattie, J. A., & Brown, G. T. L. (2004, September). *Cognitive processes in asTTle: The SOLO Taxonomy* (asTTle Tech. Rep. 43). Auckland, NZ: University of Auckland, Project asTTle.

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

Kline, P. (2000). *Handbook of psychological testing* (2nd ed.). London: Routledge.

Limbrick, L., Keenan, J., & Girven, A. (2000). *Mapping the English curriculum.* (asTTle Tech. Rep. 4). Auckland, NZ: University of Auckland, Project asTTle.

Limbrick, L., Keenan, J., & Girven, A. (2001). Mapping the English Curriculum—Assessment Tools for Teaching and Learning: The development of literacy and numeracy tools for years five to seven in English. *Reading Forum N.Z.,* (1), 5-12.

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

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

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Murphy, H., & Keegan, P. J. (2002). *Te Reo Māori literacy curriculum map. Levels 2-4* (asTTle Tech. Rep. 23). Auckland, NZ: University of Auckland/Ministry of Education.

Murphy, H., & Gray, A. (2003). *Review of Māori literacy framework for koeke 2-6 panui/tuhituhi of the Māori language curriculum statement, Te Reo Māori i roto i ngā Marautanga o Aotearoa* (asTTle Tech. Rep. 39). Auckland, NZ: University of Auckland/Ministry of Education.

Nicholls, H. (2003). *English reading curriculum framework and map: Levels 2-6* (asTTle Tech. Rep. 34). Auckland, NZ: University of Auckland/Ministry of Education.

Philipp, R. A., Flores, A., Sowder, J. T., & Schappelle, B. P. (1994). Conceptions and practices of extraordinary mathematics teachers. *Journal of Mathematical Behavior, 13*, 155-180.

Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders*. (3rd ed.). Boston: Allyn & Bacon.

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.

Torrance, H., & Pryor, J. (1998). *Investigating formative assessment: Teaching, learning and assessment in the classroom*. Buckingham, UK: Open University Press.

Thomas, G., Holton, D., Tagg, A., & Brown, G. T. L. (2002). *Numeracy item signature study: A theoretically derived basis.* (asTTle Tech. Rep. 25). Auckland, NZ: University of Auckland, Project asTTle.

Thomas, G., Holton, D., Tagg, A., & Brown, G. T. L. (2003). *Mathematics curriculum framework and map: Levels 2-6* (asTTle Tech. Rep. 36). Auckland, NZ: University of Auckland/Ministry of Education.