

Using authentic
assessments for
statewide
accountability:
Issues related to
reliability and
validity of scores



Cornelia Taylor Bruckner, Megan
Cox, Mary McLean, Patricia
Snyder

Purpose of session

- Explore the magnitude of context, rater and reporter facets across different types of early childhood assessments
- Consider methods for measuring the effect of these facets.
- Discuss the feasibility of using context and rater dependent scores to measure progress across time when children will likely change contexts and raters.



Statewide Accountability

IDEA Improvement Act of 2004

Increases accountability requirements of
IDEA: States must submit State
Performance Plans (SPP) and Annual
Performance Reports (APR)

Federal Accountability Initiatives

- 1993 Government Performance and Results Act
- 2001 No Child Left Behind
- 2002 Good Start, Grow Smart
- 2002 OMB Budgeting Process: PART Assessment
- 2004 IDEA Improvement Act

IDEA 2004 and Accountability for 0-5

The new developmental progress indicator:

- Report developmental progress between entry and exit for all infants/toddlers and preschoolers with disabilities in three broad outcome areas
- Report progress in relationship to same age peers
- Establish baseline data and targets for six years
- Report to the public about the indicators for the state and for each program.

Three child outcomes

Percent of children who demonstrate improved:

1. Positive social-emotional skills (including social relationships)
2. Acquisition and use of knowledge and skills (including language/communication and, for preschool, literacy)
3. Use of appropriate behaviors to meet their needs

Positive socio-emotional skills



Involves:

- Relating with adults
- Relating with other children
- For older children, following rules related to groups or interacting with others

Includes:

- Attachment/separation/autonomy
- Expressing emotions and feelings
- Learning rules and expectations
- Social interactions and play

Acquisition and use of knowledge and skills



Involves:

- Thinking, reasoning, remembering, and problem solving
- Understanding symbols
- Understanding the physical and social worlds

Includes:

- Early literacy, pictures, numbers, classification, spatial relationships
- Imitation
- Object permanence
- Expressive language and communication

Use of appropriate action to meet needs



Involves:

- Taking care of basic needs e.g. showing hunger, dressing, feeding, toileting
- Contributing to own health and safety e.g. follows rules, assists with hand washing, avoids inedible objects (if older than 24 mo.)
- Getting from place to place and using tools e.g. forks, pencils, strings attached to objects

Includes:

- Integrating motor skills to complete tasks
- Self-help skills
- Acting on the world to get what one wants

Authentic Assessment



Authentic Assessment

Authentic Assessment – the systematic recording of developmental observations over time by familiar and knowledgeable caregivers about the naturally occurring competencies of young children in daily routines

(Bagnato & Yeh-Ho, 2006)

Embedded assessment

The science of the strange behavior of children in strange situations with strange adults for the briefest possible period of time



The best way to understand the development of children is to observe their behavior in natural settings while they are interacting with familiar adults over prolonged periods of time.

(Bronfenbrenner, 1977)

The Birth to 6 Child Outcome System

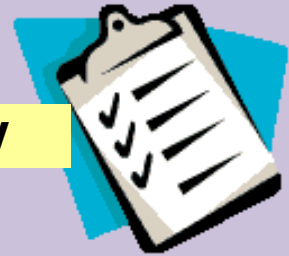
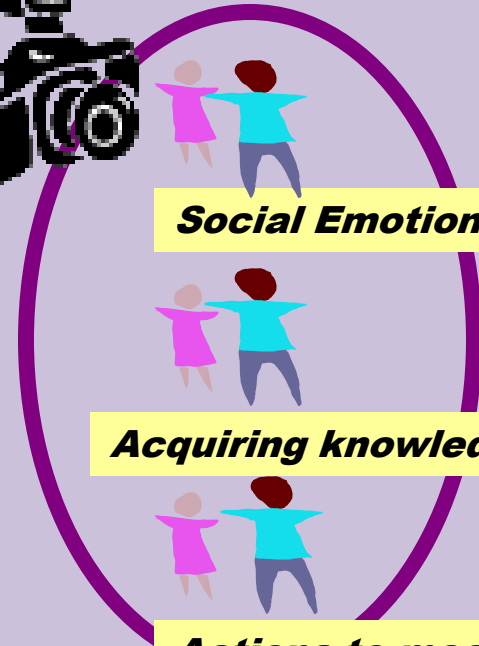
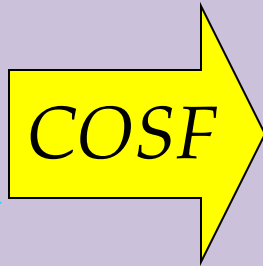
1. Utilize on-going, authentic assessment practices.

2. Determine status ratings at entry and exit

3. Provide this information to the state.

4. The state collects this data and reports to OSEP: percentages of children meeting measurement for the indicator

5. The state determines goals and improvement activities.



Assessments in Use for Statewide Accountability



Data from KY

- **Populations assessed:** All 3-5 year old children receiving preschool special education services
- **Frequency of assessment:** 2 times yearly
- **Instruments used:**
 - *Preschool Child Observation Record (COR)*
 - *Creative Curriculum*
 - *Brigance Inventory of Early Development-II (IED-II)*
 - *Assessment, Evaluation, & Programming System (AEPS)*
 - *Learning Accomplishment Profile-3 (LAP-3)*
 - *Transdisciplinary Play Based Assessment (TPBA)*
 - *Hawaii Early Learning Profile (HELP)*
 - *Carolina Curriculum for Preschoolers with Special Needs (CCPSN)*
 - *Work Sampling System (WSS)*
- **Metrics evaluated:** converted scores linked to KY early childhood learning standards and benchmarks

Data from CA

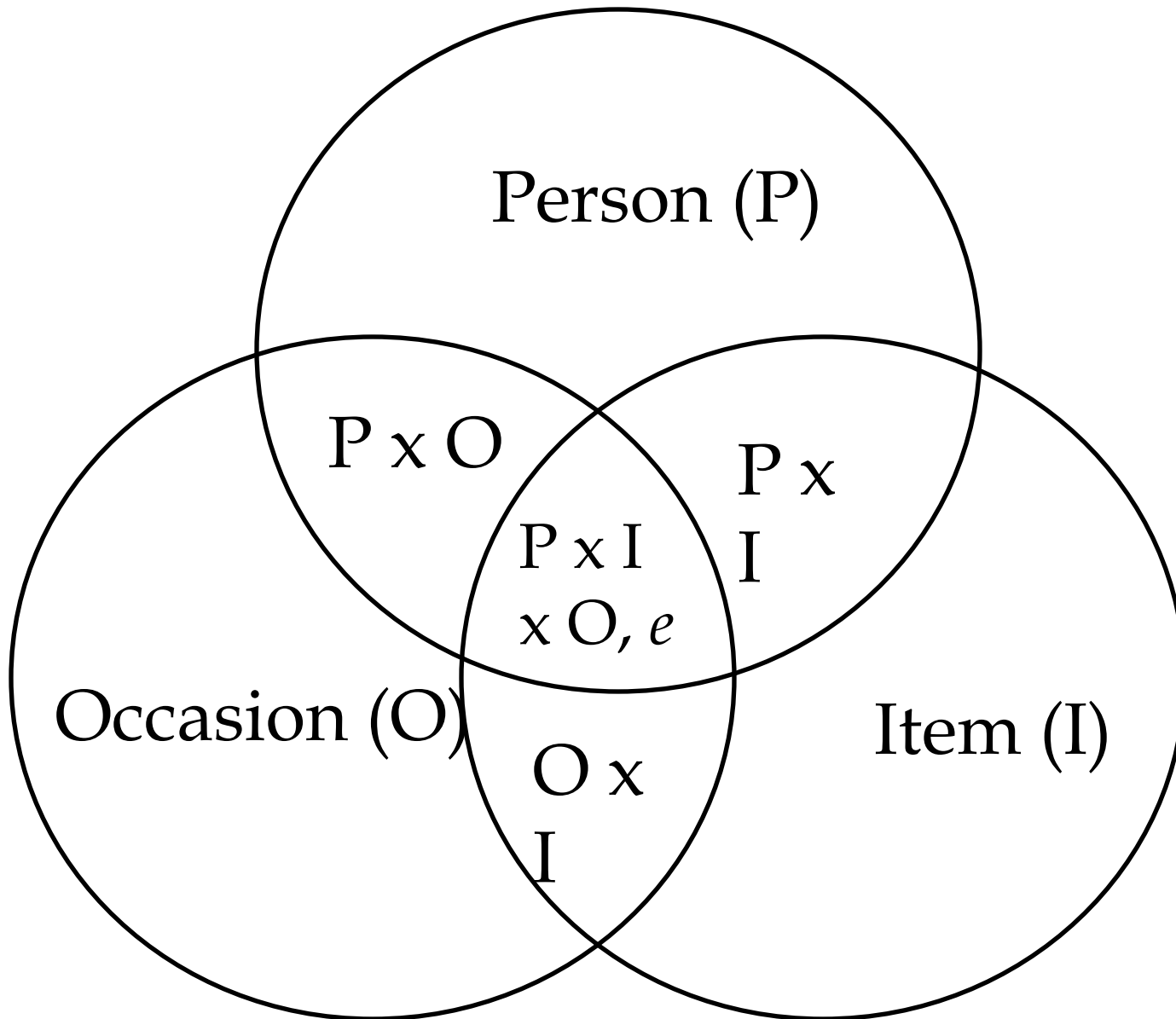
- **Populations assessed:** All 3 – 5 year old children receiving in special education services
- **Frequency of assessment:** twice a year
- **Instruments used:**
 - The Desired Results Developmental Profile *access* (DRDP *access*)
 - the Preschool Desired Results Developmental Profile Revised (PS DRDP-R)
- **Metrics evaluated:** Scale scores by Indicator Group

Data from CO

- **Populations assessed:** Part B 619, Part C, state funded preschool, Head Start and Child Care
- **Frequency of assessment:** 4 times a year (Fall, Winter, Spring, Summer)
- **Instruments used:**
 - *Creative Curriculum Developmental Continuum*®
 - the Work Sampling System
 - the High/Scope COR
 - AEPS
- **Metrics evaluated:**
 - Raw scores by OSEP outcome
 - Raw scores by developmental domain

Dependability of scores

- Scores can be considered one of a universe of scores
- We are rarely interested in a response given to a particular item or probe in a particular context at a particular time
- The ideal score would be a mean score over all acceptable observations.



Important measurement in the assessment of young children

- **Person** - The portion of the variance that is due to the construct that we are trying to measure
- **Item** - Consistency of scores across items
- **Occasion** - Consistency of scores across occasions
- **Context** - Consistency of scores across changes in context
- **Rater** - Consistency of scores across persons applying a rating

Generalizability study using CO statewide assessment data from *Creative Curriculum Developmental Continuum*®

- Extracted a random sample of **100** children from the CO Results Matter data measured using the Creative Curriculum Developmental Continuum®
- Total number of children assessed in 2008-09 - **21,576**
- **54%** Percent Male
- **26%** Percent with an IEP
- **50.4 (8.2)** Mean age in months

Design of Analysis

- Fully crossed item x occasion design
- Items grouped into three subscales
 - Social (10 items)
 - Knowledge and Skills (25 items)
 - Actions to meet Needs (6 items)
- Two occasions separated by 3 months

Analysis

- Used SAS PROC VARCOMP to estimate variance components

```
proc varcomp
```

```
method = mivque0;
```

```
class person items occasions;
```

```
model SOC = person items occ items*person
```

```
        occ*person items*occ
```

```
        items*occ*person;
```

```
run;
```

Percent of variance accounted for by each component in the items x occasions design

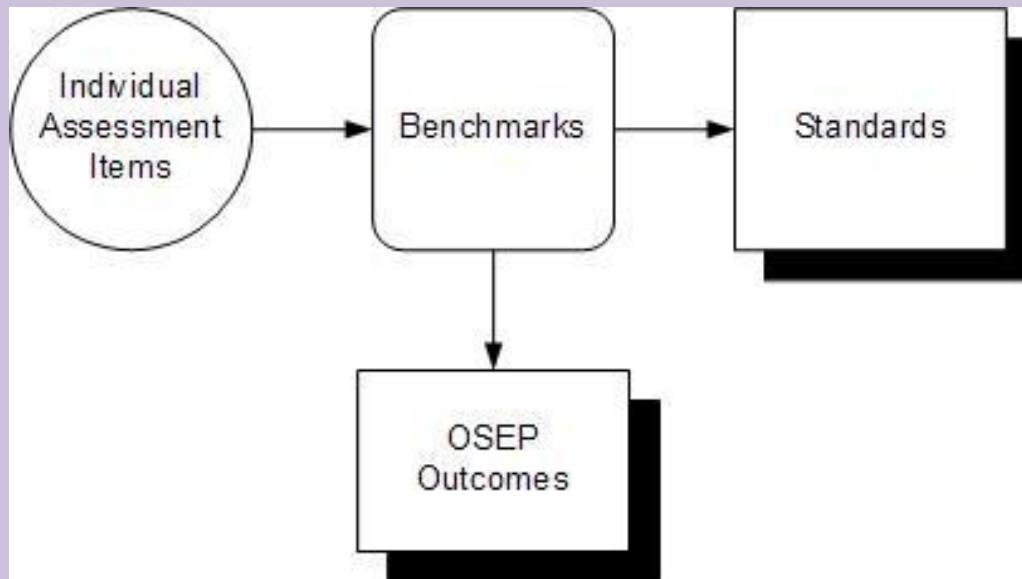
Variance Components	Social	Knowledge and Skills	Actions to meet needs
Person	51.53%	49.97%	46.06%
Person x Item	10.87%	14.44%	20.24%
Person x Occasion	9.07%	5.35%	15.42%
Person x Item x Occasion, error	11.61%	21.63%	9.40%

Controlling effects of the context facet

- Context facet - Consistency of scores across changes in context (e.g. home and school)
 - Write items that can be observed the same way across a variety of contexts
 - Combine information across contexts
 - E.g. school, home, other
 - Compare information from different metrics

Understanding the Context– KY

- Kentucky Early Childhood Data System (KEDS) Conceptual Framework



- Provides a common metric for multiple assessments and allows measurement of context effects

Testing context effects– KY

- Comparative study of KY early learning standards and benchmarks to assessment scores
 - Online survey completed every Fall and Spring (n = 410, Fall 2007)
 - Teachers responded based on children's mastery of skills on KY benchmarks
- Internal consistency of benchmarks using assessment scores

KY context

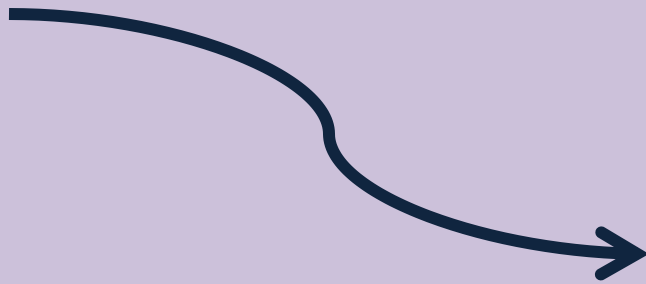
- Sample of 358 children assessed using the *Creative Curriculum Developmental Continuum*®
- Sample Characteristics
 - Mean Age in Months: **49.91**
 - 53%** Male
 - 21.8%** Non-White
 - 69.6%** At-Risk
 - 9.5%** ELL

Comparing Context–KY

- **Are Kentucky's Standards measuring the same construct across context?**
- Internal Consistency Study

50 Items on Creative Curriculum

- Recoded as mastery (met/not met) according to age and crosswalk



10 Standards

- Measured as Mastery (met/not met) based on Teacher Report of child's progress

Results: Internal Consistency

Standard	Number of Benchmarks	Number of Items	Internal Consistency (α)
Arts & Humanities	4	5	.89
Language Arts 1	3	6	.91
Language Arts 2	2	5	.86
Language Arts 3	6	6	.93
Language Arts 4	3	5	.90
Health & Mental Wellness	4	15	.97
Math	4	8	.94
Physical Development	5	9	.93
Science	5	7	.93
Social Studies	6	9	.94

Context Facets– Reliability

- Reliability was equivalently high for other Context Facets

Standard	Male (α)	Female (α)	Non- white (α)	White (α)	English Learner (α)	English Speaker (α)
Arts & Humanities	.89	.88	.91	.88	.90	.89
Language Arts 1	.92	.91	.93	.91	.93	.92
Language Arts 2	.89	.87	.91	.88	.91	.88
Language Arts 3	.94	.92	.93	.93	.94	.93
Language Arts 4	.91	.89	.90	.90	.91	.90
Health & Mental Wellness	.97	.96	.97	.97	.97	.97
Math	.94	.95	.95	.94	.96	.94
Physical Development	.93	.93	.90	.93	.91	.93
Science	.93	.92	.94	.92	.94	.92
Social Studies	.94	.93	.95	.94	.95	.94

Differences in Teacher Perceptions – Context Facets

- Age
 - Differences between CC and KY Standards were expected and confirmed
 - Majority of children under 4 did not meet standards
 - High percentages of children under 4 met age band requirements
 - CC age bands become increasingly difficult
- Gender
 - Differences in gender existed in Teacher perception of children's mastery
 - Teachers scored females higher on all benchmarks except:
 - Mathematics
 - Comparison & Patterning
 - Sense of Purpose

Differences in Teacher Perceptions – Context Facets

- English Learners
 - English learners scored higher on arts and listening benchmarks
 - Scored lower on communication, comprehension, science and social order (rules) benchmarks
- Minority Status
 - Non-white children scored higher on arts, social and exploration benchmarks
 - Scored lower on reading, comprehension and social order (rules) benchmarks

Controlling the effects of the rater facet

- Rater facet– Consistency of scores across persons applying a rating
- Training to standardize the interpretation of observed behavior
 - Identify raters who implement with fidelity
 - Control for differential populations in rating
 - Identify lenient/stringent raters
 - Analyze systematic bias in instrument and rater

Controlling Rater Effects– KY

- Master teacher study
 - Trainers identified rater with most fidelity
 - Comparable matching classrooms chosen for comparison
 - Testing of bias included:
 - Item level analysis
 - Comparison of domain scores

Rater Effects– KY cont.

- 6 classrooms (3 master teachers)
 - 73 students in anchor classrooms
 - 54 matched students
- Comparative analysis of the *Assessment, Evaluation and Programming System*®
- Analysis completed
 - Comparison of population served
 - Independent t-tests for domain scores

Rater Effect

- Item Analysis
 - Non-anchor teachers scored equivalently to anchors across items
- Domain Scores
 - Non-anchor teachers scored equivalently to anchors on all domains except Fine Motor
 - Fine motor scores were significantly lower for non-anchor teachers

Challenges– KY

- Contextual effects
 - Online data doesn't distinguish source of scoring
 - Not enough data to validate crosswalks
- Rater effects
 - Turnover
- Combination
 - Children changing contexts and raters

Changes in reliability with changes in measurement conditions–CA

- In CA the test used for statewide accountability (DRDP) was developed and scaled by the state of CA during a *Calibration Study*
- The calibration study was conducted with a subset of the population and the assessors attended multiple training and those training were required for participation in the study
- In 2006, the same tool was used by the population of Part B preschool service providers in CA for *Statewide Collection*

Changes to the measurement context between the two conditions

- Purpose is accountability
- Assessors vary across time
- Not all assessors participate in training
- Characteristics of children in accountability system might differ from children who participated in studies reported in test manual

Facets of Score Reliability Examined

Reliability Statistic	Measurement Condition	Range of Values Across 8 Indicators	Average
Alpha, the consistency of ratings across Measures within an Indicator. (Criterion Value =.80)	Calibration study (n = 1644)	.88 - .97	.94
	Statewide Collection (n = 16,105)	.89 - .96	.93
Test-retest (Pearson <i>r</i>), the stability of scores across time (Criterion Value = .60)	Calibration study (n = 887)	.86 - .92	.90
	Statewide Collection (n = 8065)	.56* - .78	.74

* The test-retest reliability of Literacy scores was .56; this was .20 lower than any other Indicator

Sensitivity across measurement conditions

Indicator (** p < .05)	Statewide Collection 95% CI (n = 8065)		Calibration 95% CI (n = 707)	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound
Language**	7.41	7.98	8.30	10.00
Literacy	9.94	10.89	8.60	10.20
Learning and Cognitive Competence**	7.79	8.40	9.60	11.20
Math**	8.46	9.02	9.40	11.00
Motor Skills**	6.79	7.38	9.00	10.50
Self-Regulation**	8.25	8.93	9.40	11.30
Self-Concept and Social and Interpersonal Skills**	8.05	8.59	9.10	10.70
Safety and Health**	7.23	7.74	7.90	9.60

Time 2 - Time 1; 6 month interval between assessments

Conclusions

- All assessments vary across facets of measurement
- There is very little research on the magnitude of effect of important facets of measurement for authentic assessment.
- Preliminary data support the reliability of these instruments across time and contexts.
 - However, more research is needed to determine the effects of context and rater facets on these scores.