A NATIONAL CERTIFICATION EXAMINATION FOR CHILD AND YOUTH CARE WORKERS: PRELIMINARY RESULTS OF A VALIDATION STUDY

Dale Curry
Kent State University

Basil Qaqish
Kent State University

Jean Carpenter-Williams
The University of Oklahoma National Resource Center for Youth Services

Frank Eckles
Child and Youth Care Certification Institute

Martha Mattingly
University of Pittsburgh

Carol Stuart
Ryerson University

David Thomas
Bryan's House

ABSTRACT: A 100-item situational judgment examination was administered to 775 child and youth care workers from 29 sites in six states in the United States (Maryland, Pennsylvania, Ohio, Oklahoma, Texas, and Wisconsin) and two Canadian provinces (Ontario and British Columbia). The examinees also completed a questionnaire providing feedback regarding face validity, suggestions for improvement, and other relevant feedback to the examination. The supervisors of the examinees also completed a six-item assessment of worker performance (80% of the examinees had supervisors who completed an assessment). An item analysis of the examination was conducted, and the individual examination scores were correlated with the supervisory ratings of worker performance. The item analysis procedures primarily consisted of (1) reliability analysis, (2) difficulty analyses, (3) discrimination analyses, (4) distracter analyses, and (5) differential item functioning (DIF) analyses. The study provided a substantial amount of useful information to help facilitate successful implementation of child and youth care worker certification.

Key words: situational judgment, certification examination, competencies, validation.

BACKGROUND AND INTRODUCTION TO THE NORTH AMERICAN CERTIFICATION PROJECT (NACP)

This study, sponsored by the Association for Child and Youth Care Practice (ACYCP), is the result of years of work by many North American Child and Youth Care Professionals. In 1992, North American child and youth care leaders established the International Leadership Coalition for Professional Child and Youth Care (ILCPYC) and identified professional certification as a major goal. A second meeting of the ILCPYC with additional leaders from the field of youth development in 1999 and a third meeting in 2003 resulted in the development of a plan to develop a certification process. Subsequently, the development of a scenario-based, situational judgment examination became a major component of the planned certification process. Other components of the certification process include a supervisory assessment of worker performance and portfolio analysis. This study pertained to the certification examination. The goals of the study were the following:

(1) Administer a pilot certification examination for child and youth care workers and acquire feedback from the examinees regarding face validity and suggestions for improvement.

(2) Conduct an item analysis of the certification examination.

(3) Examine the relationship between the test scores and supervisory assessment ratings of worker performance (concurrent validity).

(4) Explore possible differential performance results according to gender ethnic background, age, education, and type of work setting.

(5) Determine a cutoff score (pass/fail) for future (nonpilot) examinees.

This article will provide preliminary information regarding the progress achieved by the NACP pertaining to the above goals.

Methods

Participants and Procedures

A 100-item examination was administered to 775 child and youth care workers from May through July 2006. The examinees were recruited from 29 sites in six states in the U.S. (Maryland, Pennsylvania, Ohio, Oklahoma, Texas, and Wisconsin) and two Canadian provinces (Ontario and British Columbia). No time limit was placed upon the pilot examinees for completing the examination. The mean completion time was 146 minutes, ranging from 40 to 360 minutes. More than 98% completed within 240 minutes (4 hours).

The sample was very diverse, representing various segments of the child and youth care worker population. However, the most frequent characteristics of the sample were that they were female (61%), African American, (45%), spoke English as a first language (97%), practiced in a residential treatment setting (46%), worked as a direct care worker (49%), considered themselves professional child and youth care workers (95%), and held a baccalaureate degree (36.2%). The average examinee's age was 37 (ranging from 17 to 76) and 10 years was the average number of years of experience as a child and youth care worker (see table 1 for a summary of the sample's demographic characteristics).

Table 1: Demographic Characteristics

	Frequency	Percentage
Sex		
Male	301	39
Female	470	61
Race		
African American	337	44.9
American Indian or American Indian First	5	.7
Asian	7	.9
Caucasian	320	42.7
Hispanic	56	7.5
Multi-ethnic (more than one race)	23	3.1
Other	2	.3
First Language (English)	749	97
Country		
U.S.A.	735	95.3
Canada	36	4.7
Practice Setting (Education)		
Early Childhood	127	16.5
Public and Private Schools	109	14.1
Practice Setting (Out-of-Home Care)		
Foster Homes	37	4.8
Residential Treatment	355	46
Psychiatric Hospitals	21	2.7
Medical Hospitals/Clinics	12	1.6
Physical Disabilities	10	1.3
Juvenile Corrections	58	7.5
Emergency Shelters	96	12.5
Basic Residential Care	127	16.5
Transitional Living	58	7.5
Developmental Disabilities	19	2.5
Practice Setting (Community-Based Services)		
After School Programs	50	6.5
Prevention/Intervention Programs	122	15.8
Street Outreach	35	4.5
Developmental Disabilities	22	2.9
Early Intervention	45	5.8
In-home Detention Programs	6	.8

Table 1: Demographic Characteristics

	Frequency	Percentage
Physical Disabilities	13	1.7
Recreation	38	4.9
In-home Family Care & Treatment Services	45	5.8
Organizations (YMCA, Scouts, etc.)	38	4.9
Clinic-based Day Treatment Services	26	3.4
Practice Settings (Other)	57	7.4
Type of Position		
Direct Care Worker	370	48.7
Educator	38	5.0
Supervisor	102	13.4
Administrator	62	8.2
Counselor	84	11.1
Therapist	6	.8
Foster Parent	1	.1
Other	93	12.2
Professional CYC		
Yes	729	95.0
No	33	4.3
Education		
None	99	13.6
Associate	97	13.4
Baccalaureate	263	36.2
Masters	87	12.0
Doctorate	2	.3
No degree but coursework	177	24.4
	Mean	SD
Age	37.35	10.95
Years of Experience	10.43	8.05

N=775; Settings are not mutually exclusive. Respondents may have selected more than setting.

The supervisors of the examinees also completed a six item assessment of worker performance (80% of the examinees had supervisors who completed an assessment). An item analysis of the examination was conducted, and the individual examination scores were correlated with the supervisory ratings of worker performance. Exploration of a possible cutoff score (pass/fail) for future examinees using a modified Angoff method is currently being conducted.

Analysis

A set of procedures that explores the effectiveness of individual items regarding whether they function as intended is known as item analysis. These procedures are often used to increase the reliability and validity of a test by individually evaluating each item in relation to the overall test. For example, if examinees who score in the upper portion of the test score distribution tend to answer an item correctly and examinees who score in the lower portion tend to answer incorrectly, then the item is viewed as having positive discriminatory power, contributing to the overall reliability and validity of the test. (Peterson & Fox, 2001).

The item analysis procedures primarily consisted of (1) reliability analysis, (2) difficulty analyses, (3) discrimination analyses, (4) distracter analyses, and (5) differential item functioning (DIF) analyses.

Chronbach's alpha was examined as a measure of reliability. An overall test reliability score was obtained, and each item was examined for its effect on the overall test (change in alpha if the item was removed).

The first difficulty analysis indicator measured the difficulty of items for participants by calculating a difficulty index for each item (the percentage of examinees who answered the item correctly). In addition, the examinee population was divided into five segments based on their total test score, and a difficulty value was calculated by adding the total number of examinees who chose the correct answer in the top 20% with the total number of examinees who chose the correct answer in the bottom 20% divided by the total number of examinees in both groups.

Discrimination analysis included the point biserial correlation between the item score and total test score. In addition, the examinee population was divided into five segments based on their total test score, and a difficulty difference score was determined (difference between the top 20% and bottom 20%). This value was calculated by subtracting the proportion of examinees in the lower group who answered the item correctly from the proportion in the upper group who answered the item correctly. A fourth analysis involved the visual examination of each item choice across five "ability range" groups using a line graph that displayed the percentage of examinees choosing each of the responses in each item. With the five segments from low to high on the X axis and the percentage correct on theY axis, an upward trend from left to right for the correct answer, and a downward trend from left to right was expected for items with good discriminatory power (see figure 1 for an example of an item with good discriminatory power).

Differential item functioning for race and gender was explored with the computer program SIBTEST. SIBTEST takes the groups of examinees who have equivalent raw total test scores (males vs. females, or African American vs. Caucasian) and analyzes their response patterns to see if they are statistically significantly different from each other. The rationale behind using this test is to see if groups of examinees who have similar abilities respond to the same test question in the same manner, regardless of whether they were male or female and regardless of whether they were African American or Caucasian. In addition, the Mantel Haenszel procedure was also conducted.

Because of missing data points, a number of examinee responses were eliminated to be able to have complete item response vectors for males, females, African American, and Caucasian. Once this process was done, the program SIBTEST was run to compare males vs. females on the item level. The same process was done to compare African Americans vs. Caucasian.

Procedures to determine a certification cut point have been initiated but not yet completed. This involved the use of a modified Angoff procedure that entailed several steps. First, an expert panel of 10 was selected and oriented to the selection procedure. Prior to making probability ratings, panelists reviewed a brief written description of the five major competency areas and an overview of the modified Angoff procedure. Next, the panelists participated in a discussion pertaining to characteristics of a"minimally competent"child and youth care worker at the professional level. Panelists then received instructions to independently estimate the probability that a minimally competent child and youth care worker (at the professional level) will get the answer correct for each of the 100 items. The individual ratings were averaged across raters for each item and then the averages were averaged to obtain a tentative cut score recommendation. After this initial probability estimation, panelists received information regarding the actual difficulty level of the items (percentage of pilot examinees who answered each item correctly), a table displaying the probability ratings of the other panel members, and information regarding the percentage of examinees that would have passed or failed if the recommended cut point was used. A group discussion of the results was followed by a second independent probability estimation for each item (currently in-progress). Similar to the first round of ratings, the individual ratings will be averaged across raters for each item and then the averages averaged to obtain a second tentative cut score recommendation.

Correlational analysis of the total examination scores with supervisory assessments of worker performance (a six-item survey questionnaire assessing the five major competency areas and an overall performance assessment item) was conducted to provide support for concurrent criterion validity.

Measures/Instruments

Situational Judgment Examination

A predominantly situational judgment examination was developed that requires practice judgments from the examinee based on case studies elicited from the field. The instrument construction first involved defining child and youth care practice. The current description of the field as described by Mattingly, Stuart, and VanderVen (2002) was used in the development of the examination as well as inclusion of participants in the study.

> Professional Child and Youth Care Practice focuses on infants, children, and adolescents, including those with special needs, within the context of the family, the community, and the life span. The developmental-ecological perspective emphasizes the interaction between persons and their physical and social environments, including cultural and political settings. Professional practitioners promote the optimal development of children, youth, and their families in a variety of settings, such as early care and education, community-based child and youth development programs, parent

> education and family support, school-based programs, community mental health, group homes, residential centers, day and residential treatment, early intervention, home-based care and treatment, psychiatric centers, rehabilitation programs, pediatric health care, and juvenile justice programs. Child and youth care practice includes assessing client and program needs, designing and implementing programs and planned environments, integrating developmental, preventive, and therapeutic requirements into the life space, contributing to the development of knowledge and practice, and participating in systems interventions through direct care, supervision, administration, teaching, research, consultation, and advocacy.

The examination addressed child and youth care worker competencies that were identified through a meta-analysis of the field's articulation of competencies that also involved the development of new competencies where gaps were identified. Included in the review of many of the competency sets were competencies that were determined by formal job analyses. The competencies included what workers currently value, know, and do as well as what best practice standards indicate that they should value, know, and do. This includes competencies pertaining to the Code of Ethics for North American child and youth care workers. Determination of the final competency list involved the work of several work groups and several years of discussion and refinement. The competencies were organized into the following five domains: (1) professionalism, (2) cultural and human diversity, (3) applied human development, (4) relationship and communication, and (5) developmental practice methods. The reader is referred to Mattingly, Stuart, and VanderVen (2002) for a more detailed description of the competency and Code of Ethics development process.

After identification of the competencies, the test plan included a process to identify the most appropriate assessment measure (supervisor assessment, portfolio, or examination). Six panel members made independent recommendations for each competency area. After discussing areas of disagreement, the panel members came to consensus on the most appropriate measure for each competency area.

Once the items that could be assessed by a situational judgment examination were identified, the expert panel was asked to independently prioritize the competencies according to importance to determine the number of items assigned to each competency area. The panel was requested to determine the number of items for each competency (1, 2, or 3 items) based on their importance. The panel responses were then averaged to determine the number of items. During the same period of time, case studies from the field were solicited to serve as content for the scenario-based examination. A "call for case studies" was posted on the website of the Association of Child and Youth Care Practice and CYC-net. The case studies were compiled and bound into a booklet for the test construction team to review prior to attending training in Texas on the construction of situational judgment items. The test construction team was comprised of seven child and youth care experts from both the United States and Canada.

Following the training, the test construction team began the process of developing scenario-based items pertaining to each competency area. As the items were constructed, they were reviewed by the entire test construction team and team members were independently polled to determine the most correct answer for a newly constructed item and whether it addressed the targeted competency area. Discussion regarding item modification occurred when one or more raters disagreed with the majority. When the initial entire test was completed, team members reviewed each item for a second time indicating the correct answer and if it addressed the targeted competency area.

Subsequent to completion of the draft examination, another expert panel reviewed the examination with the goal of identifying cases and/or items to eliminate or modify due to possible cultural bias. The panel also made recommendations (suggestions for change) regarding the readability of the case studies and examination items. The test construction team incorporated many of the extended panel's suggestions and began to plan for the pilot validation study of the examination. While the pilot examination was being administered, another expert panel was established that included several of the test construction team members along with additional expert members. The panel was charged with two tasks. First, independently choose the "most correct" answer for each item, and later estimate the probability for each item of the "minimally competent" child and youth care worker at the professional level answering the item correctly. Nine experts assessed "correct" answers (overall agreement percentage of 87.2%). Ten panel members were involved in the modified Angoff probability ratings to help determine a recommended cut score for the examination (discussed earlier in more detail).

Supervisory Assessment of Worker Performance Survey.

Supervisors of each of the child and youth care worker examinees were requested to complete a six-item, five-choice survey assessing the worker's competence on-the-job. One item pertained to each of the five major competency domains, and one item referred to the workers' overall competence. The item anchor descriptors ranged from "consistently demonstrates competence" to "does not demonstrate competence." A composite competence score (the sum of the six items) was used as a concurrent criterion measure of job performance (Chronbach's alpha of .94 for the six items).

Face Validity, Motivation, and Feedback Survey.

A survey instrument was constructed to receive feedback from the examinees regarding their perception of the validity of the examination (four items), their degree of motivation for taking the examination (two items), and suggestions for improvement of the examination. In addition, examinees were asked to assess their perception of how well they performed on the examination (single item). The four face validity items were summed to obtain an overall face validity measure (Chronbach's alpha = .77). Similarly, the two motivation items were summed to obtain a test-taking motivation indicator (Chrobach's alpha = .87).

Preliminary Results and Discussion

Face Validity

The extent to which examinees view an assessment as suitable for its intended purpose has been described as face validity. Central to face validity is the examinees' perception concerning whether the assessment measure and/or process actually measures what it is intended to assess (e.g., important aspects of youth work) (Drummond & Jones, 2006; Mosier, 1947; Nevo, 1985).

Since face validity is considered one of the weakest indicators of validity, it is not frequently reported in studies. However, face validity has been shown to affect the reactions or attitudes of those being assessed in several areas including performance motivation, employee evaluation of organizational attractiveness, attribution of responsibility for task success or failure, and employee burnout (Anastasi & Urbina, 1997; Chan, Schmitt, DeShon, Clause, & Delbridge, 1997; Gabris & Ihrke, 2001; Nevo, B. 1985; Tweed & Cookson, 2001). An examinee's perception does not have to be a correct indicator of true validity to influence the credibility of the assessment and thus influence motivation to perform and take the assessment results in a serious manner (Tweed & Cookson, 2001). Furthermore, the acceptance of the certification process by the field as a valid measure of child and youth care competence is crucial for the certification process to make a significant impact on the profession.

Almost all (97%) of the examinees completed the face validity and feedback questionnaire. Results indicate that the vast majority of respondents (90%) perceived that the items in the examination accurately assess important aspects of child and youth care work and the case examples provide realistic samples of child and youth care work. This strongly endorses a belief that the examination seems to be measuring the essential elements of child and youth care work. Somewhat less (80%) indicated that the content in the examination is similar to their actual job duties, and only 59% stated that they believe that their performance on the examination is an accurate indicator of their actual performance on the job (34% indicated that they neither agreed or disagreed). Apparently, the examinees viewed the examination as an excellent indicator of child and youth care practice. However, they appeared to have less confidence that how they performed on the examination is indicative of their job performance. (See table 2.) Since the examination covered areas from a variety of settings and ages, some of the participants may have perceived their job duties in a more limited manner (e.g., confined to working with children ages 3 to 5 in a day care setting only). In addition, the examinees did not have access to their test results. So, the relatively large number of examinees who indicated that they neither agreed nor disagreed that the examination is an excellent indicator of their job performance tended to lower this item rating relative to the other face validity items.

Item Analysis

The item analysis includes the test results from 775 examinees, meeting the generally recommended standard of 5 to 10 subjects per item (Nunally, 1967). The test contained 100 4-choice multiple choice items. Examinees' scores ranged between 25 and 89. The average score was 60.01 with a median score of 61. The test's standard deviation was 13.65. The item analysis procedures included reliability analysis, item difficulty analysis, discrimination analysis, distracter analysis, and differential item functioning (DIF) analysis.

Table 2: Examinee Feedback Survey

SD= Somewhat Disagree	D= Disagree	N= Neither		A= Agree		SA = Somewhat Agree		
Item #	Description	% SD	% D	% N	% A	% SA	M	SD
1	Similar to my actual job duties	1.1	7.1	12.2	54.5	25.1	4.0	.87
2	Accurately assess important aspects of cyc	.3	1.3	8.6	58.2	31.6	4.2	.67
3	Accurate indicator of my job performance	4.1	12.6	34.1	37.6	11.5	3.4	.99
4	Case examples provide realistic samples	.1	.9	9.4	57.3	32.3	4.2	.65
5	Doing well on this exam was important	.5	1.5	18.3	46.5	33.2	4.1	.78
6	Extremely motivated to do well	.7	2.8	24.1	45.1	27.3	4.0	.83
7	I performed well	.7	3.6	25.9	53.7	16.1	3.8	.77

n = 748 (97% of total sample)

Reliability analysis and standard error of measurement

Cronbach's alpha was used as a reliability indicator. The test's reliability value was 0.90, and the standard error of measurement (SEM) was 4.4. Reliability is a measure of the consistency and the "replicability" of the test results. It is an indication of the quality of the instrument. A value of .90 is considered excellent reliability. The standard error of measurement, an expression of the degree of inaccuracy in the reported score, was relatively small. The lower the SEM value, the better the instrument and the more accurate are the reported scores. If the SEM value is high (e.g., 15%), then this means there is a wide range that the reported score can take. For example, if someone's score is 70% with a SEM value of 15%, then we are 68% confident that her score is between 55% and 85%. The examination's relatively small SEM value of 4.4 is another positive indication of the quality of the instrument.

Individual items were analyzed in terms of their contribution to the total instrument reliability. Reliability takes on a value between 0 and 1. The higher this value the better the instrument. A high reliability value indicates that the results can be replicated within a small error range. Table 3 presents the reliability analysis (alpha figures) on the item level.

The first column in table 3 represents the item number. Q1 means question 1 in the test, and so on. The second column represents the instrument reliability if the item is taken out. For instance, if question 6 is taken out of the instrument, the test's reliability will increase to 0.898. The third column represents the change in the test's reliability value if the item is taken out. Bolded numbers means that the test's reliability will increase if the item is taken out. The analysis shows that items 6, 9, 15, 34, 35, 40, 52, 53, 60, 61, 62, 70, 77, 89, 90, 92, 94, and 98 may need to be taken out. Those 18 items need to be looked at carefully by content experts to see if the instrument can be run without them. If content experts want to keep any or all of them, then a detailed look on the items' language and distracters need to be done to see which possible changes need to be done to such questions before using them again in the instrument.

Table 3: Item Analysis Summary

Item	Alpha without Item	Change in alpha without item	Overall diff.	Upper-Lower diff.	Point Bi-serial disc	Upper-Lower disc
Q1 Case 1	0.897	0.000	0.20	0.22	0.17	0.23
Q2	0.896	-0.002	0.74	0.73	0.39	0.47
Q3	0.896	-0.002	0.70	0.61	0.39	0.54
Q4	0.896	-0.001	0.59	0.62	0.31	0.50
Q5 Case 2	0.896	-0.002	0.77	0.74	0.40	0.52
Q6	0.898	**0.000**	0.64	0.59	0.09	0.18
Q7	0.896	-0.002	0.37	0.38	0.38	0.53
Q8	0.897	-0.001	0.73	0.73	0.25	0.34
Q9 Case 3	0.899	0.001	0.71	0.71	0.04	0.09
Q10 Case 4	0.896	**-0.002**	0.77	0.71	0.43	0.54
Q11	0.897	0.000	0.48	0.50	0.21	0.31
Q12	0.897	0.000	0.58	0.54	0.21	0.28
Q13	0.897	-0.001	0.92	0.87	0.24	0.20
Q14 Case 5	0.897	-0.001	0.45	0.51	0.28	0.46
Q15	0.898	**0.000**	0.55	0.58	0.10	0.17
Q16	0.897	-0.001	0.90	0.86	0.27	0.23
Q17	0.896	-0.001	0.71	0.66	0.33	0.44
Q18	0.897	-0.001	0.28	0.35	0.22	0.32
Q19	0.897	-0.001	0.73	0.68	0.29	0.32
Q20	0.897	-0.001	0.91	0.89	0.23	0.16
Q21	0.896	-0.001	0.79	0.76	0.35	0.43
Q22	0.896	-0.001	0.65	0.68	0.33	0.52
Q23	0.896	-0.001	0.62	0.62	0.31	0.43
Q24	0.896	-0.001	0.81	0.76	0.36	0.43
Q25 Case 6	0.896	-0.002	0.64	0.60	0.41	0.59
Q26	0.897	0.000	0.81	0.78	0.18	0.25
Q27	0.897	-0.001	0.63	0.64	0.23	0.34
Q28	0.897	-0.001	0.33	0.40	0.29	0.42
Q29	0.897	-0.001	0.76	0.77	0.26	0.34

Table 3: Item Analysis Summary

Item	Alpha without Item	Change in alpha without item	Overall diff.	Upper-Lower diff.	Point Bi-serial disc	Upper-Lower disc
Q30	0.897	-0.001	0.71	0.65	0.25	0.34
Q31	0.896	-0.001	0.72	0.67	0.34	0.46
Q32	0.895	-0.003	0.68	0.65	0.47	0.65
Q33	0.898	0.000	0.92	0.92	0.15	0.13
Q34 Case 7	0.898	**0.001**	0.43	0.41	0.09	0.19
Q35	0.898	**0.000**	0.59	0.60	0.11	0.21
Q36 Case 8	0.898	0.000	0.28	0.31	0.16	0.30
Q37 Case 9	0.897	-0.001	0.49	0.56	0.29	0.51
Q38	0.897	-0.001	0.70	0.67	0.27	0.31
Q39	0.896	-0.002	0.80	0.71	0.42	0.48
Q40	0.899	**0.001**	0.54	0.55	0.06	0.20
Q41	0.896	-0.002	0.74	0.68	0.41	0.50
Q42 Case 10	0.896	-0.001	0.38	0.44	0.31	0.48
Q43	0.898	0.000	0.67	0.63	0.17	0.30
Q44	0.897	-0.001	0.73	0.66	0.23	0.27
Q45	0.896	-0.001	0.44	0.45	0.33	0.51
Q46	0.896	-0.002	0.72	0.70	0.35	0.48
Q47	0.897	-0.001	0.63	0.60	0.23	0.34
Q48	0.896	-0.002	0.73	0.66	0.43	0.55
Q49 Case 11	0.898	0.000	0.76	0.70	0.16	0.24
Q50	0.897	-0.001	0.44	0.48	0.28	0.44
Q51	0.896	-0.002	0.73	0.70	0.35	0.46
Q52	0.899	**0.002**	0.21	0.24	-0.10	- 0.10
Q53	0.899	**0.001**	0.48	0.45	0.03	0.11
Q54	0.897	-0.001	0.90	0.83	0.27	0.25
Q55 case 12	0.895	-0.003	0.69	0.65	0.54	0.70
Q56	0.896	-0.002	0.84	0.77	0.39	0.41
Q57	0.897	-0.001	0.64	0.65	0.26	0.37
Q58	0.896	-0.002	0.70	0.66	0.42	0.58
Q59	0.896	-0.002	0.54	0.56	0.43	0.59
Q60 Case 13	0.899	**0.001**	0.37	0.37	0.01	0.03
Q61	0.898	**0.000**	0.59	0.59	0.13	0.19
Q62	0.898	**0.001**	0.19	0.26	0.05	0.13
Q63	0.898	0.000	0.39	0.42	0.16	0.32
Q64	0.897	0.000	0.53	0.55	0.21	0.37
Q65	0.896	-0.002	0.73	0.69	0.36	0.50
Q66 Case 14	0.897	-0.001	0.90	0.85	0.23	0.23
Q67	0.896	-0.002	0.68	0.65	0.43	0.59
Q68	0.896	-0.002	0.57	0.55	0.42	0.60
Q69 Case 15	0.897	0.000	0.24	0.28	0.20	0.30
Q70	0.898	**0.000**	0.14	0.20	0.03	0.06
Q71	0.896	-0.001	0.55	0.55	0.33	0.54

Table 3: Item Analysis Summary

Item	Alpha without Item	Change in alpha without item	Overall diff.	Upper-Lower diff.	Point Bi-serial disc	Upper-Lower disc
Q72	0.897	-0.001	0.67	0.61	0.30	0.44
Q73	0.896	-0.001	0.83	0.79	0.35	0.40
Q74	0.897	-0.001	0.75	0.72	0.29	0.37
Q75	0.896	-0.002	0.71	0.62	0.39	0.54
Q76	0.896	-0.002	0.73	0.69	0.36	0.48
Q77	0.898	**0.001**	0.57	0.55	0.09	0.15
Q78	0.897	-0.001	0.45	0.46	0.25	0.43
Q79 Case 16	0.897	-0.001	0.47	0.49	0.30	0.51
Q80	0.897	-0.001	0.37	0.44	0.27	0.44
Q81	0.897	-0.001	0.51	0.51	0.25	0.37
Q82	0.896	-0.002	0.53	0.54	0.35	0.54
Q83 Case 17	0.896	-0.002	0.61	0.58	0.38	0.56
Q84	0.896	-0.002	0.62	0.56	0.43	0.61
Q85	0.896	-0.001	0.50	0.54	0.32	0.52
Q86	0.895	-0.002	0.53	0.54	0.43	0.61
Q87 Case 18	0.895	-0.003	0.62	0.60	0.50	0.73
Q88	0.895	-0.002	0.70	0.65	0.45	0.63
Q89	0.898	**0.000**	0.25	0.24	0.14	0.19
Q90	0.898	**0.000**	0.23	0.25	0.11	0.14
Q91	0.897	-0.001	0.48	0.47	0.25	0.39
Q92	0.899	**0.001**	0.48	0.46	0.05	0.11
Q93	0.895	-0.002	0.57	0.54	0.45	0.66
Q94 Case 19	0.898	**0.000**	0.70	0.64	0.15	0.23
Q95	0.897	-0.001	0.77	0.76	0.29	0.35
Q96	0.897	0.000	0.48	0.45	0.20	0.30
Q97	0.896	-0.002	0.54	0.55	0.43	0.64
Q98	0.898	**0.001**	0.45	0.41	0.07	0.14
Q99	0.897	-0.001	0.75	0.71	0.27	0.37
Q100	0.896	-0.002	0.64	0.67	0.34	0.51

Overall Alpha = .8977; Table does not include DIF analysis results.

Item Difficulty and Item Discrimination

Item difficulty represents the proportion of respondents who answered the item correctly. It is calculated by dividing the number of examinees who chose the correct answer over the total number of examinees. An item difficulty of 0.10 indicates that the item is a very difficult one. Only 10% of the examinees chose the right answer. An item difficulty of 0.90 tells that the item is a very easy one, since most examinees (90%) answered it correctly.

Table 3 presents the overall item difficulty (proportion correct) and the item discrimination (the point biserial correlation between the item score and total test score). The lower the item difficulty value, the more difficult the item is (because less people chose the correct answer). The higher the difficulty value, the easier the item (because more people chose the correct answer). In general, one does not want to have items that almost everyone, or almost no one, answers correctly. The idea is not to have very easy or very hard items. Item difficulty may also have implications for the sequencing of examination questions. Generally, it is recommended that more difficult items should be placed after easier items. Since the item location is also dependent on individual case studies, exploring the difficulty of items grouped by case study and making appropriate sequencing changes is recommended in the final instrument.

Difficulty and discrimination values were also calculated by comparing the top 20% of examinees to the bottom 20% of examinees, with regard to the total raw test score. Difficulty value was calculated by dividing the total number of examinees who chose the correct answer in both groups by the total number of examinees in both groups. Measurement literature suggests using items with difficulty ranges between 0.10 and 0.8 (Hopkins, 1998, Chapter 10). Discrimination value is calculated by subtracting the proportion of examinees in the lower group who answered the item correctly from the proportion in the upper group who responded to the item correctly. Hopkins goes as low as 0.10 for item discrimination value as an acceptable level to have in a test. Mehrens and Lehman (1991, Chapter 8) regard a discrimination value of 0.20 as acceptable.

Table 4 divides the examinees into 5 groups of 155 examinees each, and reports statistical descriptives for each group. Each of the lower group and upper group included 155 examinees.

Table 4: Descriptive Statistics for Five Group Division

Summary group statistics	n	avg.	avg%	s.d.	min.	mdn.	max.
Upper 20 %	155	77.7	78%	3.5	73	77	89
2nd 20%	155	69.0	69%	2.2	65	69	73
3rd 20%	155	61.5	61%	2.2	57	61	65
4th 20%	155	52.2	52%	3.0	47	52	57
Lower 20%	155	39.7	40%	5.3	25	41	47
Total Sample	775	60.0	60%	13.7	25	61	89

This approach reveals that six items do not meet the recommended criteria for item difficulty (13, 16, 20, 33, 54, & 66). Using .15 as the minimally acceptable discrimination score, nine items do not meet the criteria (33, 52, 53, 60, 62, 70, 90, 92, & 98).

Distracter Analysis

The proportion of people who chose each response for each question were examined. The generally accepted procedure is to identify responses where no one, or a negligible proportion of people, chose that alternative. For example, only 1% of the examinees chose the fourth response in item 13 (not shown in table). Such question responses were looked at carefully for possible amendments or changes. As stated previously, a visual analysis of items was also conducted using a line graph of the four responses to each item (see figure 1). To ensure test security during the initial phases of the certification and test revision processes, statistics regarding the percentage of responses for each item alternative are not reported.

Differential Item Functioning Analysis

DIF is conceptualized as a difference in the probability of endorsing a keyed item response, when individuals with the same levels of ability possess different amounts of supplemental abilities that affect their responses to the item (Shealy, R., & Stout, W., 1993)

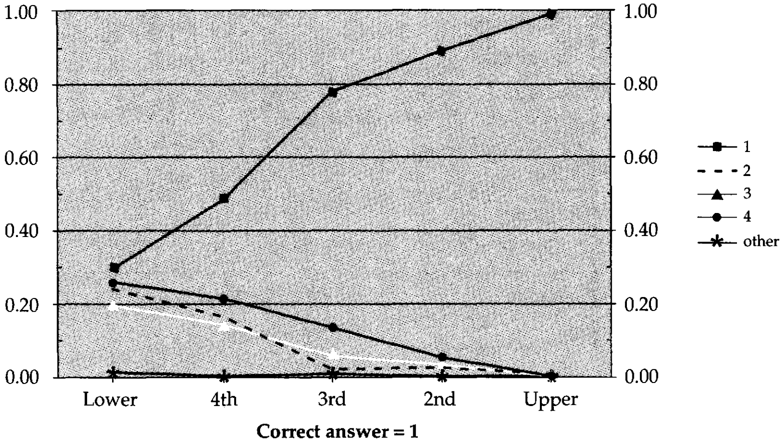

Figure 1. Visual display-line chart of percentage of responses for each item alternative from lower 20% to upper 20% of sample.

The number of examinees used in the SIBTEST procedures is shown in table 5. It is worth mentioning that the available sample sizes in the compared groups were less than 350 examinees in each case. It is preferable, when running a program like SIBTEST, to have more than 500 examinees in each group. This was not possible due to the relatively small sample size available. Because of this, the authors of this paper think that it is more appropriate to take a conservative approach in looking at the SIBTEST output results and consider each item as exhibiting DIF at the 0.01 confidence level instead of the 0.05 confidence level, which is a level that many researchers like to adopt as an agreed upon critical"cut-point"level. In any case, we have reported the possible number of items exhibiting DIF at the 0.05 level as well as at the 0.01 level. Table 6 contains a summary of the SIBTEST results in comparing items by race and by gender (Mantel Haenszel procedure results are also reported. It is another method to identify items that are exhibiting DIF).

Table 5: Number of Examinees used in the SIBTEST Procedure

Race\Gender	Male	Female	Total
African American	112	186	298
Caucasian	112	162	274
Total	224	348	572

Table 6: Items Exhibiting DIF by Gender and Race

	SIBTEST Results		Mantel - Haenszel Results	
Group\Confidence Level	0.05	0.01	0.05	0.01
Gender				
(Male vs. Female)	1, 7, 34, 36, 46, 48, 49, 78, 82, 83, 86, 92, 99 (13 items)	1 (1 item)	1, 48, 52, 78, 80, 83, 85, 92, 99 (9 items)	1 (1 item)
Race				
(African American vs. Caucasian)	13, 17, 21, 22, 27, 31, 42, 71, 82, 84, 86, 87, (12 items)	17, 22, 42, 82, 87 (5 items)	4, 17, 22, 31, 42, 71, 82, 84, 87 (9 items)	17, 31, 42, 71 (4 items)

Supervisory Assessment of Worker Performance

Table 7 summarizes the supervisors' assessments of the examinees competence on the job. On the whole, supervisors viewed their workers as competent practitioners. Approximately half indicated that their workers consistently demonstrate overall competence. Similar ratings were also characteristic of the five competency areas, ranging from 46.8% to 51.1%. Ratings indicating that workers failed to demonstrate competence were almost nonexistent. Therefore, variability of the data mostly ranged from ratings of 3 (inconsistently demonstrates competence) to 5 (consistently demonstrates competence). The composite supervisory rating (sum of six items) had a correlation of .26 (p<.000) with the total examination score, providing evidence of concurrent validity for the certification examination. The lack of variability in the supervisor ratings probably somewhat attenuated the correlation between the examination and the on-the-job criterion.

Table 7: Supervisor Assessment of Worker Competence on the Job

CD= Consistently Demonstrates		ID= Inconsistently Demonstrates			DD= Does not Demonstrate		
Competency Area	CD		ID		DD	Mean	SD
	(5)	(4)	(3)	(2)	(1)		
	%	%	%	%	%		
Professionalism	49	38.1	12.7	.2	0	4.36	.70
Culture	51.1	38	10.6	.2	.2	4.40	.69
Human Development	48.3	39.5	11.7	.2	0	4.36	.70
Relationship & Communication	48.1	37.4	13.7	.8	0	4.33	.74
Developmental Practice Methods	46.8	40.2	12.4	.7	0	4.33	71
Overall	50.5	39.2	10.2	0	0	4.40	.67

N=775

Further Discussion and Future Directions

The piloting of the certification examination has provided a substantial amount of useful information to help facilitate successful implementation of child and youth care worker certification. Although continued improvements to the examination and further validation efforts are indicated, the preliminary results from the pilot study indicate potential for the examination to be a highly reliable and valid component of a comprehensive certification process for child and youth care workers. The examination was constructed with significant involvement from child and youth care experts addressing competencies agreed-upon by child and youth care leaders in both the United States and Canada, and based on actual case studies elicited from the field. This integral connection to the field most likely contributed to the overwhelming majority of the child and youth care examinees viewing the examination as accurately assessing important facets of child and youth care work.

Although the supervisor assessments of the examinees' performance on-the-job positively correlated with the examination scores, further evidence of criterion validity should be explored. For example, the more extensive supervisory assessment that is a component of the full certification process could be correlated with the examination. Future efforts involving supervisory ratings should include strategies to increase the variability of the supervisory ratings of child and youth care worker performance. Since poorly performing workers are probably less likely to apply for certification (and some may be terminated due to poor performance), range restriction may continue to be an obstacle for future validation studies.

Significant differences of total test scores by race/ethnicity and gender were found and should be continued to be monitored in the future. However, the number of items exhibiting DIF is relatively small if a 0.01 confidence level is chosen. This is a conservative approach which we think is suitable here. The number of items exhibiting DIF is higher, as one would expect, if a 0.05 confidence level is chosen.

Three of the items exhibiting DIF influence reliability negatively (items 34, 52, and 92). Such items are candidates for possible elimination or language amendments. In addition, all the items exhibiting DIF must be looked at carefully by content experts to try to discern if any language or content bias exists in such items. If this is in fact the case, such an item should be eliminated from the instrument or go through some language amendments. It is not always possible to discern why an item exhibits DIF, but it is always a good idea to try to understand why DIF occurs in a certain item. If such an item is deemed"biased"because of one factor or another, it should be taken out of the test instrument or changed as necessary. We think that the best way to proceed with the examination is to first eliminate the items that contribute negatively to reliability. After that, content experts can look at DIF exhibiting items. Because of the relatively small sample sizes used in the DIF analysis, we think the experts' eye may help us in understanding why DIF occurred in some of the items. Proceeding on what to do with these items is probably best answered by content experts. However, once the test is restructured and administered for a while, DIF analysis can be done again on larger samples (we like to see more than 500 examinees in each data set being used for DIF analysis). In addition, DIF analysis should also be conducted with other races/ethnicities, when larger samples are obtained.

Determination of a cut point (pass/fail) must still be established prior to full implementation of the certification process. The potential adverse effects on specific populations (e.g. minorities) must be explored. Ongoing monitoring will need to continue after certification implementation.

Although substantial work remains, the North American Certification Project has taken another significant step toward professionalizing the field of child and youth care work. Conducting this pilot study required a collaborative effort among many contributors, including the 775 child and youth care worker examinees and their supervisors. It is our hope that the interim results of this study provide sufficient evidence to build support for the project among the varied child and youth care settings and direction to further improve the examination and process in support of increased standards of practice and care for children, youth, and families.

References

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Chan, D., Schmitt, N., DeShon, R.P., Clause, C.S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology, 82,* 300-310.

Drummond, R.J., & Jones, K.D. (2006). *Assessment procedures for counselors and helping professionals.* (6th ed.). Upper Saddle River, NJ: Pearson/Merrill Prentice-Hall.

Gabris, G.T., & Ihrke, D.M. (2001). Does performance appraisal contribute to heightened levels of employee burnout? The results of one study. *Public Personnel Management, 30,* 157-172.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Mattingly, M., Stuart, C., & VanderVen, K. (2001). North American Certification Project (NACP): Competencies for professional child and youth work practitioners. *Journal of Child and Youth Care Work, 17,* 16-49.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). London: Holt, Rinehart and Winston.

Mosier, C.L. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement, 7,* 191-206.

Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement, 22,* 287-293.

Nunnaly, J. C. (1967). *Psychometric theory.* New York: McGraw-Hill.

Peterson, N.A., & Fox, T.C. (2001). Knowledge test construction and validation: Evaluation of New Jersey's interdisciplinary training for public agency caseworkers to improve child welfare services. *Proceedings of the Fourth Annual National Human Services Training Evaluation Symposium.* 32- 41.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects bias/DTF as well as item bias/DIF. *Psychometrika, 58* (No. 2).

Tweed, M., & Cookson, J. (2001). The face validity of a final professional clinical examination. *Medical Education, 35,* 465-473.