Reliability Evidence for the Hawaii Early Learning Profile Birth-3 Years: Interrater Agreement of Child Assessment Crediting - Final Report

Michael D. Toland, PhD[1]

Caroline Gooden, MS

Zijia Li, MS

Department of Educational, School, & Counseling Psychology

University of Kentucky

September 2015

Recommended citation:

Toland, M. D., Gooden, C., & Li, Z. (2015, September). *Reliability evidence for the Hawaii Early Learning Profile Birth-3 Years: Interrater agreement with child assessment crediting - final report*. Lexington, KY: University of KY.

[1]Development of this technical report was completed at the University of Kentucky in the Department of Educational, School, and Counseling Psychology, funded by a grant awarded to Dr. Michael Toland (PI) to provide VORT Corporation with a psychometric project from 2014-2015. The statements made herein are those of the researchers and are not meant to represent opinions or policies of the funding agency.

Abstract

This study provides preliminary evidence on the interrater agreement of the Hawaii Early Learning Profile Birth-3 (HELP® 0-3). Interrater agreement refers to the degree to which item crediting from independent providers are interchangeable, that is, the extent to which individual providers provide essentially the same crediting. This study aimed to measure the degree of agreement birth to three providers had with each other and with a HELP® 0-3 expert, and, whether this agreement varied based on providers with or without formal HELP® 0-3 training. Eighty-two providers observed and credited 36 video-recorded clips of children exhibiting specific HELP® skills using the HELP Strands' definitions and credit criteria from the *Inside HELP*® (Administration and Reference Manual for HELP Birth – 3 Years; Parks, 1992, 2006). Most interrater agreement statistics ranged from 90% to 100%. Specifically, it was impressive that 32 of 36 (88%) video-recorded cases had agreement statistics that were above 90%. Results did not show differences in crediting based on provider training. Overall, these findings suggest that assessment crediting for the HELP® Strands when used with the *Inside HELP*® administration manual are highly dependable and consistent across providers.

Reliability Evidence for Hawaii Early Learning Profile Birth-3 Years: Interrater Agreement with Child Assessment Crediting - Final Report

Introduction

The psychometric properties of instruments designed for child assessment and planning are important considerations in appropriate use (Bisceglia, Perlman, Schaack, & Jenkins, 2009; Colwell, Gordon, Fujimotoa, Kaestner, & Korenman, 2013; Lambert, Kim, & Burts, 2013). The HELP® 0-3, which is the focus of this study, is a criterion-referenced, curriculum-based assessment designed for use with children aged birth to 3 years. The HELP® Strands 0-3 with the administration manual, *Inside HELP®*, have been widely used in early childhood education. When a birth to three professional assesses a child for progress monitoring and individualized planning, it is important that results be stable and trustworthy. Reliability analyses can provide evidence for the stability, consistency, and agreement of assessment results across sets of providers, items, and time.

Background Information on the Development of HELP® 0-3

The original Hawaii Early Learning Profile materials [*HELP® Activity Guide* (Furuno et al.,1985-2005); *HELP® Charts* (Furuno et al., 1985-2004); *HELP® Checklist* (Furuno et al., 1985-2014)] were developed through a federal demonstration and training project that began in the 1970's. A multidisciplinary pediatric team at the Hawaii School for Public Health selected 685 skills and behaviors for assessment and curricular activities, based on numerous developmental scales and standardized tests. Newer HELP® materials have been subsequently developed with multidisciplinary teams, including *Inside HELP®* (Parks, 1992, 2006); *HELP® Strands 0-3* (Parks, 1992-2013), *HELP® Family Centered Interview* (Parks, 1994-2006)*; HELP® at Home* (Parks, 1988, 2006); and *HELP® When the Parent Has Disabilities* (Parks,

1984, 1999). The original 685 skills and six domains were maintained but restructured in the *HELP® Strands* and accompanying *Inside HELP®* administration manual by sub-dividing the major six domains into 58 developmentally-sequenced, conceptual strands and by adding a Regulatory/Sensory Organization section. Definitions, suggested assessment process, and credit criteria were added during the restructuring process to promote consistency among users. There are no prior published reports of interrater reliability for using HELP® 0-3 as a curriculum-based assessment tool (Andersson, 2004).

Purpose

The purpose of this study was to investigate interrater agreement of HELP® Strands 0-3 skill assessment crediting among birth to three providers when assessing children on HELP® 0-3 skills using skill definitions and credit criteria from *Inside HELP®*, the administration manual. Interrater agreement refers to the degree to which crediting from independent providers are interchangeable or the extent to which individual providers provide essentially the same crediting. Provider creditings were compared to each other and to expert crediting (as defined by a birth to three professional and author of *Inside HELP®*). To understand the procedures used when providers conducted HELP® crediting, participants with varied levels of experience and training with the HELP® Strands 0-3 were solicited. The primary research questions for this study were:

How reliable are the credits assigned by birth to three providers for the assessment of infants and toddlers using HELP® Strands 0-3?

Are there differences in the degree of agreement for birth to three providers based on the presence, absence, or type of training and experience with the HELP® Strands 0-3?

Method

Participants

In this study, 82 birth to three providers from a national sample of providers (with varied disciplines, backgrounds, training, and experience working with children and administering the HELP®) agreed to participate (see Table 1). Of the 82 participants, most (96%) identified as White and female. Ninety percent identified as non-Hispanic. Fifty-eight percent received HELP® Strands 0-3 formal training (including 35% from the University of Kentucky's online Introduction to HELP® course and 15% from onsite training with verified HELP® trainers); 43% had no formal training with the HELP® Strands 0-3. Most participants (68%) identified their work location as a birth to three agency or early Head Start program (12%). Fifty-six percent of the full sample identified as developmental interventionists or teachers; 30% were therapists. Most providers had a Master's degree (70%), and represented a wide geographic area including 24 states; KY and IL had the largest representation. Most (79%) used the HELP® Strands 0-3 to conduct direct assessments for at least one continuous year; they varied in their length of experience in administering the HELP® 0-3 Strands from 0 to 25 years (*Mean* = 6.13) and for the number of children assessed with HELP® 0-3 Strands in the last 24 months (*Mean* = 62). All interrater reliability data were collected in January 2015.

Measure

The HELP® Strands 0-3 is an early childhood curriculum-based tool used to assess child functioning from the ages of birth to 3 years. HELP® Strands 0-3 includes developmental skills and behaviors across six major domains with a section on child regulatory/sensory organization. The administration manual provides definitions and credit criteria for each skill and behavior. Credit options include: 0 [(-) = *not observed or reported in any situation*], 1 [(+/-) = *emerging,*

not considered as fully part of repertoire; may have learned during assessment period in imitation; needs reminders; for partially displayed skill], or 2 [(+) = *present as defined in Inside HELP® by observation or caregiver report as part of the child's typical functioning across familiar settings*]. Skills can also be credited as atypical [(A) = *Atypical, dysfunctional, or quality concerns that interfere with development and everyday functioning*] or not applicable [(N/A) = *not appropriate to assess due to age, disability, cultural or functional relevance, or family preference*].

Development of video recording of skills

Initial study design included the development of videos for each HELP® 0-3 strand, with each video capturing a single skill or behavior during daily activities in natural environments. The use of videos of children gave the advantage of providing all providers with the same evidence, making the videos useful for measuring differences in professional judgement (Peabody, Luck, Glassman, Dresshaus, & Lee, 2000). A total of 36 skill video assessments were selected that represented skills across the HELP® 0-3 six major domains and Regulatory/Sensory Organization section. Limiting the number of videos to 36 allowed providers time to view the skills displayed in each video and minimized respondent burden.

A pilot version of the video crediting tasks was tested with 15 graduate students enrolled in a measurement course at a large university. None of the volunteer students were experts in early childhood research or served as providers. Their feedback was helpful for the perspective of persons with little to no knowledge of the HELP® Strands 0-3. Students provided feedback about the videos with respect to clarity of instructions, navigation of the tasks, and audio and visual quality. Overall, the feedback indicated no major concerns with task instructions or clarity of videos. However, one of the gross motor skill videos included 2 children which may have

created confusion on which child to credit. The item was retained, but flagged during analysis for closer inspection. Feedback was provided on how to clarify directions, and was incorporated into the final crediting videos. After pilot testing, full development of the video crediting tasks was completed.

Careful selection of videos ensured variation in child gender, age, level of functioning, and race (see Table 2). Videos included children with familiar adults in everyday activities and in targeted intervention sessions. The 36 videos included 21 different children, 52% of whom were girls and 75% of whom were White. The ages of the children ranged from birth to 1 year (24%), 1 to 2 years (38%), and 2 to 3 years (38%). All videos were recorded in the child's home setting. The majority of children (67%) had an Individual Family Service Plan (IFSP; i.e., a delay or disability). The author of *Inside HELP*® credited each video to provide a standard for correct responses by providers.

Procedure

All providers credited the same 36 videos independently. Each provider was emailed a unique link via Survey Monkey to complete the HELP® Strands 0-3 interrater agreement task. The survey did not provide training about use of the HELP® Strands 0-3; rather, the instructions included the types of browsers recommended, tips for moving through the videos efficiently, and procedures for crediting a child within each video. For each video-recorded skill, two online pages were presented. At the beginning of the survey, participants answered several demographic questions. On the first page for each skill, participants were introduced to the skill and its definition from *Inside HELP*®, within the context of the applicable HELP® Strand (i.e., Cognitive Symbolic Play). On the second page, providers watched a brief video of a child and then selected the appropriate credit for the skill being assessed. After completion of the online

crediting tasks, participants were thanked for their time and mailed (at no charge) an honorarium for their participation. The honorarium consisted of either a check for $50 or a copy of *Inside HELP*® (valued at $65).

Data Analysis

Interrater agreement of item crediting was examined by correlating birth to three provider results among the providers and with the expert crediting for each item, so that the degrees of agreement could be determined. This analysis was done for each item within each domain for the entire sample, and examined for participants who did and did not have formal HELP® Strands 0-3 training. Statistics are reported for percent agreement, as well as for 95% bootstrap corrected accelerated (k = 1,000) 95% confidence intervals (CIs).

Results

We first report the overall interrater agreement among providers and according to provider training status (formal or no formal training on HELP® Strands 0-3) for each of the 36 video recorded skills. Secondly, we report the provider agreement with the expert credit and according to provider training status.

The degree of item interrater agreement was represented by the percentage of agreement among 82 providers on each item; the degree of domain interrater agreement was represented by the average percentage of items that were sampled from the corresponding domain (Table 3). Specifically, at the item level, the crediting agreement ranged from a high of 100% (for Fine Motor skill #s 4.74, 4.78; Social-Emotional skill # 5.04; Self Help skill #s 6.19, 6.62; Regulatory/Sensory Organization skill #s 1.68, 1.69, 5.60) to a minimum of 70.7% (for Gross Motor skill # 3.79) for the overall sample (N = 82). The vast majority of statistics showed provider credits had high agreement regardless of domain. Over 88% of the sampled items had

agreement levels greater than 90%. At the domain level, the average crediting agreement was excellent, ranging from 100% (Regulatory/Sensory Organization) to 88.3% (Gross Motor). Four domains, including Fine Motor, Social-Emotional, Self-Help, and Regulatory/Sensory Organization (section), had almost perfect agreement (greater than 98%). Two domains (Cognitive and Language) had agreement levels of 93%. The only domain with average agreement below 90% was the Gross Motor domain (88.3%), which is considered desirable in agreement studies.

With regard to provider agreement with the expert, the average numbers of items (*M*) for which the providers agreed with the expert credit were examined. The results in Table 3 show that the agreement level was high at 33.29 across all 36 case video skills, with 95% bootstrap corrected accelerated CI [32.95, 33.63]. Moreover, the results showed no significant agreement differences for trained vs. untrained providers according to the amount of provider training (i.e., with formal training $M = 33.06$ out of 36; 95% CI [32.39, 33.63]; without formal training $M = 33.37$; 95% CI [32.96, 33.78]).

Discussion

This is the first interrater agreement study of the HELP® Strands 0-3. The study included 82 providers with a wide diversity of professional backgrounds, geographic locations, amount of training, and years' experience with the HELP® Strands 0-3. The main findings are that interrater agreement among providers was high to moderate for a diverse sample of video-recorded skills as credited by birth to three providers with and without formal HELP® Strands 0-3 training (see Table 3). It is impressive that 32 of 36 (88%) video-recorded skills had agreement levels that were above 90%. Only the Gross Motor domain had two skills with agreement statistics below 80%. A possible explanation is that one of the gross motor skill

videos presented 2 children which may have caused some confusion on which child to credit, indicating a production error and not a problem with provider crediting. Excluding this video, all agreement statistics were well above 80%, which is often deemed the minimum level of agreement to suggest interrater agreement (McMillan & Schumacher, 2001, p. 66). This trend was observed for providers with and without formal HELP® Strands 0-3 training (see Table 3). Overall, the results suggest that provider ratings in this study are highly dependable, and that we can expect providers rating children with the HELP® Strands 0-3 to provide essentially the same credit for a given skill.

Furthermore, while the results show no difference for the amount of training providers received on the HELP® Strands 0-3, this finding does not suggest that formal training should stop. The *Inside HELP*® administration manual outlines clear procedures including definitions, credit criteria, adaptations, and materials for assessment with the HELP® Strands 0-3. Training should continue to improve consistency in crediting; to better familiarize providers with the HELP® Strands 0-3; and to improve assessor skill development, relationship building, and development of functional outcomes as part of the family centered HELP® 0-3 curriculum-based assessment process. Finally, there are not any previous studies with which to compare these results, as this is the first formal study to look at interrater agreement for the HELP® Strands 0-3. Further studies are needed to show that these results are not sample-specific and are replicable.

References

Andersson, L. L. (2004). Appropriate and inappropriate interpretation and use of test scores in early intervention. *Journal of Early Intervention*, *27*(1), 55-68.

Bisceglia, R., Perlman, M., Schaack, D., & Jenkins, J. (2009). Examining the psychometric properties of the infant-toddler environmental rating scale-revised edition in a high-stakes context. *Early Childhood Research Quarterly, 24,* 121-132. doi:10.1016/j.ecresq.2009.02.001

Colwell, N., Gordon, R. A., Fujimotoa, K., Kaestner, R., & Korenman, S. (2013). *Early Childhood Research Quarterly*, *28*, 218– 233.

Furuno, S., O'Reilly, K. A., Hosaka, C. M., Inatsuka, T. T., Zeislofr-Falbey, B. (1985-2005). *Hawaii Early Learning Profile (HELP) Activity Guide*. Palo Alto, CA: VORT.

Furuno, S., O'Reilly, K., Inatuka, T., Hosaka, C., Allman, T., & Zeisloft-Falbey, B. (1985-2004). *HELP® Charts.* Palo Alto, CA: VORT.

Furuno, S., O'Reilly, K., Hosaka, C. M., Inatsuka, T., & Zeisloft-Falbey, B. (1985-2014). *HELP® Checklist.* Palo Alto, CA: VORT.

Lambert, R. G., Kim, D-H., & Burts, D. C. (2013). Technical manual for the Teaching Strategies GOLD® assessment system. *Center for Educational Measurement and Evaluation.* Charlotte, NC: University of North Carolina.

McMillan, J., & Schumacher, S. (2001). *Research in education* (5th ed.). New York, NY: Longman.

Parks, S. (1988, 2006). *HELP® at Home.* Palo Alto, CA: VORT.

Parks, S. (1994-2006). *HELP® Family Centered Interview.* Palo Alto, CA: VORT.

Parks, S. (1992-2013). *HELP® Strands.* Palo Alto, CA: VORT.

Parks, S. (1984, 1999). *HELP® When the Parent Has Disabilities*. Palo Alto, CA: VORT.

Parks, S. (1992, 2006). *Inside HELP® (Administration and reference manual for HELP birth-3 years).* Palo Alto, CA: VORT.

Peabody, J. W., Luck, J., Glassman, P., Dresselhaus, T. R., & Lee, M. (2000). Comparison of vignettes, standardized patients, and chart abstraction: A prospective validation study of 3 methods for measuring quality. *Journal of the American Medical Association 283,* 1715–1722.

Table 1

Demographic Statistics for Participating Providers (N = 82)

Variable		Number (Percentage)
Gender	Female	79 (96%)
	Male	3 (4%)
Race	White, Non-Hispanic	79 (96%)
	African American	2 (2%)
	Native American	1 (1%)
Ethnicity	Non-Hispanic Origin	74 (90%)
	Hispanic Origin	6 (7%)
	No Response	2 (2%)
HELP® 0-3 Training	UK Online Course	35 (43%)
	Verified HELP Trainers	12 (15%)
	No formal training	35 (43%)
Work Location	Birth to Three Program	68 (83%)
	Early Head Start	10 (12%)
	Other Program	4 (5%)
Discipline	Developmental Interventionist/Teacher	46 (56%)
	Speech/Language Pathologist	10 (12%)
	Physical Therapist	9 (11%)
	Occupational Therapist	6 (7%)
	Other	11 (13%)
Degree	Master's	57 (70%)
	Bachelor's	21 (26%)
	Associate's/some college	4 (5%)
State Represented	Illinois	12 (15%)
	Kentucky	12 (15%)
	California	9 (11%)
	Wisconsin	6 (7%)
	Colorado	5 (6%)
	Missouri	4 (5%)
	Others (CT, FL, HI, IN, KS, ME, MN, MT, NM, NV, NY, OH, OR, PA, TX, UT, VA, WV)	34 (41%)
Use HELP continuously/1 year	Yes	65 (79%)
	No	16 (20%)
Experience using HELP® Strands (in years)	Mean	6.13
	Median	5
	SD	5.63
Number children assessed with HELP® Strands in last 2 years	Minimum	0
	Maximum	800
	Mean	62
	SD	113

Table 2

Demographic Statistics for Videotaped Children (N = 21)

Variable		Number (Percentage)
Gender	Female	11 (52%)
	Male	10 (48%)
Age	Birth to 1 year	5 (24%)
	1-2 years	8 (38%)
	2-3 years	8 (38%)
Special Education Services	Yes	14 (67%)
	No	7 (33%)
Race	White	16 (76%)
	Non-White	5 (24%)

Table 3

Overall Interrater Agreement among Providers, with Expert, and According to Training Group

			Overall		Formal HELP® Strands 0-3 training			
		Expert	(*n* = 82)		No (*n* = 35)		Yes (*n* = 47)	
Domain	Skill #	credit	%	95% CI	%	95% CI	%	95% CI
			Agreement among providers					
Cog	1.64	-	97.6	[95.1, 100]	100		95.7	[89.4, 100]
Cog	4.66	-	91.5	[85.4, 96.3]	91.4	[81.1, 100]	91.5	[81.8, 98.0]
Cog	1.42	+	97.6	[95.1, 100]	97.1	[89.9, 100]	97.9	[92.9, 100]
Cog	1.72	+	91.5	[86.6, 96.3]	88.6	[75.7, 97.5]	93.6	[85.4, 100]
Cog	1.117	-	89.0	[82.9, 95.1]	88.6	[76.4, 97.2]	89.4	[79.9, 97.6]
Cog	Mean		93.4	[88.6, 98.3]	93.1	[86.7, 99.6]	93.6	[89.5, 97.8]
Lang	2.34	+	98.8	[97.6, 100]	100		97.9	[92.7, 100]
Lang	2.37	-	81.7	[73.2, 89.0]	88.6	[77.1, 97.5]	76.6	[64.6, 88.0]
Lang	2.67	-	90.2	[84.1, 96.3]	94.3	[84.6, 100]	87.2	[76.6, 96.0]
Lang	1.91	+	100		100		100	
Lang	1.97	+	97.6	[95.1, 100]	97.1	[89.5, 100]	97.9	[92.9, 100]
Lang	Mean		93.7	[84.1,100]	96.0	[90.1, 100]	91.9	[79.6, 100]
GM	3.58	+/-	93.9	[90.2, 97.6]	94.3	[84.8, 100]	93.6	[86.0, 100]
GM	3.61	+/-	97.6	[95.1, 100]	100		95.7	[88.6, 100]
GM	3.03[a]	-	74.4	[64.6, 84.1]	68.6	[52.4, 84.4]	78.7	[66.0, 90.0]
GM	3.95	+/-	96.3	[92.7, 98.8]	94.3	[84.4, 100]	97.9	[92.6, 100]
GM	3.79	A	70.7	[62.2, 80.5]	71.4	[57.1, 86.7]	70.2	[56.3, 84.1]
GM	3.84	+	92.7	[98.8, 97.6]	91.4	[80.3, 100]	93.6	[85.4, 100]
GM	3.77	-	92.7	[87.8, 97.6]	97.1	[80.0, 100]	93.6	[85.1, 100]
GM	Mean		88.3	[78.2, 98.5]	88.2	[76.4, 99.9]	89.0	[79.4,98.6]
FM	4.38	-	98.8	[97.6, 100]	100		97.9	[92.1, 100]
FM	4.74	+	100		100		100	
FM	4.61	-	96.3	[92.7, 100]	100		93.6	[84.4, 100]
FM	4.78	+	100		100		100	
FM	4.54	A	98.8	[97.6, 100]	100		97.9	[92.1, 100]
FM	Mean		98.8	[96.9, 100]	100		97.9	[94.6, 100]
SE	5.70	A	98.8	[97.6, 100]	97.1	[90.3, 100]	100	
SE	5.53	+	98.8	[97.6, 100]	100		97.9	[92.1, 100]
SE	1.67	+	98.8	[97.6, 100]	97.1	[90.0, 100]	91.5	[82.5, 100]
SE	5.38	+	98.8	[97.6, 100]	97.1	[90.3, 100]	100	
SE	5.42	+	98.8	[97.6, 100]	97.1	[90.3, 100]	100	
SE	5.04	+	100		100		100	
SE	Mean		99.0	[98.5,99.5]	98.1	[96.5, 99.6]	98.2	[94.7, 100]
SH	6.67	+	97.6	[93.9, 100]	97.1	[89.7, 100]	97.9	[92.6, 100]
SH	6.19	+	100		100		100	
SH	6.62	+	100		100		100	
SH	6.57	+	98.8	[97.6, 100]	100		97.9	[92.8, 100]
SH	6.38	+/-	96.3	[92.7, 100]	94.3	[93.3, 100]	97.9	[91.8, 100]
SH	Mean		98.5	[96.6, 100]	98.3	[95.1, 100]	98.7	[97.3, 100]
R/SO	1.69	+	100		100		100	
R/SO	1.68	A+	100		100		100	
R/SO	5.60	+	100		100		100	
R/SO	Mean		100		100		100	
			Agreement with expert					
M			33.29	[32.95, 33.63]	33.37	[32.96,33.78]	33.06	[32.39, 33.63]

Note. 95% CI = 95% bootstrap confidence interval (*k* = 1,000); Cog = cognition; Lang = language; GM = gross motor; FM = fine motor; SE = social-emotional; SH = self help; R/SO = regulatory/sensory organization; % = percent agreement. [a]In this case video skill there were 2 children shown, one of whom had disabilities; some providers noted confusion on which child to credit; Mean = the average percentage of agreement of sampled items from the corresponding domain; *M* = the average number of items where providers agreed with expert. **Results were examined excluding providers with 0 years' experience and findings did not change from those reported.**