Educational Assessment, 15:1–26, 2010

ISSN: 1062-7197 print/1532-6977 online
DOI: 10.1080/10627191003673208

Routledge
Taylor & Francis Group

Articulating Assessments Across Childhood: The Cross-Age Validity of the Desired Results Developmental Profile–Revised

Tzur M. Karelitz
University of California, Berkeley

Deborah (Montgomery) Parrish
American Institutes for Research

Hiroyuki Yamada and Mark Wilson
University of California, Berkeley

Assessment systems that track children's progress across time need to be sensitive to the variegated nature of development. Although instruments are commonly designed to assess behaviors within a specific age range, some children advance slower or faster than others and, as a result, often show behaviors from a younger or older age group. This situation is particularly common when the assessment is carried out close to the intersection of two age groups (e.g., just before preschoolers enter elementary school). Assessments that measure development across multiple age groups need to be able to capture behaviors across these cross-age transitions to make valid inferences about children. In other words, one needs to evaluate the extent to which constructs are measured reliably and without discontinuities across the whole age range. This article describes a process for establishing validity of cross-age inferences by using a set of age-consecutive developmental assessments. Identifying and validating an alignment structure between instruments is an essential component of the assessment development process. We present the methodology for establishing cross-age validity using the Desired Results Developmental Profile–Revised (DRDP–R). The DRDP–R assessment consists of a set of instruments designed to track children's socio-emotional, cognitive, physical, and behavioral development across three age groups from kindergarten through 12 years old. We discuss the principles, methodology, and challenges involved in constructing such complex systems of developmental assessments.

Developmental assessment is a comprehensive evaluation of children's progress in multiple domains over time. Assessments can be used to identify areas of need and support for a child,

Correspondence should be sent to Tzur M. Karelitz, Learning and Teaching Division, Education Development Center, Inc., 55 Chapel Street, Newton, MA 02458. E-mail: tzamak@gmail.com

screen for developmental delays or various health problems, monitor developmental trends within and across populations, inform improvement of child care curriculum or parenting practices, and provide evaluation for accountability purposes (Shepard, Kagan, & Wurtz, 1998). The overall goal is to use information gathered to enhance child development and learning.

Developmental assessments are often based on the understanding that child development is mostly sequential in nature, consisting of a hierarchy of milestones or developmental markers (e.g., Fischer, 1980). From a measurement perspective, researchers design assessments based on trajectories that represent a meaningful and useful order of how children progress through a domain (Masters & Forster, 1996). For example, Wilson (2005) provided a framework for creating developmental assessments based on *construct maps*, which explicate the performances and behaviors associated with increasing levels of competency on a specific construct.

Developmental assessments are complex because what is important to assess may be different at various age groups. Moreover, within each domain there could be several developmental stages, and the degree of granularity (i.e., the number of levels) also changes throughout life. Consequently, developmental assessments are commonly designed to assess and be used within a prespecified age group (e.g., the Bayley Scales of Infant and Toddler Development). In these cases, the constructs being measured are bounded by age-appropriate basal and ceiling points, which represent the lowest and highest levels of development to be measured within that age range. However, in some situations there is a need to assess the same construct across consecutive age groups. For example, when children move from Infant/Toddler (I/T) care to Preschool (PS), it is important to articulate their development across constructs assessed in the different programs. Specifically, developmental assessments should be sensitive to children near the intersection points between adjacent age groups. This can be achieved by having similar developmental levels represented on consecutive assessment instruments. In this way, "early bloomers" and "late bloomers" who fall outside their age group's ceiling and basal points can be reliably and validly assessed. Moreover, a cross-instrument perspective (i.e., a wide age span) is needed for assessment of children receiving special education services (although this is more complex and often includes the measurement of older children at lower levels of performance). Thus, alignment between age-specific assessments can improve their sensitivity to the wide range of developmental trajectories that children may exhibit. We propose that articulating successive developmental assessments can increase their *validity as a set*, as the aligned items allow instruments to better assess child development on the intended construct across time, and thus allow for valid cross-age inferences.

This article describes a methodology for examining the cross-age validity of a set of developmental assessments. The validation process targets a general question: To what extent can assessments designed for different age groups be aligned across groups in a meaningful and useful way? The process of validating assessments involves an evaluation of the extent to which theoretical considerations and empirical evidence support the interpretations and uses of assessment outcomes (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999; Messick, 1989). The extensive literature on validity provides various strategies for developing and evaluating a validity argument for different tests. However, there is no research that we are aware of on dealing directly with the validity of age-consecutive developmental assessments. Specifically, we propose that although each instrument must be reliable and valid for its target age group, a *separate* validity argument should be sought with respect to the cross-

age continuity of constructs measured by the full set of instruments. This article describes a method to construct and evaluate such a validity argument.

Our research is conducted within the framework of the Desired Results Developmental Profile–Revised (DRDP–R). The DRDP–R is a set of assessment instruments designed to track children's progress in three age groups: I/Ts, PS children, and School-Age (SA) children from kindergarten through 12 years old. The introduction provides the background for the DRDP–R assessment system, the principles of validity research that guided our research, and the construction of our validity argument. The Method section describes how evidence was collected to evaluate the validity of the DRDP–R. We report findings based on reliability and validity analyses of the instruments.

BACKGROUND OF THE DRDP–R

Components of the Desired Results System

The California Department of Education (CDE), Child Development Division, in collaboration with the CDE Special Education Division, has developed the Desired Results (DR) system to foster high-quality programs for children. The DR system[1] serves as an accountability framework, designed to help programs determine effective strategies for improvement based on assessments, surveys, rating scales, quality standards, foundations, and curriculum frameworks.

The development of the DR system was an iterative effort to address the complex theoretical, practical, and political considerations needed for a successful accountability system. A broad range of researchers, practitioners, parents, and state and local program administrators, and a national panel of experts in child development, education, and assessment worked together to develop the DR system. Four DRs—or conditions of well-being—were defined for all children. Once the DRs were agreed upon, more specific indicators were identified for each DR:

Indicators of DR1: Children are personally and socially competent.

1. SELF—Children show self-awareness and a positive self-concept.
2. SOC—Children demonstrate effective social and interpersonal skills.
3. REG—Children demonstrate effective self-regulation of their behavior.
4. LANG—Children show growing abilities in communication and language.

Indicators of DR2: Children are effective learners.

5. LRN—Children are interested in learning new things.
6. COG—Children show cognitive competence and problem-solving skills through play and daily activities.
7. MATH—Children show interest in real-life mathematical concepts.
8. LIT—Children demonstrate emerging literacy skills.

[1]More information about the DR system can be found at http://www.cde.ca.gov/sp/cd/ci/desiredresults.asp.

Indicator of DR3: Children show physical and motor competence.

9. MOT—Children demonstrate an increased proficiency in motor skills.

Indicator of DR4: Children are safe and healthy.

10. SH—Children show an emerging awareness and practice of safe and healthy behavior.

Based on the set of DRs and indicators, multiple overarching constructs were identified to develop assessment instruments for children served by CDE programs. The intent was to choose constructs that (a) represent meaningful and important aspects of child development based on the current literature, (b) are relevant and appropriate to the needs and challenges of child care programs, (c) can be assessed through observation in natural settings, and (d) can inform programs' curricular decisions.

The next step was to use the overarching constructs to derive age-specific developmental measures. The measures were combined into the DRDP–R, which is the primary assessment component of the DR system. The DRDP–R is a curriculum-embedded observational assessment tool, firmly grounded in research on early childhood education and designed to measure progress of children from birth through 12 years old. This age span was divided into three groups corresponding to different models of educational organization used by CDE: I/T, birth through 2 years old; PS, 3 and 4 years old; and SA, 5 through 12 years old. Thus, the DRDP–R system consists of three age-specific assessment instruments that represent a continuum of development for each of the four DRs. The version of the DRDP–R that we used was implemented in California in fall 2006 (Metzuk, Parrish, & Mangione, 2006).

Function and Structure of the DRDP–R

The DRDP–R is designed for teachers[2] in child development programs. During daily interactions and program activities, teachers have multiple opportunities to observe children while they naturally exhibit behaviors associated with different developmental domains. Teachers are trained to take anecdotal observation records and collect evidence (e.g., photos, checklists, quotes, and examples of children's work). The DRDP–R helps teachers reflect on each child and rate him or her on a comprehensive set of age-specific measures of development of individual children. The rating is done twice a year for each child and is supported by the evidence collected in the preceding weeks. Because teachers' ratings are based on observations, special emphasis is given to proper training of raters (the issue of rater reliability is discussed later).

Each DRDP–R instrument consists of multiple measures. Overall there are 105 measures in the three instruments (35 measures in I/T, 39 in PS, and 31 in SA).[3] Every measure is associated with 1 of the 10 indicators and, in turn, one of the four DRs. For each indicator, 2 to 7 measures that define the indicator more specifically are included in each DRDP–R. For

[2]For simplicity, the term "teachers" is used to refer to anyone who uses the DRDP–R, such as caregivers, preschool teachers, youth center staff, and administrators.

[3]The complete set of DRDP–Rs can be found in http://www.cde.ca.gov/sp/cd/ci/drdpforms.asp.

example, Identity of Self, Self-Expression, Recognition of Own Abilities, and Awareness of Diversity were the four measures of the SELF indicator included in the I/T instrument. The choice of which measures to include was motivated by the criteria detailed earlier in the text for choosing the overarching constructs.

An example of a DRDP–R measure format, using the PS SELF measure Identity of Self, is shown in Figure 1. Each measure describes a range of developmental levels that the teacher may choose from in characterizing the progress of an individual child. For each developmental level, the measure gives a brief description that explicates the type of behavior expected from a child at this level. Each level also contains several examples that provide observable behaviors a child might exhibit in the program setting. Teachers rate children at a particular level if they observe them performing the behaviors described for that level easily, confidently, consistently, and in different settings. Teachers are encouraged to provide written examples to support their ratings on each measure. If the child does not exhibit behaviors from the lowest level of the instrument, he or she can be rated as "Not Yet" at the first level.[4] If a child shows consistent behaviors at a certain level and emerging behaviors from the subsequent level, he or she can be rated as emerging to the next level.

Three criteria guided the construction of each measure in terms of its focus and the number of developmental levels and their content. First, measures were grounded in the relevant child developmental theory. For example, the Identity of Self measure is based on research about the development of self-awareness in young children (e.g., Harter, 1999; Thompson, 2006). Preschool in particular is an important transitional period from the rudimentary focus on physical and dispositional attributes to a deeper psychological self-awareness of the early grade-schooler (e.g., Eder, 1990; Marsh, Ellis, & Craven, 2002; Measelle, Ablow, Cowan, & Cowan, 1998). By the time children leave preschool, some begin to incorporate social comparisons into their self-perceptions (Pomerantz, Ruble, Frey, & Greulich, 1995).

Second, the descriptions and examples of behaviors were designed to fit observational situations within the programs' settings. In other words, the chosen behaviors did not need to be elicited from children by means of a particular predefined task. For instance, the examples in the Identity of Self measure describe common ways in which children talk about themselves in a preschool setting. The behaviors in each measure were identified based on input from child development experts and practitioners, and examples given by teachers in pilot studies.

Third, the measures were designed to assess aspects of child development that can be supported by the child care program. The measures correspond to curricular frameworks within which programs operate, and thus the information obtained from the DRDP–R could provide meaningful feedback for program improvement. For example, the PS instrument contains more measures of language, literacy, and math compared to the SA instrument (16 vs. 10). The reason for this imbalance is that teachers who use the PS instrument need information about these constructs to ensure the school-readiness of children under their care, whereas school-aged children are commonly assessed on these constructs in their schools.

The measure in Figure 1 has four developmental levels, labeled as Exploring, Developing, Building, and Integrating. All of the measures within a single indicator have the same number of levels. However, different indicators may have a different number of levels in their measures.

[4]The "Not Yet" level applies only to the PS and SA instrument. It is considered irrelevant to the I/T instrument—the first level in I/T describes reflexive behaviors that are primary in the development of infants.

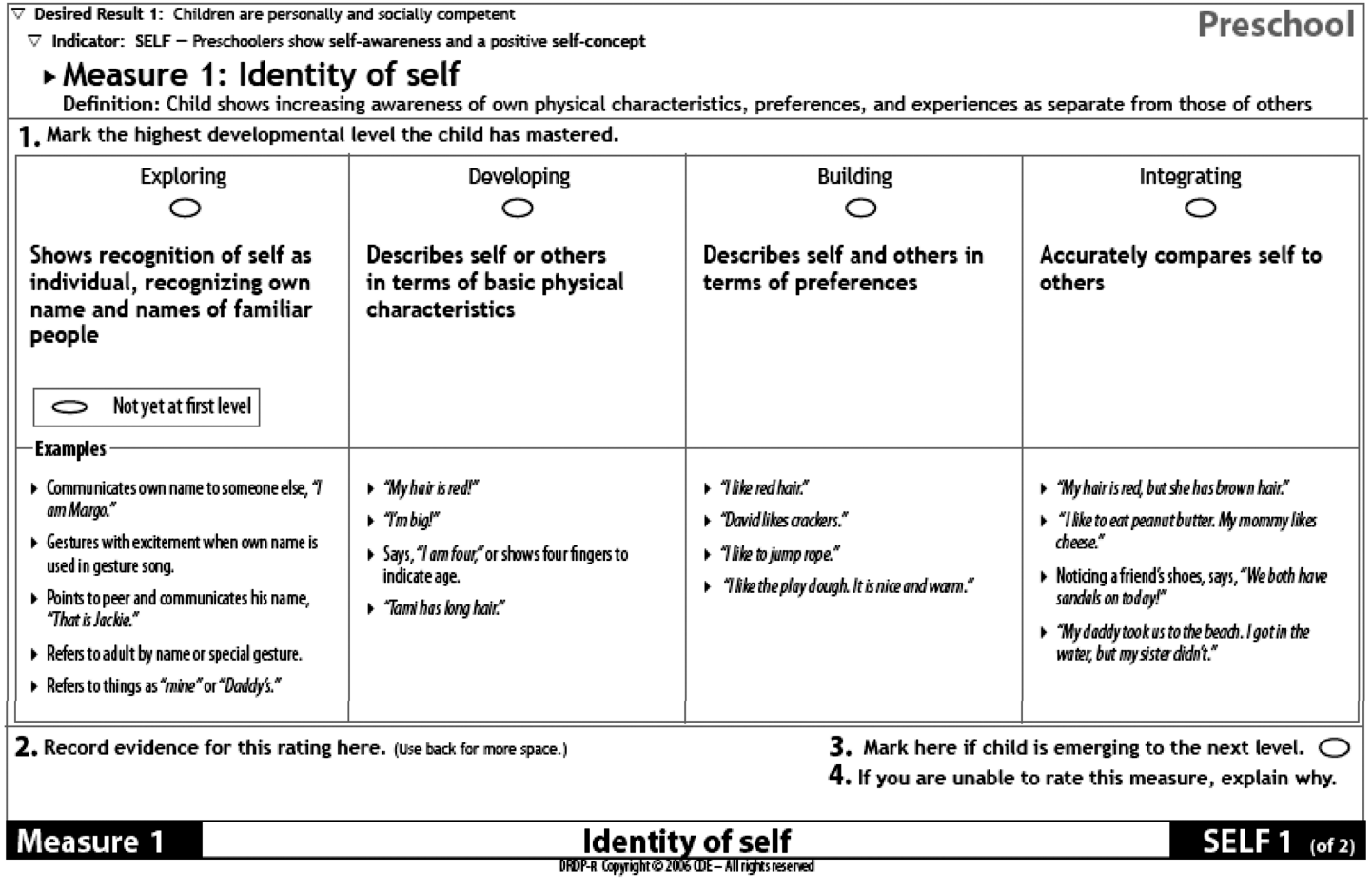

FIGURE 1 An example of a measure from the Preschool Desired Results Developmental Profile–Revised.

Moreover, different age-group instruments may have a different number of levels for measures of the same indicator. For example, the SELF measures in the I/T and SA instruments have five and four levels, respectively. Thus, there is no uniform number of levels across all of the DRDP–R instruments and measures. The number of rating categories in each measure depends only on the three criteria for measures' content described earlier.

Purposes of the DRDP–R

The main goal of the DRDP–R is for teachers to internalize the process of collecting information and reflecting about children under their care. Teachers use this information to support the growth of individual children and communicate with parents during parent–teacher meetings. Moreover, the intent is for programs to use information from the DRDP–R to evaluate and improve their curriculum to enhance services they provide for children. Child care programs are *not required* to share individual children's data with the state, although CDE supervisors may inspect DRDP–Rs during routine visits.

In addition to assisting teachers in documenting the unique developmental pathways of individual children, the DRDP–R is intended to serve as a training tool for practitioners. The measures form a research-based framework for supporting development in core domains within high-quality programs. Regular use of the instrument enables teachers to track changes over time as children grow and to share this information with parents and program administrators. Thus, the DRDP–R was designed not to be used as a screening or selection device but rather to help staff enhance the development of children in their program.

The DRDP–R is related to other developmental assessments (e.g., the Bayley Scales of Development, the Batelle Developmental Inventory, or the Kaufman Assessment Battery) in the sense that it is based on similar theories of child development and measurement. However, the DRDP–R differs from these assessments in its structure, the age span of the target population, and its intended uses. Many of the commonly used developmental assessments are designed for screening purposes, are based on direct assessment rather than embedded observation, and cover a narrower range of ages or developmental domains.

THE MEANING OF CROSS-AGE VALIDITY OF DEVELOPMENTAL ASSESSMENTS

Validity is an argument used to evaluate the adequacy of proposed usage and interpretation of an assessment (AERA, APA, & NCME, 1999; Messick, 1989). The validity argument provides the rationale for how observations should be interpreted and provides a critical evaluation of the proposed interpretation (Kane, 2001). Establishing validity requires an analysis of the assumptions underlying inferences from test scores and evaluating evidence in favor of or against the plausibility of the proposed, as well as alternate, interpretations. The evidence for a validity argument is based on the assessment's content and internal structure, the underlying response processes, the consequences of testing, and relations with other variables of interest (AERA, APA, & NCME, 1999).

Although any assessment instrument must be reliable and valid for its target population, an assessment system that targets successive age groups, such as the DRDP–R, must also show

validity across instruments. This article discusses only one component of the complete validity analysis conducted for the DRDP–R. Specifically, we focus on the validity argument concerning the assumptions needed to make cross-age inferences based on the DRDP–R. In this respect, ability scores at the indicator or DR level are used for tracking and documenting progress for individual children and identifying trends within child care centers. For instance, consider a program consisting of both an I/T program and a PS program. The administrator wishes to prepare the curriculum for children entering preschool the following year, meaning toddlers who are about to leave child care. She can use the I/T instrument to gather information about the toddlers' current state and identify areas that need further attention. She may find that the majority of toddlers have not yet reached the highest levels of the literacy indicator, which may lead her to adjust the frequency or quality of the program's literacy-based activities. Consequently, the next time the toddlers are rated (this time on the PS instrument), the administrator would be able to see whether they progressed as expected (i.e., all or most children are rated at least at the lowest level of the PS instruments). To make such inferences, the underlying assumptions are that the measures in the I/T and PS instruments target the same constructs without discontinuities or divergences.

The cross-age aspect of successive developmental assessments should be maintained by multiple measures spanning the target age groups. The levels on such measures should represent a "road map" of expected growth in terms of developmental markers or milestones throughout childhood. Where one age group ends and another begins, the intent should be to bridge individual measures so as to represent a continuous progression of developmental constructs while maintaining age-specific distinctions. If the DRDP–Rs are successful in attaining this dual goal, then cross-age inferences can be made with confidence. Thus, our cross-age validity argument focuses on the assumptions underlying the *articulation* of age-consecutive assessments. Articulation in this sense means similarity of behaviors described by measures from consecutive assessments.

The process of developing articulated developmental assessments culminates with an *alignment table* that delineates which measures target the same construct, which of their levels align (i.e., overlap), and the theoretical basis for this alignment structure. The alignment table is helpful for establishing the system's content validity, as its creation reveals discontinuities and divergences in the conceptualization of development in specific domains. Therefore, the alignment table is akin to an interpretive argument (see Kane, 2001), as it specifies the assumptions underlying possible cross-age inferences from the DRDP–Rs. The alignment table allows researchers to identify where assumptions are sensible and straightforward and where they seem problematic and warrant more attention. Once the alignment table is established, the next step in the validity argument is to evaluate whether the hypothesized alignment between instruments is empirically supported. In this article, therefore, we report analyses of data that were collected with the intention of evaluating the cross-age validity of the DRDP–R. We investigate the following three questions:

1. Are teachers using the instruments reliably?
2. To what extent do teachers rate children in accordance with expectations?
3. To what extent are the difficulties of developmental levels of the three instruments ordered according to the alignment structure?

Reliability is investigated because a valid instrument must be reliable. Teachers' rating patterns and the difficulty of levels can be used as indicators of validity, if they support the assumptions and expectations laid out by the development of the alignment table. The next section describes the process of representing cross-age assumptions in the form of an alignment table and producing expectations based on this structure.

CONSTRUCTING A CROSS-AGE VALIDITY ARGUMENT

Establishing an Alignment Structure for Successive Developmental Assessments

This section describes the process of constructing assumptions for cross-age interpretations from consecutive developmental assessments, using the DRDP–R as an example. We elaborate on the challenges one can expect to encounter when articulating assessments of this kind. Finally, we summarize the findings from our literature review in the form of an alignment table for measures across successive instruments.

During the development phase of the DRDP–R, cycles of content analysis and revisions were carried out, in accordance with the iterative nature of any validation process, but particularly following the "Four Building Blocks" approach described by Wilson (2005). Following the identification of overarching constructs, preliminary age-specific measures were developed by teams consisting of experts in developmental theory, experts in psychometric theory, and practitioners working with children at each age group. Next, another team examined the articulation (or lack thereof) of all developmental constructs across the instruments. Several principles guided this process:

1. Constructs should follow a coherent (though not necessarily uniform) continuum from birth through age 12.
2. There should be appropriate basal/ceiling cut-points for measures in each age range.
3. The developmental levels for measures targeting the same construct should align appropriately at the intersection points between consecutive instruments.
4. The alignment structure should be based on relevant child development theories and research.

The process of articulation began by asking, Which measures are conceptually similar (i.e., based on a common construct), and is the alignment between them supported by the developmental literature? The articulation team looked for logical connections between successive age groups by identifying behaviors that were described in the literature for both groups. For example, the team identified a theoretically based progression of the Identity of Self construct from infancy through preschool, and up to school age, as can be seen in Figure 2, where the levels of the three corresponding measures are ordered vertically. The intersections between the age groups were supported by literature on development of self-awareness. For instance, building on the recognition of one's separateness as a person (Kagan, 1991), older toddlers begin to communicate about themselves as individuals distinct from others (e.g., Case, 1991; Stern,

I/T	PS	SA
		Expending
		Describes self in terms of a role in a community that includes people he or she may not know (the whole school, the town where he or she lives)
		Integrating
		Describes self in terms of roles within one or more groups of people he or she knows
	Integrating	**Understanding**
	Accurately compares self to others	Describes physical characteristics, preferences and things he or she can do in relation to another person
	Building	**Developing**
	Describes self and others in terms of preferences	Accurately describes self in terms of physical characteristics, preferences, and things he or she can do
	Developing	
	Describes self or others in terms of basic physical characteristics	
Developing Ideas	**Exploring**	
Expresses ideas about self and his or her connection to other people and things	Shows recognition of self as individual, recognizing own name and names of familiar people	
Discovering Ideas		
Communicates own name and names of familiar people and things		
Acting with Purpose		
Recognizes self, familiar people, and familiar things		
Expanding Responses		
Uses senses to explore self and others		
Responding with Reflexes		
Communicates needs and attends to caregiver with reflexes		

FIGURE 2 Alignment between Desired Results Developmental Profile–Revised measures of the Identity of Self construct. *Note.* I/T = Infant/Toddler; PS = Preschool; SA = School-Age.

1985). Communicating one's own name and the names of others is an observable example of this developmental stage that is typical for children around 3 years of age, hence the alignment between the I/T and PS instruments. During early childhood, self-descriptions are mostly based on physical characteristics (Damon & Hart, 1988). However, by the time children enter grade school, psychological characteristics and early notions of social comparisons become more pronounced in children's self-descriptions (e.g., Eder, 1990; Marsh et al., 2002; Measelle et al., 1998; Pomerantz et al., 1995). Thus, what the literature identifies as typical in late toddlerhood overlaps with the early preschool years, and behaviors typical to the late preschool years overlap with early school-age years.

This articulation process highlighted natural discontinuities between particular constructs that were either more important for one age group than another or, in some cases, even nonexistent in one group. The constant tension between maintaining fidelity to different developmental theories and accommodating variations among them led to a series of deliberations by the articulation team and the different age-level teams. The main issues were the challenges in capturing the variegated nature of development and the appropriateness of construct discontinuities across age groups. In some cases, it was determined that a theoretically sound case could be made for the discontinuity of a construct. For example, physical development is highly differentiated in infancy as babies and toddlers gain increasing control over fine and gross motor skills, yet by school age, motor development consists mostly of refinement of coordination and dexterity (e.g., Delahunt, 2002). In other cases, the team concluded that differences could be bridged by identifying a common ground among age-level experts with regard to the representation of research-based developmental constructs. For example, in the I/T instrument there is no measure of the Writing construct because precursor writing behaviors in infancy are essentially subsumed under fine motor skill development rather than literacy development.

These articulation discussions led to further content refinement of DRDP–R measures, with the overall purpose of achieving a continuity of constructs based on theoretically valid assumptions. In other words, the content validity of the DRDP–R has been established through the development process of the assessment system. The result was that each DRDP–R instrument targets the developmental progress typical of the majority of children within each age range while also allowing for some overlap between the highest levels of a "younger" instrument and the lowest levels of the next, "older" instrument. The established alignment structure between instruments serves as an interpretive argument for validity purposes in the sense that it lays out the assumptions that are the basis for empirical investigation and subsequent uses of the assessment system. In addition, the work laid out by the articulation team helped pave the road to the development of the California Preschool Learning Foundations (CDE, 2008).

Table 1 presents the cross-age alignment table for the DRDP–R. Because of space limitations, names and descriptors of developmental levels are not shown and we do not report all of the literature underlying the assumptions of the interpretive argument. We identified 33 aligned constructs across the DRDP–Rs, which are measured by 88 instrument-specific measures (84% of all DRDP–R measures). Each row in Table 1 contains a bridge between two or more measures that assess the same construct. An empty cell indicates no bridged measures for that construct. Values from 1 to 6 represent which levels are aligned between the measures on each pair of instruments. For example, the first construct in the SELF indicator, Identity of Self, has "4/5" under I/T and "1" under PS, meaning the last two levels in the I/T measure (Levels 4 and 5) are aligned with the first level in the PS measure (Level 1). The alignment between PS and SA is represented as "2/3,4" and "1,2," respectively. This means that Levels 2 and 3 in the PS measure are aligned with Level 1 in the SA measure and Level 4 in PS is aligned with Level 2 in SA.

Considering the 33 constructs in Table 1, 13 constructs were related to DR1 measures, and 15, 3, and 2 constructs were related to DR2, DR3, and DR4, respectively. As shown in Table 1, two thirds of the aligned measures bridge all three instruments and one third connect only two instruments. The typical alignment pattern has two levels aligned on both instruments.

Not all alignments are the same. In the first typical case (the most common—about two thirds of the cases), measures assess the same aspects of development (i.e., they are conceptually

TABLE 1
The DRDP–R Alignment Table

		Level Alignment			
Indicator	*Aligned Measures*	*I/T→*	*PS*	*PS→*	*SA*
Self-Concept (SELF)	Identity of Self	4/5	1	2/3,4	1,2
	Self-Esteem	4,5	1,2	4	1
Social Interpersonal Skills (SOC)	Empathy	4,5	1,2	3,4	1,2
	Relationships w/ Familiar Adults	4,5	1,2	—	—
	Interactions with Peers	3,4,5	1,2,3	—	—
	Friendship	3,4/5	1,2	4	2
	Conflict Negotiation	—	—	3,4	1,2
	Awareness of Diversity	5	1/2	2,3,4	1,2,3
Self-Regulation (REG)	Impulse Control	—	—	2,3/4	1,2
Language (LANG)	Comprehends Meaning	4/5,6	1,2	1/2,3,4	1,2,3
	Responsiveness to Language	4/5/6	1	1,2,3/4	1,2,3
	Expression of Oral Language	5/6	1	3,4	1,2
	Uses Language in Conversation	5,6	2,3	—	—
Learning (LRN)	Curiosity and Initiative	3,4/5	1,2	3/4	1
	Engagement and Persistence	3,4,5	1,2,3	2/3/4	1/2
Cognitive Competence (COG)	Memory	4,5	1,2	3,4	1,2
	Cause and Effect	3,4,5	1,2,3	3/4	1/2/3
	Problem Solving	4/5	1	4	1/2
	Socio-Dramatic Play	—	—	4	1
Math (MATH)	Number Sense	5	1	4	1
	Math Operations	—	—	4	2
	Shapes	5	1	4	1
	Classification	3/4,5	1,2	—	—
	Measurement	—	—	2,3,4	1,2,3
	Time	5	1	4	1
Literacy (LIT)	Interest in Literacy	4,5	1,2	3,4	1,2
	Letter and Word Knowledge	5	1	4	1/2
	Writing	—	—	3,4	1,2
Motor Skills (MOT)	Gross Motor	6	1	3,4	1,2
	Balance	6	1	—	—
	Fine Motor	6	1/2	3,4	1,2
Safety and Health (SH)	Personal Care Routines	4,5	1,2	1,2,3,4	1,2,3,4
	Safety	4,5	1,2	1,2	1/2,3

Note. Values represent aligned levels between pairs of instruments. Commas separate aligned levels. For example 4,5 under Infant/Toddler (I/T) and 1,2 under Preschool (PS) means that Level 4 in I/T aligns with Level 1 in PS, and Level 5 in I/T aligns with Level 2 in PS. A slash between levels means they share a connection. For example, Levels 4/5 both align with Level 1. DRDP–R = Desired Results Developmental Profile–Revised; SA = School-Age.

similar) and use similar language to describe behaviors; thus, the alignment is *complete*. In the remaining cases, measures are conceptually similar, but the description of behaviors in each measure may not show the alignment in a straightforward way, meaning there is only a *partial* alignment. For instance, this occurs when the alignment spans several levels (e.g., one level aligned with two levels as in the Identity of Self construct in the SELF indicator).

Out of the 105 DRDP–R measures, 17 measures did not form alignments across instruments. Some measures did not align because they assess age-specific aspects of development that are less relevant or less important in other age groups. For example, the REG measure, Self Comforting, existed only in the I/T group. Other measures existed in more than one instrument but emphasize age-specific aspects of development that could not be aligned across instruments. For example, the I/T COG measure, Symbolic Play, does not align with the PS COG measure, Socio-Dramatic Play, because the emphasis in PS is on play that involves social interactions and language skills (e.g., Bergen, 2002).

The alignment table lays out the assumptions underlying any cross-age inferences that can be made based on the DRDP–R. Each pair of aligned measures explicates a set of expectations about teachers' ratings on two instruments. For example, consider the Identity of Self alignment between I/T and PS (illustrated in Figure 2 and shown in the top row of Table 3). Levels 4 and 5 on I/T align with Level 1 on PS (i.e., 4/5 $\rightarrow$ 1). If a teacher rates a child at one of the aligned levels of the younger age instrument (i.e., 4 or 5 on I/T), we expect the teacher to rate the child at the appropriately aligned level on the older age instrument (i.e., 1 on PS). Naturally, if the child is rated below the lowest aligned level on the younger instrument (e.g., 1, 2, or 3 on I/T), we expect the rating on the older instrument to also be below the lowest aligned level (e.g., the "Not Yet" level in PS). The goal of our validity analysis is to assess the extent to which the data supported these expectations, as reflected in teachers' ratings and the psychometric properties of DRDP–R measures.

Collecting Evidence for Cross-Age Validity

The methodology we follow in validating the assumptions and expectations laid out by the alignment table is derived from methods for linking assessment instruments. In educational measurement, "linking" refers to various approaches for comparing outcomes from different assessments, such as *equating*, *calibration*, *statistical moderation*, and *projection* (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999; Linn, 1993; Mislevy, 1992). The general goal of these linking approaches is to find a set of transformations between students' scores on two related tests, so as to facilitate comparisons along a common scale.

The DRDP–R instruments target increasingly higher levels of development on the same constructs. This design calls for a *vertical scaling* approach to analysis, where data from tests that target increasing difficulties are calibrated to have the same scale. When assessments are vertically scaled, a content framework is needed to ensure that the same construct is measured by both assessments (Kolen, 2004). The DRDP–R articulation process and the resulting alignment table constitute such a framework. In this article, vertical scaling is used *not* to transform or equate ability scores but rather to calibrate the difficulty of measures' developmental levels across age-consecutive assessments. In other words, the goal of our study is to perform *level alignment* across instruments, rather than linking test scores. Validity evidence is then obtained by inspecting the extent to which the DRDP–R measures are aligned as expected and the ways in which they might deviate from these expectations.

A common design for vertical scaling studies is a "same rater, two tests" design, where a group of examinees take both tests so that both abilities and item difficulties can be scaled together (Kolen, 2004; Wright & Masters, 1981). For example, if a group of preschoolers is rated on both I/T and PS instruments, the ratings could be calibrated in a single analysis

under an Item Response Modeling framework. Consequently, difficulties of being rated at different developmental levels on the I/T instrument are ordered on the same scale as the PS difficulties. One would not expect the difficulties of the levels to perfectly align. Specifically, we expected the difficulties of the SA instrument to be higher overall than the PS difficulties, which in turn would be higher than I/T. However, because particular levels were designed to bridge across instruments, we expected to find alignment between difficulties of levels that are assumed to measure the same behaviors, as explicated in Table 1. By vertically scaling the DRDP–Rs we provide empirical evidence to support or refute these expectations. This analysis targets the third research question: Are difficulties ordered as in the alignment table? The same data can be analyzed to answer the second question: Are teachers rating as expected by the alignment table? Both of these research questions are addressed in the validity analysis results.

An assessment cannot be valid without being reliable. Because the DRDP–R data are obtained by teachers' observation-based ratings, the reliability of the instrument must be carefully studied. Some inconsistencies are expected because ratings are affected by many sources of error (e.g., rater bias, level of training and effort, clarity of the instrument, time of DRDP–R completion, experience). To answer the first question—Are teachers using the instruments reliably?—we collected data to evaluate the interrater agreement of the DRDP–Rs. The Method section describes the sample and procedures for data collection for all research questions. The Results section provides the DRDP–R reliability analysis, followed by the validity analysis.

METHOD

Procedure

In the spring of 2005 we conducted a study to establish the psychometric properties of the DRDP–R (for more details see *DRDP–R Technical Manual*; Berkeley Evaluation and Assessment Research [BEAR] Center, in press). Data for the study were collected by teachers from 140 state-founded programs, sampled based on their socioeconomic status within regions throughout California. More than 700 staff personnel were trained in observing children and completing the DRDP–R in several day-long sessions. Many of the teachers were already familiar with the DR system, with the previous version of the DRDP, and with methods for conducting naturalistic observations. The training seminars were carried out by WestEd, California Institute on Human Services, Mathematica Policy Research, American Institutes for Research, and BEAR personnel, and supervised by the Child Development Division of the CDE.

Following the training seminars, teachers returned to their programs to observe and evaluate a sample of the children under their care (the median number of rated children per teacher was three). Caregivers and teachers were asked to observe each child for at least 2 weeks before they began rating (but no longer than 60 days) and to complete the DRDP–R only for children who attended the program for at least 10 hours per week and who were under the rater's direct care for at least 30 days. This way, we ensured that observers had ample time to know the child before completing the assessment. Following the training sessions, the teachers had to

observe the child, complete the DRDP–Rs, and return them in prepaid envelopes. The teachers were compensated for their time and effort.

Participants

Two samples from the DRDP–R 2005 Calibration Study were used to analyze the reliability and validity of the instruments. First, participants in the "two raters, same test" condition were pairs of teachers who rated the same child using the same instrument, allowing an investigation of their level of agreement. An ideal setup for testing reliability would include several teachers rating multiple children under their care. However, child care program settings make such a design practically impossible to implement. Overall, 71 pairs of staff members rated an average of about 3 children per pair, for a total of 193 double-rated children in this condition (105 in I/T, 76 in PS, and in 12 SA). In the next section, we report on the interrater agreement analysis of these reliability data. This analysis included all of the measures from the three DRDP–Rs—105 in total.

For cross-age validation purposes, 263 children whose chronological age was close to the limits of their age group were rated by their primary caregiver in a "same rater, two tests" condition. Each child was assessed by the same rater using the relevant two instruments (e.g., preschool teachers used the I/T and PS instruments to rate the 3-year-olds, and the PS and SA instruments to rate the 5-year-olds). The teachers who participated in this condition were trained to use both instruments. The younger sample included ratings for 144 children, all of whom were approximately 3 years old. The older sample included ratings for 119 children, all of whom were approximately 5 years old.

To achieve better estimates of model parameters in the psychometric analysis, the data were augmented with 488 children who were rated only on the PS instrument. We refer to this sample as the *augmented* data. Overall, this data set included 751 children (51% female) with an approximately normal age distribution between the ages of (almost) 3 and (a little older than) 5. For 51% of the children, the language spoken at home was English, for 27% the language was Spanish, and for 14% it was both English and Spanish; other languages were spoken at the homes of the remaining 8%. The ethnic distribution was typical for child care programs in California: 55% Hispanic American, 13% European American, 13% African American, and 5% Asian American; the remaining 14% were categorized under other ethnicities. This analysis included only the aligned measures from the three DRDP–Rs, 88 in total.

RESULTS

Reliability Analysis

To evaluate interrater agreement levels, we first compared the ratings of each pair of teachers across all measures of the same child. This allowed us to evaluate how often and to what extent teachers agreed. An instance of perfect agreement occurred when the two teachers assigned the same rating on a single measure for the same child. An instance of adjacent agreement occurred when the ratings deviated by only one level. Any other situation was considered disagreement. The average proportion of perfect, adjacent, and no agreement in each instrument is presented

TABLE 2
Average Proportion (and Standard Deviation) of Agreement Types Across Teacher Pairs

	Perfect Agreement	*Adjacent Agreement*	*Total Agreement (Perfect + Adjacent)*	*No Agreement*
I/T[a]	0.58 (0.22)	0.35 (0.17)	0.93 (0.13)	0.07 (0.13)
PS[b]	0.54 (0.20)	0.38 (0.15)	0.92 (0.09)	0.08 (0.09)
SA[c]	0.67 (0.27)	0.24 (0.21)	0.92 (0.13)	0.08 (0.13)
Overall[d]	0.57 (0.21)	0.36 (0.17)	0.93 (0.11)	0.07 (0.11)

Note. I/T = Infant/Toddler; PS = Preschool; SA = School-Age.
[a]$n = 105$. [b]$n = 76$. [c]$n = 12$. [d]$N = 193$.

in Table 2. The bottom line shows that, overall, teachers tended to agree more than disagree. When disagreements did occur, the ratings were rarely more than one developmental level apart. The level of disagreement was about the same across the three DRDP–Rs, whereas users of the SA instrument had more instances of perfect agreement than I/T or PS. However, these differences should be regarded with caution due to the small SA sample.

To further analyze the instruments' reliabilities, we calculated the intraclass correlation coefficient (ICC) for each of the 105 measures in the DRDP–Rs. The average ICC (and standard deviation) was 0.66 (0.05), 0.62 (0.07), and 0.68 (0.14) for I/T, PS, and SA, respectively. The correlations for all of the I/T and PS measures were significant at $p < .00038$ (Bonferroni adjustment applied). Although the SA proportions of perfect agreement and ICC were relatively high, the small sample sizes inflated the ICC standard errors, causing only 23% of the SA measures to show significant correlations. The remaining SA measures were significant at the .001 level. Overall, these results support the claim that the DRDP–Rs can be used reliably by teachers.

Validity Analysis Based on Teachers' Rating Tendencies

We begin by summarizing the extent to which teachers in the "same rater, two tests" sample tended to rate children according to the cross-age expectations. The second section summarizes results from psychometric analyses within each indicator using the augmented data. The overall purpose of this analysis was to compare the order of difficulty of levels across instruments to their expected order predicted from the alignment table.

Teachers in this condition had to decide on which developmental level to place a child on two successive instruments. First, they rated the child using the age-appropriate instrument, and then they rated the child on the adjacent instrument. Therefore, the data came from two pairs—the I/T and PS instruments, and the PS and SA instruments. To study teachers' rating patterns, we categorized ratings based on their level of support for the expected alignment. Specifically, each teacher's pair of ratings was labeled as showing full support, partial support, or no support. We explain what these support categories mean using the Self-Esteem construct as an example, where the expected alignment between I/T and PS is 4,5 → 1,2.

Cases labeled as full support were instances where the ratings on two bridged measures were consistent with the expected alignment, such as (a) the pair of ratings were identical to one of

the expected alignments in Table 1 (e.g., 4 on I/T and 1 on PS, or 5 on I/T and 2 on PS) or (b) both ratings were below the aligned levels, in a manner consistent with expectations (e.g., 3 or below on I/T and "not yet" on PS). A pair of ratings would be considered as partially supporting the expected alignment if they conformed to one of two conditions: (a) the two ratings were within the range of aligned levels but did not exactly match one of the specified alignments (e.g., 4 on I/T and 2 on PS, or 5 on I/T and 1 on PS), or (b) the two ratings were one level apart from forming a perfect alignment (e.g., 3 on I/T and 1 on PS, or 5 on I/T and 3 on PS). Any other combination of ratings would be considered as not supporting the expected alignment (e.g., 3 on I/T and 2 on PS, or 4 on I/T and 4 on PS).

In the case of the Self-Esteem construct, there are 25 possible pairs of ratings overall because both instruments have five rating categories. One fifth of the pairs of ratings reflect full support of expectations, another fifth reflect partial support of expectations, and three fifths of the pairs reflect no support. Across all constructs, the average proportion of pairs of ratings that could contribute to the "no support" category was 60%. Thus, our analysis of validity support is conservative in the sense that there were potentially more instances that would help reject than support our expectations.

The percentage of observed ratings that fell in each of the three support categories was calculated separately for each indicator in each pair of instruments. The results were very similar between the two pairs of instruments. On average, 40% ($SD = 8\%$) of the ratings within each indicator showed full support, and another 41% ($SD = 10\%$) showed partial support for an overall total support of 81% ($SD = 9\%$). The remaining 19% ($SD = 9\%$) showed no support. These results suggest that teachers were at least twice as likely to rate children in a manner consistent with the alignment expectations than to rate them in a way that was inconsistent with these expectations. The only exceptions were the Motor Skills indicator for the I/T and PS pair, and the Safety and Health indicator for the PS and SA pair. In these cases, teachers' rating patterns tended to show no support rather than full support for expectations. Overall, the analysis of rating tendencies suggests that, for the most part, teachers interpreted the meaning of successive instruments' levels in ways that were similar to expectations based on the cross-age validity argument. Therefore, the theoretical alignment between adjacent instruments is empirically supported.

Validity Analysis Based on Instrument Scaling

Psychometric analysis. The augmented "same rater, two tests" sample allows for a common scaling of the three DRDP–R instruments. The psychometric analysis was used to produce estimates of the difficulty of attaining different developmental levels. Support for a cross-age validity argument was found if difficulty estimates of developmental levels from successive instruments were aligned as predicted by the alignment structure. That is, for measures that target the same construct, the probability of rating a given child at a pair of aligned levels should be similar across pairs of instruments.

The psychometric analyses of the "same rater, two tests" sample were done separately for each of the nine indicators,[5] using all of the relevant aligned measures across the three DRDP–

[5]We did not perform this analysis on the REG indicator because it had only one set of aligned measures and therefore did not conform to psychometric requirements.

Rs. We fitted the Random Coefficients Multinomial Logit Model (Adams, Wilson, & Wang, 1997) using *ConQuest* (Wu, Adams, & Wilson, 1998). The Random Coefficients Multinomial Logit Model is an extension of the Rasch family of item response models, which enable estimation of abilities and item difficulties on a common scale. Within this formulation, a Partial Credit model (Masters, 1982) was employed, so the difficulty of attaining each developmental level could be estimated separately for each measure (as opposed to estimating only one general difficulty per measure). The partial credit model allows for variation between difficulties of developmental levels of different measures of the same indicator. Each unidimensional analysis produced ability estimates for the children and step difficulty estimates for the developmental levels of all measures.

Due to the probabilistic nature of the model, it is expected that occasionally a child will be rated higher or lower than his or her true ability. Therefore, we calculated item-fit statistics to measure the extent to which the psychometric model captures the observed variability in the data. The weighted mean square error (WMSE) compares the variability in teachers' ratings to the variability expected by the model, given the distribution of ability scores. Values of WMSE larger than 1 indicate more variability than expected, meaning that over- or underrating of children abilities is too common. Values lower than 1 indicate less variability than expected, meaning that teachers rate children in a manner that is too consistent with the model's prediction. Generally, values between 0.75 and 1.33 are considered to indicate satisfactory item fit (Adams & Khoo, 1996). Overall, the DRDP–R measures showed good fit within the desired range. The average WMSE was 0.96 ($SD = 0.14$), and only 6 out of 85 measures had WMSE slightly outside the range.

Alignment of developmental level difficulties as evidence for cross-age validity. The psychometric analysis orders the developmental levels of measures on a single continuum across age groups. For a specific construct, if Level 5 in the I/T measure is approximately as difficult to reach as Level 1 in PS, then the two measures are aligned at those levels. The difficulty of reaching each level is best represented by the Thurstonian threshold. For a measure with *K* developmental levels, there are *K-1* thresholds (one between each successive pair of levels). Threshold *k* is defined as the point on the ability continuum where there is a probability of .5 for achieving level *k* or more on the measure (Wu et al., 1998). Within each measure, the thresholds are ordered with respect to their logit values,[6] as are the children's abilities. In Figure 3, children's abilities are represented by triangles aligned with each threshold. The probability that a teacher would rate a child at any of the levels is represented by the curved line in the figure. For example, the child at the *k*th threshold has a probability of .5 to be rated *at least* at level *k*. The probability that the child will be rated at lower levels (e.g., $k - 1$, $k - 2$) is larger than .5. The probability that the child will be rated at higher levels (e.g., $k + 1$, $k + 2$) is smaller than .5.

The thresholds define the bandwidths of developmental levels within each measure. The *k*th threshold represents the upper bound of the *k*th developmental level. For example, Figure 4 juxtaposes the developmental bandwidths of the aligned measures of the Identity of Self construct for the I/T, PS, and SA measures. The steps in each "ladder" in the figure represent

[6]Technically, logit is the log of an odds ratio. In this case, the logit is the log of the ratio between probabilities of correct and incorrect response.

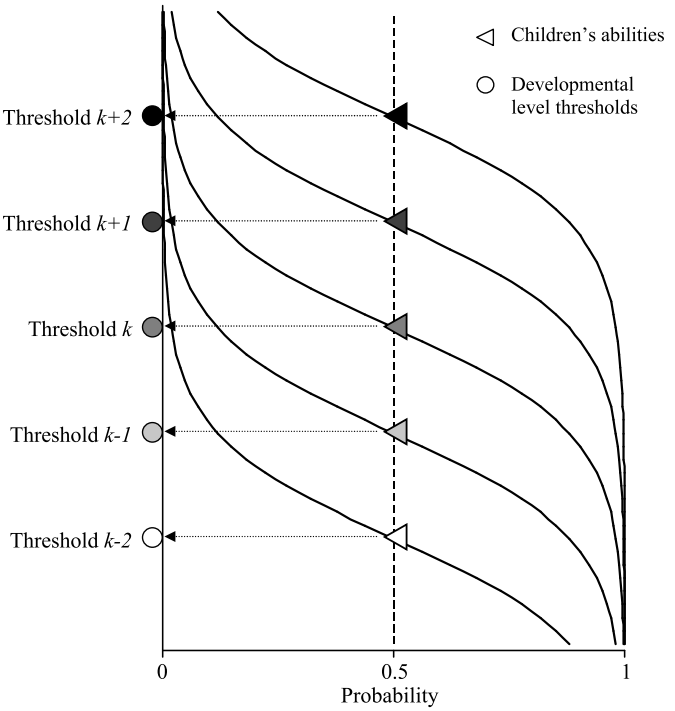

FIGURE 3 Example of Thurstonian thresholds.

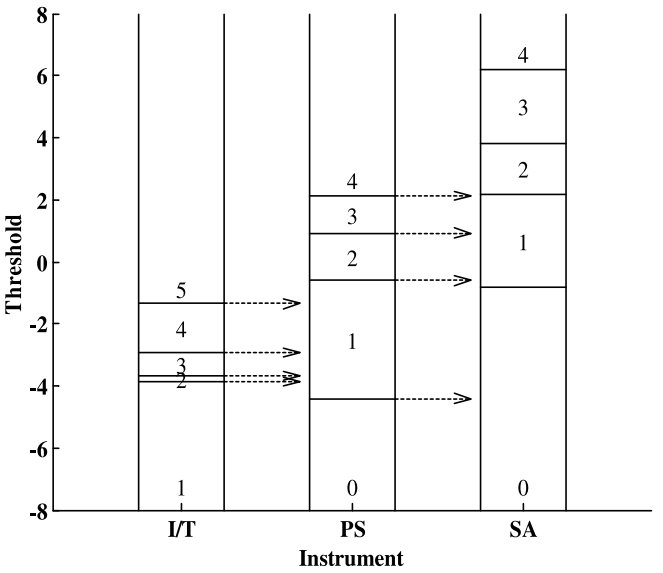

FIGURE 4 Developmental levels' bandwidths for the Identity of Self construct across the three Desired Results Developmental Profiles–Revised. *Note.* I/T = Infant/Toddler; PS = Preschool; SA = School-Age.

the thresholds obtained from fitting the partial credit model to aligned measures of the SELF indicator. For example, in the PS instrument, the line between the 0 and 1 labels (roughly −4.4 logits) indicates where the "not yet" level ends and the first developmental level begins.[7]

In Figure 4, the highest threshold in PS is at 2.1 logits and defines where the third developmental level ends and the fourth begins. Because thresholds define the upper bound of their corresponding levels, the bandwidth for the last level has no upper bound. This makes sense because no matter what value of ability the child possesses above the last threshold, he or she will have to be rated at the highest developmental level on this measure. We discuss several examples of how levels of aligned measures show different degrees of overlap, and thus provide different degrees of support for the expectations specified by the validity argument.

We inspected the match between levels' overlap and expected alignment in order to determine whether expectations were fully, partially, or not supported. For example, the SA instrument targets self-identity conceptions typical for school-aged children. We expected the SA levels to be harder to attain than the PS levels, yet align in accordance with the specifications in Table 1. Specifically, Levels 2 and 3 in PS should overlap with Level 1 in SA, and Level 4 in PS should overlap with Level 2 in SA. Indeed, this expectation was fully supported by the analysis results, as shown in Figure 4. In addition, lower levels overlapped as expected (i.e., Level 1 and "not yet" in PS overlap with "not yet" in SA).

When all of the levels between a pair of instruments aligned as expected, the results were considered as fully supporting the construct's cross-age validity argument. When some levels aligned as expected and some levels did not, the results were considered as providing partial support for the validity argument. Figure 4 shows that the expected alignment between the I/T and PS instruments (4/5 → 1) was only partially supported by the results. Although Level 4 was well within the range of Level 1, Level 5 only overlapped with a small portion of Level 1. Contrary to expectations, levels below the expected alignment (e.g., I/T Levels 2 and 3) also overlapped with Level 1. Overall, these findings provided partial support for our predictions.

Figure 5 shows the developmental bandwidths for the Self-Esteem construct. The expected alignment (from Table 1) is 4,5 → 1,2 between I/T and PS, and 4 → 1 between PS and SA. Here, the results partially supported the first alignment. Although Levels 4 and 5 overlapped with most of Levels 1 and 2, respectively, there was also some overlap between Level 4 and Level 2. In addition, Levels 2 and 3 in I/T also overlapped with Level 1 in PS, contrary to the "below alignment" expectations. Moreover, the results showed no support for the alignment between PS and SA. Level 4 in PS aligned with Level 2 in SA and had almost no overlap with Level 1. Therefore, the expected alignment between these two instruments was not supported. Across the three instruments, the Self-Esteem construct received inconsistent support. About half of the constructs had a complex support pattern across the measures from different age groups, where some of the expected alignments were fully supported and others were either partially supported or not supported.

For the constructs in Figures 4 and 5, the I/T and PS measures showed more overlap than expected, meaning the PS instrument's levels seemed easier than they were supposed to be. Other constructs showed less overlap than expected, meaning the PS instrument's levels seemed too difficult. For example, for the Uses of Language in Conversation construct, the expected

[7]For the sake of clarity, all graphs shown in this article range between −8 and 8; thus the "Not Yet" level usually seems to span a wide range of logit values.

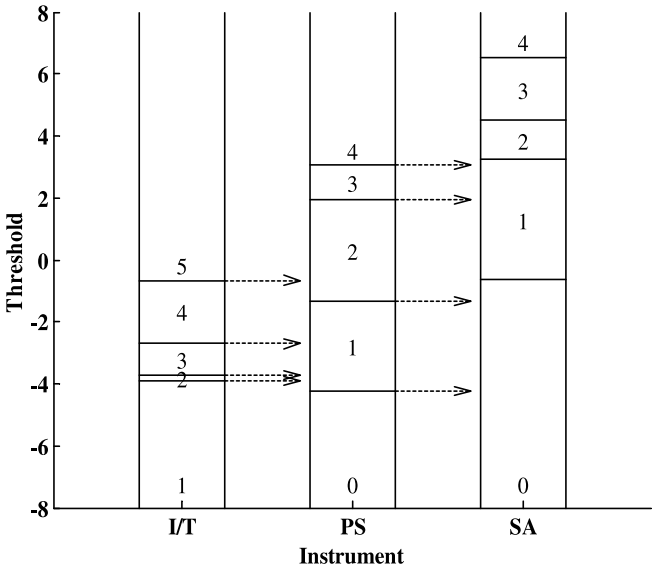

FIGURE 5 Developmental levels' bandwidths for the Self-Esteem construct across the three Desired Results Developmental Profiles–Revised. *Note.* I/T = Infant/Toddler; PS = Preschool; SA = School-Age.

I/T and PS alignment was 5,6 → 2,3 (there was no alignment with SA). However, as shown in Figure 6, Level 5 in I/T overlapped with Level 1 in PS, and Level 6 overlapped with Level 2. The measures showed less overlap than expected, indicating no support for this alignment.

Table 3 presents the patterns of validity support across consecutive DRDP–R instruments within each indicator. For example, the SELF indicator has two measures bridging each pair of instruments, for a total of four alignments. The table shows that expectations regarding the two alignments between I/T and PS were partially supported—one alignment between PS and SA received full support and the other received no support. Overall, 33% of the expected cross-age alignment structure was fully supported by the psychometric analysis, 54% of the structure was partially supported, and 13% was not supported at all. In total, about 87% of the cases showed either full or partial support for the cross-age validity argument.

Another way to inspect the results is to consider the support for individually aligned levels (as opposed to measures). For example, in Figure 5 the overall alignment between I/T and PS was only partially supported, as explained earlier. However, if we consider individual aligned levels, we find that one level was fully supported (Level 5) and the other was partially supported (Level 4). Overall, 50% of individual levels received full support and 27% received no support. The remaining 23% of levels received partial support, meaning they overlapped with the expected level *and* with the level below or above it. The number of cases that overlapped above the expected level was approximately equal to the number of cases that overlapped below it. This implies that across the age groups, some measures were harder than expected, and others were easier than expected, but there was no specific tendency of DRDP–R measures as a whole to be either too easy or too difficult.

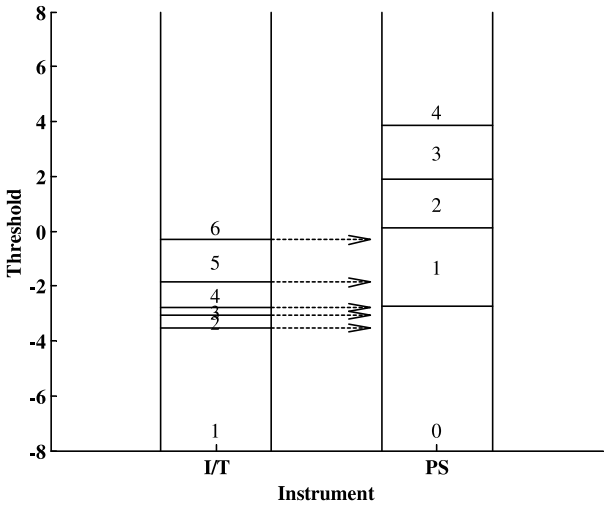

FIGURE 6 Developmental levels' bandwidths for the Uses of Language in Conversation construct across the three Desired Results Developmental Profiles–Revised. *Note.* I/T = Infant/Toddler; PS = Preschool.

TABLE 3
Types of Cross-Age Validity Support Based on the Psychometric Analysis

	Support Type					
	Full		*Partial*		*None*	
Indicator (No. of Measures)	*I/T → PS*	*PS → SA*	*I/T → PS*	*PS → SA*	*I/T → PS*	*PS → SA*
SELF (2)	—	1	2	—	—	1
SOC (6)	2	1	3	3	—	—
LANG (4)	2	1	1	2	1	—
LRN (2)	—	1	2	1	—	—
COG (4)	1	1	2	2	—	1
MATH (6)	1	3	2	1	1	1
LIT (3)	1	1	1	2	—	—
MOT (3)	—	1	1	1	2	—
SH (2)	1	—	1	2	—	—
Overall	8	10	15	14	4	3
Proportion	0.33		0.54		0.13	

Note. I/T = Infant/Toddler; PS = Preschool; SA = School-Age; SELF = Self-Concept; SOC = Social Interpersonal Skills; LANG = Language; LRN = Learning; COG = Cognitive Competence; MATH = Math; LIT = Literacy; MOT = Motor Skills; SH = Safety and Health.

CONCLUSIONS

Articulation of successive developmental assessments is important for enhancing the sensitivity of assessment systems to various developmental paths, and increasing their usefulness in child care programs. The DRDP–R was presented as an example of a complex assessment system that was designed to track children's development while maintaining continuity of constructs across childhood. This article describes how articulation was achieved by aligning developmental levels from adjacent instruments using theoretical and practical considerations. The alignment table, which specifies assumptions about which levels should align, was used to form expectations about how children who were transitioning between programs were expected to be rated. The instruments' alignment structure is akin to an interpretive validity argument which explicates the cross-age inferences that can be made using the DRDP–R. We used "two raters, same test" and "same rater, two tests" data to analyze the instruments' reliability and cross-age validity. Our analyses show that, overall, the three DRDP–Rs have reasonable levels of reliability and validity evidence, although variations exist among indicators and among instruments.

The results of the "same rater, two tests" analysis indicate that when teachers complete the adjacent-age instrument, in the main, they tend to rate children according to the expected alignment between the instruments. That is, most children at the transition age are rated either at or below the aligned levels. However, in some cases, the expected alignment is not empirically supported. For example, a child may be rated above the highest aligned level on the older instrument and below the lowest aligned level on the younger instrument. Such a rating pattern is inconsistent with the assumption that the instruments are articulated at the aligned levels. The results indicate general support for the DRDP–R alignment structure presented in Table 1, but also quite a bit of variation from expectations—probably due to variation in how teachers rate children, and perhaps due to experimental effects (e.g., the situation of rating a child twice using different, but related, instruments, is relatively unusual).

In Figures 4 through 6 we presented results from the psychometric analysis, which sheds light on the underlying unidimensional latent variables that aligns measures of the same indicator across successive instruments. The analysis shows that for the majority of the constructs, the DRDP–R alignment structure is either supported or partially supported by the data. In other words, levels that we identified to mean the same thing were estimated to be similarly difficult to attain. In the rest of the cases, the observed overlap was either lower or higher than the expected alignment, usually by no more than one developmental level. When expectations were not met by the data, the overlap usually shifted, so the first level in the older-age instrument was aligned with one level lower or higher than expected in the younger-age instrument (although the DRDP–Rs did not show any specific bias to either direction of misalignment).

There could be many reasons why this shift occurred. First, we may have identified the wrong alignment structure. Although this might be true in some cases, it is probably not true in others. Specifically, some aligned levels were worded in exactly the same way but were still estimated to have somewhat different difficulties. In this case, it may be the presence of the other categories, or the age-specific examples in the measure, that made the difference. Second, teachers' ratings on different constructs showed various patterns of children's locations on the constructs—that is, a relatively large presence of either low- or high-achieving children in the sample. These sample biases could have had a downward or upward "pulling" effect on the

thresholds, causing level misalignment. In most cases, the misalignment was weak, and the thresholds showed partial support. In a few cases, the misalignment was too strong to warrant support for validly. Finally, teachers may have misunderstood the behaviors described in the levels or misunderstood the purpose of the common-purpose study. Such misunderstanding may have caused some teachers to rate children in the same manner (i.e., at the exact same numerical level) on both instruments. These instances were less than 1% for the alignments between I/T and PS and about 14% between PS and SA, indicating that teachers in those programs may have had greater difficulty working with two adjacent instruments.

One limitation of our study design is that some cases that could indicate even greater agreement within and between raters have been intentionally ignored. Specifically, natural variation among the rated children implies that some teachers are likely not to find suitable levels to describe the child on one of the instruments. When a child performs behaviors below the lowest aligned level on the younger instrument, the teacher can always rate him or her at the "not yet" level on the older instrument, which we considered as supporting our expectations. However, for a child that performed behaviors above the highest aligned level on the older instrument (e.g., rated 4 on SA when the highest aligned level is 2), the teacher would not be able to find an appropriate rating level on a younger instrument. The teacher would then be forced to choose the highest level on the younger instrument (e.g., rated 4 on PS). Such a rating pattern does not seem to support the expected alignment. However, this lack of support is misleading, because the teacher did recognize, accurately, that the child should not be rated below the highest level on the younger instrument. Unfortunately, it is impossible to distinguish between this supportive case and a case where the teacher simply rated the child on both instruments, without paying attention to what the levels mean. Therefore, to maintain a conservative analysis of validity, these cases were considered as providing no support. In fact, some teachers opted to mark the "emergent" box for advanced children that were rated at the highest level on the younger instrument, in this way indicating that the child should be rated using an older instrument. However, the data about children's emergent ratings were also not used in our analyses, because these ratings were optional. In this respect as well, our analyses produced conservative estimates of the true level of full support.

The level alignment procedure described in the previous text has been applied to a "typical" assessment context. Where a similar question is being investigated in a special education context, a different sample of children would be needed—a sample spanning the range of children in special education programs. As these programs are often quite different from one another, it would be imperative to apply the level alignment procedures within appropriate subcategories of such children, as the results may differ significantly. This has not been done for the DRDP–R.

The main conclusion from this study is that the three parts of the DRDP–R are aligned so that within each age group, children at various levels of development can be readily assessed. In other words, each instrument provides levels of assessment for the typical children in the age group it was designed to assess, as well as relatively low-achieving and relatively high-achieving children in this age group. Although this is very reassuring for the current version of the DRDP–R, there is a need for further research on the measures that do not align as expected. Moreover, future versions of the DRDP–R will include alignment to state standards, newer research findings, and feedback from practitioners. It is crucial that any revisions to one instrument be applied to the aligned levels of the other instruments, to the extent possible.

Establishing the theoretical and empirical alignment structure of the three instruments is an important contribution to the validity of the DRDP–R assessment system as a unified whole.

REFERENCES

Adams, R. J., & Khoo, S.-T. (1996). *Quest: The Interactive Test Analysis System*. Melbourne: Australian Council for Educational Research.

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Bergen, D. (2002). The role of pretend play in children's cognitive development. *Early Childhood Research & Practice, 4*. Retrieved May 1, 2009, from http://find.galegroup.com.proxy.bc.edu/itx/start.do?prodId=AONE

Berkeley Evaluation and Assessment Research Center. (in press). *Desired Results Developmental Profile–Revised Technical Manual*. Berkeley: University of California, Berkeley.

California Department of Education. (2008). *California preschool learning foundations* (Vol. 1). Sacramento, CA: Author. Retrieved May 1, 2009, from http://www.cde.ca.gov/sp/cd/re/documents/preschoollf.pdf

Case, R. (1991). Stages in the development of the young child's first sense of self. *Developmental Review, 11*, 210–230.

Damon, W., & Hart, D. (1988). *Self-understanding in childhood and adolescence*. New York: Cambridge University Press.

Delahunt, J. Z. (2002). Motor development. In N. J. Salkind (Ed.), *Macmillan encyclopedia of child development* (pp. 279–282). New York: Macmillan.

Eder, R. (1990). Uncovering children's psychological selves: Individual and developmental differences. *Child Development, 61*, 849–863.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press.

Fischer, K. W. (1980). A theory of cognitive development: the control and construction of hierarchies of skills. *Psychological Review, 87*, 477–531.

Harter, S. (1999). *The construction of the self: A developmental perspective*. New York: Guilford.

Kagan, J. (1991). The theoretical utility of constructs for self. *Developmental Review, 11*, 244–250.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342.

Kolen, M. J. (2004). Linking assessments: Concept and history. *Applied Psychological Measurement*, *28*, 219–226.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, *6*, 83–102.

Marsh, H., Ellis, L., & Craven, R. (2002). How do preschool children feel about themselves? Unraveling measurement and multidimensional self-concept structure. *Developmental Psychology, 38*, 376–393.

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

Masters, G. N., & Forster, M. (1996). *Progress Maps: Assessment Resource Kit*. Camberwell, Australia: Australian Council for Educational Research.

Measelle, J., Ablow, J., Cowan, P., & Cowan, C. (1998). Assessing young children's views of their academic, social, and emotional lives: An evaluation of the self-perception scales of the Berkeley Puppet Interview. *Child Development, 69*, 1556–1576.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.

Metzuk, B., Parrish, D., & Mangione, P. (2006, April). *Overview of the desired results for children and families system*. Paper presented at the AERA annual meeting, San Francisco, CA.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: ETS Policy Information Center.

Pomerantz, E. M., Ruble, D. N., Frey, K. S., & Greulich, F. (1995). Meeting goals and confronting conflict: Children's changing perceptions of social comparison. *Child Development, 66*, 723–738.

Shepard, L., Kagan, S. L., & Wurtz, E. (1998). *Principles and recommendations for early childhood assessments* (Goal 1 Early Childhood Assessments Research Group). Washington, DC: National Education Goals Pane.

Stern, D. N. (1985). *The interpersonal world of the infant: A view from psychoanalysis and developmental psychology*. New York: Basic Books.

Thompson, R. A. (2006). The development of the person: Social understanding, relationships, conscience, self. In W. Damon & R. M. Lerner (Eds.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (6th ed., pp. 24–98). New York: Wiley.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Wright, B. D., & Masters, G. N. (1981). *The measurement of knowledge and attitude*. Chicago: Department of Education, University of Chicago.

Wu, M., Adams, M. J., & Wilson, M. (1998). *ACERConQuest* [Computer program & manual]. Hawthorn, Australia: ACER Press.