

Technical Issues in RTI: CBM and Standard Error

Presented at the
Curriculum-Based Measurement Research and
Innovations Conference in St. Paul, MN on May 5, 2006

Theodore J. Christ, Ph.D.

UNIVERSITY OF MINNESOTA



School of Educational Psychology
College of Education and Human Development

tchrist@umn.edu

Contact Information

Please contact me by
e-mail (tchrist@umn.edu) or
phone (612.624.7068)
if you are interested in
collaboration or
local assistance

Introduction

- For what purpose do we use CBM?
 - progress monitor? (self-referenced, or criterion)
 - screening? (norm-referenced)
 - benchmarking? (criterion-referenced)
 - eligibility?
- How many data points?
- How do you report data?

- For what purpose do we use CBM
 - high stakes decisions
 - low stakes decisions

Introduction

- **Assessment**
 - refers to the procedures and outcomes that are used to compile information that describes phenomena. Assessment data may be qualitative or quantitative.
- **Measurement**
 - refers to the procedures and outcomes that are used to quantify a phenomenon. Well established measurement procedures and metrics have the potential to communicate information with greater precision and efficiency
- **Evaluation**
 - is the process of consuming data along with the results of interpretation that guide educational decisions

Introduction

- **psychometrics** is the study of measurement procedures and outcomes ... the science that guides the development, selection, and compilation of procedures and instrumentation to quantify phenomena
 - (a) tasks and/or stimuli which invoke behavior for measurement
 - (b) procedures that translate responses/behavior into numerical quantities (frequency, proportion, rate, duration, and latency);
 - (c) transformation and contextualization of those quantities onto common scales/distributions;
 - (d) the establishment of procedures to facilitate interpretation; and
 - (e) evidence to justify each proposed interpretation and use of outcomes

“CBM is a valid and reliable ...”

...for what purpose?

Introduction

- Cone (1981), measures of behavior should document
 - (a) its occurrence,
 - (b) its repeated occurrence,
 - (c) its occurrence in different settings,
 - (d) its measurability by different assessment methods, and
 - (e) its relation to other responses”
- Notice: no recommendation for **consistency** – which is typical of a behavioral assessment paradigm where **sensitivity** is emphasized
 - behavioral assessment has ecological focus
 - psychometric assessment has internal trait focus

Introduction

behavioral assessment has ecological focus

psychometric assessment has internal trait focus

Introduction

*Within a behavioral assessment paradigm,
there is only a minimal expectation that
measurement outcomes are consistent
across measurement occasions ...*

if consistency is expected at all

Introduction

- **accuracy** “how faithfully a measure represents objective topographic features of a behavior of interest” (Cone, 1981, p. 59)
- **reliability** “the consistency with which repeated observations of the same phenomena yield equivalent information” (Cone, 1981, p.59)
- **reliability of effect** – consistency of response to treatment (and withdrawal) across replications (within case) (Baer, Wolf, & Risely, 1986)

Introduction

- True Score Model: $X = T + E$
 - (a) the true scores and error scores are uncorrelated [$\rho(TX) = 0$];
 - (b) the error scores across parallel tests are uncorrelated [$\rho(EE') = 0$]; and
 - (c) the mean of error scores is zero whenever there are a sufficient number of responses [$M(E) = 0$]
- (Hambleton & Jones, 1993).

Introduction

Should we establish an expectation of consistency (reliability) in measurement?

What is the potential impact on RTI?

- Does CBM have measurement error?

Yes...

and we should consider SEM & CIs

CBM-R: Standard error of measurement (SEM)

Theodore J. Christ, Ph.D.
University of Minnesota
Benjamin Silbergitt
St. Croix River Education District

- Think of SEM as a proxy for range...

Standard Error of Measurement

- Standard Error of Measurement (SEM)
 - amount an observed score is likely to fluctuate
 - Inversely proportional to the reliability of a test

$$SEM = SD_x \sqrt{1 - r_{xx}}$$

Confidence Interval

- Confidence Interval
 - band or range around the observed score in which the true score is most likely to fall
 - $CI \sim N(0, SEM)$
 - $CI(68\%) = 1.00 * SEM +/- x$
 - $CI(85\%) = 1.44 * SEM +/- x$
 - $CI(90\%) = 1.65 * SEM +/- x$
 - $CI(95\%) = 1.96 * SEM +/- x$
 - $CI(99\%) = 2.58 * SEM +/- x$

Purpose

- Can we develop generic estimates of SEM for CBM-R?

Method: Participants & Setting

- Data collection in the Fall, Winter, and Spring from 1996 to 2004
- 1st through 5th Grade (N = 8,200)
 - five rural and suburban school districts in the upper mid-west
 - Sex
 - 52% males
 - 48% females
 - Ethnicity:
 - 3% Native American,
 - 1% Asian,
 - 1% Hispanic,
 - 1% Black,
 - 94% White
 - Free and Reduced Lunch
 - A: 3500 (6%),
 - B: 900 (19%),
 - C: 1100 (16%),
 - D: 1700 (9%),
 - E: 1000 (9%)

Results: Descriptives (collapsed across year)

	N	M	Mdn	SD	Skew	Kurtosis
First Grade						
Winter	4196	32	23	27	1.85	4.09
Spring	4356	60	55	33	.78	.55
Second Grade						
Fall	4325	50	46	32	.86	.61
Winter	4339	77	75	37 ^b	.44	.08
Spring	4544	95	92	38	.37	.18
Third Grade						
Fall	4288	74	71	37	.57	.14
Winter	4340	95	92	41 ^b	.31	.00
Spring	4691	108	105	43	.23	-.09
Fourth Grade						
Fall	4455	93	91	39	.42	.17
Winter	4470	113	111	41 ^b	.23	.21
Spring	4877	125	123	43	.23	.15
Fifth Grade						
Fall	4635	112	110	40	.26	-.07
Winter	4655	130	130	42 ^b	.12	.07
Spring	5036	139	137	45	.16	.05

Grade	Academic Year								Range		
	96-97	97-98	98-99	99-00	00-01	01-02	02-03	03-04	Mdn	Min	Max
First											
Winter	24	25	29	26	27	27	27	30	27	24	30
Spring	29	31	34	34	35	32	30	33	33	29	35
Second											
Fall	30	30	34	31	30	34	32	32	32	30	34
Winter	33	37	41	36	35	37	36	34	36	33	41
Spring	38	33	40	38	37	37	37	34	37	33	40
Third											
Fall	39	32	35	37	36	38	38	39	38	32	39
Winter	42	34	38	41	39	41	40	40	40	34	42
Spring	36	38	39	43	41	42	42	43	42	36	43
Fourth											
Fall	35	33	35	39	39	39	39	39	39	33	39
Winter	40	36	36	40	39	41	40	40	40	36	41
Spring	40	38	37	43	43	43	42	43	43	37	43
Fifth											
Fall	39	39	33	37	41	43	41	40	40	33	43
Winter	37	43	40	38	41	44	43	41	41	37	44
Spring	43	41	37	40	43	45	46	44	43	37	46

SD = 30 - 40

Recall...

$$SEM = SD_x \sqrt{1 - r_{xx}}$$

notice, there are two variables in the equation that will determine SEM

Grade	SD ^b	Expected Levels Reliability								
		Lower r _{xx}			Middle r _{xx}			Higher r _{xx}		
		.89	.90	.91	.92	.93	.94	.95	.96	.97
Homogenous										
1 st	24	8	8	7	7	6	6	5	5	4
2 nd	30	10	9	9	8	8	7	7	6	5
3 rd	32	11	10	10	9	8	8	7	6	6
4 th	33	11	10	10	9	9	8	7	7	6
5 th	33	11	10	10	9	9	8	7	7	6
Typical										
1 st	30	10	9	9	8	8	7	7	6	5
2 nd	34	11	11	10	10	9	8	8	7	6
3 rd	39	13	12	12	11	10	10	9	8	7
4 th	39	13	12	12	11	10	10	9	8	7
5 th	41	14	13	12	12	11	10	9	8	7
Heterogenous										
1 st	35	12	11	11	10	9	9	8	7	6
2 nd	41	14	13	12	12	11	10	9	8	7
3 rd	43	14	14	13	12	11	11	10	9	7
4 th	43	14	14	13	12	11	11	10	9	7
5 th	46	15	15	14	13	12	11	10	9	8

Notice: SEMs might typically be within the range of 8 – 12 WRCM

- CBM is sensitive ... to what?
 - the administrator (Derr-Minneci, 1990; Derr-Minneci & Shapiro, 1992; Derr & Shapiro, 1987)
 - setting/locations (Derr-Minneci, 1990),
 - presentation of directions (Colon & Kranzler, 2006), and
 - passage difficulty (Christ, 2003; Fuchs & Deno, 1992; Hintze, 1995; Hintze & Christ, 2004; Hintze et al., 1994)
- and instructional effects...

Conclusion

- there is error associated with CBM-R – just like any other measurement
- *SEM* is likely to approximate 8 – 12 WRCM
- Implication: report either a range of values from repeated assessment or an estimate (i.e., *SEM*)

Conclusion

- Assessment outcomes should always be reported in a manner that communicates the potential for error (AERA, APA, & NCME, 1999)

Yes, this applies to CBM

Conclusions

- Develop *SEMs* from local norms
 - simple excel spreadsheet function
 - see Christ, Davie, & Berman (Communiqué, in submission)

$$SEM = SD_x \sqrt{1 - r_{xx}}$$

Is there CBM-R error associated with growth estimates?

Theodore J. Christ, Ph.D.
University of Minnesota

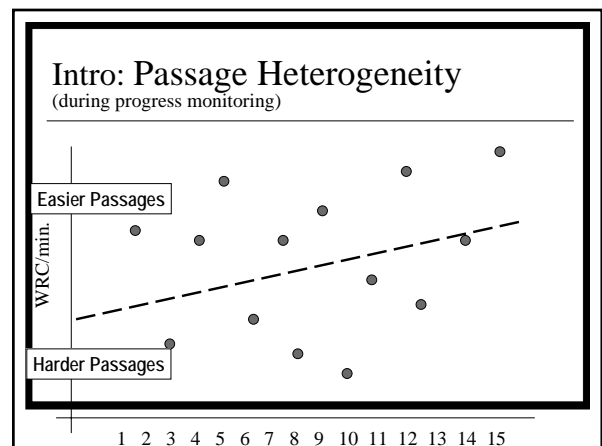
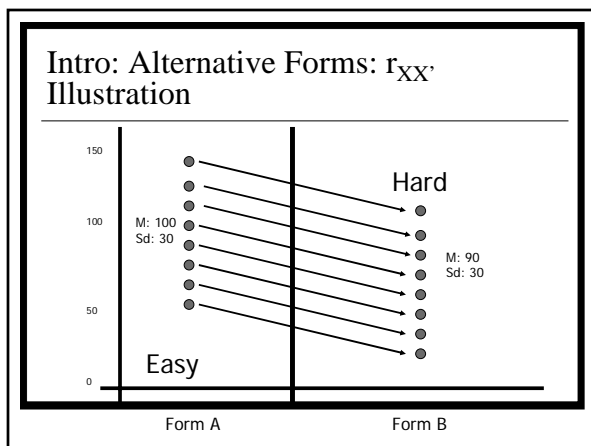
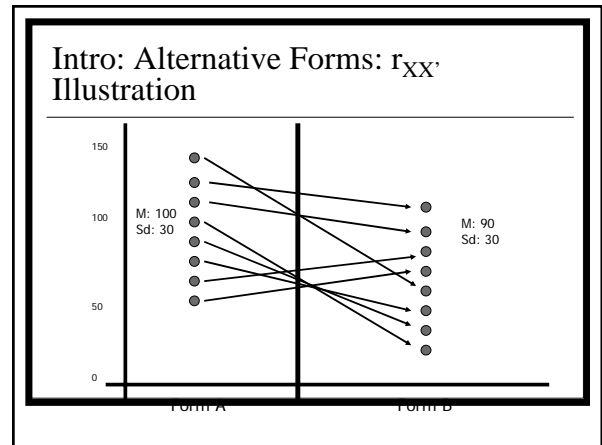
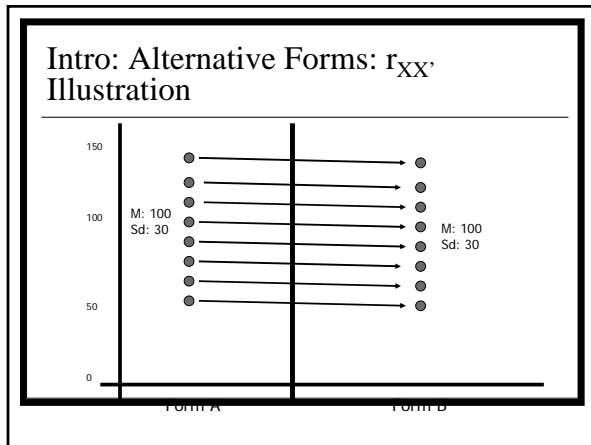
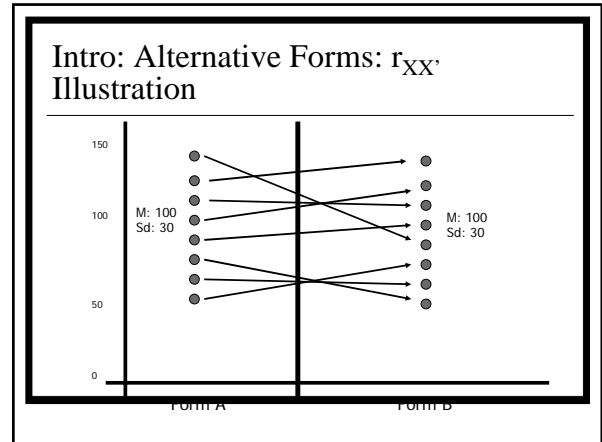
Intro: Progress Monitoring

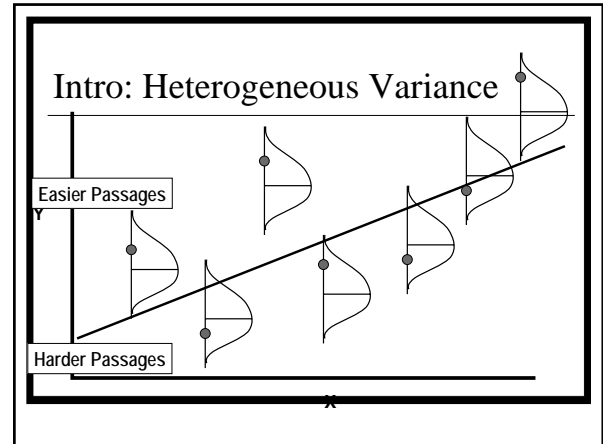
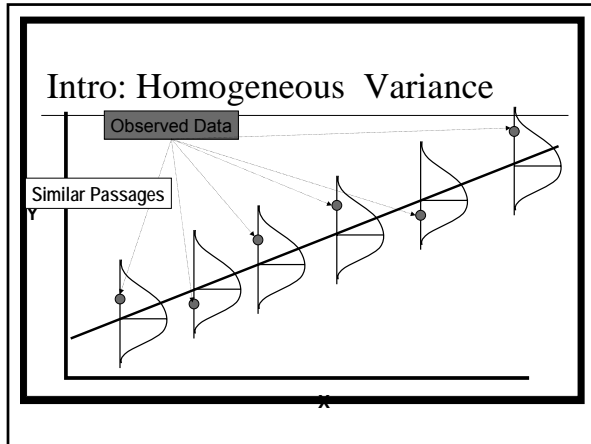
- **Progress Monitoring** – repeated measurement across time to evaluate the rate of skill development
- Issues
 - Alternate v. Parallel Test Forms
 - Relative v. absolute reliability
 - how many measurements? how precise (SEb)?

Intro: Alternate Form

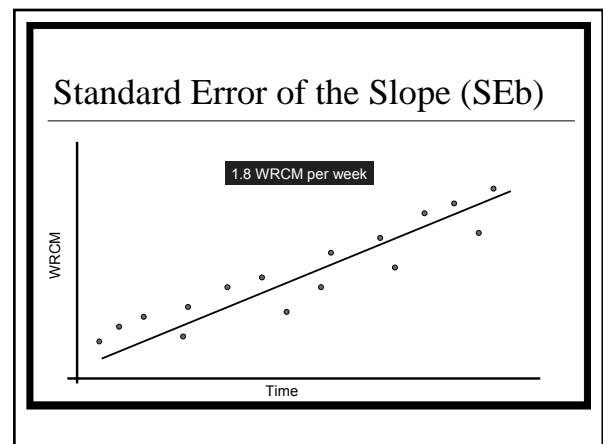
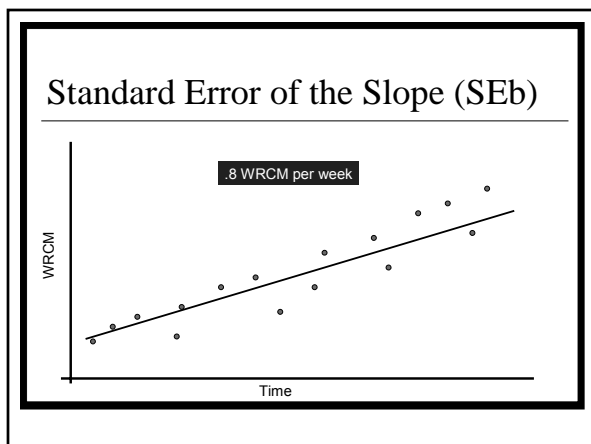
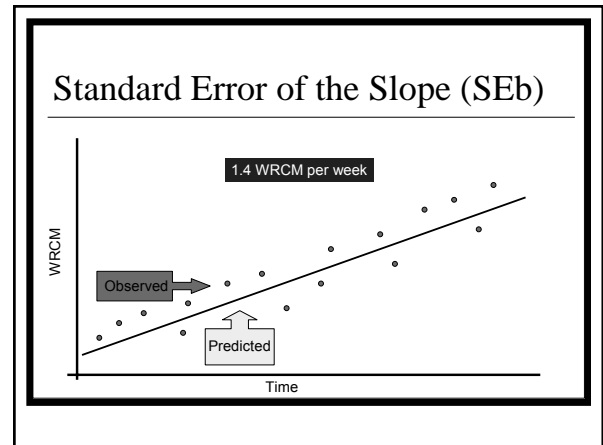
- Alternate form v. Parallel Form
 - **Parallel** (strictly speaking) Assumptions
 - $M_1 = M_2$
 - $SD_1 = SD_2$

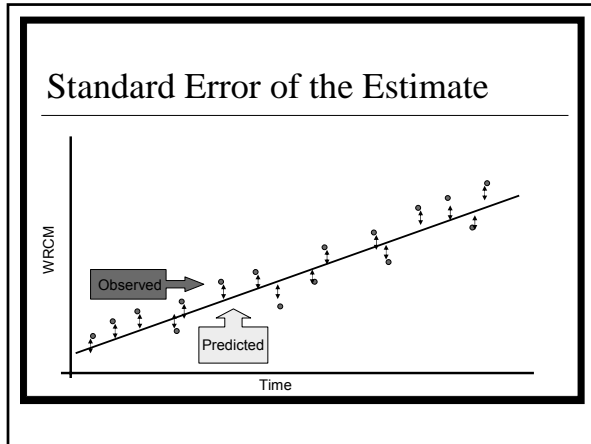
Alternate \neq Parallel





- Linear Model
 - $\hat{Y} = \beta_0 + \beta_1 X_i$
 - b - slope
 - *SEE* – standard error of the estimate
 - measure of the accuracy for the point predictions of the regression line
 - *SEb* – standard error of slope
 - $SEE = \sqrt{[\sum(Y_i - \hat{Y}_i) / n]}$





	Participants		CBMRs Collected ^a		
	N	Grade	Weekly	Study	SEE ^b
		Range	Range	Total	Range
Hintze and Christ (2004)	99	2-5	2	16-20	12-18
Parker, Tindal, and Stein (1992)	45	4-5	2-3	24-36	10-13
Skiba, Deno, Marston, and Wesson (1986)	61	2-5	0-5	21-131	8-12

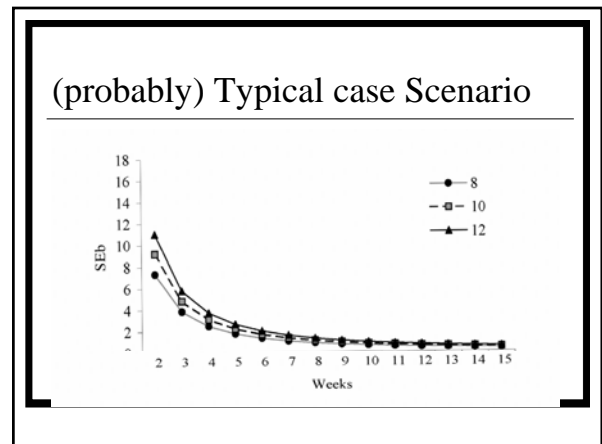
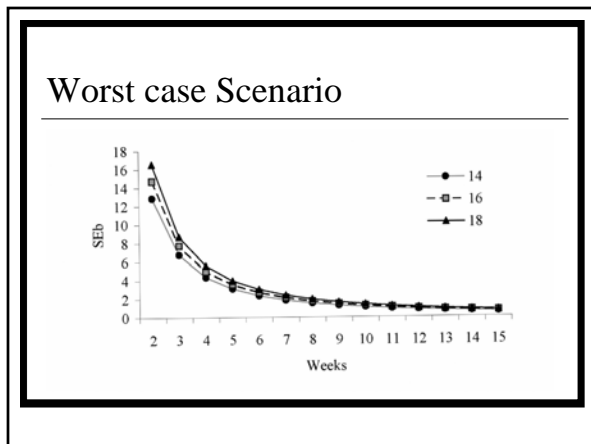
Standard Error of the Slope

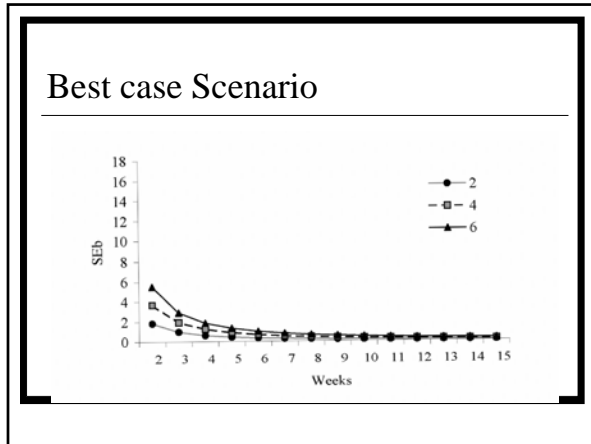
$$SE(b) = SEE / (SD_{days} \sqrt{n})$$

- SE(b) – standard error of the slope
- SEE – standard error of the estimate
- SD_{days} – standard deviation for x-axis
- n – number of data points

Assumptions

- growth is linear
- two data points collected per week
- SEE – 2 to 18 WRCM
 - recall SEMs & published estimates of SEE





Standard Error of the Slope (SEb)

Weeks of Progress Monitoring^a

SEb ^c	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Optimal^d														
2	1.84	.97	.62	.44	.34	.27	.22	.18	.16	.13	.12	.10	.09	.08
4	3.68	1.93	1.24	.88	.67	.53	.43	.36	.31	.27	.24	.21	.19	.17
6	5.51	2.90	1.86	1.33	1.01	.80	.65	.55	.47	.40	.35	.31	.28	.25
Moderate^d														
8	7.35	3.87	2.48	1.77	1.34	1.06	.87	.73	.62	.54	.47	.42	.37	.34
10	9.19	4.84	3.11	2.21	1.68	1.33	1.09	.91	.78	.67	.59	.52	.47	.42
12	11.03	5.80	3.73	2.65	2.01	1.59	1.30	1.09	.93	.81	.71	.63	.56	.51
Poor^d														
14	12.87	6.77	4.35	3.10	2.35	1.86	1.52	1.27	1.09	.94	.83	.73	.66	.59
16	14.71	7.74	4.97	3.54	2.68	2.13	1.74	1.46	1.24	1.08	.94	.84	.75	.68
18	16.54	8.71	5.59	3.98	3.02	2.39	1.96	1.64	1.40	1.21	1.06	.94	.84	.76

- ### Conclusions
- reliability matters
 - measurement conditions matter
 - we need more data than the literature suggests

- ### Other Conclusions
- we need to rely on visual analysis
 - take three and use the median (or mean)

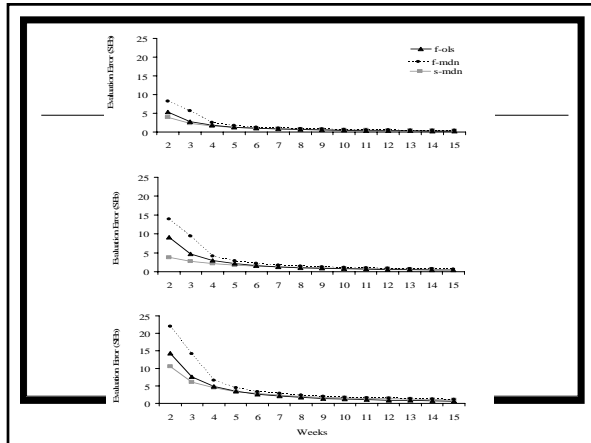
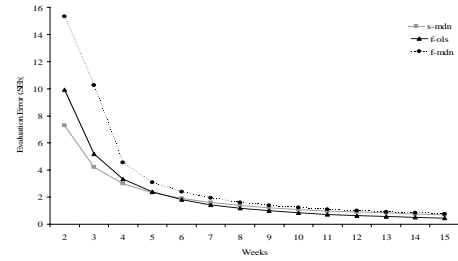
- ### Passage Development
- Goal: examine alternate approaches to the development of passages
 - Options
 - grade level – minimal control
 - readability – modest improvement
 - SEE improves by 2 to 4 WRCM
 - mean performance –
 - equating & scaling – (requires software)

If we have time...

Introduction

- How should we evaluate the data?
 - Least Squares Regression (visual analysis)
 - Pre-post formative median method
 - Pre-post summative median method

Collapsed Across SEE (6 – 12)



Thoughts?
Questions?

Future Directions...

Other Issues

- scaling & equating
- controlled passages
- combined SMM & GOM
- Item Response Theory (IRT)