DOCUMENT RESUME

ED 478 078 TM 035 019

AUTHOR	Wang, Shudong; Witt, Elizabeth
TITLE	Validation and Invariance of Factor Structure of a National Licensing Examination across Gender and Race.
PUB DATE	2002-04-00
NOTE	20p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleams, LA, April 1-5, 2002).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Factor Structure; *Licensing Examinations (Professions); *Racial Differences; *Real Estate Occupations; *Sex Differences; *Validity
IDENTIFIERS	*Invariance

ABSTRACT

In the context of licensure testing, this study addressed the importance of supplementing the usual content-related validity evidence (job analysis) with empirical validation. Evidence supporting the validity and fairness of the Real Estate National Licensing Examination (RENSE) is provided. Confirmatory factor analysis (CFA) with structural equation modeling (SEM) was used to investigate the internal structural validity of the RENSE across gender and race. Study data were sampled from raw scores of 21,301 real estate sales licensure candidates. For the purpose of cross-validation, the fit of two competing models was examined for a base calibration and a validation sample. Evidence of the invariance of factor structure of RENSE scores across race and gender was found in all fit statistics when model structure, factor loading, latent variable variance, and unique variance are constrained to be equal across groups. Results contribute to the body of evidence supporting the validity and fairness of the RENSE.(Contains 4 tables and 29 references.) (Author/SLD)

Validation and Invariance of Factor Structure of a National Licensing Examination Across Gender and Race

Shudong Wang, Ph.D.

Elizabeth Witt, Ph.D.

CAT*ASI

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April 2002. All correspondence should be sent to Shudong Wang, CAT*ASI, Three Bala Plaza West, Bala Cynwyd, PA 19004 -3481. E-mail: *Shudong_wang@asisvcs.net*.

Abstract

In the context of licensure testing, this paper addresses the importance of supplementing the usual content-related validity evidence (job analysis) with empirical validation. Evidence supporting the validity and fairness of the Real Estate National Licensing Examination (RENSE) is provided. Confirmatory factor analysis (CFA) with structural equation modeling (SEM) is used to investigate the internal structural validity of the RENSE across gender and race. For the purpose of cross-validation, the fit of two competing models is examined for both a base calibration and a validation sample. Evidence of the invariance of factor structure of RENSE scores across gender and race group is found in all fit statistics when model structure, factor loading, latent variable variance, and unique variance are constrained to be equal across groups. Results contribute to the body of evidence supporting the validity and fairness of the RENSE.

Validation and Invariance of Factor Structure of a National Licensing Examination Across Gender and Race

Introduction

Licensing tests exist to protect the public by ensuring that entry-level professionals possess the relevant knowledge and skills in sufficient degree to perform their jobs competently. Like other credentialing tools, licensing tests are intended to help the public, employers, and government agencies identify practitioners who have met a particular standard. In most states, mandatory licensure programs are among the most restrictive regulatory programs (Nelson, 1994). Licensing organizations have a responsibility not only to candidates--to ensure that all licensure procedures are fair and consistent--but also to the consumer--to ensure the validity of the licensure process so that individuals who are licensed are indeed competent. Like any high-stakes tests, licensing tests must satisfy the legal requirements of validation and fairness. Validity, according to the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), is the most important consideration in test development and evaluation; fairness is also required by the *Standards* (AERA et al., 1999) as well as by federal laws and regulations (Equal Employment Opportunity Commission [EEOC], Civil Service Commission, Department of Labor and Department of Justice, 1978; Mehrens & Popham, 1992; Mehrens, 1994).

Validity refers to the degree to which empirical evidence and theoretical rationale support the inferences and actions based on test scores (Messick, 1989). Traditionally, test validation has focused on three aspects of validity evidence: content-related, criterion-related, and construct-related evidence. Credentialing exams, including those used in licensure, rely primarily on content-related validity evidence. Typically the central support of licensure test validity is a job analysis that identifies the knowledge and skills required for competent performance and weights them according to their importance in protecting the public (Stocker & Impara, 1995). Criterion-related validity studies are often not feasible in licensure testing, and predictive validity is not a concern, since the exams are neither designed nor employed to predict future professional success (Kane, 1982, 1992, 1994). However, as Cronbach (1971) and Messick (1980, 1989) argue, validity should be considered a unitary concept; all aspects of validity evidence ultimately serve to support construct validity. The more evidence collected, the better. The *Standards* (AERA et al., 1999) also confirm the unitary concept of validity and recommend integrating evidence from a variety of sources.

The test validation process in licensure and certification has come to rely heavily on content validation procedures based on a job analysis or practice analysis (AERA et al., 1999; Kane, 1997; Knapp & Knapp, 1995), but the most carefully collected job analysis data does not, in itself, constitute definitive evidence of the validity of test scores. Rather the job analysis data must be viewed as contributory evidence in the "interpretive argument" for the validity of the test scores, as one piece of evidence supporting a specific interpretation or use of test scores (Newman, Slaughter, & Taranath, 1999). Although job analysis is necessary for the development of valid credentialing examinations, it does not sufficiently address all aspects of validity. A job analysis can provide strong evidence that a test measures primarily relevant knowledge, skills, and abilities (KSAs), yet it does not guarantee the absence of irrelevant constructs (Raymond, 1995). Nor does a job analysis detect or prevent item or test bias. Thus additional evidence of validity is desirable. Factorial validity (Guilford, 1946), or the investigation of the factor structure underlying a test, can be a valuable component of validity evidence (Messick, 1995) and can be used to support the fairness of the tests. The *Standards* (AERA, APA, NCME, 1999) advise providing evidence of structural validity when relevant to the purpose and use of the exam (see Standards 1.11 and 1.12, p. 20). The *Standards* also point out the relationship between fairness and the construct being assessed:

> Regardless of the purpose of testing, fairness requires that all examinees be given a comparable opportunity to demonstrate their standing on the construct(s) that test is intended to measure. (p. 74)

In seeking evidence of test fairness, the researcher should address questions such as whether the test measures the same construct in all relevant populations. Fairness is closely related to the factor structure validity of the test. Factorial validity may be used not only to evaluate the dimensionality of an exam, but also to provide evidence of fairness. Similarity of factor structure across gender groups, for example, suggests that the test is measuring the same construct(s) for males and females. Different factor structures could imply that different constructs are being measured for the two groups. Of course, if evidence of differential validity is found, further investigation is needed. Factorial validity procedures, in and of themselves, cannot tell for which group validity is higher, nor can they explain why group differences occur. They can only serve as a flag to identify where psychological constructs may be structured differently over different subpopulations.

Despite emphasis on factor structure validity by some researchers, little attention has been devoted to structural validations of standardized tests (Stevens, 2001), especially in licensure testing (LaDuca, 1994). Job analysis plays a vital role in validating credentialing examinations by ensuring that test content specifications are closely tied to the job itself. Indeed, a job analysis may be essential to conform to professional standards and legal requirements in licensure testing. However, as Raymond (1995) expressed, a job analysis "tells us very little about the nature of test scores." (p. 32). Job analysis should be the beginning of the validation process for credentialing exams rather than the end of it. We need to expand our practical validation procedure -- one that is currently based solely on job analysis -- to focus more on evidence and theory related to the internal structure of the test whose scores are intended for a specific use and interpretation. In licensure settings, an investigation of factor structure is often both feasible and useful in providing additional evidence of validity.

The purposes of this study are: (a) To investigate the factorial (structure) validity of a major national licensing test. (b) To apply confirmatory factor analyses (CFAs) to cross-validate the resulting structural models across a second independent sample within each gender and racial group. (c) To investigate the invariance of measurement model structure, latent variable variance, factor loading, and unique variance across gender and racial groups. In so doing we will (d) supplement the usual content-related evidence and support the overall construct validity of the instrument, extending validity evidence beyond the methodology typically used in licensure testing.

Method

Sample

The study data were sampled from raw test scores of a total of 21,301 real estate sales licensure candidates who took a real estate licensure exam in the years 1998 to 2000. The test was administered on computer at 75 different test centers in 19 states. Among the participants, 11,893 (56%) were women and 9408 (44%) were men. Among the female participants, there were 10,243 White, 835 Black or African-American, 422 Hispanic or Latin-American, and 393 Asian, Asian-American or Pacific Islander candidates; among the male participants, there were 8059 White, 643

Black or African-American, 328 Hispanic or Latin-American, and 378 Asian, Asian-American or Pacific Islander candidates.

Instrument

The CAT*ASI real estate examinations consist of two parts: The first covers general topics and is administered nationally; the second is a state-specific test covering state laws. This study focuses on the national exam only. One examination is administered for test takers seeking licensure as salespeople; a broker level examination is administered for test takers who want to become brokers. This study focuses only on the salesperson examination that we will refer to in this paper as the Real Estate National Salesperson Examination (RENSE). This national test for salespeople consists of 80 scored questions and 5 pretest questions. Test forms are developed according to a content outline based on a rigorous job analysis (Newman & Joseph, 1998). The job analysis identified the most important tasks performed by real estate salespersons and the knowledge required to perform each task. After screening out tasks unrelated to the protection of the public, the job analysis committee classified knowledge statements by content area, and assigned proportionate weightings to each content area. Five major content areas were defined: (I) Real property characteristics, definitions, ownership, restrictions, and transfer (16 items – 20%); (II) Assessing and explaining property valuation and the appraisal process (12 items – 15%); (III) Contracts, agency relationships with buyers and sellers, and federal requirements (20 items – 25%); (IV) Financing the transaction and settlement (20 items - 25%); and (V) Leases, rents, and property management (12 items - 15%). For convenience, we shall refer to these five areas as: (I) Terminology; (II) Valuation; (III) Contracts; (IV) Finance; and (V) Property. This study examines data from Form X, which is one of three equivalent forms used in practice.

Data Analyses

All analyses were conducted using AMOS 4.0 (Arbuckle & Wothke, 1999). First, for the purpose of validating the factorial structure of the test, two competing models were investigated. Model A was a single-factor model, the factor being defined as "essential real estate sales ability." Model B comprised essential real estate sales ability and a second factor labeled, for lack of a better term, as "facility." Ultimately, the strength of the validity argument for the use of occupational tests in licensing depends upon evidence that the test scores are related to competence in the profession (Downing & Haladyna, 1997; Kane, 1997, 1994; Raymond, 1995; Nelson, 1993; Harvey, 1991). The rationale for using underlying essential real estate sales ability in both models reflects the structure of real estate salespersons' professional competence as defined by the job analysis. This essential real estate sales ability can be defined as a person's grasp of the legal and technical knowledge necessary to perform his/her job competently within regulatory guidelines. It does not include sales techniques, persuasive skills, and the like. The second factor is quite similar to the first, except that it is less closely related to Finance and more closely related to Property Management. As a sample, Models A and B for the White female group are graphically represented in Figures 1 and 2. All groups exhibited this same pattern of loadings and correlations. Confirmatory factor analysis (CFA) was conducted to determine the adequacy of fit of both models via structural equation modeling (SEM).

Second, for the purpose of cross-validation, subjects were grouped according to gender and race, and each group was separately and randomly split into two to form a base calibration sample and a validation sample. One of the purposes of 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 same population. However, assuming the model fits well for the base 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.

To evaluate the adequacy of the one-factor model to fully account for the relationships among subtests, a CFA using SEM with maximum likelihood estimation was conducted on the calibration sample for each gender/ethnic group. Once the best-fitting model for each base sample was determined, the validity of the model structures for the validation samples was investigated.

As a third step, the appropriate models were compared across gender to determine whether gender invariance was supported. Finally, similar comparisons were made across race to investigate whether the RENSE measures the same construct(s) for different racial groups. Both tests of invariance began with a global test of the equality of covariance structures across groups (Joreskog, 1971b) and the data for all groups were analyzed simultaneously to obtain efficient estimates (Bentler, 1995). In both steps, a series of nested constraints were equally applied to the same parameters across validation samples (gender or race) for subsequent testing of increasingly restrictive hypotheses. This was done in an effort to identify the source of departures from invariance, if any. In other words, these constraints were applied to ascertain whether invariance held for certain parameters (model structure, factor loading, and unique variance) of measurement

Figure 1. One Factor Model for White Female (Original Sample)

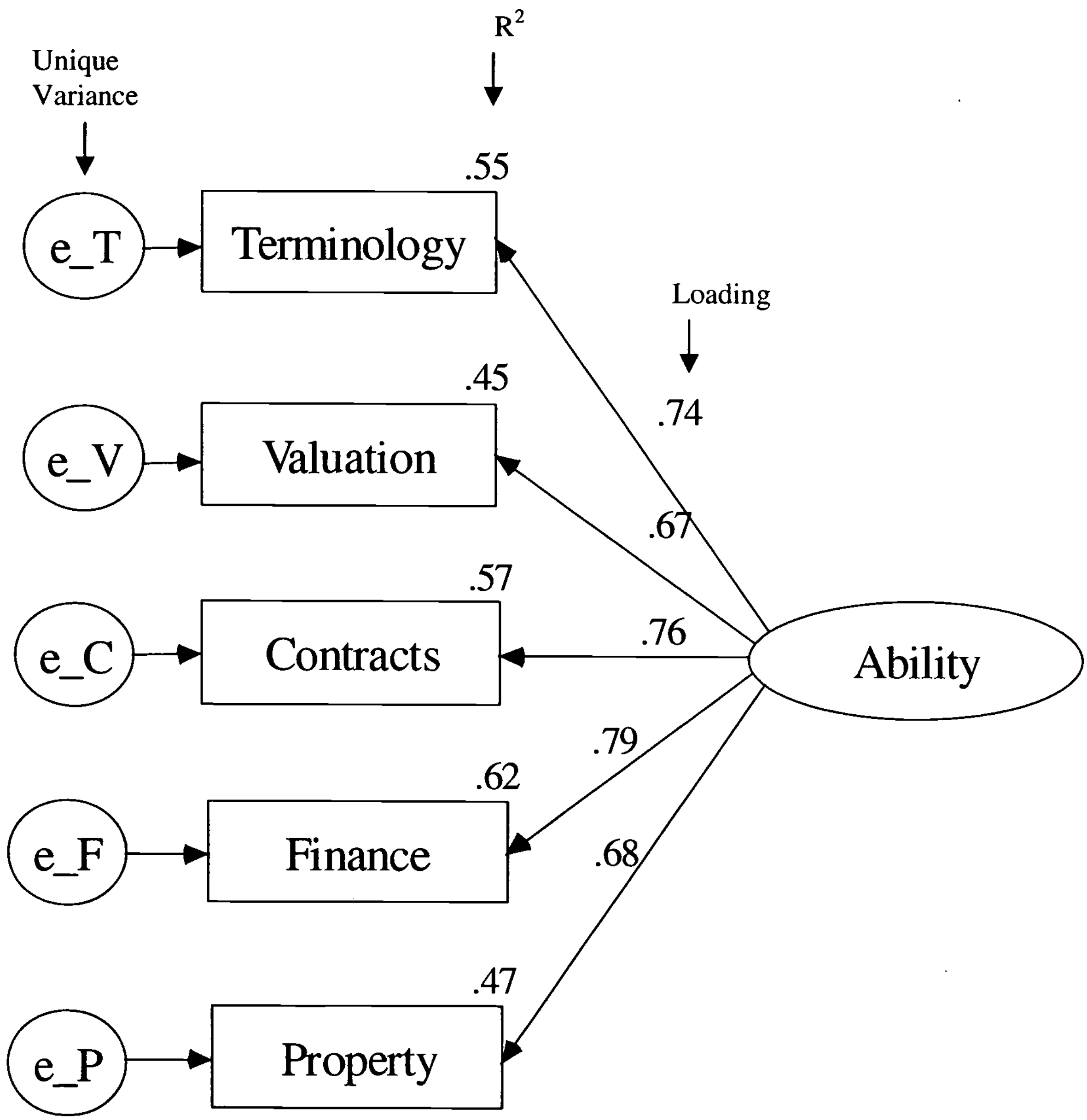
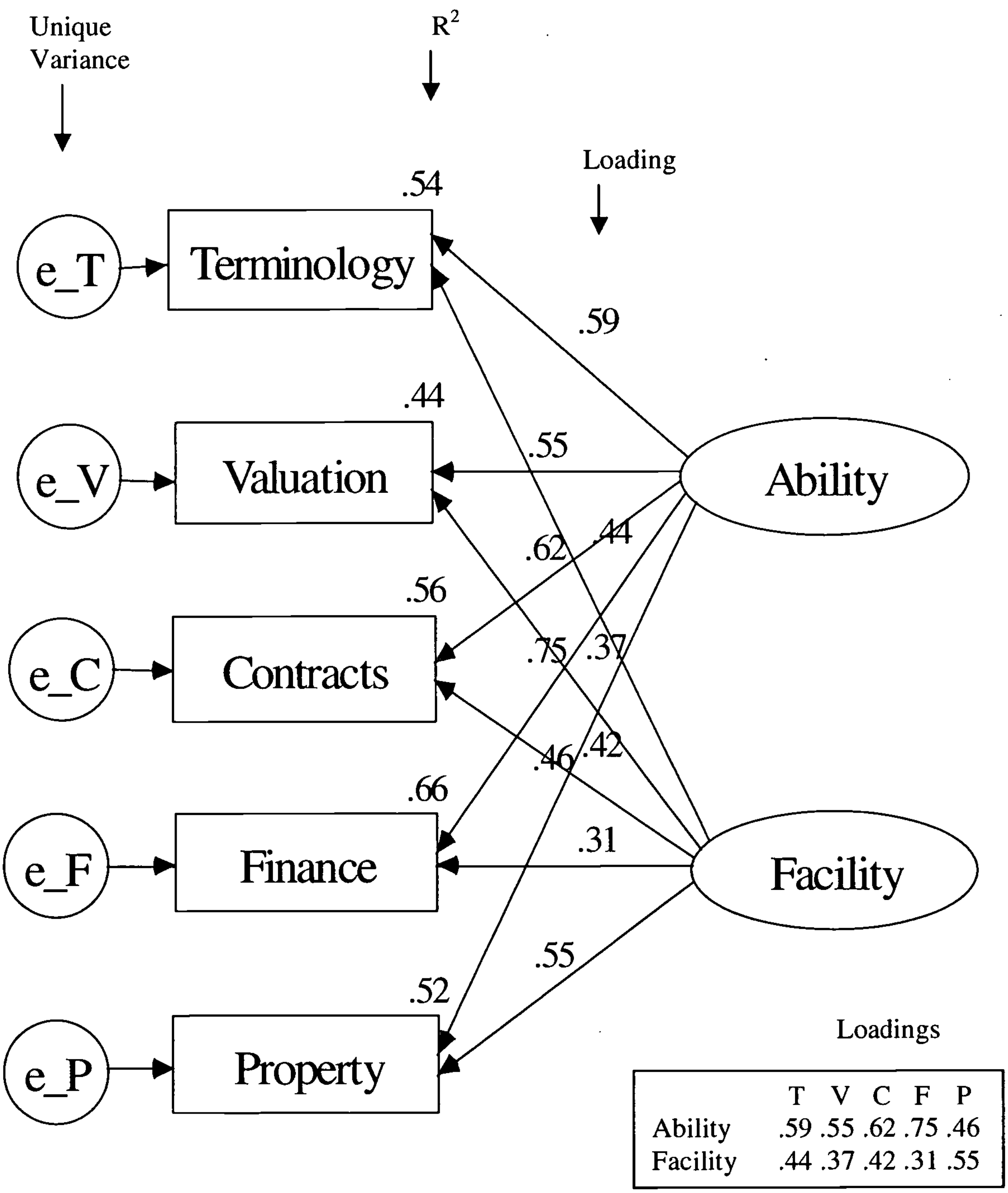


Figure 2. Two Factor Model for White Female (Original Sample)



	T	V	C	F	P
Ability	.59	.55	.62	.75	.46
Facility	.44	.37	.42	.31	.55

models across both gender and race. The constraints used in the tests include, from weaker to stronger, (1) model structure, (2) model structure and factor loadings, and (3) model structure, factor loadings, and unique variance. All models were identified by fixing the factor(s) variance at 1.0; for Model B, two regression weights (for Terminology and Property Management) were also fixed at 1.0 for the purpose of model identification (Arbuckle, 1999). Changes in goodness-of-fit statistics were examined to detect differences in structural parameters. Several well-known goodness-of-fit indexes were used to evaluate model fit: the chi-square χ^2, the comparative fit index (CFI), both the unadjusted and adjusted goodness-of-fit indexes (GFI and AGFI), 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). For the group comparisons with increased constraints, the χ^2 value provides the basis of comparison with the previously fitted model. A non-significant difference in χ^2 values between nested models reveals that all equality constraints hold across the groups. Therefore, the measurement model remains invariant across groups as the constraints are increased. Sample size must be taken into account, however, in interpreting a significant χ^2. A significant χ^2 does not necessarily indicate a departure from invariance when the sample size is large.

Results

Evaluation of Model Fit

Table 1 shows the fit indexes for both the one- and two-factor solutions of the RENSE for the different gender and race groups in the base calibration sample. 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 good fit are CFI>0.95, TLI>0.95, RMSEA<0.06, and SRMR<0.08. For the one-factor model, nearly all values satisfy the Hu and Bentler criteria for these four fit statistics. The only exception is the RMSEA value of 0.076 for Black females. For the two-factor model, all values satisfy the criteria. Chi-squares are significant only where the sample size is very 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 both the one-factor and the two-factor model provide reasonably close fits to the data. Since the adequacy of both models is supported, the judgement as to which model should be chosen ultimately rests on the substantive meaningfulness of the underlying theory. In this case, the one-factor model is more interpretable than the two-factor model; the second factor is difficult to describe or even name.

Although factors should not be considered the equivalent of dimensions, the results of these initial CFAs do provide some support for the unidimensionality of the test.

Evaluation of Equality Across Base Calibration and Validation Samples

Because both models were reasonably good fits, both were examined for cross-validation. Table 2 displays the four main fit indexes for cross-validation of both models using the base calibration and validation samples. Within each race and gender group, equal numbers of subjects were randomly assigned to the calibration or validation sample; the counts shown in Table 2 indicate the number in *each* sample. All goodness-of-fit indexes satisfy the Hu and Bentler (1999) criteria well and are quite comparable for the base calibration and validation samples. (The GFI, AGFI, and NFI are omitted here to save space. The values for these indexes are even closer for the two groups than the figures shown in Table 2.) The most notable difference is that the one group (Black females) that did not show a good fit of RMSEA for the base calibration sample (RMSEA is 0.076) does in fact meet the criteria for fit in the validation sample (RMSEA is 0.042).

Evaluation of Equality Across Gender Samples

The goodness-of-fit indexes across gender under both the one- and two-factor models in a nested series of tests are presented by racial group in Table 3. Because the difference in sample sizes between the genders is small, equal sample sizes were obtained across gender by randomly trimming the larger sample to match the smaller sample size. For each race, the specified parameters for each constraint condition were constrained to be equal for both genders. Although not listed in Table 3, for all races, the chi-square *differences* among the nested models are statistically nonsignificant at the 0.01 level except for the White group. (This was expected because of large white sample size.) All other fit indexes also indicate no gender differences under a variety of model constraint conditions. This suggests that the factor structure of the RENSE is the same for males and females within racial group.

Evaluation of Equality Across Race Samples

The specified parameters for each condition were constrained to be equal across the four racial groups for each gender. The goodness-of-fit indexes across race under both the one- and two- factor models in a nested series of tests are presented in Table 4. In terms of the χ^2 difference values, the fit between each nested model is significantly different. However, this is undoubtedly an artifact of the large sample size. In terms of the other fit indexes, the factor structures appear to be the same for all four racial groups within gender.

Table 1

Summary of Fits Indexes of One- and Two-Factor Models for RENSE Structure by Gender and Race (Base Calibration Sample)

Model and Group	N	χ^2	p for χ^2	CFI	GFI	AGFI	NFI	TLI	RMSEA	SRMR
One Factor Model (df = 5)										
Female										
White	5075	31.275	<.001	.997	.998	.993	.997	.995	.032	.010
Black	418	16.933	.005	.985	.985	.955	.980	.971	.076	.024
Hispanic	209	4.007	.548	1.000	.992	.977	.992	1.004	.000	.014
Asian	191	3.295	.655	1.000	.991	.972	.990	1.001	.000	.016
Male										
White	3983	29.181	<.001	.997	.997	.991	.996	.993	.035	.011
Black	317	7.839	.165	.996	.990	.969	.989	.992	.042	.018
Hispanic	153	7.737	.171	.993	.982	.945	.980	.986	.060	.022
Asian	188	4.327	.503	1.000	.991	.972	.991	1.003	.000	.016
Two Factor Model (df = 2)										
Female										
White	5075	16.723	<.001	.999	.999	.990	.998	.992	.038	.007
Black	418	1.581	.454	1.000	.999	.991	.999	1.000	.000	.008
Hispanic	209	1.244	.537	.998	.982	.998	.999	1.007	.000	.007
Asian	191	.358	.836	1.000	.999	.993	.999	1.016	.000	.004
Male										
White	3983	9.182	.010	.999	.999	.993	.999	.995	.030	.006
Black	317	1.93	.381	1.000	.998	.986	.998	1.000	.000	.009
Hispanic	153	3.224	.223	.997	.993	.945	.992	.986	.057	.014
Asian	188	2.990	.224	.997	.993	.950	.994	.989	.000	.013

Table 2

Comparison of the Cross-validation Samples by Race and Gender for Both One- and Two- Factor Models

			Base Calibration Sample					Validation Sample				
Race/Gender Group		df	χ^2	CFI	TLI	RMSEA	SRMR	χ^2	CFI	TLI	RMSEA	SRMR
White	Female (N = 5075)											
	One Factor Model	5	31.275	.997	.995	.032	.010	25.298	.998	.996	.028	.009
	Two Factor Model	2	16.723	.999	.992	.038	.007	13.054	.992	.993	.033	.006
	Male (N = 3983)											
	One Factor Model	5	29.181	.997	.993	.035	.011	14.324	.999	.999	.022	.008
	Two Factor Model	2	9.182	.999	.995	.030	.006	4.975	.999	.998	.019	.004
Black	Female (N = 418)											
	One Factor Model	5	16.933	.985	.971	**.076**	.024	7.839	.996	.992	.042	.018
	Two Factor Model	2	1.581	1.000	1.000	.000	.008	1.931	1.000	1.000	.000	.009
	Male (N = 317)											
	One Factor Model	5	7.839	.996	.986	.042	.018	9.363	.994	.999	.046	.020
	Two Factor Model	2	1.930	1.000	1.000	.000	.009	1.455	1.000	1.004	.000	.007
Hispanic	Female (N = 209)											
	One Factor Model	5	4.007	1.000	1.004	.000	.014	7.781	.994	.987	.052	.023
	Two Factor Model	2	1.244	.998	1.007	.000	.007	3.849	.996	.978	.067	.015
	Male (N = 153)											
	One Factor Model	5	7.737	.993	.986	.060	.022	3.537	1.000	1.000	.000	.014
	Two Factor Model	2	3.004	1.000	.987	.057	.014	2.737	.998	.991	.049	.012
Asian	Female (N = 191)											
	One Factor Model	5	3.295	1.000	1.000	.000	.016	5.233	.999	.999	.016	.017
	Two Factor Model	2	.358	1.000	1.016	.000	.004	.818	1.000	1.012	.000	.007
	Male (N = 188)											
	One Factor Model	5	4.327	1.000	1.003	.000	.016	6.946	.996	.991	.046	.022
	Two Factor Model	2	2.990	.997	.989	.000	.013	.421	1.000	1.017	.000	.005

Table 3

Goodness-of-Fit of Invariance of Models Constraints* Across Gender by Race (Validation Sample)

Model and Constraint	N_F, N_M	*df*	χ^2	*p*	CFI	TLI	RMSEA	SRMR
One Factor Model								
White	3983							
Constraint I		10	37.991	<.000	.998	.996	.019	.010
Constraint II		15	65.500	<.000	.997	.995	.021	.016
Constraint III		20	157.364	<.000	.991	.991	.029	.012
Black	317							
Constraint I		10	11.719	.070	.995	.990	.031	.020
Constraint II		15	20.876	.141	.995	.994	.025	.027
Constraint III		20	27.602	.119	.994	.994	.025	.030
Hispanic	153							
Constraint I		10	7.357	.691	1.000	1.000	.000	.020
Constraint II		15	15.904	.388	.999	.998	.014	.039
Constraint III		20	23.597	.260	.995	.995	.024	.051
Asian	188							
Constraint I		10	11.719	.304	.998	.996	.021	.017
Constraint II		15	13.891	.534	1.000	1.002	.000	.022
Constraint III		20	20.788	.410	.999	.999	.010	.027
Two Factor Model								
White	3983							
Constraint I		7	16.166	.003	.999	.996	.020	.007
Constraint II		12	43.791	<.000	.998	.996	.018	.015
Constraint III		17	134.688	<.000	.992	.991	.029	.010
Black	317							
Constraint I		7	11.900	.104	.996	.989	.033	.019
Constraint II		12	16.927	.152	.996	.994	.025	.023
Constraint III		17	23.113	.146	.995	.994	.024	.026
Hispanic	153							
Constraint I		7	5.305	.623	1.000	1.000	.000	.016
Constraint II		12	14.127	.293	.997	.995	.024	.038
Constraint III		17	21.927	.188	.993	.992	.031	.049
Asian	188							
Constraint I		7	11.708	.111	.995	.986	.042	.017
Constraint II		12	12.515	.405	.999	.999	.011	.020
Constraint III		17	20.458	.252	.996	.995	.023	.027

Note. N_F and N_M represent female and male sample size.
* The levels of model constraints that were restricted to be equal across race 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.

Table 4

Goodness-of-Fit for Invariance of Models Constraints* Across Race by Gender (Validation Sample)

Model and Constraint	*df*	χ^2	*p*	CFI	TLI	RMSEA	SRMR
Female							
One Factor Model							
Constraint I	20	47.990	<.000	.998	.995	.015	.009
Constraint II	35	115.912	<.000	.993	.992	.020	.015
Constraint III	50	227.781	<.000	.984	.987	.024	.012
Two Factor Model							
Constraint I	8	20.305	<.000	.999	.995	.016	.007
Constraint II	23	104.875	<.000	.994	.992	.020	.014
Constraint III	38	217.770	<.000	.984	.987	.025	.010
Male							
One Factor Model							
Constraint I	20	37.272	<.011	.998	.996	.014	.008
Constraint II	35	115.483	<.000	.991	.990	.022	.017
Constraint III	50	202.085	<.000	.984	.987	.025	.013
Two Factor Model							
Constraint I	8	9.948	.269	.999	.999	.007	.004
Constraint II	23	63.893	<.000	.996	.992	.019	.010
Constraint III	38	130.948	<.000	.941	.990	.023	.007

* The levels of model constraints that were restricted to be equal across race 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.

Summary and Discussion

The present study examined the comparability of RENSE scores across gender and race for base calibration and validation samples that were randomly drawn from same population. Results show that factor structure validities of the RENSE are well supported for both one-factor and two-factor models, but the one-factor model of essential real estate sales ability (Model A) was preferred because it better describes the underlying theory of salespersons' professional competence. For White examinees, 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 CFA 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) criteria. Thus the overall pattern of fit statistics for the RENSE data indicates a reasonable fit even when the chi-square test suggests rejection of both the one-factor and two-factor models when sample sizes are large. Exceptions occur for the Black female with small base calibration sample size under the one-factor model, where the χ^2 and RMSEA are significant. These exceptions are not enough to void the conclusion of a reasonable fit in light of the overall pattern of evidence; however, they may suggest continued monitoring of the fit for this group. The evidence of fit holds for both the base calibration and validation samples for all race and gender groups. Further evidence of the invariance of factor structure of the RENSE scores across gender and race groups is found in all fit statistics when model structure, factor loading, latent variable variance, and unique variance are constrained to be equal across groups. Thus the data suggest not only that the RENSE measures a single construct, but also that this construct is similarly structured (fair) across gender and racial groups.

In summary, this study underscores the importance of empirical validation of licensure 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 gender or ethnicity. 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.

References

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.

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.

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

Downing, S. M., & Haladyna, T. M. (1996). A mode for evaluating high-stakes testing programs: Why the fox should not guard the chicken coop. *Educational Measurement: Issues & Practices, 15(1)*, 5-12.

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.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6* 427-439.

Harvey, (1991). Job analysis. In M. Dunnette and L. Hough (Eds.), *Handbook of industrial and organizational psychology (2nd Ed).* Palo Alto: Consulting Psychologists Press.

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.*

Jonson, J. L., & Plake, B. S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement, 58,* 736-753.

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

Joreskog, K. C. & Sorbom, D. (1989). LISREL 7: *A guide to the program and applications.* Chicago: SPSS Inc.

Kane, M. T. (1982). The validity of licensing examinations. *American Psychologist, 37*, 911-918.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112(3),* 527-535.

Kane, M. T. (1994). Validating interpretive arguments for licensing and certification examinations. *Evaluation & the Health Professions, 17(2)*, 133-159.

Kane, M. T. (1997). Model-based practices analysis and specifications. *Applied Measurement in Education, 10*, 5-18.

Knapp, J., & Knapp, L. (1995). Practice analysis: Building the foundation for validity. In J. C. Impara (Ed.), *Licensure testing: Purpose, procedures, and practices* (pp. 93 - 116). Lincoln, NE: Buros Institute of Mental Measurements.

LaDuca, A. (1994). Validation of professional licensure examinations. *Evaluation & the Health Professions, 17*, 178-198.

Messick, S. (1980). Test validity and ethics of assessment. *American Psychologist, 35*, 1012-1027.

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

Messick, S. (1995). Standards-based score interpretation: Establishing valid grounds for valid inferences. In *Proceedings of the joint conference on standard setting for large-scale assessments of the National Assessment Governing Board (NAGB) and the National Center for Education Statistics (NCES)* (Vol. II, pp. 291-305). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.

Mulaik, S.A., James, L. R., Alstine, J. V., Bennett, N., Lind, S., & Stilwell, D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105*, 430-445.

Nelson, D. S. (1994). Job analysis for licensure and certification exams: Science or politics? *Educational Measurement: Issues and Practices, 13(3)*, 29 -15.

Newman, L., & Joseph, M. J. (1998). *A national analysis of the occupational tasks and activities of real estate professionals* (ASI research report). Bala Cynwyd, PA: Assessment Systems, Inc.

Newman, L., Slaughter, R. C., & Taranath, S. N. (1999). *The selection and use of rating scales in task surveys: A review of current job analysis practice*. Paper presented in the symposium "Establishing Validity of Licensure and Certification Examinations: Job Analysis Methodologies and Developing Test Specifications" at the 1999 annual meeting of the National Council of Measurement in Education, Montreal, Canada.

Raymond, M. R. (1995, April). *Job analysis and the development of test specifications for licensure and certification examinations*. Paper presented at the annual meeting of the National Council for Measurement in Education, San Francisco, CA.

Raymond, M. R. (2001). Job analysis and the specification of content for licensure and certification examinations. *Applied Measurement in Education, 14(4),* 369-415.

Stevens, J. (2001, April). *Confirmatory factor analysis of the TerraNova-CTBS/5.* Paper presented at the annual meeting of the American Educational Research Association, Seattle.

Stocker, H. W., & Impara, J. C. (1995). Basic psychometric issues in licensure testing. In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices.* (pp. 165-186). Nebraska-Lincoln, Buros Institute of Mental Measurements.

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

ERIC Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title:
Validation and Invariance of Factor Structure of a National Licensing Examination Across Gender and Race

Author(s): Shudong Wang, Elizabeth Witt

Corporate Source:
CAT*ASI

Publication Date:
4/2002

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: CAT*ASI Three Bala Plaza West Cynwyd, PA 19004-3481	Telephone: 210-339-5535	FAX: 210-339-5973
	E-Mail Address: Shudong_wang@Harcourt.com	Date: 9/29/03

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

EFF-088 (Rev. 2/2003)