Desired Results Access Project

DRDP (2015)
2014-2015 Interrater Agreement Study Report

Introduction

The DRDP (2015) is a judgment-based, authentic assessment instrument. Assessors use observations and other documentation to inform their ratings of developmental continua measures organized under eight domains:

1. Approaches to Learning—Self-Regulation (ATL-REG),
2. Social and Emotional Development (SED),
3. Language and Literacy Development (LLD),
4. English Language Development (ELD),
5. Cognition, including Math and Science (COG),
6. Physical Development—Health (PD-HLTH),
7. History-Social Science (HSS), and
8. Visual and Performing Arts (VPA).

The instrument is used for children ranging in age from birth to kindergarten entry. The domain-specific content on the DRDP (2015) is based on developmental research and constructs specified in the California Infant/Toddler Learning and Development Foundations and Preschool Learning Foundations. The content reflects the knowledge, skills, or behaviors important for infants, toddlers, and preschool children to learn (California Department of Education, 2015). Observation-based assessments, such as the DRDP (2015), are completed by assessors (e.g., teachers, special education service providers) who have the opportunity to interact regularly with children. The DRDP (2015) has two versions or views: an infant/toddler view and a preschool view. The DRDP (2015) has been under development since 2011. The Calibration Study version of the instrument was used for the present study. In the calibration version, the Infant/Toddler view was comprised of 27 measures and the Preschool view had 29 additional measures, for a total of 56 measures.

Interrater Agreement Study

The 2014-15 Interrater Agreement Study was conducted by the Desired Results Access Project in collaboration with the Special Education Division (SED) and Early Education and Support Division (EESD) of the California Department of Education in Fall 2014 and Spring 2015. The focus of the study was to gather evidence about rating agreements between pairs of special education assessors who independently rated the same child on the same DRDP (2015) measures within the same time period. This research was supported by the Special Education Division of the California Department of Education as part of the ongoing development of the DRDP (2015).

At the onset of the development of the DRDP (2015), the instrument developers (a collaborative of agencies serving under the direction of the California Department of Education) outlined a series of assessment specifications. Within these specifications, adherence to the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014) was established. Evidence of interrater agreement is desired in many educational and psychological measurement contexts, particularly for judgment-based, authentic assessment instruments in which observations are used to inform ratings. Standard 2.7 of the *Standards for Educational and Psychological Testing* states:

> *When subjective judgment enters into test scoring, evidence should be provided on both inter-rater consistency in scoring and within-examinee consistency over repeated measures. A clear distinction should be made among reliability data based on (a) independent panels of raters scoring the same performances or products, (b) a single panel scoring successive performances or new products, and (c) independent panels scoring successive performances or new products* (p. 44).

For the present study, interrater consistency takes the form of interrater agreement and part (a) of the standard is interpreted as independent assessors rating measures for the same children on the DRDP (2015). Interrater agreement is the extent to which observers produce similar ratings when using the DRDP (2015). The interest in examining interrater agreement was the absolute value of the ratings, recognizing agreement estimates do not reflect variance among units of analysis.

Interobserver agreement estimates are important to obtain for the DRDP (2015) given the rating levels on each measure involve subtle discriminations. Evidence that assessors who have comparable knowledge about children apply similar ratings when they complete the DRDP 2015 independently but concurrently provides important information about interrater consistency in scoring at a single point in time.

Method

Instrument

The DRDP (2015) is comprised of 56 items (measures) that are contained within one of eight groupings of measures (developmental domains). The developmental domains, the affiliated domain abbreviation and the number of measures assigned to each are show below.

Table 1: Developmental Domains of the DRDP (2015)

Developmental Domain	Abbreviation	No. Measures
1. Approaches to Learning and Self Regulation	ATL-REG	6
2. Social-Emotional Development	SED	5
3. Language and Literacy Development	LLD	10
4. English Language Development	ELD	4
5. Cognition, Including Math and Science	COG	12
6. Physical Development and Health	PD-HLTH	10
7. History-Social Science	HSS	5
8. Visual and Performing Arts	VPA	4
	Total	56

The preschool view of the DRDP (2015) is comprised of all eight domains while the view for use with infants and toddlers includes five of the eight domains (ATL-REG, SED, LLD, COG, and PD-HLTH). To assign measure-level ratings, assessors are asked to observe the child over time, in a variety of situations, in different activities, and interacting with familiar people.

Once enough evidence has been obtained, assessors then assign a judgment-based rating for each of the DRDP (2015) measures. The rating should reflect what the observer judges the child's level of mastery to be for that measure and is to be based on both cumulative observations and documentation. When rating the measures, an assessor considers a developmental level mastered if the child demonstrates the knowledge, skills, and behaviors defined at that level consistently over time and in different situations or settings.

As shown in Table 2, measures on the DRDP (2015) are presented to the rater on an ordinal scale and the number of possible rating levels varies from five to nine, depending on the measure. Ratings are assigned to one of the possible developmental levels listed below.

Table 2: Developmental Rating Levels of the DRDP (2015)[1]

Responding		Exploring			Building			Integrating
Earlier	**Later**	**Earlier**	**Middle**	**Later**	**Earlier**	**Middle**	**Later**	**Earlier**

Procedures Used by Assessor Pairs

Researchers associated with the Desired Results Access Project recruited "assessor pairs" defined as two independent assessors able to concurrently assess the same child with the DRDP (2015). The two independent assessors were required to have a close-to-comparable understanding of the developmental competencies of the children they would be assessing.

Assessor pairs were recruited from public school early intervention and preschool special education programs across California and each pair met the following five criteria:

1. Assessors were early interventionists or preschool special education teachers or service providers serving children with individualized family service plans (IFSPs) or preschool individual education programs (IEPs) in their respective setting in California;
2. Assessors selected children for whom they had sufficient knowledge of the children's abilities and skills in order to make informed and confident ratings;

[1] Not all developmental levels are available as rating options across all measures.

3. Assessors were trained on the DRDP (2015) instrument and received additional instructions related to the inter-rater study;
4. Assessors followed the practices recommended for observation, documentation, and rating of the instrument; and
5. Assessor pairs were required to conduct all ratings for the DRDP (2015) independent of their partner. Assessors were encouraged to engage in their typical observational and documentation gathering activities as well as collaborative interactions (including those with their partner) to conduct their DRDP (2015) assessments. The exception to this rule was the exclusion of any discussion or information referencing their final DRDP (2015) rating determinations.
6. To ensure ratings would be concurrent, assessors were asked to make ratings for each measure within a 6-week study assessment period in Fall 2014 or Spring 2015.

Participants

Children. Assessors completed the DRDP (2015) assessment for 79 children. The sample of children included 18 infants or toddlers and 61 preschool-aged children. Table 3 shows demographic information for child participants.

Table 3: Demographic Information for Study Children

Child Participants	Fall 2014		Spring 2015		Total	
Gender	**%**	**n**	**%**	**n**	**%**	**n**
Male	78	51	71	10	77	61
Female	22	14	29	4	23	18
Age	**%**	**n**	**%**	**n**	**%**	**n**
0 – 12 months	5	3	0	0	4	3
12 – 23 months	6	4	21	3	9	7
24 – 35 months	6	4	36	5	11	9
36 – 47 months	29	19	21	3	28	22
48 – 59 months	46	30	14	2	41	32
60+ months	8	5	7	1	8	6
Disability	**%**	**n**	**%**	**n**	**%**	**n**
Intellectual Disability	11	7	21	3	13	10
Hard of Hearing	5	3	7	1	5	4
Deafness	2	1	0	0	1	1
Speech or Language Impairment	23	15	0	0	19	15
Visual Impairment	6	4	0	0	5	4
Orthopedic Impairment	3	2	21	3	6	5
Other Health Impairment	2	1	7	1	3	2
Deaf-Blindness	0	0	7	1	1	1
Multiple Disability	3	2	14	2	5	4
Autism	46	30	21	3	42	33
Ethnicity	**%**	**n**	**%**	**n**	**%**	**n**
Hispanic	35	23	71	10	42	33
Not Hispanic	63	41	29	4	57	45
Not Specified	2	1	0	0	1	1

Assessors. Fifty-one individual teachers or service providers completed the assessment in Fall 2014 and Spring 2015. Thirty-one unique assessor pairs completed the DRDP 2015 assessment for 1 to 4 children. Table 4 shows additional information about assessors and assessor pairs.

Table 4: Details for Participating Assessors

Term	Assessors	Pairs
Fall 2014 Only	34	24
Spring 2015 Only	8	4
Both Fall and Spring	9	3
Total	51	31

The assessors completed their ratings of the measures within the standard assessment window of six weeks in the Fall and Spring. Most ratings (85%) occurred within a week of the corresponding partner's ratings.

As part of electronically entering their DRDP rating data for each child, assessors were asked for additional information about how well they knew the child and how long they worked with their partner rater. The questions asked of assessors included the length of time the child had been enrolled in the program, the length of time the assessor had worked with the child, the number of hours spent per week with the child, length of time working with partner rater in same program, length of time working together, serving the same children, and length of time working together specifically with the child who was assessed. Table 5 shows a summary of this information.

Table 5: Description of Assessor Variables

Assessor Variables	N	Percent
Length of time child enrolled in early intervention or preschool special education		
Less than 2 months	2	3%
2 months to one year	38	48%
One year or more	39	49%
Length of time assessor had worked with child		
Less than 2 months	2	3%
2-6 months	30	38%
7-12 months	21	26%
More than one year	26	33%
Hours per week spent with child		
Less than 2 hours	23	29%
2 to 6 hours	10	13%
6 to 10 hours	4	5%
11 to 20 hours	12	15%
21 to 35 hours	29	37%
More than 35 hours	1	1%
Length of time working with partner rater in same program		
Less than one year	10	13%
About one year	15	19%
2 years or more	54	68%
Length of time working together, serving same children		
We do not overlap at all	14	18%
Less than half the time	33	42%
At least half of the time	32	40%

Assessor Variables	N	Percent
Length of time working together for the child who was assessed		
Less than 2 hours	41	52%
2 to 6 hours	5	6%
6 to 10 hours	1	1%
11 to 20 hours	3	4%
21 to 35 hours	29	37%
More than 35 hours	0	0%

Data from Fall 2014 and Spring 2015 administrations were combined for analyses. Ratings made by the two assessors were collected and compared measure-by-measure. Assessors providing an exact rating match (exact agreement) were assigned a value of 0. Assessor pairs providing ratings that differed in either direction by one rating level were assigned a value of 1 (agreement within one point).

For the present study, interrater agreement percentages were calculated for both exact agreement and agreement within 1 point. Exact agreement combined with agreement within one point were considered absolute agreement in the present study. When using percentage of agreement, values from 75% to 90% are deemed acceptable levels of agreement (Hartmann, 1977; Stemler, 2004).

Results

For the 6-measure Approaches to Learning—Self-Regulation domain (ATL-REG), exact inter-rater agreement ranged from 48%-68% and averaged 59.8%. The inter-rater agreement within one level was between 92%-98% and averaged 93.7%.

For the 5-measure Social-Emotional domain (SED), exact inter-rater agreement ranged from 59%-64% and averaged 61.0%. The inter-rater agreement within one level was between 90%-94% and averaged 92.2%.

For the 10-measure Language and Literacy (LLD) domain, exact inter-rater agreement ranged from 55%-71% and averaged 63.8%. The inter-rater agreement within one level was between 83%-95% and averaged 90.4%.

For the 12-measure Cognition domain (COG), exact inter-rater agreement ranged from 54%-76% and averaged 66.6%. The inter-rater agreement within one level was between 83%-97% and averaged 91.9%.

For the 6-measure Physical Development and Health domain (PD-HLTH), exact inter-rater agreement ranged from 53%-68% and averaged 62.8%. The inter-rater agreement within one level was between 87%-95% and averaged 91.2%.

For the 5-measure History and Social Science (HSS), exact inter-rater agreement ranged from 53%-78% and averaged 69.2%. The inter-rater agreement within one level was between 93%-97% and averaged 94.6%.

For the 4-measure Visual and Performing Arts (VPA), exact inter-rater agreement ranged from 54%-71% and averaged 64.0%. The inter-rater agreement within one level was between 88%-95% and averaged 91.8%.

For the 4-item English Language Development (ELD), exact inter-rater agreement ranged from 62%-81% and averaged 71.0%. The inter-rater agreement within one level was between 90%-95% and averaged 93.8%. This exceeds the standard put forth above of 80%.

For the entire 56-measure instrument, exact inter-rater agreement ranged from 48%-78% and averaged 64.1%. The inter-rater agreement within one level was between 83%-98% and averaged 92.0%.

Table 6 in the Appendix shows inter-rater agreement findings for each of the 56 measures of the DRDP (2015). The shaded columns indicate the standard of agreement used in this study: exact agreement and agreement within one level. Exact agreement is noted in the column labeled "0" (no levels off of exact agreement) and columns to the right are labeled, "1," "2," or "3" in relation to the number of levels from exact (within one level, within two levels, and within three levels).

Discussion

Findings from the present study show that exact inter-rater agreement across the eight DRDP domains averaged 64.1% and agreement within one level averaged 92%. This latter value exceeds the 80% standards put forth for absolute agreement (Hartmann, 1977; Stemler, 2004). As shown in Table 6 in Appendix A, measures ATL-REG 4 (Self-Control of Feelings and Behaviors) and ATL-REG 5

(Engagement and Persistence) in the Approaches to Learning—Self-Regulation domain had the lowest exact agreement percentages at 49% and 48%, respectively; while HSS 2 (Sense of Place) in the History-Social Science domain had the highest percentage of exact agreement (78%), followed by COG 11 (Number Sense of Quantity) at 76% and COG 5 (Documentation and Communication of Inquiry) in the Cognition domain at 74% exact agreement. HSS 3 (Ecology) in the History-Social Science domain also had an exact agreement percentage of 74%.

In examining the combination of exact agreement and within one level, the measures with the lowest agreement are LLD 9 (Letter and Word Knowledge) and Cog 4 (Classification) at 83%, LLD 7 (Concepts about Print) at 85%, PD-HLTH 3 (Gross Motor Manipulative Skills) and PD-HLTH 10 (Nutrition) at 87%, and VPA 4 (Dance) at 88%. The remaining 50 measures had agreement at 90% or above. The measures with the highest agreement are ATL-REG 5 (Engagement and Persistence) at 98%, COG 10 (Inquiry through Observation and Investigation) at 97%, COG 6 (Number Sense of Math Operations) at 96%, and LLD 4 (Reciprocal Communication and Conversation) at 95%. By domain, all agreement percentages were above 90%. The lowest agreement was in the Language and Literacy domain at 90.2% and the highest agreement 93.7% in Approaches to Learning—Self-Regulation.

In general, these findings demonstrate adequate to excellent consistency in scoring across this sample of rater pairs and children. These data are promising given the DRDP (2015) instrument involves raters making subtle discriminations across rating levels.

Future Research

For the present study, raw score measure-level ratings were used to calculate the agreement statistics. During calibration of the DRDP (2015), a Rasch Partial Credit Model (Masters, 1982) was used to develop the scaled scores that are assigned based on performance on groups of measures within a domain. Under the partial credit model, each measure has a unique rating scale structure that takes into consideration levels assigned on other measures within the domain. The domain-level ratings are converted from ordinal-level values into interval-level values (provided in logits) that can be used to determine the consistency in assessors' scores using reliability statistics (e.g., correlation coefficients, generalizability coefficients). These statistics provide estimates of interrater reliability for each domain of the instrument, which will be explored in future studies.

In addition to on-going interrater agreement studies, studies will continue to explore interrater agreement as well as investigate other types of score reliability and validity as part of a comprehensive research agenda designed to ensure the psychometric integrity of the DRDP (2015). The present study was conducted only with special education teachers and service providers. Future interrater agreement studies will include data collected by the broader DRDP (2015) Collaborative (the developers of the DRDP (2015)) who represent both the EESD and SED divisions of the California Department of Education. These studies will include data from infant, toddler, and preschool teachers as well as early intervention and early childhood special education teachers or providers.

Conclusion

For the purposes of the present study, interrater agreement was defined as having been met when 80% of paired ratings were within 1 rating level for a DRDP measure. Across each developmental domain of the DRDP (2015), interrater agreement within 1 level was found to exceed 80%. Given that this version of the DRDP (2015) had not been used previously by these assessor pairs, the interrater agreement data are promising. Interrater agreement results will likely improve as assessors become more familiar over time with the measures of the DRDP (2015) and use the instrument regularly. In addition, revisions were made to the Calibration Study version of the DRDP (2015), which were intended to improve clarity of rating descriptors and associated examples as well as scoring procedures. Future studies will explore score reliability and validity using the revised version of the DRDP (2015).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

California Department of Education. *Desired Results Developmental Profile (2015): A developmental continuum from early infancy to kindergarten entry*. Sacramento: Author.

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability measures. *Journal of Applied Behavior Analysis*, 10, 103–116.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47 (2), 149-174.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating inter-rater reliability. *Practical Assessment, Research & Evaluation, 9*(4), 1-.19.

Appendix

The table below shows percent agreement achieved by assessor pairs, by each measure and developmental domain of the DRDP (2015).

The shaded columns indicate the standard of agreement used for this study: exact agreement and agreement within one rating level. Exact agreement is noted in the column labeled "0". The columns to the right are labeled, "1," "2," or "3" in relation to the number of rating levels away from exact agreement.

Table 6: Percent Agreement by Measure and Domain

Measure	n	% Agreement				% Agreement within 1 Rating Level	
		0	1	2	3+	Measure	Domain
ATL-REG 1	79	68	24	6	1	92	93.7
ATL-REG 2	78	65	27	5	3	92	
ATL-REG 3	79	65	28	6	1	93	
ATL-REG 4	76	49	45	5	1	94	
ATL-REG5 (PS Only)	58	48	50	2	0	98	
ATL-REG 6 (PS Only)	59	64	29	3	4	93	
SED 1	77	64	30	3	4	94	92.2
SED 2	78	62	31	6	1	93	
SED 3	76	59	34	5	1	93	
SED 4	75	60	31	8	1	91	
SED 5	78	60	30	5	5	90	
LLD 1	78	58	36	4	3	94	90.4
LLD 2	78	69	21	8	3	90	
LLD 3	78	64	27	6	3	91	
LLD 4	78	63	32	4	1	95	
LLD 5	77	55	38	8	0	93	
LLD 6 (PS Only)	59	63	25	10	2	88	
LLD 7 (PS Only)	59	71	14	14	2	85	
LLD 8 (PS Only)	60	63	27	8	2	90	
LLD 9 (PS Only)	59	66	17	7	10	83	
LLD 10 (PS Only)	58	66	29	2	3	95	
ELD 1 (PS Only)	21	81	14	5	0	95	93.8
ELD 2 (PS Only)	21	76	14	10	0	90	
ELD 3 (PS Only)	21	62	33	5	0	95	
ELD 4 (PS Only)	20	65	30	5	0	95	

Measure	n	% Agreement				% Agreement within 1 Rating Level	
		0	1	2	3+	Measure	Domain
COG 1	78	67	27	5	1	94	91.9
COG 2	76	54	38	7	1	92	
COG 3	77	57	34	8	1	91	
COG 4	77	62	21	12	5	83	
COG 5	77	74	18	7	1	92	
COG 6 (PS Only)	60	63	33	3	0	96	
COG 7 (PS Only)	58	71	21	9	0	92	
COG 8 (PS Only)	59	61	29	7	3	90	
COG 9 (PS Only)	60	73	22	2	3	95	
COG 10	67	72	25	3	0	97	
COG 11 (PS Only)	59	76	14	7	3	90	
COG 12	68	69	22	6	3	91	
PD-HLTH 1	76	65	25	5	5	90	91.2
PD-HLTH 2	78	62	27	6	5	89	
PD-HLTH 3	76	61	26	9	4	87	
PD-HLTH 4	78	60	30	8	3	90	
PD-HLTH 5	77	61	34	3	3	95	
PD-HLTH 6	76	68	25	7	0	93	
PD-HLTH 7	77	66	29	5	0	95	
PD-HLTH 8	76	67	28	5	0	95	
PD-HLTH 9 (PS Only)	60	53	38	8	0	91	
PD-HLTH 10 (PS Only)	60	65	22	10	3	87	
HSS 1 (PS Only)	59	73	22	5	0	95	94.6
HSS 2 (PS Only)	59	78	19	2	2	97	
HSS 3 (PS Only)	58	74	21	5	0	95	
HSS 4 (PS Only)	60	68	25	5	2	93	
HSS 5 (PS Only)	60	53	40	2	5	93	
VPA 1 (PS Only)	58	71	24	5	0	95	91.8
VPA 2 (PS Only)	60	67	27	5	2	94	
VPA 3 (PS Only)	58	64	26	10	0	90	
VPA 4 (PS Only)	59	54	34	9	3	88	

Produced by the **Desired Results Access Project** – Napa County Office of Education
Funded by California Department of Education, Special Education Division – ©2015 California Department of Education