DOCUMENT RESUME

ED 477 925 TM 035 021

AUTHOR Wang, Shudong; Wang, Ning; Hoadley, David

TITLE Construct Equivalence of a National Certification Examination that Uses Dual Languages and Audio Assistant.

PUB DATE 2003-04-00

NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE EDRS Price MF01/PC01 Plus Postage.

DESCRIPTORS Audio Equipment; *Certification; *Construct Validity; English; *Nurses Aides; Second Languages; Spanish; *Test Format

ABSTRACT

This study examined the comparability of scores on the National Nurses Aides Assessment Program (NNAAP) test across language and administration condition groups for calibration and validation samples that were randomly drawn from the same population. A sample of 20,568 candidate responses to 1 test form was used. This examination is given in English or Spanish, with or without audio equipment assistance. Results show that factor structure validities of the NNAAP are well supported. Statistically significant chi square (or difference of chi square) statistics occur because of the large sample sizes. For this reason, it is frequently appropriate to conclude that a SEM model fits the data even if "p" is significant. The overall pattern of NNAAP data indicates a reasonable fit even when the chi square test suggests rejection of factor models when sample sizes are large. The evidence of fit holds for the calibration and validation samples for language and administration condition groups. Data suggest that the test is fair across administration groups. (Contains 3 tables and 27 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 477 925

Construct Equivalence of a National Certification Examination That Uses Dual Languages and Audio Assistant

Shudong Wang, Ph.D.
Harcourt Educational Measurement

Ning Wang, Ph.D.
David Hoadley
Promissor

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

S. Wang

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

□ Minor changes have been made to improve reproduction quality.

● Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM035021

Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 2003. All correspondence should be sent to Shudong Wang, Harcourt Educational Measurement, 19500 Bulverde Road, San Antonio, Texas 78259-3701. E-mail: Shudong_wang@Harcourt.com

BEST COPY AVAILABLE

Construct Equivalence of a National Certification Examination That Uses Dual Languages and Audio Assistant

Objective

The purposes of this study are: (a) investigate the factorial (structure) validity of a national certification examination; (b) assess the construct equivalence of a national certification examination across different languages with or without audio assistant; and (c) provide an example of how to extend validity evidence beyond the methodology typically used in certification testing.

Perspective

Certification tests are designed to assess professional competence. Like other credentialing tools, certification tests are intended to help the public, employers, and government agencies identify practitioners who have met a particular standard. Certification organization have a responsibility not only to candidates--to ensure that all certification procedures are fair and consistent--but also to the consumer--to ensure the validity of the certification process so that individuals who are certified are indeed competent. Like any high-stakes tests, certification tests must satisfy the legal requirements of validation and fairness. Validity (and fairness), according to the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), is the most important consideration in test development and evaluation. Validity and fairness of the certification tests are also required by federal laws and regulations (Equal Employment Opportunity Commission [EEOC], Civil Service Commission, Department of Labor and Department of Justice, 1978; Mehrens, 1994; Mehrens & Popham, 1992). The degree to which different language tests (with/without audio assistant) are comparable is an important validity issue because these tests are typically interpreted as if they are equivalent or use the same cut score to determine the pass/fail status of examinees. The trend of using multi-language tests will continue to grow because of the increase in globalization of markets, linguistic diversity, and culture exchanges. Like achievement tests (Gierl, 2001; Sireci & Khaliq, 2002; Hambleton, 2001), in order to reduce construct-irrelevant variance in a candidate's test scores due to proficiency in a specific language, many certification tests are adapted for different languages. Comparing competence of candidates who take different language tests is not an easy task because the differences in test scores between different languages could be due to either

ERIC Full Text Provided by ERIC

competence differences or to psychometrical differences (Angoff & Cook, 1998; Geisinger, 1994; Hambleton, 1993, 1994; Prieto, 1992; Sireci, 1997). According to the Guidelines (ITC, 2002) and Gierl (2001), the comparability between constructs of dual language tests is the prior consideration before any other psychometric attempts to link or use IRT methods, such as IRT equating or differential item functioning (DIF), because of assumptions about latent trait of IRT models. Although the DIF analyses can be used to identify problematic items at item level, the factor structure of the tests can only be evaluated at the total test score level (Sireci, 1997; van der Vijver & Tanzer, 1998).

If the comparison between tests using different languages is meaningful, the construct measured by the tests must be equivalent (Gierl, Rogers, & Klinger, 1999; Hambleton, 1994; Hulin, 1987; van de Vijver & Hambleton, 1996; van de Vijver & Poortinga, 1997). The construct underlying a test is a theoretical representation of the underlying trait, concept, attribute, process, or structures the test is designed to measure (Cronbach, 1971; Messick, 1989). The construct equivalent will be achieved if the same construct is measured across different groups over different factors such as language and administration.

The National Nurse Aide Assessment Program (NNAAP) consists of written examination forms that use different languages (English and Spanish) under different administration conditions (with or without audio equipment assistant.) Therefore, determining test validity and fairness for the various examination modes was undertaken.

In seeking evidence of test validity and fairness, the research should address questions such as whether the test measures the same construct for all relevant populations. The difference in test scores of the NNAAP among examinees who use a different language with or without audio assistant could be due to language differences, administration condition (audio assist) differences, or true competence difference. The additional administration mode option makes comparisons between the NNAAP forms even more difficult than a comparison of different languages. Although countless studies using structural equation modeling (SEM), scaling, and exploratory factor analysis have been conducted to assess the structural equivalency of tests across language and cultural group (Gierl, 1999; Reise, Widaman, & Pugh, 1993; Robin, Sieci, & Hambleton, 2000; Sireci & Allalouf, in press), none of these studies using SEM has been done to evaluate the equivalence across language under different administration conditions.

This study investigates the structure of the certification test for the National Nurse Aide Assessment Program with regard to psychometric equivalence across dual languages and administration conditions.

Method and Data

Instrument

The National Nurse Aides Assessment Program is a nationally administered certifying examination program that is based on the activities and knowledge required for competent performance by nurse aides in long-term care, acute care, and home health care settings. The NNAAP consists of two components: the first is a knowledge test that is referred to as the written examination, and the second one is a skill demonstration that is called a skill evaluation. This study focuses on the written exam only.

The written exam forms are created according to a content outline based on the results of a job analysis conducted by the National Council of State Boards of Nursing (1995). The job analysis identified the most important activities performed by nurse aides across all settings and the knowledge required for performing each activity. Using the job analysis results, the subject matter experts developed the content outline and assigned proportionate weightings to each content area. Three major content areas are defined: (I) Physical care skills (47%), (II) Psychosocial care skills (22%), and (III) Role of the nurse aide (31%). Each of the major content area includes different subcontent areas. Physical care skills include: activities of daily living, basic nursing skills, and restorative skills. Psychosocial care skills include: emotional and mental health needs, and spiritual and culture needs. Role of the nurse aide include: communication, client right, legal and ethical behavior, and being a member of the health care team. Each written exam form consists of 70 multiple-choice items. Sixty are used to determine a candidate's test score, and the remaining ten items are pre-test items.

There are three formats for the NNAAP written exam. The standard format consists of test items written in English. For candidates with limited reading proficiency, the written English test items are used and administered together with cassette tapes that present directions and test questions orally. For Spanish speakers, the English version is translated into Spanish and administered with Spanish language tapes in the same mode as the English language oral administration.

Sample

In the years 1998 to 2000, a total of 273,492 nurse aide candidates from 31 states took the national nurse aides examination. A sample of 20,568 candidate test responses to one examination form created in 1998 was used in this study. These candidates were selected from 4 geographically representative states: Colorado, Florida, New Jersey, and Texas. Among the candidates, 10,908 (53%) took the English version of the exam form (E), 6,140 (30%) took the English with audio-tape version (EA), and 3,520 (17%) took the Spanish with audio-tape version (SA) of the same exam form. Table 1 shows background characteristics of NNAAP test for different examinee groups.

Table 1

Background Characteristics of the NNAAP Test for Different Groups

Background Characteristics	Written English (E)	Written English + Audio Tape (EA)	Written Spanish + Audio Tape (SA)
Sample Size	10908	6140	3520
Raw Score Mean	51.79	47.38	48.48
Raw Score SD	5.89	6.66	5.53
Reliability (KR-20)	0.84	0.79	0.76

Data Analyses

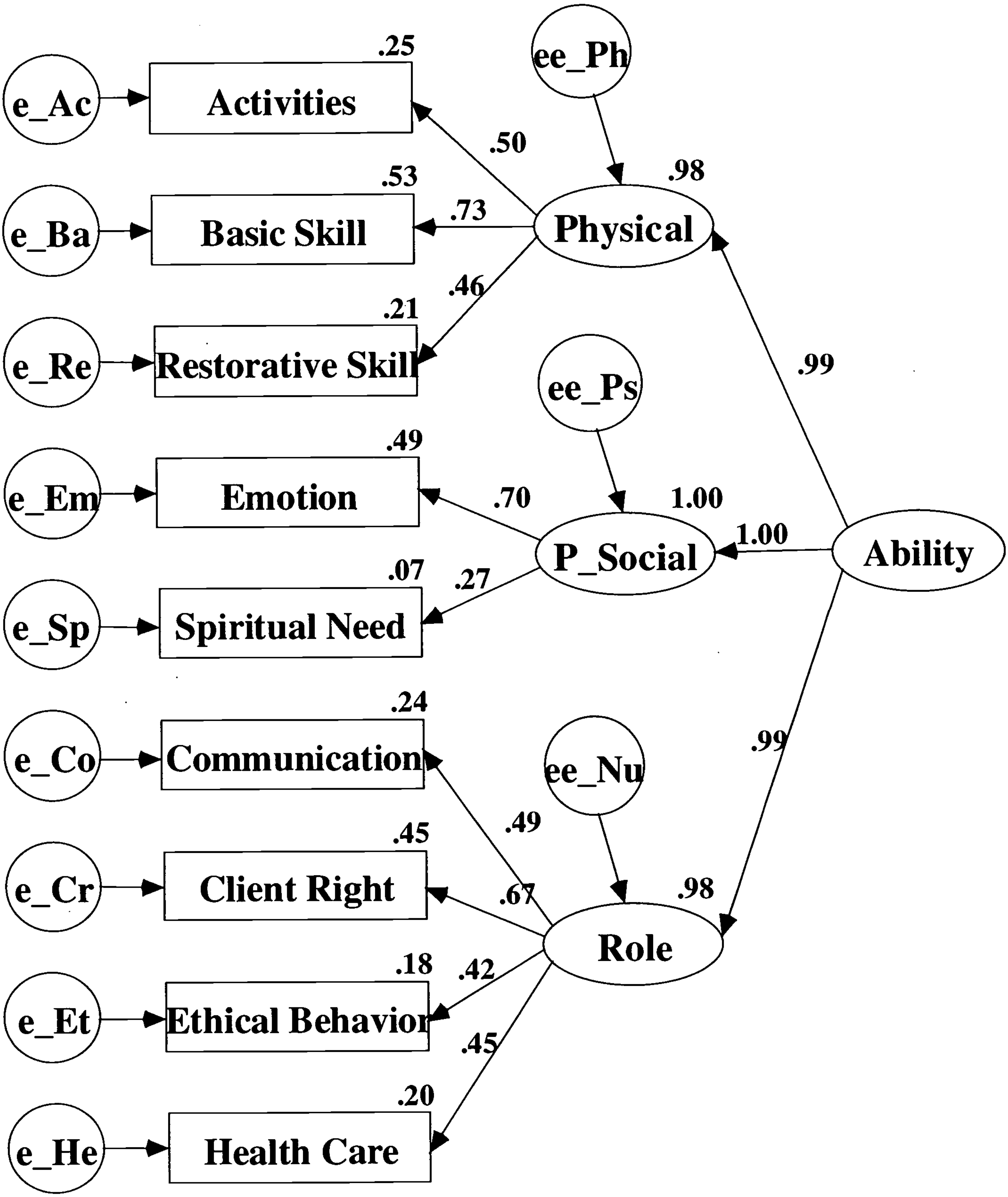
A series of structural equation modeling (SEM) procedures were conducted for this structure invariance study. For the purpose of cross-validation, subjects were randomly split into two samples to form a calibration and a validation sample (Byrne, 2001). One of the purposes for using a cross-validation strategy here is to assess the reliability of model fit. Having chosen a SEM model that is best for a particular sample of data, one may not automatically assume that this SEM model can be reliably applied to other samples of the same population. However, assuming the model fits well for the calibration sample, if the model also fits well for the validation sample, a different sample from the same population of interest, then we may say that this SEM model is reliable.

The NNAAP model (Figure 1) is a structural model with three endogenous latent variables as first-order factors. The model represents the 1995 NNAAP job analysis content outline (National Council of State Boards of Nursing, 1995.) Nurse aide ability was defined as a person's grasp of the basic knowledge and skills necessary to provide care to patients as a nurse aide, within regulatory guidelines. The first endogenous variable was measured by 3 subtests.

The second variable was measured by 2 subtests. The third endogenous variable was measuremd by 4 subtests. The 3 endogenous variables matched the 3 first level content areas of the outline. Nine subtests (observable variables) formed the second content levels. It was assumed that each of the groups' subtests measured a unique aspect of the NNAAP test. One exogenous variable was a second-order factor called competence or ability. Candidate ability was hypothesized to account for all variance and covariance related to the first-order factors. For identification purposes, all three first-order factor variances were set equal, and first factor loadings from each of three endogenous variables and the variance of ability were scaled to 1.0.

In order to evaluate the adequacy of the NNAAP factor model to fully account for the relationships among subjects, a series of SEM with the maximum likelihood estimation were conducted on the calibration sample, for each language and administration group. Once the fitting of model for each calibration sample was determined, the invariance of the model structure for the validation samples was investigated across dual languages (English and Spanish) and administration conditions (with or without audio assistant). All tests of invariance began with a global test of the equality of covariance structures across groups (Joreskog, 1971b.) The data for all groups were analyzed simultaneously to obtain efficient estimates (Bentler, 1995). Then, a series of nested constraints were equally applied to the same parameters across E, EA, and SA groups in order to detect the configuration and factor pattern difference across groups. The constraints used include, from weaker to stronger, (1) model structure, (2) model structure and factor loadings, and (3) model structure, factor loadings, and unique variance. Changes in goodness-of-fit statistics were examined to detect differences in structure parameters. Several well-known goodness-of-fit indexes were used to evaluate model fit: the Chi-square χ^2, the comparative fit index (CFI), unadjusted goodness-of-fit indexes (GFI), the normal fit index (NFI), the Tucker-Lewis Index (TLI), the root mean square error of approximation (RMSEA) and the standardized root mean square error residual (SRMR). All analyses were conducted by using AMOS 4.0 (Arbuck & Wothke, 1999). For the group comparisons with increased constraints, the value provides the basis of comparison with the previously fitted model, however, a significant value of χ^2 does not necessarily indicate a departure from invariance when the sample size is large because a chi-square test is correlated with sample size and will detect even minute differences between the hypothesized model and the data (Bollen & Long, 1993; Brown & Cudeck, 1993).

Figure 1. The NNAAP SEM Model of EA Test for Calibration Sample



Results

Evaluation of Model Fit Across Calibration and Validation Samples

Table 2 shows the fit indexes of the NNAAP model of cross-validation samples for different languages and administrations. Hu and Bentler (1999) recommend using combinations of goodness-of-fit indexes to obtain a robust evaluation of model fit. The criterion values they list for a model with a good fit are CFI>0.95, TLI>0.95, RMSEA<0.06, and SRMR<0.08. For this model, nearly all values satisfy the Hu and Bentler criteria for these four fit statistics. Other than the Chi-square, all values satisfy the Hu and Bentler criteria for these four fit statistics. Chi-squares are significant because the sample size is large. All the figures for GFI, AGFI, and NFI also support the evidence of fit for all groups. All factor loadings are reasonable and statistically significant. The overall picture suggests that the model provides reasonably close fits to the data and is cross-validated.

Evaluation of Equivalence Across Language

The goodness-of-fit indexes across languages in a nested series of tests are presented in Tables 3. Because both EA and SA groups, in both calibration and validation samples, used audio tapes, the only different factor between the two groups is the language factor (English and Spanish). For each of the EA and SA groups, the specified parameters for each constraint condition were constrained to be equal for both languages. For the calibration sample, the differences of χ^2 between the EA + SA baseline and the Constraint I nested models are not statistically significant at the 0.05 level even given the large sample size. The differences of χ^2 between Constraint II and I, Constraint III and II are significant and were expected for such large sample sizes. All other fit indexes are well under the Hu and Bentler (1999) criteria except for NFI and TLI for constraints II and III, and CFI for constraint III. For the validation sample, the fit indexes of GFI, NFI, TLI, RMSEA, and SRMR are all under Hu and Bentler criteria. This suggests that the factor structure, latent variances, and factor loadings of the NNAAP are the same for English and Spanish speakers. But the chances of unexplained unique variances varying across languages are still high.

Table 2

Summary of Fit Indexes of NNAAP Structure Model for Independent Groups

Sample/Group	N	*df*	χ^2	GFI	NFI	TLI	CFI	RMSEA	SRMR
Calibration Sample									
Written English (E)	5454	26	108.48	1.00	.99	.99	.99	.02	.01
Written English + Audio (EA)	3070	26	67.35	1.00	.99	.99	.99	.02	.02
Written Spanish + Audio (SA)	1760	26	75.00	.99	.97	.97	.98	.03	.02
Validation Sample									
Written English (E)	5454	26	74.47	1.00	.99	.99	1.00	.02	.01
Written English + Audio (EA)	3070	26	88.56	.99	.98	.98	.99	.03	.02
Written Spanish + Audio (SA)	1760	26	250.67	.99	.97	.96	.97	.04	.03
Total Sample									
Written English (E)	10908	26	135.99	1.00	.99	.99	1.00	.02	.01
Written English + Audio (EA)	6140	26	130.00	1.00	.99	.99	.99	.03	.01
Written Spanish + Audio (SA)	3520	26	124.10	.99	.97	.97	.98	.03	.02

Table 3

Goodness-of- Fit Indexes of Invariance of Model Constraints* Across Groups Based on Calibration and Validation Samples

Sample/Group	*df*	χ^2	GFI	NFI	TLI	CFI	RMSEA	SRMR
Calibration Sample								
E + EA Baseline	50	175.83	1.00	.99	.99	.99	.02	.01
Constraint I	53	176.74	.99	.99	.99	.99	.02	.01
Constraint II	62	253.87	.99	.98	.99	.99	.02	.03
Constraint III	71	1752.59	.95	.89	.90	.90	.05	.03
EA + SA Baseline	50	142.35	.99	.98	.98	.99	.02	.01
Constraint I	53	143.59	.99	.98	.98	.99	.02	.02
Constraint II	62	409.38	.98	.94	.94	.95	.03	.02
Constraint III	71	680.85	.97	.91	.91	.91	.04	.06
E + SA Baseline	50	183.48	.99	.99	.99	.99	.02	.01
Constraint I	53	189.15	.99	.99	.99	.99	.02	.02
Constraint II	62	462.36	.99	.97	.97	.97	.03	.07
Constraint III	71	1229.09	.96	.91	.91	.91	.05	.09
E + EA + SA Baseline	74	234.12	.99	.99	.99	.99	.01	.01
Constraint I	80	256.76	.99	.99	.99	.99	.01	.02
Constraint II	98	624.91	.99	.97	.97	.97	.02	.07
Constraint III	116	2491.02	.95	.87	.88	.87	.04	.09
Validation Sample								
E + EA Baseline	50	163.03	1.00	.99	.99	.99	.02	.02
Constraint I	53	163.39	1.00	.99	.99	.99	.02	.02
Constraint II	62	289.76	.99	.98	.98	.99	.02	.04
Constraint III	71	1710.53	.95	.90	.90	.91	.05	.03
EA + SA Baseline	50	339.23	.99	.97	.97	.98	.03	.02
Constraint I	53	341.16	.99	.97	.97	.98	.03	.03
Constraint II	62	425.55	.99	.97	.97	.97	.03	.03
Constraint III	71	514.49	.99	.96	.96	.96	.03	.03
E + SA Baseline	50	325.14	.99	.98	.98	.99	.02	.03
Constraint I	53	326.50	.99	.98	.98	.99	.02	.02
Constraint II	62	551.15	.99	.97	.97	.97	.03	.04
Constraint III	71	2405.79	.95	.87	.88	.88	.06	.05
E + EA + SA Baseline	74	413.7	.99	.98	.98	.99	.02	.02
Constraint I	80	415.90	.99	.98	.98	.99	.02	.03
Constraint II	98	723.95	.99	.97	.97	.97	.02	.04
Constraint III	116	2990.66	.95	.88	.89	.88	.04	.05

* The levels of model constraints restricted to be equal across language (or administration) are:

I. Model structure and latent variable variance.

II. Model structure, latent variable variance, and factor loading.

III. Model structure, latent variable variance, factor loading, and unique variance.

Evaluation of Equivalence Across Administration Conditions

Also from Table 3, for both calibration and validation samples, the goodness-of-fit results for E + EA groups show if the administration condition factor affects the equivalence of factor structure of the NNAAP because both groups used the same language (English) and the language factor can be canceled out. For both calibration and validation samples, the differences of χ^2 between the E + EA baseline and the Constraint I nested models are not statistically significant at the 0.05 level even given the large size of the samples. The differences of χ^2 between Constraints II and I, Constraints III and II are significant and were expected for such large sample sizes. All other fit indexes are well under the Hu and Bentler (1999) criteria except NFI, TLI, and CFI for Constraint III. This suggests that the factor structure, latent variances, and factor loadings of the NNAAP are same for English and English with audio assistant. However, the chances of unexplained unique variances varying across administration condition are still high.

Evaluation of Equivalence Across Language and Administration Condition

The goodness-of-fit results for the E + SA and E + EA + SA groups for both the calibration and validation samples are shown in Table 3. Both conditions mixed language and administration conditions together. Any differences of nested models across language and administration condition could be due to the either language factor or the administration condition factor, or even both. For both the calibration and validation samples, the differences of χ^2 between the E + SA baseline and the Constraint I nested models are not statistically significant at the 0.05 level even given the large size of the samples. The differences of χ^2 between Constraints II and I, Constraints III and II, are significant and were expected for such large sample sizes. All other fit indexes are well under the Hu and Bentler (1999) criteria except NFI, TLI, CFI, and SRMR for constraint III. The differences of χ^2 between the E + EA + SA baseline and the Constraint I nested models are not statistically significant at the 0.05 level even given the large size of the samples. The differences of χ^2 between Constraints II and I, Constraints III and II, are significant and were expected for such large sample sizes. All other fit indexes are well under the Hu and Bentler (1999) criteria except NFI, TLI, CFI, RMSEA and SRMR for Constraint III. This suggests that the factor structure, latent variances, and factor loadings of the NNAAP are the same for English, English with audio assistant, and Spanish with audio groups. However, the chances of unexplained unique variances varying across administration condition are still high.

Practical Implication

The present study examined the comparability of NNAAP scores across language and administration condition groups for calibration and validation samples that were randomly drawn from the same population. Results show that factor structure validities of the NNAAP are well supported. Statistically significant χ^2 (or difference of χ^2) statistics occur because of the large sample sizes. For this reason, it is frequently appropriate to conclude that a SEM model fits the data even if *p* is significant (Joreskog & Sorbom, 1989; Mulaik, James, Alstine, Bennett, Lind, & Stillwell, 1989). The values of all other fit statistics (CFI, AGFI, NFI, TLI, RMSEA, and SRMR) fall within the bounds of Hu and Bentler's (1999). Thus, the overall pattern of fit statistics for the NNAAP data indicates a reasonable fit even when the chi-square test suggests rejection of factor models when sample sizes are large. The evidence of fit holds for both the calibration and validation samples for language and administration condition groups. Further evidence of the invariance of factor structure of the NNAAP scores across language and administration groups is found in all fit statistics when model structure, factor loading, and latent variable variance are constrained to be equal across groups except unique variance. Thus, the data suggest that this construct is similarly structured (fair) across different language and administration condition groups.

In summary, this study underscores the importance of empirical validation of certification exams and provides evidence supporting the validity and fairness of a widely used national exam. It carries the validation process beyond the content-related evidence (job analysis) that often serves as the sole documented support of validity for credentialing exams. By publicizing the results of this study, we hope to encourage the credentialing community to strengthen the validity of its exams by investigating their factor structure and making modifications, if warranted, to ensure that the same constructs are measured regardless of language and administration condition. We also hope to encourage the practice of providing evidence of validity from a variety of sources, thus strengthening the defensibility of licensure and certification exams across the board.

Reference

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

Angoff, W. H., & Cook, L. L. (1988). *Equating the scores of the Prueba de Aptitude Academica and the Scholastic Aptitude Test (Report No. 88-2).* New York: College Entrance Examination Board.

Arbuckle, J. L., & Wothke, W. (1999). *Amos 4.0 User's Guide.* Chicago, IL: SmallWaters Corporation.

Bentler, P. M. (1995). *EQS: Structural equations program manual.* Encino, CA: Multivariate Software, Inc.

Bollen, K. A., & Long, J. S. [Ed.] (1993). *Testing structural equation models.* Newbury Park, CA: Sage.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In Bollen, K. A. and Long, J. S. [Ed.] *Testing structural equation models.* Newbury Park, California: Sage, 136-162.

Byrne, B. M. (2001). *Structural equation modeling with AMOS.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Chornick, N. L., Yocom, C. J., Jacobson, J., & Ginsburg, K. (1995). *Job Analysis: nurse aides.* National Council of State Boards of Nursing, Inc. Chicago, IL.

Chronbach, L. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed). Washinton DC: American Council on Education.

Equal Employment Opportunity Commission [EEOC], Civil Service Commission, Department of Labor and Department of Justice. (1978, August 25). Uniform guidelines on employee selection procedures. *Federal Register, 43*, 38290-38315.

Geisinger, K. F. (1994). Cross-cultural normative assessment: translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6,* 304-312.

Gierl, M. J. (2001). Construct equivalence on translated achievement tests. *Canadian Journal of Education.*

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999, April). *Consistency between statistical procedures and content reviews for identifying translation DIF.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229-244.

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guideline. *European Journal of Psychological Assessment, 17,* 164-172.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6, 1-55.*

Joreskog, K. G. (1971b). Simultaneous factor analysis in several populations. *Psychometrika,* 36, 409-426.

Mehrens, W. A. (1994, Winter). The validity of licensing and certification exams. *CLEAR Exam Review*, 19-21.

Mehrens, W. A., & Popham, W. J. (1992). How to evaluate the legal defensibility of high-stakes tests. *Applied Measurement in Education, 5(3),* 265-283.

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

Robine, F., Sireci, G. S., & Hambleton, R. K. (2000). *Evaluating the equalence of different language versions of a credentialing exam* (LR-359). Unpublished manuscript, University of massachusetts Amherst, Laboratory of Psychometrics and Evaluative Research.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invarinace. *Psychological Bulletin, 114,* 552-556.

Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practices, 16,* 12-19.

Sireci, S. G., & Allalouf, A. (in press). Appraising item equivalence across multiple languages and cultures. *Language Testing.*

Sireci, S. G., & Khaliq, S. N. (2002*). An analysis of the psychometric properties of dual language test forms*. Paper presented at the annual meeting of NCME, New Orleans, LA.

van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*, 89-99.

van der Vijver, F., & Tanzer, N. K. (1998). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47*, 263-279.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Construct Equivalence of a National Certification Examination That Uses Dual Languages and Audio Assistant	
Author(s): Shudong Wang, Ning Wang	
Corporate Source: Harcourt Educational Measurement Promissor	Publication Date: 4/2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☒	Level 2A ☐	Level 2B ☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Shudong Wang	Printed Name/Position/Title: Shudong Wang, Ph.D.	
Organization/Address: Harcourt Educational Measurement 19500 Bulverde Road San Antonio, Texas 78259-3701	Telephone: 210-339-5535	FAX: 210-339-5973
	E-Mail Address: Shudong_wang@Harcourt.com	Date: 9/29/03

ERIC Full Text Provided by ERIC

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone:	**301-552-4200**
Toll Free:	**800-799-3742**
FAX:	**301-552-4700**
e-mail:	**info@ericfac.piccard.csc.com**
WWW:	**http://ericfacility.org**