Findings on Literacy Learning Environment and Learning Record™ Validity, 2001 at Combined *Learning Record™* Sites

P.J. Hallam, Ph.D.
Center for Language and Learning

Analysis of data collected from Learning Record™ (Barr, Craig, Fisette, & Syverson, 1999; Barr & Syverson, 1999) (LR) score reports and moderations from a combination of sites where the majority of students had Learning Records provides insights into instruction, learning and assessment practices at LR schools. In school year 2000-01, 72 teachers from five Combined Learning Record Sites (CLRS) completed records for 3,089 kindergarten through 12th grade students. Several sources were used for this report: (a) LR scores, (b) norm referenced test (NRT) scores, (c) LR proficiency levels, (d) CLRS teachers' post-moderation survey responses and (e) interrater reliability data from May 2001 LR moderations.

Major findings are as follows:

1. CLRS students significantly increased their scores from 1999 to 2001.
2. The majority of CLRS students attained grade level proficiency in school year 2001, 64% in reading, 60% in writing. Almost all primary students attained grade level proficiency in school year 2001, 87% in reading, 89% in writing.
3. While the LR and the norm-referenced tests are very different measures of literacy, their scores correlated significantly in both reading and writing.
4. Disaggregation of LR scores by gender, ethnicity, and Title 1 status revealed areas of strengths (gender) and areas that could be improved.
5. CLRS teachers reported that they utilize a wide variety of indicators linked to literacy theory and best practice, especially in regard to writing processes and reading comprehension.
6. CLRS teachers reported that the LR had a positive impact on instruction and learning.
7. Teachers at CLRS scored records consistently, resulting in good interrater reliability correlations between their scores and moderation final scores (r = .76 reading; r = .75 writing).

Findings that lead to these conclusions are presented and discussed after a brief discussion of CLRS background information. In the first two findings section, CLRS students' literacy growth over time and current proficiency percentages are presented. The next three sections analyze data related to aspects of assessment validity: (a) correlation between LR and norm-referenced test scores, (b) disaggregation by gender, ethnicity and home language, and (c) consequential validity. The final data analysis section explores CLRS teachers' interrater reliability.

CLRS Background

Analysis of background data for the seven sites in this study reveals that the majority of students in this study are high school students from nonmainstream ethnic groups. Table 1 displays key background information.

Table 1
CLRS Background Information for School Year 2000-01

Site #	Site Name	Grade Level	*n LR records*	% of CLRS records at Site	% European Am. Students	% ELL Students	*n Teachers*	% Teachers new to LR
All	All	K - 12	*3089*	100%	50%	23%		
1	Primary School	K - 2	*433*	15%	64%	21%		
2	Elementary School	K - 6	*209*	6%	35%	25%		

3	BIA[1] School	K - 12	*74*	2%	0%	68%		
4	Charter School	K - 8	*157*	5%	50%	12%		
5	Continuation High School	9 - 12	*85*	3%	40%	16%		
6	BIA High School	9 - 12	*144*	5%	0%	0%		
7	Comprehen-sive High School	9 - 12	*1987*	64%	50%	25%		

The data in Table 1 reveal that the majority of students, 64%, are from Comprehensive High School (n = 1987). Three of the remaining schools have high school students (BIA[1] School, Continuation High School, and BIA High School), so high school students represent the majority of subjects in this study (73%). Primary School has the next highest percentage of total students, 15% (n = 433). Three sites have primary students (K – 2) as well (Elementary School, BIA School, and Charter School) so the next most represented age group is primary students. Upper elementary and middle school students are least represented by this sample since they are from schools with the lowest percentages of students (Elementary School 7%, BIA School 2%, and Charter School 5%). A proportionately higher percentage of students at the high school level does not impact this report greatly since the majority of analyses are separated by LR Scale Levels, where each level represents a different age group of students. LR Scale 1 is for kindergarten through third graders, Scale 2 is for fourth through eighth graders, and Scale 3 is for high school students.

Half of students in the study are from European American ethnic backgrounds and almost one-fourth (23%) are English Language Learners (ELL). Students from nonmainstream ethnic and language backgrounds are proportionally higher in this study than they are in the American student population (Hoagland, 2000). Since the LR is an assessment system that is ethnographic in nature (Miserlis, 1993), and this approach is not commonly encountered in American education, teachers and administrators who implement the LR school-wide must be highly motivated. To be at this point in implementation, the schools have committed to three years of LR staff development and infrastructure modifications. It can be conjectured that teachers and administrators who chose to use the LR are motivated by the need for assessments that counteract the negative side effects on nonmainstream students that are frequently associated with the more commonly used norm referenced testing (Linn, 2000). The high numbers of nonmainstream students in this study support that hypothesis.

Two of the schools in the study are Bureau of Indian Affair (BIA) schools. BIA policy supports using the LR because format features, such as parent conferencing and multiple listening and speaking contexts, build on cultural and home literacies (Flores & Diaz, 1991; Gifford, 1993). The LR's emphasis on documenting what students know and can do rather than on deficits defined by mainstream texts and tests also permits diverse ways of knowing. These schools are located on tribal reservations in rural areas and all students qualify for Title 1 services. Students at BIA High School speak only English, while the students at BIA School get instruction in their native language as well as in English.

Comprehensive High School and Continuation High School are from the same central California school district. They are located in the residential section of a large city that struggles with problems commonly

[1] Bureau of Indian Affairs

encountered in urban areas of America, such as poverty, transiency, and violence. Title 1 data was not available for every student at these schools, but 25% of the students are immigrants who speak English as a second language. School data also indicates that the majority of these ELLs are immigrants from Southeast Asia who came to America for economic reasons, so a substantial number of students are from low-income families.

Primary School and Elementary School are located in towns in rural Northern California. Approximately one-fourth of their students are ELL, mostly from families who recently emigrated from Mexico. At Elementary School, 43% of students are Hispanic and 19% are Native American. All of the students at Primary School, and 46% of the students at Elementary School, qualify for Title 1 funding.

While all sites had at least two years of LR staff development, not all teachers were experienced LR teachers. The number of teachers using the LR for the first time was not available. This information might be useful when analyzing LR data. The next section begins investigation of LR data by analyzing LR scores over time.

CLRS Students' Literacy Growth from 1999 to 2001

LR score data for the past three years indicates that students significantly improved their reading and writing skills each year. Students who had scores for at least two years in a row are analyzed in this section. Longitudinal data was available for students scored with LR Scale 1 and Scale 3, but not Scale 2, since all sites that used Scale 2 were first-year sites.

Table 2 displays LR reading and writing scoring data for LR Scale 1 students who had LR scores for at least two of the past three years. These findings are presented as mean scores for graduating class cohorts. For example, students in kindergarten in 2000 and 1st grade in 2001 will graduate in 2012 (12th grade), so this cohort's findings are presented as the class of 2012 (projected year of high school graduation). Table 2 displays LR scores from the two cohorts for which longitudinal data were available, 2012 and 2011.

Table 2
Comparison of CLRS **Scale 1** Students' Mean LR Scores from May 1999 to May 2001 in Reading and Writing by Graduating Class Cohorts[2]

Cohort	*n*	Reading			Writing		
		1999	**2000**	**2001**	**1999**	**2000**	**2001**
Class of 2012 K to 1st	*210*	NA	1.5	2.8**	NA	1.6	2.7**
Class of 2011 K to 1st to 2nd Grade	*189*	1.3	2.7**	3.5**	1.3	2.8**	3.7**

The data in Table 2 indicate that CLRS students' overall mean scores increased significantly each successive year. The class of 2012's mean reading score increased from 1.5 in 2000 to 2.8 in 2001. The class of 2011 moved from a mean score of 2.7 in 1999 to 2.7 in 2000, and then to 3.5 in 2001. In writing, the data indicate similar increases (from 1.6 to 2.7 for the class of 2012, and from 1.3 to 3.7 for the class of 2011). These findings

[2] LR data from other classes were unavailable.
** The difference between means successive years is significant at the .0.01 level.

indicate that Scale 1 CLRS students' reading and writing skills improved significantly from 1999 to 2001, a positive reflection on CLRS' literacy learning environment.

Table 3 displays LR reading and writing scoring data for students assessed by Scale 3 who had longitudinal scoring data. Since the only site that had longitudinal data at Scale 3 is Comprehensive High School, the data in Table 3 are all from this site. Mean scores over time are again presented by cohorts. Students who were 9th graders in 1999, 10th graders in 2000 and 11th graders in 2001 are in the class of 2002, while students who were 10th graders in 1999, 11th graders in 2000 and 12th graders in 2001 are in the class of 2001. The cohorts of 2001 and 2002 are the only ones with longitudinal data for Scale 3.

Table 3
Comparison of CLRS Students' Mean LR Scores from May 1999 to May 2001 in Reading and Writing by Graduating Class Cohorts[3]

Cohort	*n*	Reading			Writing		
		1999	**2000**	**2001**	**1999**	**2000**	**2001**
Class of 2002 9th, 10th, 11th Grade	*346*	2.9	3.0**	3.4**	2.9	3.0**	3.4**
Class of 2001 10th, 11th, 12th Grade	*335*	3.0	3.3**	3.6**	3.0	3.3**	3.5**

The data in Table 3 indicate that CLRS students' overall mean scores increased significantly each successive year. For example, in reading, the class of 2002's means scores moved from 2.9 in 1999 to 3.0 in 2000, and to 3.4 in 2001. The class of 2001 moved from a mean score of 3.0, in 1999 to 3.3 in 2000 to 3.6 in 2001. In writing, the data indicate similar increases. These findings indicate that CLRS students' literacy skills improved significantly from 1999 to 2001, a positive reflection on CLRS' high school students' literacy learning environment.

CLRS Students' Proficiency Levels

It is expected that at schools with quality learning environments, students' literacy skills progress developmentally. The Center for Language and Learning (CLL), the administrative arm of the LR, designated which scores on the LR developmental scales indicate grade level achievement, or grade level proficiency. Since LR Scale descriptors provide information about key indicators of literacy progress, stakeholders can understand what is expected of students and become aware of the rigor of the LR's cut scores by reading the scale descriptor for each proficiency level (Barr et al., 1999; Barr & Syverson, 1999). Analysis of the number of students whose scores meet or exceed grade level proficiency provides information about the efficacy of instruction and learning at LR schools.

[3] LR data from other classes were unavailable.

The following tables provide data on students' proficiency percentages at CLRS. Each LR Scale is represented, Scale 1 for grades K – 3, Scale 2 for grades 4 – 8, and Scale 3 for grades 9 – 12.

Table 4
Percentage of CLRS Students' 2001 LR **Scale 1** Reading and Writing Scores that Meet or Exceed Grade Level Proficiency

Grade	Proficiency LR Cut Score	*n*	% Proficient	
			Read	Write
All	varies	*645*	**87%**	**89%**
K	1	*179*	100%	100%
1st	2	*210*	94%	95%
2nd	3	*189*	89%	94%
3rd	4	*67*	64%	66%

The data in Table 4 indicate that overall, 87% of CLRS students met or exceeded proficiency levels in reading and 89% in writing. Analysis of proficiency percentages by grade level reveals that K, 1st and 2nd grade students' proficiency percentages, with cutoff levels of 1, 2, and 3 were high, with almost all students meeting or exceeding the cutoff score in reading and writing. Lower percentages of proficiency were found in 3rd grade (64% reading; 66% writing), where cutoff level 4 is used.

The main differences between the highest proficiency scale point, 4, and lower scale points 1 and 2 center on readers' experiences with a variety of genres and deeper levels of interpretation. The following scale descriptor for Level 4 on Scale 1 demonstrates how it represents the more challenging and accomplished stages of reading.

> Level 4, Scale 1
> A capable reader who now approaches familiar texts with confidence but still needs support with unfamiliar materials. Beginning to draw inferences from books and stories. Reads independently. Chooses to read silently. (Barr, Craig, Syverson and Fisette, 1999)

It may be conjectured that as students move up the LR scale, standards become more rigorous and difficult to attain. This finding perhaps reveals an area of instruction that calls for supporting the orchestration and consolidation of strategies for reading and writing more complex texts. The kinds of scaffolding necessary to widen and deepen comprehension of a broad array of genre could be targeted for future staff development and budgeting decisions. While these findings on proficiency attainment reveal possible areas for improvement, the majority of 645 CLRS students did attain reading and writing proficiency. This finding provides evidence of positive literacy learning environment for CLRS primary students.

Upper grade and middle school students using Scale 2 displayed a more divergent pattern of proficiency attainment, as displayed in Table 5.

Table 5
Percentage of CLRS Students' 2001 LR **Scale 2** Reading and Writing Scores that Meet or Exceed Grade Level Proficiency

Grade	Proficiency LR Cut Score	*n*	% Proficient	
			Read	Write

All	varies	*214*	**49%**	**35%**
4th	2	*51*	82%	71%
5th	3	*64*	66%	56%
6th	4	*64*	38%	20%
7th	4	*15*	13%	0%
8th	4	*20*	45%	30%

The data in Table 5 indicate that overall, 49% of CLRS students met or exceeded proficiency levels in reading and 35% in writing. Analysis of proficiency percentages by grade level reveals that 4th and 5th grade students' proficiency percentages, with cutoff levels of 2 and 3 were higher, with the majority of students meeting or exceeding the cutoff score in reading and writing. Lower percentages of proficiency were found in the upper grades where cutoff level 4 is used.

The main differences between the higher scale point of 4 and lower scale points 1 and 2 center on readers' experiences with a variety of genres, deeper levels of interpretation, and responsibility for independent learning. The following scale descriptor for Level 4 on Scale 2 demonstrates how it represents the more challenging and accomplished stages of reading.

> Level 4, Scale 2
> A self-motivated, confident, and experienced reader who may be pursuing particular interests through reading. Capable of tackling some demanding texts and can cope well with the reading of the curriculum. Reads thoughtfully and appreciates shades of meaning. Capable of locating and drawing on a variety of sources in order to research a topic independently.

The small sample size should be kept in mind when interpreting data in Table 5. One school, Charter School, provided all 35 records for 7th and 8th grade students and thus is limited as a representative sample. The pattern of difficulty in reaching proficiency at the higher LR scale scores is repeated, however, with high school students, as indicated in Table 6.

Table 6
Percentage of CLRS Students' 2001 LR **Scale 3** Reading and Writing Scores that Meet or Exceed Grade Level Proficiency

Grade	**Proficiency LR Cut Score**	***n***	**% Proficient**	
			Read	**Write**
All	varies	*2230*	**57%**	**55%**
9th	2	*638*	88%	85%
10th	3	*553*	65%	65%
11th	4	*560*	33%	30%
12th	4	*479*	42%	39%

The data in Table 6 indicate that overall, 57% of CLRS high school students met or exceeded proficiency levels in reading and 55% in writing. Analysis of proficiency percentages by grade level reveals that 9th and 10th

grade students' proficiency percentages, with cutoff levels of 2 and 3 were higher, with the majority of students meeting or exceeding the cutoff score in reading and writing. Lower percentages of proficiency were found in the higher grades where the cutoff level is 4.

The main differences between the higher scale point of 4 and lower scale points 1 and 2 also center on readers' experiences with a variety of genres, deeper levels of interpretation, and responsible application of skills and strategies to achieve personal and academic goals is required on LR Scale 3. The following scale descriptors for Level 4 on Writing Scale 2 demonstrates how it represents the more challenging and accomplished stages of high school literacy.

Level 4, Writing Scale 3
•Organizes texts to support intended effects.
•In final drafts consistently uses text conventions, e.g. reader friendly punctuation, preferred spellings, standard usage.
•Makes thoughtful word choice.
•Is in control of own composing process, from the generating of topics through the collection of data and the drafting of text to the editing for readability by specific audiences.
•Integrates information from multiple or varied sources into own paper.
•Uses criteria to evaluate both conventions and rhetorical aspects in own work.

The consistency of lower proficiency percentages when the cut scores is at Scale Level 4 supports consideration for emphasizing reading and writing strategies for more complex texts as well as opportunities to practice composing for public audiences. The kinds of scaffolding necessary to widen and deepen comprehension of a broad array of genre could be targeted for future staff development and budgeting decisions.

While these findings on proficiency attainment reveal possible areas for improvement, they also indicate that the majority of 3,069 CLRS students did attain reading (64%) and writing (60%) proficiency, as shown in Table 7.

Table 7
Summary of CLRS Students' 2001 LR Reading and Writing Scores
that Meet or Exceed Grade Level Proficiency

Scale	*n*	% Proficient	
		Read	Write
All	*3089*	**64%**	**60%**
1	*645*	87%	89%
2	*214*	49%	35%
3	*2230*	57%	55%

These findings in Table 7 provide evidence of positive literacy learning environments for CLRS students, especially at the beginning grade levels of the LR Scales, where 87% of students attained proficiency in reading and 89% of students attained proficiency in writing. The next section explores the validity of LR scores by comparing students' LR scores to norm-referenced test (NRT) scores.

Correlation between CLRS Students' LR and NRT Scores

A traditional method for establishing validity of an assessment is to compare test results with those of comparable, established tests. Assessments similar to the LR are not commonly used in American education, so comparison of LR scores to scores from tests currently administered to students is the best that can be done at this time. California CLRS used the SAT9, while BIA High School used the CTBS. Both are norm-referenced, multiple-choice tests (NRT) of academic achievement. While these assessments differ markedly in format and content, the SAT9 and CTBS scores most closely aligned to LR reading scores are the reading comprehension national percentile scores, and the NRT scores mostly closely aligned to LR writing scores are the language national percentile scores. Table 8 presents the results of comparing mean NRT national percentile scores (NP) with mean LR scores in both reading and language/writing at grade levels with enough scores from both assessments for the purpose of comparison (n > 20).

Table 8
Correlations[4] between CLRS Students' 2001 NRT Mean National Percentile (NP) Scores and Mean LR Scores in Reading and Language/Writing by Grade Level

Grade Level[5]	**Reading**		**Language/Writing**	
	n[6]	$r_{NP, LR}$	*n*	$r_{NP, LR}$
2nd	*139*	.65**	*157*	.57**
9th	*594*	.48**	*598*	.50**
10th	*522*	.58**	*521*	.64**
11th	*522*	.65**	*520*	.61**

The data in Table 8 indicate students' NRT national percentile (NP) scores in reading and LR reading scores as well as NRT language NP scores and LR writing scores are positively correlated. These correlations are strong (the closer the correlation score is to 1, the stronger the correlation) and indicate that the scores have a relationship that is far more than what could be expected from random assigning of scores (the correlations were significant at the 0.01 level for two-tailed tests).

These correlational findings provide some validating evidence for the LR, an assessment system designed to be student-centered and useful in classroom instruction and learning. The following section exp lores another aspect of validity, fairness, through analysis of LR scoring data disaggregated by demographic categories.

Scoring Data Disaggregated by Student Demographic Categories

This section examines CLRS students' LR scoring data disaggregated by the demographic characteristics of students. Analysis of scoring data from standardized, norm-referenced literacy tests over the past half century

[4] Pearson's formula was used for calculating correlations; Spearman results were not statistically different, so the more commonly used formula was chosen.
[5] Students below 2nd grade and over 11th grade are commonly excluded from norm-referenced testing. Other grade levels did not have large enough sample sizes for comparison (n < 20).
[6] The differences in totals are due to some students not having scores from both CTBS subtests and LR scores in both reading and writing.
** Correlation (Pearson, 2-tailed) is significant at the 0.01 level.

indicate that European American, boys, students from high income families and students whose home language is English tend to score higher than do non-white, female, and poor students (Linn, 2000). These findings prompt concerns about equity issues in testing (Hilliard, 1991). Sufficient demographic data was available in this study for analyzing students' gender, and ethnicity at all three LR Scale levels. Title 1 data, commonly used as an indicator of low-income background, was available only for Scales 1 and 2. Gender differences are presented in Table 9.

Table 9
CLRS Students' LR Reading and Writing Means Disaggregated by Gender and Scale

Scale	***n***		**LR Read**		**LR Write**	
	all	*n Female*	Female	Male	Female	Male
all	*3086*	*1599*	2.9	2.7*	2.9	2.6*
1	*645*	*309*	2.7	2.8	3.0	2.7*
2	*214*	*105*	2.8	2.7	2.5	2.3
3	*2230*	*1185*	3.0	2.7*	2.9	2.7*

The data in Table 9 indicate that when LR mean scores were disaggregated by gender, significant differences were indicated in some areas (p-values less than .05 are indicated by *). Scale 2 mean scores in writing and reading were not significantly different for males and females, but they were for Scales 1 and 3. Male's mean reading scores were significantly different only for Scale 3, which indicates female and male readers did not have significantly different mean scores in elementary and middle school grades. The significant differences between female and male scores for both reading and writing at Scale 3 flags this as a topic for further consideration for LR high school instructional practices. Similarly, gender differences in Scale 1 mean writing scores are noteworthy.

Disaggregation by ethnicity reveals that sample sizes were large enough to explore the difference between European American (EA), Asian American (AsA), Hispanic (H), and Native American students' LR mean scores. Table 10 displays this data.

Table 10
CLRS Students' LR Reading and Writing Means Disaggregated by Ethnicity and Scale

Scale	***n***					**LR Read**				**LR Write**			
	all	*EA*	*AsA*	*H*	*NA*	*EA*	*AsA*	*H*	*NA*	*EA*	*AsA*	*H*	*NA*
all	*3086*	*1426*	*546*	*596*	*307*	3.0	2.8*	2.5*	2.6*	3.0	2.7*	2.6*	2.6*
1	*645*	*345*	*11*	*181*	90	2.9	2.9	2.5*	2.8*	3.0	3.0	2.6*	2.7*
2	*214*	*84*	*1*	*61*	*57*	3.0	NA	2.5*	2.5*	2.7	NA	2.2*	2.0*
3	*2230*	*997*	*534*	*354*	*171*	3.1	2.8*	2.6*	2.5*	3.0	2.7*	2.6*	2.6*

The data in Table 10 indicate that when LR mean scores were disaggregated by ethnicity, European American scores were significantly higher than those of other ethnicities in most cases (p-values were less than .05). The only exception to this pattern was with Asian American students' scores at Scale 1, that is, with Asian American

* T-test of independent variables indicates the p-value is less than .05.
* T-test comparing mean scores of European Americans with this ethnicity indicates the p-value is less than 0.05.

primary aged students. Their mean LR scores were the same as those of European American students. Older Asian American, Hispanic, and Native American students' mean scores were significantly lower in both reading and writing.

The significantly lower scores of non-white students prompts investigation into strategies and structural changes that LR teachers could implement in efforts to improve these students' literacy skills. Concerns about bias based on student ethnicity can also be investigated by stakeholders, since the openness of the LR assessment system supports such scrutiny, an aspect that differs markedly from standardized, norm-referenced testing practices.

Parents, for example, not only have the opportunity to review LR portfolio contents and challenge any decisions they perceive as biased, they also are contributors to the process, so their input about equity and/or specific cultural considerations are incorporated into the LR assessment practices from the beginning. Also, since randomly selected samples of LRs are re-scored by teachers from other districts in the moderation process (see section below on interrater reliability), concerns about equity and/or cultural biases can emerge and prompt investigation at this part of the process. Analysis of comments made anonymously by teachers this year and the past three years of LR moderations reveals that scorers have not raised any concerns about teacher bias (Hallam, 1999; Hallam, 2000a; Hallam, 2001).

Title 1 (T1) data was available for students at LR Scales 1 and 2, and disaggregation by Title 1 status is displayed in Table 11.

Table 1
CLRS Students' LR Reading and Writing Means Disaggregated by Title 1 (T1) Status and Scale

Scale	***n***		**LR Read**		**LR Write**	
	all[7]	*not T1*	Not T1	T1	Not T1	T1
all	*701*	*309*	3.1	2.5*	2.9	2.5*
1	*565*	*249*	3.1	2.4*	3.0	2.6*
2	*136*	*60*	3.0	2.5*	2.5	2.2*

The data in Table 11 indicate that when LR mean scores are disaggregated by gender, significant differences are indicated for both Scale 1 and Scale 2 (p-values less than .05 are indicated by *). Students whose family income level qualified for Title 1 funding scored significantly lower than students whose parents had higher income levels. This finding indicates that investigation into learning strategies and structures that better support literacy achievement by students from lower income families. Evidence of bias on the part of teachers towards students from lower income levels could also be investigated, as discussed in the previous section on ethnic bias. The difficulties that lower income students have with literacy growth have been documented for decades, and closing this achievement gap is a complex problem that requires continued special effort and funding for solution.

The following section explores data related to another important aspect of assessment quality, consequential validity.

[7] only 701 records had Title 1 information, and these were all for LR Scale 1 and 2.
* T-test of independent variables indicates the p-value is less than .05.

LR Validity as Reported by CLRS Teachers on Post-moderation Surveys

One avenue for investigating an assessment's consequential validity is to gather data on participants' opinions of the assessment. To this end, all LR teachers were surveyed on their view of the LR's validity after site moderations in April. This survey population includes teachers at first-year LR sites that complete only 3 records per teacher, so their LR scores were not included in analysis of scoring data above, where only sites who completed records for the majority of students were included in CLRS. The surveys they completed are considered in this report in an effort to hear as many voices as possible. Table 12 displays background information on the 188 LR teachers who responded to the surveys.

Table 12
2001 Site Survey Teachers' Background Information

	Surveys	Female	White	Nat Am	Scale 1	Scale 2	Scale 3	New to LR
n	188	149	108	53	84	52	52	88
%	100%	79%	57%	28%	45%	28%	28%	47%

According to the data in Table 12, teachers who returned surveys were mostly white females who were new to the LR and taught primary grades. Native American teachers and upper grade and secondary teachers were also represented in the sample.

Table 13 deals with the levels of CLRS teachers' reported satisfaction with the LR's *consequential* validity, i.e. to what extent the assessment supported students' learning, teacher's instructional decisions, and other aspects considered to be important in the field of educational assessment (Messick, 1989).

Table 13
LR 2001 CLRS Moderation Scorers' Survey Results: Uses of LR

(Scale labels: 1 *not useful*, 2 *occasionally*, 3 *sometimes*, 4 *often* to 5 *very useful*)

n	*158*
Mean Uses	2.99
Communicate w/ other teachers	3.39
Support academic growth	3.38
Support reform efforts	3.02
Communicate with parents	2.90
Public accountability	2.69
Identify for programs	2.57

The data in Table 13 indicate that CLRS teachers ranked the LR as being "sometimes useful" for all uses on the survey. The overall mean score was 2.99, which rounds to 3, as do all the mean scores for each use (from 2.57 to 3.39). Teachers rated the LR as being slightly more useful for communicating with other teachers and promoting academic growth than it was for supporting reform efforts, communicating with parents, public accountability, and identifying students for special programs. These rounded-off mean scores of 3 indicate that CLRS teachers found the LR to be useful in a variety of ways in the classroom setting. This finding differs

markedly from teacher reports on standardized, norm-referenced testing, where reported usefulness decreases as it moves away from the classroom context (Linn, 1995; Mabry & Stake, 1994).

Another aspect of consequential validity, the extent to which an assessment specifically supports teaching and instruction, was also included on the CLRS teachers' survey. Table 14 displays the findings from this part of the survey.

Table 14
LR 2001 CLRS Moderation Scorers' Survey Results:
Extent to which LR has Positive Impact on Teaching and Learning

(Scale labels: 1 *no effect,* 2 *occasionally,* 3 *some effect,* 4 *positive* to 5 *very positive*)

n	*158*
Mean Positive Effect	3.1
Professional growth	3.55
Professional self concept	3.37
Student reflection	3.27
Oral discussion opportunities	3.21
Shared authority w/ students	3.19
Inclusion of non-mainstream texts	2.79
Parental input, involvement	2.76
Time management	2.67

CLRS teachers found the LR had an especially positive effect on their professional growth. The data in Table 14 indicates that the mean score for professional growth was 3.55, which rounds to 4, "positive effect." The overall mean score was rounds off to 3, "somewhat positive" for the other aspects which support learning and instruction: professional self-concept, student reflection, oral discussion opportunities, share authority with students, inclusion of non-mainstream texts, parental input and involvement, and time management. Overall, these data indicate that CLRS teachers found the LR supportive of many aspects considered important to quality teaching.

To gauge CLRS teachers' opinions on *construct* validity, the extent to which an assessment actually assesses important aspects of the domain being assessed, teachers were asked to indicate how often they included evidence in their students' LRs from 39 activities linked to literacy learning theory and best practice (Hallam, 2000b). Survey responses asked teachers to respond to prompts on a scale of 1 to 5, with "1" indicating "rarely," "3" indicating "sometimes" and "5" indicating "very frequently." Survey results indicate that LR teachers who completed the surveys reported using 15 of the indicators "frequently" (mean scores round off to 4) and the remaining 24 indicators "sometimes" (mean scores round off to 3) (Hallam, 2000a). Since effective learning relies on a variety of indicators, it is expected that mean scores would not often round off to 5, "very frequently," and since none of the indicators were reported as being used rarely or even infrequently, these findings are positive.

For the purposes of this report, the top literacy indicators reported by CLRS teachers are displayed in Table 15 to provide insights into the breadth of literacy learning and assessment activities at CLRS.

Table 15
LR 2001 CLRS Moderation Scorers' Survey Results:
Teachers' Report of Top Thirteen of 39 Literacy Indicators Used in Learning Records

(Scale labels: 1 *rarely,* 2 *occasionally,* 3 *sometimes,* 4 *often* to 5 *very frequently*)

n	*138*
Mean Literacy Indicators	3.28
Reader response	3.76
Context cues	3.72
Sustained reading and writing	3.72
Pleasure and interest	3.71
Independent reading	3.66
Literal comprehension	3.64
*Print concepts	3.61
Independent writing	3.61
*Basic sight words	3.6
Connect to personal	3.59
Read/write connection	3.57
Group discussions	3.56
"Standard" English	3.56
*Sound-symbol patterns	3.56

The data in Table 15 indicate CLRS teachers felt they used the following key strategies "often" (mean scores round off to 4): (a) reader response to texts literature responses, (b) context cues, (c) sustained reading and writing, (d) pleasure and interest in reading and writing, (e) independent reading, (f) literal comprehension, (g) independent writing, (h) connecting text to personal experience, (i) connecting reading and writing, (j) group discussions, and (j) accurate usage of "standard" English. Primary LR teachers and teachers of students struggling with literacy growth reported that they also used following strategies often: (a) print concepts, (b) basic sight words, and (c) sound-symbol patterns. The overall mean score of 3.28 on all 39 indicators also indicates that CLRS teachers report a high level of construct validity for the LR.

Positive findings on the validity of CLRS teachers' evidence in their LRs are moot, however, if inconsistencies in teacher judgment make their scoring decision unreliable. The extent to which stakeholders such as students, parents, other teachers, and the public at large value teachers' collections of assessment data is influenced by the consistency of teachers' judgments. For this reason, teachers' LR scoring decisions are calibrated and moderated in the spring of each year. The following section provides an analysis of CLRS teachers' interrater reliability.

CLRS Teacher Consistency in Judging Student Achievement

Interrater reliability data were collected from various steps of the moderation process. The first scores collected were from originating teachers (OT), who scored the records after writing end of the year summaries. Then randomly selected records received scores at Site Moderations (SS) and Intersite Moderations (IS). If originating

teacher, site or intersite scores matched, then the matched score became the final score (FS). If the three scores differed, then these records received a score from an expert scorer (EX). If the expert score matched the originating teacher, the site or intersite score, it became the final score (FS). If the expert score did not match the originating teacher, site or the intersite score, the record was labeled "Not Enough Evidence" (NEE) and did not receive a score. Analysis of interrater scoring data, including NEEs, from the 3089 complete records provides information about the strength of correlations and scoring patterns between the various scorers.

Analysis of interrater scoring data from LR records provides information about the strength of correlations and scoring patterns between the various scorers.

Originating Teacher Interrater Correlations across Contexts

Data from the 2001 Moderations, presented below, reveal that randomly sampled originating teachers' scores in reading and writing are indeed correlated with other scores from the moderation process. In statistics, the closer the correlations are to 1, the stronger the relationship between two sets of scores is considered to be. In Table 16, Pearson[8] correlation coefficients for originating teachers and other scorers' reading scores are presented in the first row and writing correlations are presented in the second row. A total of 712 records were scored at moderations, which means that 23% of LR records were selected for moderation scoring.

Table 16
Pearson Correlation Coefficients between Originating Teachers and Site, Intersite, Expert Scorers and Final Reading and Writing Scores

	Site Score	**Inter-Site Score**	**Expert Score**	**Final Score**
Originating Teachers Reading	**.77** *n=692*	**.66** *n=658*	**.66** *n=198*	**.76** *n=665*
Originating Teachers Writing	**.76** *n=701*	**.58** *n=693*	**.66** *n=201*	**.75** *n=692*

The data in Table 16 indicate that originating teachers' (OT) scores have a positive correlation with final scores (FS) in reading (r = .76) and writing (r = .75). In the field of educational assessment, correlation coefficients in the range of .80 to .99 indicate strong positive correlations, coefficients in the .50 to .79 indicate positive correlations, coefficients in the .20 to .49 indicate some correlation, and coefficients between 0 and .19 indicate that there is no positive correlation (Harris & Martinovich-Barhite, 1998). Coefficients for reading and writing in Table 16, .76 and .75, indicate that LR teachers' scoring decisions are positively correlated with scorers who do not know the students. These coefficients indicate that scorers from both inside and outside the classroom interpreted the scale descriptors consistently. These coefficients indicate that, overall, LR teachers had reasonably high levels of interrater reliability in school year 2000 - 2001.

OT scores are also positively correlated with scorers throughout the moderation process in both reading and writing: OT and SS (reading r = .77 and writing r = .76), OT and IS (reading r = .66 and writing r = .58), and OT and EX (reading r = .66 and writing r = .66). These correlations indicate that teachers' scores from the same site (SS) were

most consistent with OT scores. This is not surprising since teachers from the same site have more contextual knowledge of the students' academic opportunities. Correlations from scorers outside the school, IS and EX, were less correlated. This is not surprising, since teachers at the same site are more likely to share background information about the learning environment, which promotes more consistency in interpreting the LR scales. For this reason, the reasonably strong OT and FS correlations (.76 reading, .75 writing) indicate LR teachers interpreted LR scales at a high level of consistency. The following section investigates OT and FS scoring decisions further.

Relationship between Originating Teacher and Final Scores

Cross tabulations of interrater agreement data provide insights into scoring patterns. Cross tabulation data for OT and FS reading scores are displayed in Table 17, and writing scores are displayed in Table 18. Appendix 1, below, provides a detailed explanation of how to interpret cross tabulation data using Table 16 data.

Table 17[9]

Cross Tabulation of Originating Teachers' Reading Scores with Final Scores

	Final Scores						
Originating Teacher	**1**	**2**	**3**	**4**	**5**	*Total*	Mismatch
1	87	24	2			113	OT lower FS
2	23	212	13	1		249	46
3	1	**87**	120	3		211	
4	2	11	31	31	3	78	OT higher FS
5		1	4	7	1	13	167*
Total	113	335	170	42	4	664	

r = .76
7% of OT scores were higher, 25% of OT scores are lower than FS.
* McNemar p-value < .01

The data in Table 17 indicate that OT had some distinct mismatch patterns. One notable mismatch pattern is that the majority of OT mismatches are from scoring higher than FS (n lower = 46, n higher = 167). At score level 1, this mismatch pattern was not found. In fact, OT scores were lower than FS, since 24 records that OT scored as 1 had FS of 2. At OT score level 2, there were a fair number of OT scores both lower (n = 23) and higher (n = 13 + 1 = 14) than final scores. At score level 3, a large number (n = 87) of OT records got FS scores of 2, indicating that the majority of OT mismatches were from OT scoring higher than FS. At score level 4, the majority of mismatched scores (n = 44) were also from OT scoring higher than FS. The large number of OT mismatches due to scoring higher than FS creates a symmetrical imbalance that is statistically significant (McNemar p-value < .01).

OT scores at levels 4 and 5 were mismatched far more often than there were matched. This indicates that OT and FS matches were particularly difficult at the higher end of the scoring scale. Score level 4 had the most non-adjacent mismatches. These findings merit further investigation by LR staff at the site level. Finally, there were very few FS scores at level 5 (n = 4 for reading and writing), and no FS of 6 for Writing Scale 1. LR teachers and support staff may wish to investigate what they can do to help students reach the highest levels of literacy accomplishment.

---

[8] Spearman rank-order correlations yielded nearly identical results. Pearson's correlation was chosen because it is more widely known.

Mismatch patterns are similar for writing, as can be seen in Table 18.

Table 18
Cross Tabulation of Originating Teachers' Writing Scores with Final Scores

	Final Scores						
Originating Teacher	**1**	**2**	**3**	**4**	**5**	*Total*	Mismatch
1	94	38	2			134	OT lower FS
2	33	181	38			252	85
3	1	**71**	137	7		216	
4	1	13	24	31	3	72	OT higher FS
5			6	8	3	17	
6			1			1	158*
Total	129	303	208	46	6	692	

r = .75
12% of OT scores were higher, 23% of OT scores are lower than FS.
* McNemar p-value < .01

Cross tabulation of OT scores for both reading and writing in Tables 17 and 18 reveal the following three patterns:

1. Positive correlations between OT and FS in reading (r = .76) and writing (r = .75) are indicated in the cross tabulation tables by the large number of exact matches (scores in the diagonal shaded cells). Correlations between OT and FS in reading and writing are also indicated since the majority of mismatches are adjacent to the shaded diagonal cells.

2. The majority of OT scorers' mismatches were due to OT scoring higher than other scorers. Both cross tabulation tables display strong diagonal asymmetry between OT and FS because the majority of mismatched scores are below the diagonal, shaded cells (see Appendix 1 for detailed explanation). For example, narrative in the rectangle at the bottom of Table 3 indicates that only 12% of OT and FS mismatches were due to OT scoring lower, while 23% of mismatches were due to OT scoring higher than FS. Results from the McNemar test of symmetry reveals that the counts of mismatches from OT scoring higher are significantly different from the counts of OT scoring lower (p-value < .01).

3. OT's mismatches with FS were proportionately large at the higher end of the scoring scales, at levels 3, 4, and 5. This pattern is revealed through the larger number of mismatches at these higher levels. Since scoring level 4 had more non-adjacent mismatches than the other scoring levels, these findings indicate that scoring decisions may be particularly difficult at score level 4.

The mismatch scoring patterns between OT and FS raise a red flag in regard to scoring decisions of LR classroom teachers, especially since OT scores were consistently higher than those of other scorers. Analysis of comments made by site and intersite scorers on moderation scoring forms indicate that missing evidence precluded placement at higher levels (Hallam, 2000b). It may be conjectured from this evidence that classroom teachers may have made accurate placement decisions, but they did not have adequate evidence to back up their decisions. Analysis of teacher comments on LR scoring forms in previous years revealed that the type of evidence reported

[9] A detailed explanation of cross tabulation Table 2 is in Appendix 1.

missing most often centered on critical thinking and integrating references in reading and writing, aspects at the higher levels of LR Scales (Hallam, 2000a; Hallam, 2000b). This finding also provides an insight into the low number of scores at the higher end of LR Scales. These findings may prompt more attention to techniques for documenting these types of evidence in future staff development at CLRS schools. Cross tabulation data for each site was provided so that CLRS staff could discuss and investigate further factors that may have contributed to these scoring patterns (Hallam, 2001).

While further investigation is merited, overall, findings on interrater reliability indicate LR teachers' scoring decisions merit respect because they indicate reasonably high levels of consistency.

Conclusion

Analysis of scoring and moderations data from the Learning Record and from NRT scores indicate some positive aspects about the learning environment at CLRS, the validity of the LR, and the interrater reliability of CLRS teachers. Findings also indicate a few areas that could be improved.

Positive Indicators about LR Learning Environments

Several findings in this report indicate that the LR students are in a positive literacy environments. Significant increases were noted in LR scores each successive year (p-values < .01). This finding reflects positively on the literacy learning environments at the sites with longitudinal evidence, Scale 1 schools and Comprehensive High School.

The majority of CLRS students attained proficiency, 57% in reading and 55% in writing, which also is a positive indicator of LR schools' learning environments. The high percentage of primary students reaching proficiency in reading (87%) and in writing (89%) is especially noteworthy.

Findings on interrater reliability indicate LR that teachers' scoring decisions merit respect because they indicate reasonably high levels of consistency with moderation final scores (r = .76 in reading and .75 in writing). This is a positive finding for LR students and parents because it increases their trust in LR teachers' scoring decisions.

Results from post-site moderation surveys indicate that teachers reported using a wide variety of literacy indicators linked to literacy theory and best practice. Since literacy learning requires utilization of a variety of techniques, the overall mean score of 3.1 "uses sometimes" is a positive finding on the literacy practices in LR schools. The "top thirteen" literacy indicators CLRS teachers reported using most often provide a picture of literacy instruction that is balanced, learning-centered, with an emphasis on scaffolding important steps in reading and writing processes. CLRS teachers reported that they paid special attention to pleasure and interest in reading and writing, an aspect whose importance is consistently emphasized in literacy research (Mathewson, 1994).

LR Validity

LR teachers' reported usage of a variety of key literacy indicators is a positive reflection on LR construct validity as well. If teachers had reported low or no usage of the 39 key indicators, evidence indicating low construct validity would have been collected. In validity research, the larger the number of findings which do not indicate negative findings, the more convincing the argument is that the assessment is valid (Cronbach, 1984; Salvia & Ysseldyke, 1991).

Analysis of LR scoring data through comparison with students' NRT scores also provides evidence of validity, in that the tests were not found to be negatively correlated or uncorrelated. Correlational analysis revealed that NRT scores are positively correlated with LR students' scores at grade levels 2 (.57), 9 (.50), 10 (.64), and 11 (.61), the grade levels where data was available for such comparison.

Survey findings indicate that the LR has moderately high levels of consequential validity. Teachers indicated, for example, that the LR is especially useful for communicating with other teachers and in supporting students' academic growth. They also indicated it was somewhat useful in supporting reform efforts, promoting student self-reflections, shared authority in literacy interpretations and oral discussion. This data provides evidence of a positive literacy-learning environment as well, since literacy research emphasizes the importance of these strategies (Ruddell, Ruddell, & Singer, 1994).

The value of these findings is most apparent when viewed from the perspective of current concerns about the consequential validity of norm-referenced, multiple-choice tests, such as the NRT. Formidable amounts of data indicate that many teachers, parents, and students are disturbed with the impact of these tests on curriculum and learning (Kohn, 2000) (Linn, 1995). For example, studies over the years indicate that rote learning is emphasized at the expense of critical thinking and thoughtful reflection when NRT scores are emphasized at schools (Shepard, 1990). According to recent reports in the media on teachers' dissatisfaction with the CTBS, the SAT9, and other NRTs, survey results from teachers who use these tests exclusively are not likely to be nearly as positive as they are with teachers who use the LR (Neil, 2000).

Indicators of Need for Future Staff Development

Findings from this report play an important part in the feedback loop for the LR assessment system for two reasons (Hallam, 2000). Positive findings highlight areas of strength, while other findings help pinpoint areas of need. Findings which revealed areas in instructional and assessment practices that could be improved are useful for making future staff development and funding decisions. The following areas were revealed as areas with potential for improvement:

1. CLRS students' proficiency levels revealed that the majority of LR students did not reach proficiency levels when the cutoff scores were at LR Scale levels of 4 and 5. Since the higher levels of LR Scales require deeper levels of comprehension and control over a wide variety of literacy genres, it may be that these areas could be targeted.

2. Interrater reliability was lower at LR Scale levels of 3, 4, and 5, and moderation comments indicate that learning how to, and finding time for, documenting the deeper levels of comprehension and control over a wide variety of literacy genres is an area of need. To promote increased trust and confidence among stakeholders CLRS teachers need continued support in learning how to document evidence that supports their placements.

3. Investigation of assessment equity issues at CLRS through disaggregation of LR scoring data by pertinent demographic categories revealed that while there are a few areas where equity in scoring data was indicated, notably in the area of gender for younger students and for Asian Americans, continued efforts need to be targeted on low-income students, non-white students, and high school boys.

Overall, analysis of data from the 2000 LR scoring and moderation data reflected positively on many aspects of the CLRS learning environment, highlighted some areas that could be improved, and supported the validity of the LR.

References

Barr, M., Craig, D., Fisette, D., & Syverson, M. (1999). *Assessing literacy with the Learning Record: A handbook for teachers, grades K - 6*. (2nd ed.). Portsmouth, NH: Heinemann.

Barr, M., & Syverson, M. (1999). *Assessing literacy with the Learning Record: A handbook for teachers, grades 6 - 12*. (2nd ed.). Portsmouth, NH: Heinemann.

Cronbach, S. (1984). *The essentials of psychological testing*. NY, NY.

Flores, B. P., & Diaz, E. (1991). Transforming deficit myths about learning, langauge and culture. *Alnguage Arts, 68*, 369-378.

Gifford, B. (Ed.). (1993). *Policy perspectives on educational testing*. Boston: Kluwer Academic.

Hallam, P. J. (1999). *Preliminary report on Learning Record 1999 moderations: Reliability and validity findings* . San Diego, CA: Center for Language and Learning.

Hallam, P. J. (2000a). *Preliminary report on Learning Record 2000 moderations: Reliability and validity findings* . San Diego, CA: Center for Language and Learning.

Hallam, P. J. (2000b). *Reliability and validity of teacher-based reading assessment: Application of "Quality Assurance for Teacher-based Assessment" (QATA) to California Learning Record moderations.* Unpublished Doctoral dissertation, University of California, Berkeley, CA.

Hallam, P. J. (2001). *Report on Learning Record 2001 moderations: Reliability and validity findings* . San Diego, CA: Center for Language and Learning.

Harris, D., & Martinovich-Barhite, D. (1998). *Practical and technical issues in the use, development, scoring and equating of performance assessments.* Paper presented at the National Council on Measurement in Education.

Hilliard, A. G. (Ed.). (1991). *Testing African American Students*. Chicago, IL: Third World Press.

Hoagland, D. (2000). *Educators favor teaching early on*, [webpage]. Fresno Bee. Available: http://www.fresnobee.com/localnews/story/0%2C1724%2C197618%2C00.html9-23-2000].

Kohn, A. (2000). *The case against standardized testing*. Portsmouth, NH: Heinemann.

Linn, R. (2000). Assessments and accountability. *Educational Researcher, 29*(March), 4-16.

Linn, R. L. (1995). *Assessment-based reform: Challenges to educational measurement* : Center for Research on Evaluation, Standards, and Student Testing at Boulder.

Mabry, L., & Stake, R. (1994). Aligning measurement with education. *Educational Researcher, 23*(2), 33-4.

Mathewson, G. C. (1994). Model of attitude influence upon reading and learning to read. In R. B. Ruddell, M. R. Ruddell, & H. Singer (Eds.), *Theoretical models and processes of reading* (4th Edition ed., pp. 1131-1161). Newark, DE: International Reading Association.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (third ed., pp. 13-103). New York: Macmillan.

Miserlis, S. (1993). *The classroom as an anthropological dig: Using the California Learning Record (CLR) as a framework for assessment and instruction.* Paper presented at the The Claremont Reading Conference 57th Yearbook, Claremont, CA.

Neil, M. (2000). *FairTest: National center for fair and open testing*, [web page]. Available: http://www.fairtest.org/.

Ruddell, R. B., Ruddell, M. R., & Singer, H. (Eds.). (1994). *Theoretical models and processes of reading* (4th ed.). Newark, DE: International Reading Association.

Salvia, J., & Ysseldyke, J. E. (1991). *Assessment: 5th edition*. Boston, MA: Houghton Mifflin Company.

Shepard, L. (1990). Inflated test score gains: Is the problem old norms or teaching to the test? *Educational Measurement: Issues and Practices, 9*, 15-22.

Appendix 1
Detailed Interpretation of Cross Tabulation of
Originating Teachers' Reading Scores with Final Scores

	Final Scores						
Originating Teacher	**1**	**2**	**3**	**4**	**5**	*Total*	Mismatch
1	87	24	2			113	OT lower
2	23	212	13	1		249	46
3	1	**87**	120	3		211	
4	2	11	31	31	3	78	OT higher
5		1	4	7	1	13	167*
Total	113	335	170	42	4	664	
r = .76 7% of OT scores were higher, 25% of OT scores are lower than FS. * McNemar p-value < .01							

This detailed explanation starts on the top left side and moves across to the right side of the table. When the end is reached, the next level down is explained, and continues until the bottom is reached.

Table 3 shows that 87 OT reading scores of "1," were matched by FS of "1" (matching scores are shaded). Next to the 87 in the "OT 1" score row is "24." This means that in 24 cases, OTs gave a score of "1," but the FS was a "2." In these 24 cases, OTs scored records LOWER than other scorers. Likewise, 2 OT scored records as a 1, but their FS were 3. Since there were no FS of "4" or "5," the remaining spaces in the OT score of 1 row are blank. Next to the blank score of 5, is the number 113 under the word "Total", which indicates that a total of 113 records received a final score of "1."

The next row, "Originating Teacher 2," shows that 23 records scored as "2" by OT, ended up with a FS of "1." This means that in 23 cases, the OT score was HIGHER than the majority of other scorers. The next space on the "OT 2" row is shaded, and shows that 212 of OT "2" scores were matched exactly by FS. Thirteen records, as indicated by the "13" in the next space, were scored as "3" and one record was scored as a "4." Fourteen records, therefore, received lower scores from OT than they did from FS. The "Total" of records OT scored as "2" is 249.

The "OT 3" row shows that there was one case where OTs scored a record as a 3, but the FS was a 1. There were 87 cases where OTs gave records a score of "3," but the FS was a "2." So in 87 cases, OT scores of "3" were higher than the FS of "2." The shaded space shows that 120 of OT scores were matched exactly by FS. The "3" in the "FS 4" column indicates that three times an OT score of "3" was lower than the FS of "4."

The next row, "OT 4," shows how often OT scores of "4" were matched by FS. One record was assigned a FS of "1," eleven records were assigned FS of "2," and 31 received a "3." All of these mismatches were from the OT scoring lower than the FS. The shaded space reveals that 31 records matched exactly, while the last score column indicates that 3 of the records scored as a 4 ended up with a FS of "5."

The final scoring row, "OT 5," shows that one record scored as a "5" by OT received a FS of "2." Eleven other records were also scored lower than the OT, with four receiving a FS of "3" and seven receiving the adjacent score of "4." These 7 lower scores mean that 47% of the time, OT scores at the 5 level were not matched by FS.

The last column, "Total" indicates the number of scores given by FS at each score level. The "664" where the vertical "Total" row meets the horizontal "Total" column indicates the total number of records scored for reading.

The last column, "mismatch," provides insights about OT scoring higher or lower than FS. Table 2 shows that 2001 moderation scorers had significantly more mismatches from scoring higher than FS than they did from scoring lower than FS.

Under the cross tabulations are additional pieces of information. First, the correlation coefficient of the two scores being compared is provided. For Table 5, the OT-FS reading correlation is .76 (r = .76). Since most of the scores fall in the shaded areas, or are adjacent, the information displayed in cross tab Table 2 supports this statistic.

The percentage of mismatches due to OTs scoring higher and lower than the other scorers is also given. In Table 2, 7% of OT scores were higher than FS, while 25% were lower. The McNemar test of significance produces

a p-value lower than .01, which indicates that the percentage of OT mismatches due to scoring higher is disproportionately larger than the number of mismatches due to scoring lower.