MARK H. MOULTON, PH.D.
EDUCATIONAL DATA SYSTEMS

ONE RULER, MANY TESTS

A PRIMER ON TEST EQUATING

ABSTRACT

Given the variety of languages, cultures, and curricular priorities across APEC countries, it would seem difficult to unite around a common set of teaching and learning standards for purposes of international communication. Yet the nascent field of "psychometrics" offers practical solutions to such problems by its ability to "equate" tests that differ in difficulty and even content, and by its ability to set standards that have the same meaning across tests and countries. After summarizing the principles behind classical and modern educational measurement, the paper discusses several technologies that can make it possible for APEC countries to jump the language barrier without sacrificing local imperatives. These technologies include: local, national and international item banks, computer adaptive testing, and the Lexile framework.

15650A VINEYARD BLVD.
#161
MORGAN HILL,CA 95037
markm@eddata.com

THE PROBLEM

It need hardly be said that the ability to conduct business and interact socially across national boundaries requires a common language, a fact nowhere more apparent than in the 13 exceedingly diverse economies that make up the APEC. By accident of historical and geographic circumstance, the *de facto* common language appears to be English, though other languages are used to bind together smaller subsets of the Pacific rim. APEC members, naturally, seek ways to make communication in English across national boundaries more easy and reliable than it is now. I wish in this paper to discuss ways to make the process more straightforward using tools from the educational measurement world. I will explain where these tools came from, how they work, and what they can be used for.

In order to conduct business across national, cultural, and economic borders, employers and academic institutions are obliged to work with individuals who have demonstrated an acceptable level of fluency in an international language such as English. That means applicants need to be tested and their fluency quantified on an English fluency scale. The scale must be applicable to the four domains of language– Reading, Writing, Listening, and Speaking. The scale must include one or more cut-points or passing grades above which it is reasonable to believe that an individual is acceptably fluent in a specified domain. The Test of English as a Foreign Language (TOEFL) and the Test of English for International Communication (TOEIC), both products of U.S.-based Educational Testing Service (ETS), seek to fill this need in the academic and business realms, as does the International English Language Testing System (IELTS) based in Cambridge, England.

In theory, such tests would be sufficient to meet the English language assessment needs of the APEC. It is becoming clear that they are not:

1. There is a limited number of seats and testing dates available at test centers. Demand exceeds supply.

2. Each test is on its own scale. That means it is not possible to compare the score of a person who has taken the TOEFL, say, with the score of person who has taken another test, such as the IELTS or one of the national exams.

3. The TOEFL and TOEIC and their language scales are proprietary to ETS. Therefore, only ETS can increase the availability of tests that use those scales. While another company or a government agency can administer its own tests, such tests would be on a different scale and not comparable to the ETS tests.

4. These tests are given only to adults. They are not suitable for testing the fluency of children in the lower grades. Countries need to keep track of student growth in the lower grades and have no simple way to match student performance in lower grades with expected performance on one of the adult English fluency tests.

5. Aside from a few international tests like the TOEFL and the IELTS, English testing is largely a national affair. Because the separately developed tests of the individual APEC economies are not matched to each other – equated – it is impossible to know the relative English fluencies of examinees who take different tests.

6. This also means there is no single set of standards that is applied uniformly across APEC to students, and especially teachers, of English as a foreign language.

These and similar considerations lead me to believe that APEC countries will eventually need to rethink the English assessment problem from first principles and develop, jointly, a more robust and generalizable methodology than what is used today. Fortunately, there are straightforward solutions to many, if not all, of the issues raised above using tools that are either currently available, or that will become available over the next five years. I am thinking of such tools as:

- Test equating
- Item banking
- Computer adaptive testing (CAT)
- The Lexile scale

To understand these tools it is necessary to review a bit of the history of educational measurement as regards the problem of test equating.

CLASSICAL TEST THEORY – THE SAME TEST FOR EVERYBODY

In the old days (pre-1980, say, though Classical Test Theory is still widely used) the solution was simple. Administer the same test to everybody and compare their scores on some agreed upon metric. This comparison involved calculating how each student performs *relative* to the rest of the students who take the test. This could be done with percentiles – what percent of students score less than Student X – or with some other way of locating the relative position of Student X in the overall distribution of scores, such as standard deviation units or normal curve equivalents.

(A "score distribution" is – ideally – a bell-shaped curve reporting how many students receive a given score for each point on the scale. The "mode" of the distribution is the score that most students receive and tends to be near the center of the distribution. The "mean" of the distribution is the average score, usually pretty close to the mode. The "standard deviation" of the distribution is the average distance from the mean. When a distribution of student scores is described in standard deviation units, the mean score is assigned a value of 0.0 standard deviation units. A value of 1.0 means one standard deviation unit above the mean, and a value of -1.0 means one standard deviation below the mean. If Student X's score is 1.5, that means he scored one and a half standard deviations above the average score for that population of students. Often, the standard deviation unit is converted into some equivalent unit that is deemed more user-friendly. For instance, the famous Intelligence Quotient (IQ) evolved so that a score of 100 refers to the mean or average of the test population, and 15 points above and below, 115 and 85, refer to one standard deviation above and below the mean.)

Classical test theory works so long as several assumptions are met:

1. Same test. Everybody takes the same test, the same set of questions, generally called "items." Students who leave an item blank either have to be dropped from the test or given a zero for that item, or someone needs to guess how the student would have answered the item. Statistical analysis is performed at the level of the total test score. Scores on individual items ("right" or "wrong") do not play much of a role.

2. Representative examinee samples. The distribution of scores for a sample of examinees should be representative of what one would obtain if every student in the total population of all possible students were tested using the same test. The larger the sample of students the more likely the sample is to be "representative" though this is by no means assured unless examinees are picked at random. Therefore, classical test theory works best with large samples of students, at least above a thousand and preferably in the tens of thousands.

3. Representative items. In a manner parallel to students, we assume that the items are representative of the population of all possible items for a given content area.

4. Common person equating. To compare the relative difficulty of two tests, we need a representative sample of students to take both tests. Using one of several methods, we adjust the scores on one test to equal the scores on the other test. The function used to perform this adjustment becomes the equating function.

I want to call your attention to two things in particular. First, in classical test theory the student measures depend on the items they happen to take, and the item measures (how difficult they are) depend on the students who happen to take the exam. There is no guarantee that if the same set of students were given a different set of items they would get the same relative scores. That is why it is so important that the items be "representative" of the content matter for every test aimed at that content. The term for this is that both the student measures and the item measures in classical test theory are "sample dependent."

Second, classical test theory has no way to address the problem of missing data. This may not sound important since most students answer all the items on a test, but it is the central limitation of classical test theory. Many of the interesting psychometric problems are missing data problems. Most test equating tasks involve large blocks of missing data, as does computer adaptive testing (CAT). More complicated measurement designs that take into account rater severity, for example, are completely impractical without a way to handle missing data.

Third, classical test theory has no way to assess the quality of individual student responses, their internal consistency. Each set of responses assigned to a particular person or item is assumed to be as good as any other set of responses.

It was not long before several theoreticians (Fred Lord, Georg Rasch, Ben Wright) decided these and other limitations were impractical and unnecessary, especially when it came to equating tests for different age groups of students. Rasch in particular, and his student Ben Wright, discerned the fundamental weakness of classical test theory and derived from first principles a new model of psychological measurement that, if applied according to its specifications, would grant the measurement of psychological and educational quantities the same status as the measurement of physical quantities in physics. This measurement model became known as the Rasch model. In the 1960's, the Rasch model and related models became the basis of a field known as Item Response Theory. Despite decades of published research, these models only took hold in the educational testing arena in the 1990's. It is the Rasch model that makes most of what I am going to discuss possible.

ITEM RESPONSE THEORY – TEST LINKING

Stated simply, Item Response Theory uses probability models to calculate the probability that an examinee will get a specific item correct. It does this by counting all the correct responses associated with that examinee (ability) and all the incorrect responses associated with that item (difficulty), and adjusting the examinee abilities and item difficulties until the fit between the observed values and the model's predictions has been maximized. The general form of IRT models is:

Probability of a Correct Response = function (Person Ability – Item Difficulty)

This is the simplest form of the model; many complex variations have been derived. Within the IRT community, there has been a difference in philosophy between advocates of the Rasch model, which relies on a simple additive function that relates person ability to item difficulty, and other models such as the "3-PL model" which include parameters that take into account "item discrimination" and "guessing" to improve the fit between the observed responses and the corresponding values predicted by the model. The debate has largely run its course, but the upshot is that the more complex models like the 3-PL model do a better job of producing estimates that match the observed data for a *specific* test, whereas the Rasch model does a better job of computing person abilities and item difficulties that *generalize* across multiple tests, so long as the data set is "edited" using diagnostic statistics to maximize fit between the observed responses and predicted values. The difference is often summarized as: The 3-PL model fits the model to the data; the Rasch model fits the data to the model.

From an outside perspective the difference seems academic. Philosophically, the difference is profound and it has enabled the Rasch model to develop in directions that the 3-PL model has not. For instance, the Rasch model can be extended to take into account rater severity and any number of other factors impacting scores, as in the "Many-Facet Rasch Model," whereas models like the 3-PL model become mathematically intractable in such circumstances. The distinguishing characteristic of the Rasch model is that it specifies the unique condition under which item differences can be disentangled or "separated" from person differences, so that persons will receive the same relative ability measures regardless of the items they take, and items will receive the same relative difficulty measures regardless of the persons who take them. Unlike Classical Test Theory which is "sample dependent," Rasch models are "sample free."

There is, of course, a price to be paid for the Rasch model's "sample free" property. It is not achieved simply by running any set of test data through a computer program. The items have to be carefully written and studied to ensure that they match a clearly understood and defined construct and are not biased for or against any subset of examinees. Not any set of items will do. Instances in which there is a significant difference between the model's predictions and the observed responses (called "misfit") need to be understood and the test altered accordingly. Misfitting persons are suspended from the analysis temporarily when calibrating item difficulties. Misfitting items are dropped. Fortunately, because the Rasch and other IRT models are not compromised by missing data, it is straightforward to drop misfitting persons and items from the analysis and run the program without them.

TEST EQUATING

I just stated a very important point. Unlike Classical Test Theory, IRT models can handle *missing data*. Remember, IRT models calculate the probability that a given person will get a given item correct. This applies both in cases where we have a response for the person on that item, and in instances where a person did not answer an item. Thus, IRT automatically *completes* any test data matrix with "expected values," filling in missing values as desired.[1] That is what makes test equating so straightforward with something like the Rasch model and so hard with Classical Test Theory.

Figure 1 is a (highly idealized) example of a typical equating problem. The rows are test examinees (persons) arranged in order of English fluency. The columns are test items arranged in order of difficulty. The zeros and ones indicate whether a given person got an item correct ("1" means "correct"). There are two tests. Test X contains items 1-5 and may be considered an easy test. Test Y contains items 4-12 and may be considered a hard test. Items 4 and 5 appear on both tests. No persons took both tests. The "% Correct" column gives the percentage of items each person got correct. The "% Incorrect" row at the bottom gives the percentage of persons who got each item incorrect.

Here is my item to you: Which examinee is more fluent, Person A or Person G?

Figure 1

		Easy		**[Test X]**			*Questions*			**[Test Y]**			Difficult	
		1	2	3	4	5	6	7	8	9	10	11	12	% Correct
Less	A	0	0	1	1	0	.	.	.	.	.	.	.	***0.40***
Fluent	B	1	0	1	0	0	.	.	.	.	.	.	.	0.40
	C	0	1	0	0	1	.	.	.	.	.	.	.	0.40
	D	1	1	1	1	0	.	.	.	.	.	.	.	0.80
	E	0	1	1	0	0	.	.	.	.	.	.	.	0.40
	F	1	1	1	0	0	.	.	.	.	.	.	.	0.60
Persons	G	.	.	.	1	0	0	1	1	0	0	0	0	***0.33***
	H	.	.	.	1	0	1	0	1	1	1	0	0	0.56
	I	.	.	.	1	1	0	1	1	0	0	0	0	0.44
	J	.	.	.	1	1	1	1	1	1	1	0	1	0.89
	K	.	.	.	1	1	1	0	0	0	1	1	0	0.56
More	L	.	.	.	1	1	0	1	1	1	1	0	1	0.78
Fluent	M	.	.	.	1	1	1	0	1	1	1	1	1	0.89
	% Incorrect	0.50	0.33	0.17	0.31	0.54	0.43	0.43	0.14	0.43	0.29	0.71	0.57	

If we naively used Classical Test Theory, we would assume that both sets of items (Test X and Test Y) are equally representative of the total population of test items, and that therefore we can compare Persons A and G in terms of their scores. Person G gets a lower percent correct (0.33) than Person A (0.40), so Person A is more fluent.

If we were less naïve we would concede the obvious, that the two tests have mostly different items and that their scores cannot be compared to each other. Since there are no persons who took both tests, our only recourse is to pay for an equating study and give both tests to a representative

[1] This statement is an oversimplification. IRT models use algorithms that calculate person and item parameters without filling in missing data. However, they have the *capacity* to fill in missing data if desired and it is instructive and valid to interpret IRT models as tools for filling in missing data.

sample of persons and use the average difference in performance between the two tests to calculate their relative difficulty.

Or, we could just run the data set through the Rasch model and fill in the missing cells with expected values, then calculate a new percent correct as the average of the observed and expected values for each row (Figure 2).[2]

Figure 2

		Easy		**[Test X]**			*Questions*			**[Test Y]**			Difficult	
		1	2	3	4	5	6	7	8	9	10	11	12	% Correct
Less	A	0	0	1	1	0	0.26	0.22	0.17	0.13	0.09	0.04	0.00	***0.24***
Fluent	B	1	0	1	0	0	0.30	0.26	0.22	0.17	0.13	0.09	0.04	0.27
	C	0	1	0	0	1	0.35	0.30	0.26	0.22	0.17	0.13	0.09	0.29
	D	1	1	1	1	0	0.39	0.35	0.30	0.26	0.22	0.17	0.13	0.49
	E	0	1	1	0	0	0.43	0.39	0.35	0.30	0.26	0.22	0.17	0.34
	F	1	1	1	0	0	0.48	0.43	0.39	0.35	0.30	0.26	0.22	0.45
Persons	G	0.74	0.70	0.65	1	0	0	1	1	0	0	0	0	***0.42***
	H	0.78	0.74	0.70	1	0	1	0	1	1	1	0	0	0.60
	I	0.83	0.78	0.74	1	1	0	1	1	0	0	0	0	0.53
	J	0.87	0.83	0.78	1	1	1	1	1	1	1	0	1	0.87
	K	0.91	0.87	0.83	1	1	1	0	0	0	1	1	0	0.63
More	L	0.96	0.91	0.87	1	1	0	1	1	1	1	0	1	0.81
Fluent	M	1.00	0.96	0.91	1	1	1	0	1	1	1	1	1	0.91
	% Incorrect	0.30	0.25	0.19	0.31	0.54	0.52	0.54	0.41	0.58	0.53	0.78	0.72	

Having filled in the missing cells, we can now say that *all* the examinees, Persons A-M, have in effect taken the *same* test, a combination test called XY. It is now valid to compare their scores. When we compare Persons A and G, we see that Person G is almost *twice* as fluent as Person A – 0.42 percent correct versus 0.22 percent correct. The naïve use of Classical Test Theory was not just a little wrong, but spectacularly wrong.

This example yields three important rules:

1. Valid Comparisons. Examinees can only legitimately be compared when they can be said, either literally or theoretically, to have taken the same test.

2. Definition of Equating. Two tests X and Y are equated when from the responses on X and the responses on Y it is possible to infer the responses and total score that students *would* have received on a common test XY composed of the items from both tests.

[2] So called "expected values" are in this case the same as probabilities, the probability of getting a "1". They are more precise than observed values because they take into account all the data in the matrix. For purposes of illustration I am mixing observed and expected values in the "% Correct" column, but this is not what IRT algorithms do.

3. Test Linking. In order to infer responses on one test based on another, the tests must somehow be linked. There are three ways to link two tests X and Y:

    a. Common Persons. Tests X and Y must be administered, in their entirety, to a common sample of persons at more or less the same time. This is called common-person equating and is the only way to equate using Classical Test Theory.

    b. Common Items. Tests X and Y must have items in common. This is called common-item equating. Figures 1 and 2 show common item equating (Items 4 and 5 are on both tests) but not common person equating (no persons took both tests completely).

    c. Common objective characteristics. Tests X and Y must have "objective characteristics" in common. There must be some way to infer from the test questions themselves their likely difficulty without recourse to person responses. This third type of equating is rarely noted but is the foundation of the Lexile scale, as discussed below.

A FEW TECHNICAL CLARIFICATIONS

To avoid misconceptions, let me point out that IRT software does not actually compute the expected values for missing cells, nor does it average observed values and expected values. It arrives at similar person and item parameters (closely analogous to "percent correct") by way of a more efficient algorithm, called Joint Maximum Likelihood. Once the person and item parameters have been derived, they can be used to estimate the missing values if desired. If one averages across the expected values for all cells in a row or column, they will closely approximate a simple function of the person and item parameters calculated using Joint Maximum Likelihood. IRT software *does* calculate expected values for each cell that contains an *observed* value. It adjusts the person and item parameters until the expected values are as close as possible to, or best "fit," the observed values, given the mathematical constraints of the model.

I did not use the Rasch model to compute the expected values in this example but used theoretical values for illustrative purposes only. However, they are similar to what Rasch expected values would look like.

Psychometricians do not work in a "% correct" metric but in what is called a "logit" metric. Because percentages cannot exceed 100 or go below zero, percentage units are not all the same size and cannot be added and subtracted like regular measures. To convert percentage units into units that all have the same size, we divide the percent correct by the percent incorrect (called the odds of success), then take the log of the odds so that measures can be compared by adding and subtracting rather than by multiplying and dividing. The result is a "logit," short for "log odds unit." Logits range from negative infinity to positive infinity (though for most tests they range from around -4 to +4) and are centered at zero.

Finally, the idealized data set above includes "random error" (a random number added to each data value before rounding to 0 or 1). That is why items that I know to be more difficult (because I generated this pseudo-data set) can sometimes appear to be less difficult. Error is intrinsic to all measurement and is explicitly modeled and calculated by the Rasch model. It is also why IRT models are probability models. We are not predicting an actual value of "1" or "0"; we are calculating the *probability* that an examinee will get a "1".

Simple as they are, Figures 1 and 2 capture the important contribution made by IRT to measurement theory. Almost all measurement and equating problems can be reduced to this example.

ITEM RESPONSE THEORY REQUIRES UNIDIMENSIONALITY

Classical Test Theory, unlike IRT, does not require that all items on a test measure along a single educational construct, just that they somehow be "representative" of a specified domain. The Rasch model requires that test items measure along a single clearly defined "dimension," and no more than one dimension at a time. This is called the requirement of "unidimensionality." A math test should contain only math items. An English test should contain only English items. While it is up to test designers to be clear what they want to measure and to target their items accordingly, it is the model itself that will tell them whether they succeeded. When the observed responses associated with a particular item differ significantly from what the model predicts for those cells, it is likely that the item is sensitive to a dimension that was not the intended construct or dimension of the test.

The classic example is when a mathematically competent examinee fails to solve a simple word problem because it is written in an unfamiliar language. While the test is targeted on mathematics and generally succeeds in detecting math ability, for the word problem it behaves a bit like an English test. Rasch diagnostic statistics would inform test developers either to drop the word problem or to administer the test only to examinees who can understand the language in which the test is written. When test developers ignore these diagnostics, they are sacrificing the most important property of Rasch models, namely, the capacity to create measures that can generalize to other tests, sometimes called "objectivity" or "invariance." In this case, a mathematically competent examinee is given a lower test score merely as a result of a language issue that has nothing to do with his math ability. We say, the test is not "objective," contains "biases." Rasch fit statistics are used to identify and remove such biased items.

Figure 1 does not look like a unidimensional data set, but it is. Random error and the capture of data only as zeros and ones obscures the underlying unidimensional structure, but one can see that the more difficult items have more zeros while the more able students have more ones, and that the expected values are perfectly correlated to each other, column to column. What does one do when the underlying structure is more complex, more multidimensional?

MULTIDIMENSIONAL ITEM RESPONSE THEORY

The requirement of unidimensionality imposes a salutary discipline on item writers and forces clarity of thought while identifying items that are not likely to yield generalizable measures. It must nonetheless be recognized that there are instances – many instances -- when educational and psychological data sets are highly and intrinsically multidimensional and the Rasch model is unable to make sense out of them and should not be used. Psychological profiling for marketing purposes is a typical example. This problem has lead to the development of so-called "Multidimensional IRT" (MIRT). In the context of our discussion on test equating, a MIRT model would, or should, allow us to predict the missing cells in data sets like Figure 1

Although IRT does offer models that are called "multidimensional," the MIRT subfield is still in an early and relatively disordered stage of development despite models that have been in existence since the 1970s. For whatever reason, the models are poorly understood and software is hard to find and use. In this regard, the field of "data mining" has taken the lead in implementing methodologies and software to analyze multidimensional data sets. It is not yet clear that existing MIRT models or those used in advanced data mining do for multidimensional spaces what the Rasch model does for 1-dimensional spaces.[3] If such a model did, it would have the following properties:

1. Fit multidimensional data. It should produce cell estimates that reasonably fit the observed values even though the items are not all measuring along the same dimension or construct and are to varying degrees uncorrelated, i.e., the data set is multidimensional.

2. Predict missing values. It should accurately predict the values of missing cells, regardless of whether they are randomly missing (the student forgot to answer) or missing by design (as in test equating).

3. Person invariance. It should locate persons in space such that their relative positions remain the same regardless of the sample of items they answer, assuming that each sample of items erects the same space.

4. Item invariance. It should locate items in the same space as the persons, and the relative positions of the items should stay the same regardless of the sample of persons that responds to the items.

5. Misfit when invariance is not achieved. When person responses to particular item vary due to the influence of dimensions that are not embodied in the remaining items on the test, that person's observed responses should on average be significantly different from the values predicted for the same items. In other words, the model's predictions and the observed values should disagree. The same applies to the response associated with a given item.

6. Transferability. Person and item parameters should be transferrable across data sets of the same kind, so that the results of one analysis can be applied to another.

7. Maximal use of information. The model estimate corresponding to each cell in the data matrix should to the maximum extent possible use all the information in the entire data set and be correspondingly precise.

8. Standard errors. Each cell estimate – how a person is expected to score on an item – should be accompanied by its own standard error, and that standard error should accurately reflects the range of estimates that is likely to be produced with different samples of persons and items drawn from data sets with the same dimensionality and sample size. Person and item parameters should also have standard errors.

While practical multidimensional test-equating software meeting these requirements appears to be years in the future, there can be little doubt that multidimensional IRT will become the preferred methodology for equating tests.

[3] The author and colleagues have developed multidimensional models under the rubric of NOUS that model highly multidimensional data sets and generate parameters that can be generalized across data sets. These models have proven useful with educational data. However, they are not complete or available and have not been widely reviewed. For more information, visit www.eddata.com.

TOOLS RELEVANT TO APEC

IRT, and the Rasch model in particular, has spawned an array of tools that are highly relevant to APEC's problem of establishing a common international scale with international standards. All of these tools can be understood as variations on the theme of equating posed in Figures 1 and 2.

EQUATING THROUGH AN INTERNATIONAL ITEM BANK

The idealized data set in Figure 1 has 12 items. Imagine there were 12,000 items or more. Obviously no single student would take all 12,000 items, but if groups of examinees were administered different overlapping sets of items it is conceivable we could construct a data set that would be appropriately linked through common items. Then we could run the data set through a Rasch software program and calculate the difficulty of each item, dropping items that do not fit. The remaining set of items, with their difficulty measures, would be what is called an "item bank."

Item banks have been used very successfully for decades by such organizations as the Northwest Evaluation Association (NWEA). They make it possible to create any test on the fly, for any occasion. Items are selected for their difficulty and appropriateness, assembled onto a test form, and administered by paper or computer, or online. Because the item difficulties are calculated ahead of time, it is easy to compute each examinee's expected score, roughly equivalent to what they would have received had they taken the entire 12,000 item test (though the level of precision reflects only the number of items actually taken). All examinees are measured on the same common scale. Building an item bank involves: a) writing the test items, somewhere above 12,000, preferably 30,000; b) constructing a database to manage the items and assemble them on tests; c) gathering data on the items through overlapping test administrations; and d) removing items that behave inconsistently.

While item banks take some work, they have many benefits:

1. Different tests, a common scale. All examinees are placed on a common scale, regardless of the test they take. Each APEC nation could assemble and administer its own tests, yet place examinees on a common international language fluency scale.

2. Test Security. Because there are so many test items, it is unlikely that any one person could prepare for them, even if the person somehow gained access to the item bank. The accidental exposure of items from a particular test can easily be corrected by drawing a new set of items from the bank and retiring the exposed items for a few years.

3. Test All Grades. Because any set of items can be selected from the item bank, items can be targeted to any grade of students and any ability level. Examinees in Grade 3 can be compared to students who take the TOEFL.

4. Include Non-Bank Items. It is easy to include new items on the test that are not already in the item bank. These items can be analyzed after the test administration and added to the international item bank if they are of sufficient quality. This makes it easy to grow the item bank.

5. Freedom of Use. An international item bank would free APEC member countries from having to rely exclusively on private companies such as the publishers of the TOEFL and TOEIC.

The construction of item banks to test Reading is straightforward and is done routinely. It would not be hard to construct an item bank to test Listening. However, Writing and Speaking present technical challenges that complicate item banking. The challenges are not insurmountable, however.

COMPUTER ADAPTIVE TESTING

Another benefit of a large international item bank is that it would enable member countries to employ computer adaptive testing (CAT). The principle is simple: the examinee is administered a sample of range-finder questions that yield an approximate starting ability. The computer looks in the item bank for a suitable item that is targeted to the examinee's starting ability. If the examinee gets the item correct, he is given a harder item. Otherwise, he is given an easier item. The examinee's ability is recomputed after each item. After 20 items or so, the computer "homes in" on the student's ability to any desired degree of accuracy. Such scores tend to be *much* more precise and reliable than scores from a paper test since paper tests, by being inflexible, tend to be poorly targeted on a large percentage of examinees.

To be successful, CAT requires a large bank of carefully screened items, more than 1000 items per content area -- as many as possible. It is important that items not be over-exposed to examinees, that they range across the entire difficulty scale so that high- and low-ability examinees have as many items targeted to them as mid-range students. It is especially important that the items be carefully examined for their fit to the Rasch model. CAT only works if the items can be trusted to have constant and reliable difficulties, invariant across examinee types.

ETS abandoned its CAT program for scoring the TOEFL for undisclosed reasons. However, other companies and testing agencies have been successful with CAT in high-stakes certification exams over many years.[4] The methodology is sound and well proven, if properly applied.

In the language fluency domain, the application of CAT requires that a computer be able to score an examinee's responses on the fly. This is easily done with Reading and Listening (though Listening items require ear-phones and a microphone). Writing and Speaking are more challenging. Technology has been developed by a number of companies, including ETS, that machine-scores writing samples with a level of precision that is comparable to that of human scorers. I have little doubt that such technologies will come to dominate the writing-scoring industry. However, I have not seen any efforts to integrate this technology with computer adaptive testing. Speaking is even harder to score, but could theoretically be managed with recent voice recognition technologies. Again, I have not seen these attempted in the context of adaptive testing.

THE LEXILE SCALE

The Lexile scale was constructed by a U.S. company called MetaMetrics in North Carolina (with which I have no business relationship).[5] It relates reading difficulty, as assessed using test data analyzed with the Rasch model, to "objective characteristics" of the reading items. Most reading comprehension items involve reading a passage of text. It was found that the average difficulty of items relating to that text can to a large extent be predicted by looking at the semantic difficulty and

[4] The National Council of State Boards of Nursing offers the NCLEX computer adaptive exam to certify nurses. Go to www.ncsbn.org. The highly regarded Norwest Evaluation Association (NWEA) has shifted entirely to a computer adaptive format for its K-11 English and Mathematics tests.

[5] Go to www.metametricsinc.com or www.lexile.com and related links for information about the Lexile framework.

syntactic complexity of the text to be read, variables that can be operationalized using log transformations of word frequency (how frequently each word is used in the English language) and sentence length (the number of words in a sentence). Their relationship to item and test difficulty were used to construct a regression equation that makes it possible to predict the difficulty of a passage of text simply by looking at these two objective characteristics. A software program called the Lexile Analyzer was written to scan and analyze text and return the reading difficulty of that text on a scale called the Lexile scale. When a person reads text that is at that person's location on the Lexile scale, it means he is expected to answer 75% of hypothetical questions based on that text accurately. The Lexile scale is used widely to assess the readability of text, and many large-scale assessments include a Lexile score for each student who takes the test. This Lexile score is used to guide teachers and parents in selecting texts that are at the right difficulty level for that student.

I raise the Lexile scale because it may offer the easiest and least expensive means of equating Reading (and Writing) tests in the APEC economies. Instead of writing items and building item banks, participating testing agencies within the APEC would submit test materials to MetaMetrics for scanning and obtain Lexile measures for their materials along with information on how to calculate Lexile scores for examinees who take the test. The Lexile framework would become a *de facto* English fluency scale for APEC countries. The strategy offers a number of attractive benefits:

- Easy, inexpensive, and flexible. The work of scaling existing tests used in APEC countries would be minimal. All such tests would automatically be equated to each other since they report results on the same scale. New tests can be added to the Lexile framework without the effort of writing linking items or performing equating studies.

- Usable by teachers. The scale has been applied to a large corpus of texts. English instructors would have automatic access to a powerful tool for matching their curricular materials to student ability at all grade levels, including adults.

- Performance standards. Setting international performance standards becomes straightforward. The task of defining performance levels for different levels of English proficiency (for students, adults, and teachers) is greatly simplified by the fact that so much Lexiled data has been accumulated. Exams, texts, functional tasks, and various educational criteria can already be found on the scale. Existing performance scales such as those used on exams in APEC countries, or the Common European Framework of Reference for Languages, can be placed on the scale without undue difficulty.

- Wide applicability. TOEIC measures business language skills; TOEFL measures academic language skills. Each requires its own items and its own scale. However, both a business language test and an academic language test could be scaled to report in Lexiles – even though the constructs are different. All that is necessary is that the appropriate qualifications be added to each score. Thus, we would say based on his results on two tests that Examinee X has a Reading ability of 900 Lexiles in business English and a Reading ability of 800 Lexiles in academic English. He might have different (generally lower) Lexile scores for Writing.

- Transparency. Although developed and marketed privately, the Lexile scale is based on objective test characteristics that can be independently verified and replicated. This allows the framework to extend beyond any single exam or set of curricular materials. In this sense it differs markedly from the closed, proprietary scales that tend to be used for international assessments such as the TOEFL.

- Stability. The stability of the scales used for proprietary assessments depends on a wide variety of factors involving the selection, use, and removal of test items from tests over time. A scale based on objective characteristics like sentence length and word frequency is much more likely to retain its stability over many years, even as individual tests go in and out of use.

There are at least two psychometric challenges posed by the Lexile scale that would have to be solved:

- Listening and Speaking. So far, MetaMetrics does not appear to be offering a way to convert responses to Listening and Speaking tests into Lexiles. The problem does not seem to be insurmountable, however, and may in fact prove quite simple.

- English and Spanish. The Lexile scale has only been constructed for English and Spanish, so far as I know. It can in theory be developed for any language, including the Asian languages, but the process would require substantial time and research.

I attach a *much* reduced picture of a poster used to explain the Lexile scale, downloaded from the website for the State of Illinois.[6] Developed by MetaMetrics, it provides an excellent overview of how the scale works. Each band, moving from the bottom to the top, represents a Lexile range in increments of 100. The left column provides a list of works of literature for each band of the scale. The column in the center provides an actual sample of text from one of these works. The column to the right lists assessments and curricular materials that have been assigned to that Lexile band. (Note that the TOEFL has been measured at 1400L.)

CONCLUSION

The problem of establishing a common scale and common language standards for the APEC economies – particularly as regards the learning, teaching, and use of English – is eminently solvable. Psychometric tools developed over the last 40 years in response to the need to equate tests – to place students on a common scale despite being assessed with different instruments – offer powerful and practical ways to equate tests used in APEC countries and to increase the range of options available to national testing agencies and companies and organizations that need quick and reliable ways to assess foreign language fluency. These tools include item banks, computer adaptive testing, and the Lexile scale. They are well worth researching further in addressing this important issue.

[6] A more readable version can be accessed at www.isbe.state.il.us/assessment/pdfs/lexile.pdf. Full-sized posters can be obtained from MetaMetrics.

The Lexile Framework® for Reading

STUDENT ASSESSMENT DIVISION RESOURCE

ILLINOIS STATE BOARD OF EDUCATION
www.isbe.net/assessment/pdfs/lexile.pdf

GRADUATE SCHOOL
COLLEGE JUNIOR-SENIOR
COLLEGE FRESHMAN- SOPHOMORE
TWELFTH GRADE
ELEVENTH GRADE
TENTH GRADE
NINTH GRADE
EIGHTH GRADE
SEVENTH GRADE
SIXTH GRADE
FIFTH GRADE
FOURTH GRADE
THIRD GRADE
SECOND GRADE
FIRST GRADE

Text Level	Literature Titles	Benchmarks	Tests/Textbooks
1700L	1690 Concerning Civil Government 1680 Critique of Judgment 1660 On Abraham Lincoln 1660 On the Law Which Has Regulated the Introduction of New Species	**DISCOURSE ON THE METHOD AND MEDITATIONS ON FIRST PHILOSOPHY** To such a class of things pertains corporeal nature in general, and its extension, the figure of extended things, their quantity or magnitude and number, as also the place in which they are, the time which measures their duration, and so on. That is possibly why our reasoning is not unjust when we conclude from this that Physics, Astronomy, Medicine and all other sciences which have as their end the consideration of composite things, are very dubious and uncertain; but that Arithmetic, Geometry and other sciences of that kind which only treat of things that are very simple and very general, without taking great trouble to ascertain whether they are actually existent or not, contain some measure of certainty and an element of the indubitable. (Rene Descartes, author)	1670 The Principles of Scientific Management; Dover Publications 1630 The American Constitution: Cases, comments, questions, 7th ed.; West Publishing 1610 The Condition of Postmodernity; Blackwell Publishers□ □
1600L	1570 Aeropagitica 1550 God, Idea of the Ancients 1530 Plutarch's Lives 1520 A Modest Proposal 1510 On Human Nature 1500 The Decameron	**FUNDAMENTAL PRINCIPLES OF THE METAPHYSICS OF MORALS** In fact, it is absolutely impossible to make out by experience with complete certainty a single case in which the maxim of an action, however right in itself, rested simply on moral grounds and on the conception of duty. Sometimes it happens that with the sharpest self-examination we can find nothing beside the moral principle of duty which could have been powerful enough to move us to this or that action and to so great a sacrifice; yet we cannot from this infer with certainty that it was not really some secret impulse of self-love, under the false appearance of duty, that was the actual determining cause of the will. (Immanuel Kant, author)	1550 Culture/Power/History: A Reader in Contemporary Social Theory; Princeton University Press 1530 On Injuries of the Head; Project Gutenberg 1510 On Human Nature; Howard University Press 1500 On Liberty; Hackett Publishing 1500 The Making of Memory:From MoleculestoMind;Doubleday
1500L	1490 Ring of Bright Water 1470 Utilitarianism 1450 The Confessions of Nat Turner 1440 The Legend of Sleepy Hollow 1420 Master Humphrey's Clock□ 1410 Profiles in Courage	**ON ANCIENT MEDICINE** And as to him who had been accustomed to dinner, since, as soon as the body required food, and when the former meal was consumed, and he wanted refreshment, no new supply was furnished to it, he wastes and is consumed from want of food. For all the symptoms which I describe as befalling to this man I refer to want of food. And I also say that all men who, when in a state of health, remain for two or three days without food, experience the same unpleasant symptoms as those which I described in the case of him who had omitted to take dinner. (Hippocrates, author)	1450 Philosophical Essays; Hackett Publishing 1440 *Graduate Management Admission Test (GMAT)** 1430 *Certified Public Accountant Examination (CPA)** 1430 Criminal Justice Today; Prentice Hall 1410 Science and Education; The Citadel Press 1400 *Test of English as a Foreign Language (TOEFL)**
1400L	1380 Life in a Medieval Castle 1350 The Guns of August 1340 The Hunchback of Notre Dame 1330 The Snow Leopard 1320 The Metamorphosis 1300 Arctic Dreams	**THE SCARLET LETTER** But the point which drew all eyes, and, as it were, transfigured the wearer—so that both men and women who had been familiarly acquainted with Hester Prynne were now impressed as if they beheld her for the first time—was that SCARLET LETTER, so fantastically embroidered and illuminated upon her bosom. It had the effect of a spell, taking her out of the ordinary relations with humanity, and enclosing her in a sphere by herself. "She hath good skill at her needle, that's certain," remarked one of her female spectators; "but did ever a woman, before this brazen hussy, contrive such a way of showing it? Why, gossips, what is it but to laugh in the faces of our godly magistrates, and make a pride out of what they, worthy gentlemen, meant for a punishment?" (Nathaniel Hawthorne, author)	1390 *Graduate Record Examination (GRE)** 1380 *College Board Achievement Test in English (CBAT)** 1380 *Law School Admission Test (LSAT)** 1330 *Scholastic Aptitude Test (SAT)** 1330 *Medical College Admission Test (MCAT)** 1320 Psychology: An Introduction; Prentice Hall
1300L	1280 House of the Spirits 1270 Chronicle of a Death Foretold 1240 The Midwife's Apprentice 1240 Dragon Seed 1210 Cold Mountain 1200 The Trumpeter of Krakow	**BROWN v. BOARD OF EDUCATION: 1954** Under that doctrine, equality of treatment is accorded when the races are provided substantially equal facilities, even though these facilities be separate. In the Delaware case, the Supreme Court of Delaware adhered to that doctrine, but ordered that the plaintiffs be admitted to the white schools because of their superiority to the Negro schools. The plaintiffs contend that segregated public schools are not "equal" and cannot be made "equal," and that hence they are deprived of the equal protection of the laws. Because of the obvious importance of the question presented, the Court took jurisdiction. Argument was heard in the 1952 Term, and reargument was heard this Term on certain questions propounded by the Court. (347 US 483, 98 L ed 873, 74 S Ct 686)	1290 Understanding Sociology; Glencoe/McGraw-Hill 1290 Speech Science Primer; Williams & Wilkins 1240 Business; Prentice Hall 1230 *Armed Services Vocational Aptitude Battery (ASVAB)** 1210 *American College Testing Program (ACT)**
1200L	1190 Hiroshima 1160 The Pickwick Papers 1130 The Great Fire 1130 Abigail Adams: Witness to a Revolution 1120 Gutsy Girls: Young Women Who Dare 1100 Eleanor Roosevelt: A Life of Discovery	**WAR AND PEACE** Pierre had been educated abroad, and this reception at Anna Pavlovna's was the first he had attended in Russia. He knew that all the intellectual lights of Petersburg were gathered there and, like a child in a toyshop, did not know which way to look, afraid of missing any clever conversation that was to be heard. Seeing the self-confident and refined expression on the faces of those present he was always expecting to hear something very profound. At last he came up to Morio. Here the conversation seemed interesting and he stood waiting for an opportunity to express his own views, as young people are fond of doing. (Leo Tolstoy, author)	1170 *Scholastic Reading Inventory (SRI-Level 18)** 1160 History of a Free Nation; Glencoe/McGraw-Hill□ 1150 *National Assessment of Educational Progress (NAEP-Grade 12)** 1130 Modern Biology; Holt, Reinhart & Winston 1100 Modern Masonry; Goodheart-Wilcox Co. 1100 *Stanford Achievement Test (SAT 9-TASK 2)**
1100L	1090 Amos Fortune, Free Man 1070 All Things Bright and Beautiful 1030 Now is Your Time! 1030 Adam of the Road 1000 Island of the Blue Dolphins 1000 Parrot in the Oven: Mi Vida	**PRIDE AND PREJUDICE** Occupied in observing Mr. Bingley's attentions to her sister, Elizabeth was far from suspecting that she was herself becoming an object of some interest in the eyes of his friend. Mr. Darcy had at first scarcely allowed her to be pretty; he had looked at her without admiration at the ball; and when they next met, he looked at her only to criticise. But no sooner had he made it clear to himself and his friends that she had hardly a good feature in her face, than he began to find it was rendered uncommonly intelligent by the beautiful expression of her dark eyes. (Jane Austen, author)	1060 *Test of General Educational Development (GED)** 1050 *Test of Adult Basic Education, General Form (TABE-D)** 1040 *Scholastic Reading Inventory (SRI-Level 17)** 1040 Writing & Grammar: Gold Level; Prentice Hall 1020 African American Literature; Holt-Reinhart & Winston
1000L	970 Leon's Story 960 The Samurai's Tale 950 Bud, Not Buddy 940 All the Pretty Horses 930 The Golden Compass 920 Talking with Artists	**BLACK BEAUTY** One day, when there was a good deal of kicking, my mother whinnied to me to come to her, and then she said: "I wish you to pay attention to what I am going to say to you. The colts who live here are very good colts, but they are cart-horse colts, and of course they have not learned manners. You have been well-bred and well-born; your father has a great name in these parts, and your grandfather won the cup two years at the Newmarket races; your grandmother had the sweetest temper of any horse I ever knew, and I think you have never seen me kick or bite. I hope you will grow up gentle and good, and never learn bad ways; do your work with a good will, lift your feet up well when you trot, and never bite or kick even in play." (Anna Sewell, author)	990 *National Assessment of Educational Progress (NAEP-Grade 8)** 950 *Scholastic Reading Inventory (SRI-Level 16)** 940 World Cultures: A Global Mosaic; Prentice Hall 930 *Stanford Achievement Test (SAT 9-Advanced 2)** 910 *Test of Adult Basic Education (TABE-M)** 900 *Stanford Achievement Test (SAT 9-Advanced 1)**
900L	880 Her Stories 870 The View from Saturday 860 Julie of the Wolves 820 Maniac Magee 800 Homeless Bird 800 Scooter	**TOM SWIFT IN THE LAND OF WONDERS** Just what Tom's thoughts were, Ned, of course, could not guess. But by the flush that showed under the tan of his chum's cheeks the young financial secretary felt pretty certain that Tom was a bit apprehensive of the outcome of Professor Beecher's call on Mary Nestor. "So he is going to see her about 'something important,' Ned?" "That's what some members of his party called it." "And they're waiting here for him to join them?" "Yes. And it means waiting a week for another steamer. It must be something pretty important, don't you think, to cause Beecher to risk that delay in starting after the idol of gold?" "Important? Yes, I suppose so," assented Tom. (Victor Appleton, author)	870 Word 97; Glencoe/McGraw-Hill 860 *Scholastic Reading Inventory (SRI-Level 15)** 850 *Stanford Achievement Test (SAT 9-Intermediate 3)** 820 *National Assessment of Educational Progress (NAEP-Grade 4)** 810 *Stanford Achievement Test (SAT 9-Intermediate 2)** 800 Energy from Water; Harcourt
800L	790 Flour Babies 770 The Giver 760 Walk Two Moons 740 Henry and Ribsy 720 Some of the Kinder Planets 700 Catwings	**THE ADVENTURES OF PINOCCHIO** "Great soul!" said Pinocchio, fondly embracing his friend. Five months passed and the boys continued playing and enjoying themselves from morn till night, without ever seeing a book, or a desk, or a school. But, my children, there came a morning when Pinocchio awoke and found a great surprise awaiting him, a surprise which made him feel very unhappy, as you shall see. Everyone, at one time or another, has found some surprise awaiting him. Of the kind which Pinocchio had on that eventful morning of his life, there are but few. What was it? I will tell you, my dear little readers. On awakening, Pinocchio put his hand up to his head and there he found—Guess! He found that, during the night, his ears had grown at least ten full inches! (Carlo Collodi, author)	780 World Explorer: The U.S. & Canada; Prentice Hall 770 World Explorer: Latin America; Prentice Hall 760 *Scholastic Reading Inventory (SRI-Level 14)** 760 *Stanford Achievement Test (SAT 9-Intermediate 1)** 730 *Test of Adult Basic Education (TABE-E)** 720 Health 4; McGraw-Hill School Division
700L	670 The Girl Who Loved Wild Horses 670 Number the Stars 660 Holes 650 The Robber and Me 620 M.C. Higgins, the Great 610 Beat the Story-Drum, Pum-Pum	**BUNNICULA: A RABBIT TALE OF MYSTERY** "Of course he bites vegetables. All rabbits bite vegetables." "He bites them, Harold, but he does not eat them. That tomato was all white. What does that mean?" "It means that he paints vegetables?" I ventured. "It means he bites vegetables to make a hole in them, and then he sucks out all the juices." "But what about all the lettuce and carrots that Toby has been feeding him in his cage?" "Ah ha. What indeed!" Chester said. "Look at this!" Whereupon, he stuck his paw under the chair cushion and brought out with a flourish an assortment of strange white objects. Some of them looked like unironed handkerchiefs, and the others well, the others didn't look like anything I'd ever seen before. (Deborah and James Howe, authors) © 1979 by James Howe. Reprinted by permission of Simon & Schuster Children's Publishing Division. All rights reserved.	680 One Nation Many People, Volume One; Globe Fearon 670 Science; Addison-Wesley 660 Understanding Technology; Goodheart-Wilcox 650 *Scholastic Reading Inventory (SRI-Level 13)** 610 *Stanford Achievement Test (SAT 9-Primary 3)** 600 Community Quilt; Scholastic Inc.
600L	570 The Whipping Boy 560 Sarah, Plain and Tall 540 The Adventures of Sparrowboy 530 It's All Greek to Me 520 John Henry: An American Legend 510 Karen's Chain Letter	**A BABY SISTER FOR FRANCES** "Did you forget that I like raisins?" "No, I did not forget," said Mother, "but you finished up the raisins yesterday and I have not been out shopping yet." "Well," said Frances, "things are not very good around here anymore. No clothes to wear. No raisins for the oatmeal. I think maybe I'll run away." "Finish your breakfast," said Mother. "It is almost time for the school bus." "What time will dinner be tonight?" said Frances. "Half past six," said Mother. "Then I will have plenty of time to run away after dinner," said Frances, and she kissed her mother good-bye and went to school. After dinner that evening Frances packed her little knapsack very carefully. She put in her tiny special blanket and her alligator doll. (Russell Hoban, author) © 1964 by Russell Hoban. Reprinted by permission of HarperCollins Publishers, Inc. All rights reserved.	550 Communities; Harcourt Brace Jovanovich 540 People and Places; Silver Burdett Ginn 510 Team Spirit; Scholastic Inc. 510 *Scholastic Reading Inventory (SRI-Level 12)** 500 *Stanford Achievement Test (SAT 9-Primary 2)**
500L	490 Harold and the Purple Crayon 440 Curious George Goes to an Ice Cream Shop 420 All Tutus Should Be Pink 420 Michael Bird-Boy 410 Angel Child, Dragon Child 400 Sam the Minuteman	**THE MAGIC SCHOOL BUS INSIDE THE EARTH** But suddenly, the bus began to spin like a top. That sort of thing doesn't happen on most class trips. When the spinning finally stopped, some things had changed. We all had on new clothes. The bus had turned into a steam shovel. And there were shovels and picks for every kid in the class. "Start digging!" yelled Ms. Frizzle. And we began making a huge hole right in the middle of the field. Before long CLUNK! we hit rock. The Friz handed out jackhammers. We began to break through the hard rock. "Hey, these rocks have stripes," said a kid. Ms. Frizzle explained that each stripe was a different kind of rock. We chipped off pieces of the rocks for our class rock collection. "These rocks are called sedimentary rocks, class," said Ms. Frizzle. (Joanna Cole, author) THE MAGIC SCHOOL BUS is a registered trademark of Scholastic Inc. © 1987 by Joanna Cole. Reprinted by permission of Scholastic Inc. All rights reserved.	480 Once Upon a Hippo; Scott Foresman 470 Bears Don't Go to School; Houghton Mifflin 440 Imagine That!; Scholastic Inc. 440 Traveling Star; SRA/McGraw Hill 400 We Are All Alike; Benchmark Education
400L	370 The Drinking Gourd 370 A My Name Is Alice 370 Owl at Home 360 The Best Way to Play 330 Clifford, the Small Red Puppy 320 Miss Nelson Is Back	**FROG AND TOAD ARE FRIENDS** "That button is thin. My button was thick." Toad put the thin button in his pocket. He was very angry. He jumped up and down and screamed, "The whole world is covered with buttons, and not one of them is mine!" Toad ran home and slammed the door. There, on the floor, he saw his white, four-holed, big, round, thick button. "Oh," said Toad. "It was here all the time. What a lot of trouble I have made for Frog." Toad took all of the buttons out of his pocket. He took his sewing box down from the shelf. Toad sewed the buttons all over his jacket. The next day Toad gave his jacket to Frog. Frog thought it was beautiful. He put it on and jumped for joy. (Arnold Lobel, author) © 1970 by Arnold Lobel. Reprinted by permission of HarperCollins Publishers, Inc. All rights reserved.	390 Discover Science; Scott Foresman 390 Carousels; Houghton Mifflin 360 *Scholastic Reading Inventory (SRI-Level 11)** 350 My World; Harcourt Brace 340 *Stanford Achievement Test (SAT 9-Primary 1)** 330 Who Painted the Porcupine Purple?; Silver Burdett Ginn
300L	290 Sarah's Unicorn 270 Baseball Ballerina 270 In the Forest 260 At the Crossroads 230 The Boy Who Cried Wolf 220 Play Ball, Amelia Bedelia	**CLIFFORD'S MANNERS** Clifford loves to go visiting. When he visits his sister in the country, he always calls ahead. Clifford always arrives on time. Don't be late. Knock before you walk in. He knocks on the door before he enters. He wipes his feet first. Wipe your feet. Clifford kisses his sister. He shakes hands with her friend. Shake hands. Wash up before you eat. Clifford's sister has dinner ready. Clifford washes his hands before he eats. Clifford chews his food with his mouth closed. He never talks with his mouth full. Don't talk with your mouth full. Help clean up. Clifford helps with the clean-up. Say good-bye. Then he says thank you and good-bye to his sister and to his friend. Everyone loves Clifford's manners. (Norman Bridwell, author) © 1972 by Norman Bridwell. Reprinted by permission of Scholastic Inc. All rights reserved.	280 Too Big; Houghton Mifflin 270 *Test of Adult Basic Education (TABE-L)** 270 Parades; Houghton Mifflin 250 My Family, Your Family; Silver Burdett Ginn 240 My Pet Pup; Benchmark Education
200L			

About The Lexile Framework® for Reading

The Lexile Framework for Reading provides a common, developmental scale for matching reader ability and text difficulty. Lexile® measures enable educators, parents and students to select targeted materials that can improve reading skills and to monitor reading growth across the curriculum, in the library and at home. Lexiles are a powerful tool for linking assessment with instruction, by taking the guesswork out of selecting reading materials that meet and challenge a student's ability.

Recognized as the most widely adopted reading measure, Lexiles are part of reading and testing programs in the classroom and at the district and state levels. More than 100,000 books, 80 million articles and 60,000 Web sites have Lexile measures, and all major standardized tests can report student reading scores in Lexiles. The Lexile Framework was developed by MetaMetrics®, Inc., a privately-held educational measurement company based in Durham, N.C., after 20 years of research funded, in part, by the National Institutes of Health. For more information, call 1–888–LEXILES or visit www.Lexile.com.

NOTE: Lexiles are a measure of text difficulty. They do not address age-appropriateness, student interest or the quality of a text. Books should always be previewed by educators or parents. The Lexile map provides a graphic representation of texts and titles matched to appropriate levels of reading ability. This book list is not endorsed or recommended by the Illinois State Board of Education. This resource does not recommend a curriculum, nor necessarily endorse or recommend these particular books for study.

*The Lexile measure associated with a test describes the reading demand/readability of the test passages. It does not describe the reading ability necessary to score at the "proficient" level.

REFERENCES

For information regarding references to specific methodologies, please contact the author.