교육과정평가연구
The Journal of Curriculum & Evaluation
2007, Vol. 10, No. 1, pp. 103～119

"Types" of Classification Consistency from the Perspective of Replications

Hyun Sook Yi
(Associate Research Fellow, Korea Institute of Curriculum & Evaluation)

≪SUMMARY≫

Classification consistency is often considered as reliability of a measurement procedure involving classifications of examinees into a set of ordered categories. In order to quantify classification consistency, an investigator should define the universe of intended replications, and determine data collection designs as well as methods for estimating classification consistency that would be the most suitable for the intended replications. This process is important because classification consistency estimated from different conceptualizations about replications may produce different types of classification consistency estimates. This paper differentiates three *types* of classification consistency from the perspective of replications (test-retest classification consistency; alternate-forms classification consistency; and internal classification consistency), and introduces three model-based approaches to estimating classification consistency based on each conceptualization, with focuses being given to the intended universe of replications. An empirical analysis will be provided at the end to illustrate differences in estimation according to different conceptualizations about the universe of replications.

Key Words : Classification Consistency, Reliability, Replications

I. Introduction

Human performance, in nature, entails variability over occasions and any measurement procedures related to human behaviors must also involve some degree of inconsistency from a

variety of factors. For example, an examinee who has earned a particular mark on a math computation exam today could have earned a different mark on another day, and still a different mark if s/he had taken a different math computation exam. The inconsistencies may be due to different levels of physical or mental conditions, different levels of content familiarity, etc. For this reason, a particular score obtained on a certain measurement procedure should be interpreted merely as an instance of many possible values that the person could have obtained under various measurement circumstances, not as the constant value innate to the person. Reliability is defined as a psychometric property quantifying these consistencies and/or inconsistencies in examinee scores over replications of a measurement procedure (Feldt & Brennan, 1989; Brennan, 2001). In other words, it is the degree to which a person would obtain consistent scores over possible hypothetical replications.

Classification consistency is often considered as reliability of a measurement procedure involving classifications of examinees into a set of ordered categories such as pass/fail or proficient/average/below-average. Several authors have defined classification consistency with slightly different conceptualizations (Carver, 1970; Swaminathan, Hambleton, & Algina; 1974; Hanson & Brennan, 1990; Livingston & Lewis, 1995; Lee, Hanson, & Brennan, 2002). Although it takes many different forms, a common thread running through the various definitions is that classification consistency is the degree of consistent decisions on two independent replications of a measurement procedure. Because of the inconsistent nature of human performance, classification decisions based on characteristics of human behavior must entail some degree of inconsistency over replications. For this reason, as in the notion of reliability, classification decision of a person based on a single measurement procedure may be considered as an instance of virtually infinite number of possible replications under similar conditions.

Based on the conceptualization described above, at least two instances of possible replications are essential, in order to quantify the degree of classification consistency. It is an investigator's role to define replications and determine methods for estimating classification consistency based on the intended replications. However, it should be noted that classification consistency estimated from different conceptualizations about replications may produce different types of classification consistency estimates, just as different methods of estimating reliability may yield different types of reliability estimates such as test-retest reliability, alternate-forms reliability, and reliability as internal consistency. This point is well described by Brennan (2001, pp. 301).

> There is no such thing as the reliability; there are as many reliabilities as there are specifications of a universe of generalization that one or more investigators is (are) willing to assert as meaningful for some purpose... Different methods may estimate different types of reliability... (We need to) differentiate between the universe of replications that are of interest to an investigator and the replications (actual or contrived) that directly or indirectly characterize available data.

He added that an investigator must clearly specify the intended replications of the measurement procedure in order to understand reliability and meaningfully interpret any estimate of it, and data collection design must be considered according to the specified replications. The same arguments can be made for classification consistency. For example, suppose that a testing program has been developed to determine qualified and non-qualified physician. In order to provide evidences of reliability of a test form, the test developer needs to replicate the testing procedure in some way. If the test developer's intended universe of replications implies various occasions, s/he may want to collect data on two different occasions keeping other conditions fixed, and estimate the degree of consistent classifications over the two occasions. In this case, the test developer's interest is to measure consistency of the test form over time. As an analogue of typically used types of reliability, this kind of classification consistency can be considered as test-retest classification consistency. If the test developer specifies the universe of replications as a wide variety of test items, s/he may need to devise two alternate forms measuring proficiency as a physician and administer the two test forms to the same examinees to compute the degree of consistent classifications. In this case, other conditions are fixed with only items being varied. This kind of classification consistency can be considered as alternate-forms classification consistency. If the test developer views the classification consistency as an indicator of internal decision consistency, s/he needs to find a way to estimate the degree of consistent classifications over items within the test form. This type of classification consistency may be considered as internal classification consistency.

Although the intended universe of replications can take different types described above, unavailability of data often restricts the use of certain types of classification consistency, in practice. For example, if the test developer's interest is to know the degree to which an examinee will be classified into the same category if s/he had taken a different test form, then alternate-forms classification consistency would be the best measure of the classification consistency of interest. If two alternate test forms are available, the degree of consistent classifications can be obtained directly by computing the percentage of the same decisions over

the two measurements. However, such kinds of data are seldom available in practice.

For this reason, several researchers provided methods for estimating classification consistency based on a single administration (Huynh, 1976; Subkoviak, 1976; Hanson & Brennan, 1990; Breyer & Lewis, 1994; Livingston & Lewis, 1995; Wang, Kolen, & Harris, 2000; Lee, Hanson & Brennan, 2002), but not much of them explicitly specifies the intended universe of replications. Availability of such information would be beneficial to practitioners by helping them with making informed decisions on choosing the appropriate type of classification consistency and an appropriate method for estimating it.

Based on the logic described above, the present paper differentiates three types of classification consistency from the perspective of replications (test-retest classification consistency; alternate-forms classification consistency; and internal classification consistency), and introduces three model-based approaches to estimating classification consistency based on each of the three types, respectively, with explanations about the intended universe of replications. Three different methods selected from the literature include a single administration approach by Lee, Hanson and Brennan (2002), a method based on two equated forms suggested by Yi, Kim, and Brennan (2004), and a method based on split-halves by Breyer and Lewis (1994). An empirical analysis will be provided at the end to illustrate differences in estimation according to different conceptualizations about the universe of replications.

Ⅱ. Methods for Estimating Classification Consistency

1. Lee, Hanson, and Brennan's Method

Lee, Hanson, and Brennan (2002) proposed a method for estimating classification consistency based on a single administration of a test form and used a psychometric model to generate a hypothetical replication. Lee et al. defined the classification consistency as the probability of a randomly selected examinee being classified in the same observed category on two testing occasions. They considered a situation in which a testing procedure is developed to measure a single latent trait θ, and the measurement procedure is one in which examinees are classified into one of H categories based on H-1 cut scores (c_1, c_2, ..., c_{H-1}). They assumed that, conditional on

the latent trait of examinees, raw scores on an actual form (Form X) consisting of K dichotomously scored items and a hypothetical alternate form (Form Y) are independent and identically distributed. In other words, an examinee is assumed to have the same probability of obtaining the same raw score on two alternate forms. Then, the conditional joint distribution of scores on the two forms is given by

$$f(x, y|\theta) = f(x|\theta) \cdot f(y|\theta) = [f(x|\theta)]^2. \quad (1)$$

By the same logic, letting I_h denote the h^{th} category on both forms into which examinees with scores from $c_{(h-1)}$ to c_h-1 are classified (h=1, 2, ..., H) and letting c_0=0 and c_H=K+1, the probability that an examinee of ability θ would be classified into the same category I_h on both forms is

$$\Pr(X \in I_h, Y \in I_h | \Theta = \theta) = \left[\sum_{x=c_{(h-1)}}^{c_h-1} f(x|\theta) \right] \left[\sum_{y=c_{(h-1)}}^{c_h-1} f(y|\theta) \right]$$

$$= \left[\sum_{x=c_{(h-1)}}^{c_h-1} f(x|\theta) \right]^2, \qquad h=1,2,...,H, \quad (2)$$

and the conditional probability of consistent decisions over all categories is given by

$$P(\theta) = \sum_{h=1}^{H} \Pr(X \in I_h, Y \in I_h | \Theta = \theta). \quad (3)$$

The marginal probability of consistent decisions, P, can be obtained by integrating the above equation over the distribution of θ, $g(\theta)$.

$$P = \int_{\Omega} P(\theta) g(\theta) d\theta, \quad (4)$$

where Ω denotes a parameter space of θ.

In order to estimate the marginal probability of consistent classification, P, Lee et al. presented two different classes of psychometric models: the strong true score models (two-parameter beta-binomial model and the four-parameter beta-binomial model) and the item response theory model (the three-parameter logistic model). The procedures presented by the authors are not limited to the two classes of models, but these models were chosen because they were considered to fit empirical score distributions reasonably well. Under the strong true score model, the binomial density was used to estimate the conditional distribution, $f(x|\theta)$, in Equation 2, and the two-parameter beta distribution or four-parameter beta distribution was used to estimate the true score distribution, $g(\theta)$. The mathematical form of the two-parameter beta distribution is given by

$$g(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}, \quad (5)$$

where

$$\alpha = (-1 + 1/K R_{21})\mu_x, \tag{6}$$

$$\beta = -\alpha - K + K/K R_{21}, \tag{7}$$

and

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \tag{8}$$

where Γ is the gamma function (Lord & Novick, 1968). A procedure for estimating the four-parameter beta distribution is described in Hanson (1991) in detail.

Under the item response theory model, the three-parameter logistic model was used to estimate the probability of getting each item correctly and the recursive algorithm of Lord & Wingersky (1984) was used to obtain the conditional distribution $f(x|\theta)$. Quadrature distributions from the output of BILOG-3 (Mislevy & Bock, 1990) were used to estimate $g(\theta)$. Once $f(x|\theta)$ and $g(\theta)$ were estimated, the marginal probability of consistent classification, P, was computed using Equation 4.

As indicated in the definition by Lee et al., the intended universe of replications includes various testing occasions. The definition is silent about the variability of test forms over occasions but, as implied in the presentation of deriving P (Equations 1 through 4), the universe of replications seems to include various test forms as well, by considering the existence of alternate forms, Form X and Form Y. This notion of replications is reflected in the estimation procedure by making the assumption of the identical conditional distributions over the two independent administrations of hypothetical alternate forms. By using one of the psychometric models described above, a fitted score distribution is obtained from an actual score distribution based on a single administration of a single test form, and the fitted score distribution is used to represent identical distributions of two randomly chosen alternate forms. In this sense, the classification consistency can be considered as alternate-forms classification consistency. However, a noteworthy point here is that the degree to which estimates based on the procedure of Lee et al. reflect the alternate-forms classification consistency depends on the feasibility of the above assumption in real testing contexts. Under strong true score models, replications are based on a single instance of a randomly constructed form drawn from an infinite population of items, while replications are based on a single fixed form under IRT models (Lee, Brennan, & Kolen, 2000). Either way, effects of actual form differences due to content variations and conditions of examinees over occasions are not explicitly reflected over two replications. In other words, by using data from a

single administration of a single form, the data collection design effectively fixes the condition of testing occasions as well as test forms. In this sense, the classification consistency estimated by the procedure of Lee, Hanson, and Brennan (2002) may be considered as test-retest classification consistency without interval.

2. Yi, Kim, and Brennan's Method

It is typical of large-scale testing programs involving classification decisions that multiple forms are developed periodically and equating is conducted to establish score interchangeability over forms. Yi, Kim, and Brennan (2007) proposed a method for estimating classification consistency under this situation, reflecting actual form differences and results of equating. Yi et al. presented the estimation procedure under three commonly used equating designs: the single group design, the random groups design, and the common-item nonequivalent groups design. Under the single group design, classification consistency can be computed directly from repeated measures by counting frequencies of the same classifications. Under the other two equating designs, since two groups involved in equating studies take different alternate forms, a psychometric model was used to estimate score distributions of the form not taken by each group. The following presentation applies to the random groups design and the common-item nonequivalent groups design.

They considered a situation in which two alternate forms were developed to measure a single latent trait, θ, and a new form (Form X) was equated to an old form (Form Y) on which cut scores were determined through standard setting procedures to find cut score equivalents. It was assumed that Form X was administered to Group 1 examinees and Form Y was administered to Group 2 examinees. Under the random groups design, the two groups were assumed to be drawn from a single population, while they were assumed to be drawn from nonequivalent populations under the common-item nonequivalent groups design. They used the same notational scheme that was used in Lee, Hanson, and Brennan (2002), except for using prime ($'$) to represent quantities for the new form. Under the conceptualization of Yi et al., the probability that an examinee having ability θ will be classified into the h^{th} category on Form X and Form Y is

$$\Pr(X \in I_h', Y \in I_h | \theta) = \sum_{x=c'_{h-1}}^{c'_h-1} f(x|\theta) \sum_{y=c_{h-1}}^{c_h-1} f(y|\theta) \quad (9)$$

$$= \sum_{x=c'_{h-1}}^{c'_h-1} f_1(x|\theta) \sum_{y=c_{h-1}}^{c_h-1} f_2(y|\theta), h=1,2,...,H, \quad (10)$$

where I_h' denotes the h^{th} category on Form X into which examinees with equated cut scores from $c'_{(h-1)}$ to c_h'-1 are classified, and subscripts 1 and 2 denote group membership.

The major difference between this procedure and the procedure of Lee et al. is that examinee scores obtained from two actual alternate forms are used to estimate each conditional distribution in Equation 9 here, while the assumption of independent and identical distributions was made to estimate the conditional distributions, $f(x|\theta)$ and $f(y|\theta)$, under the Lee et al.'s procedure. Under the procedure of Yi et al., since the scores on two alternate forms are not for the same group of examinees, assumptions were also made to use score distributions of each group of examinees to estimate conditional distributions not taken by the respective group. Under the random groups design, given that the two groups were randomly sampled from a single population, if two examinees from different samples are at the same ability level, θ, they will have the same probability of being classified into the h^{th} category if both had taken the same form. Under the common-item nonequivalent groups design, since the two groups are nonequivalent with respect to the ability of interest, a synthetic population (Braun and Holland, 1982) was conceptualized so that classification consistency indices should be estimated and interpreted for a single population combined from two different populations involved in equating. Then, it was assumed that score distributions conditional on θ are the same for both groups, although two populations are nonequivalent with respect to the ability being measured (Kolen & Brennan, 2004). This assumption was used to estimate the conditional distribution of Form X and Form Y taken by the synthetic population. Mathematical proofs for these assumptions are given in Yi, Kim, and Brennan (2004). In order to estimate conditional distributions, they also used the three psychometric models used by Lee et al. to illustrate the use of the procedure, although the procedure is not model-specific. Once the conditional distributions are estimated, the probability of consistent classifications can be computed by Equations 3 and 4.

Yi et al. conceptualized classification consistency as the agreement of decisions on two alternate forms independently administered to the same group of examinees. Although the two forms are not administered to the same group of examinees, the equating context made it possible to make inferences for the same group of examinees. As implied in the definition, the intended universe of replications includes various forms. Since two alternate forms could be equated at the same time or on two independent occasions, the condition of occasions could be fixed or varied under their conceptualization. The estimation procedure reflects this conceptualization. By using the data from two independent administrations of two actual alternate forms, the data collection design effectively allows the condition of test forms to vary. In this sense, the conceptualized

classification consistency and estimates of it can be considered as alternate-forms classification consistency with or without interval.

3. Breyer and Lewis' Method

Breyer and Lewis (1994) proposed a method for estimating classification consistency based on a single test form. They defined single-administration approaches to estimating classification consistency as the estimates of the probability of consistently classifying examinees to mastery or nonmastery as if those examinees had been tested with two alternate forms. Their procedure is an anologue of split-halves reliability index. They considered a situation where a full test can be split into two comparable half-tests, each with its own cut score, so that a 2×2 contingency table can be constructed. Then, instead of counting frequencies of consistent decisions, they made an assumption of bivariate normal distributions of the two half-tests to compute a tetrachoric correlation as a measure of reliability of the half-test. The probability of failure on a half-test used to compute the tetrachoric correlation was estimated from the average proportion of failure for the two half-tests as

$$P_{f,\,half} = \frac{2X_{11} + X_{12} + X_{21}}{2N}, \tag{11}$$

where X_{11} is the frequency of examinees failing on both half-tests, X_{12} is the frequency of examinees failing half-test 1 but passing half-test 2, and X_{21} is the frequency of examinees failing half-test 2 but passing half-test 1. N is the total number of examinees. The proportion of examinees who fail both half-tests was computed by

$$P_{ff,\,half} = \frac{X_{11}}{N}. \tag{12}$$

These probabilities were then used to find normal deviates from the standard normal distribution, and tables summarized by Huynh (1976) were used to find the tetrachoric correlation associated with the normal deviates and probabilities of failing both tests. Once the tetrachoric correlation was computed from the above probabilities, Spearman-Brown prophesy formula was applied to obtain the measure of reliability of the full-test. Finally, the proportion of consistent classifications for the full-test was calculated using the following formula,

$$P = 1 - 2[\,P(z < c) - P(z < c, z' < c)], \tag{13}$$

where $P(z<c)$ is the proportion of observed scores below the cut score (c), and $P(z<c,\ z'<c)$ is

the probability from the bivariate normal distribution of the two hypothetical full-tests.

As indicated in the definition and implied in the presentation of the procedure, the method by Breyer and Lewis intends to measure consistency of classifications over items. However, since the estimation procedure uses data from half-tests of a single form administered on a single occasion, the data collection design effectively fixes the condition of occasion and allows the items to vary within a single form. In this sense, classification consistency estimated from their method can be viewed as an estimate of internal classification consistency.

Ⅲ. An Illustrative Example

Since each estimation procedure introduced in this paper measures different types of classification consistency, it is not meaningful to directly compare the magnitude among estimates. However, it may be worthwhile to show how measures of classification consistency would differ depending on estimation procedures. For the illustrative purpose, therefore, three approaches introduced in the previous section were applied to a real data set drawn from a large-scale testing program. The data set consists of two alternate forms consisting of 24 dichotomous items administered each to two randomly equivalent groups, respectively. Form *X* (new form) was administered to 2421 examinees (Group 1) and Form *Y* (old form) was administered to 2527 examinees (Group 2), and equating was conducted to find score equivalents of the two forms. Cut scores on Form *Y* were first determined at the 41^{st} percentile (C1) and the 90^{th} percentile (C2) of the observed score distribution of Form *Y*, and cut scores on Form *X* were found through equating relationships. The choice of cut scores was based on actual cut scores defined by nationally normed scores on a standardized achievement test to classify examinees into three performance levels. Table 1 summarizes descriptive statistics and cut scores of Form *X* and Form *Y*.

〈Table 1〉 Descriptive Statistics and Cut Scores on Form *X* and Form *Y*

	Sample Size	Mean	SD	C1(41^{st})	C2(90^{th})
Form *X*	2421	10.80	4.32	10	16
Form *Y*	2579	14.01	5.01	13	20

Classification consistency index *(P)* defined by each estimation procedure introduced in the previous section was then estimated based on a selected psychometric model. For procedures of Lee, Hanson, and Brennan (2002) and Yi, Kim, and Brennan (2004), the three-parameter logistic (3PL) model and the four-parameter beta-binomial (4PB) model were used to fit score distributions. For the 3PL model, item parameters and ability parameters were concurrently calibrated for two groups using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). Default options were used for BILOG-MG, except for using 31 quadrature points and posterior weights for estimating ability distributions. Using the estimated item parameters, the probability that an examinee with ability θ was computed for each item and conditional distributions were then estimated from a recursive algorithm by Lord and Wingerky (1984). Ability distributions of each group of examinees were estimated from the quadrature distributions from the output of BILOG-MG. Under the 4PB model, true score distributions and conditional score distributions were estimated using BB-CLASS (Brennan, 2004) with the EDquad option for numerical integration, where 1000 equally spaced quadrature points of the true score range (between 0 and 1) and the density in the interval for each point were used. Once conditional distributions and true score distributions were found, the marginal classification consistency *(P)* was computed by Equation 4.

For the procedure of Breyer and Lewis (1994), test form was split into even- and odd- halves and *P* was estimated based on the bivariate normal assumption of the two halves. First, observed frequencies were counted from univariate and bivariate distributions constructed from the two halves. Instead of computing the proportion of agreement in decisions over the two halves directly, a tetrachoric correlation was calculated under the assumption of bivariate normality, and Equation 13 was used to estimate the model-based probability of consistent classifications of the full-test. To estimate the probabilities in Equation 13, the tables summarized in Huynh (1976) were used.

Since the procedures of Lee, Hanson, and Brennan (LHB). and Breyer and Lewis (BL) are based on a single administration of a single form, classification consistency *(P)* defined by each estimation procedure was calculated for Form *X* and Form *Y*, respectively. For the procedure of Yi, Kim, and Brennan (YKB), *P* was estimated using data from Form *X* and Form *Y* at the same time. Classification consistency was estimated for the two predetermined cut scores (C1=41st percentile, C2=90th percentile), respectively. Table 2 shows estimates of *P* obtained under each estimation procedure.

〈Table 2〉 Classification Consistency Estimated by Three Different Approaches

Estimation Method	Model	C1		C2	
		Form *X*	Form *Y*	Form *X*	Form *Y*
LHB	3PL	.80	.86	.87	.88
	4PB	.77	.83	.84	.86
YKB	3PL	.82		.87	
	4PB	.75		.78	
BL	even-odd split	.79	.86	.86	.84

As shown in the table, estimates were fairly different from estimation method to method, from model to model. For all cases, estimates based on the 3PL model produced higher estimates than those based on the 4PB model. This observation is consistent with findings from Lee, Hanson, and Brennan (2002). Lee et al. explained this pattern in association with different model assumptions. That is, the beta-binomial model assumes randomly parallel forms of a test allowing an additional source of errors due to form variation, as opposed to the IRT assumption of strictly parallel forms, which involves a conceptual replication of a test with a set of items having identical item parameters. For most cases, estimates based on Form *Y* were larger than those based on Form *X*, simply reflecting the idiosyncrasy of the particular form chosen. Estimates based on the cut score of C2 were larger than those based on C1. Having larger estimates on the 90^{th} percentile makes an intuitive sense because fewer examinees are located around the 90^{th} percentile than 41^{st} percentile and thus, probability of misclassifications will be lower around the 90^{th} percentile.

Estimates were different depending on the estimation procedure as well as the psychometric model being applied. In general, the YKB method yielded lower estimates than the LHB method. Estimates based on the YKB method were lower than the average of the two estimates based on the LHB over Form *X* and Form *Y*, and even lower than the lowest of the two, for most cases. This pattern was consistent with the author's anticipation that, with other conditions being equal, test-retest classification consistency would yield higher estimates than alternate-forms classification consistency, because the former does not take into account actual form differences.

Ⅳ. Discussion

Classification consistency has been considered as an index quantifying the degree to which an examinee would be classified into the same category if the measurement procedure had been replicated. Depending on measurement conditions constituting the universe of intended replications determined by an investigator or a test developer, classification consistency takes different types: test-retest classification consistency, alternate-forms classification consistency, and internal classification consistency. Three model-based procedures for estimating classification consistency introduced in the literature have been considered in this paper as examples of each of the three types of classification consistency, and an illustrative example was provided to show how different approaches would yield differences in estimates.

Although the patterns found from the particular example are not generalizable and it does not make sense to directly compare the magnitude among estimation procedures, it is meaningful to see from the example that different estimation procedures produce estimates with fairly large differences. This implies that choosing an appropriate estimation procedure suited for the investigator's interest is crucial in analyses of classification consistency. However, practical factors often restricts one from choosing an ideal estimation procedure. For example, in reliability analyses, since two repeated measures are not usually available, coefficient α is often computed as an estimate of reliability, regardless of the intended universe of replications. However, as Brennan (2001) pointed out, test developers, test users, or whoever interprets the reliability index should be aware that α based on a single occasion overestimates the reliability being conceptualized, if the intended universe of replications includes various occasions. Similarly, in analyses of classification consistency, if an investigator wants to know the degree to which an examinee would be classified into the same category regardless of form taken, then using a single administration approach such as the procedure of Lee, Hanson, and Brennan (2002) would provide overestimates to the classification consistency being conceptualized. By the same logic, if an investigator is only interested in the degree to which an examinee would be classified into the same category on the same test over time, then alternate-forms classification consistency such as the one presented by Yi, Kim, and Brennan (2004) would provide lower estimates to the classification consistency being conceptualized. In sum, practitioners should choose the best estimation procedure to the type of classification consistency under consideration, but if practical factors hinder one from using the best procedure, then cautious interpretations should be made for the estimates obtained from the

selected estimation procedure.

There are several suggestions for future studies. First of all, the estimation procedures introduced in this paper are not an exhaustive list of model-based estimation procedures for classification consistency. It would provide a valuable information to practitioners if more extensive list of model-based approaches could be analyzed in terms of the universe of replications and data collection design for estimating the index. Second, the findings from the illustrative example do not provide a general description about comparisons among psychometric models or estimation procedures. A simulation study of systematic comparisons among methods would be beneficial to practitioners by having information about which method provides lower or upper limits to classification consistency being conceptualized. Lastly, as in the reliability analyses, interpretations of indices of classification consistency often accompany subjective judgments. Since the three methods introduced in this paper are relatively new and have not been used extensively in practice, guidelines for determining high or low values of classification consistency have not been established. Although interpretations of classification consistency should be made within contexts of a particular testing program and subjectivity is inevitable, it would be valuable if studies providing standards for estimates based on each estimation procedure are available.

References

Braun, H. I. & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland and D. B. Rubin (Eds.), *Test equating,* 9-49. New York: Academic Press.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*, 295-317.

Brennan, R. L. (2004). Manual for BB-CLASS: *A computer program that uses the beta-binomial model for classification consistency and accuracy*. CASMA Research Report No. 9. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Breyer, F. J. & Lewis, C. (1994). *Pass-Fail Reliability for Tests with Cut Scores: A Simplified Method.* ETS Research Report RR-94-39. Princeton, NJ: Educational Testing Service.

Carver, R. P. (1970). Special problems in measuring change with psychometric devices. *In Evaluative research: Strategies and methods*. Washington: American Institutes for Research.

Feldt, L. S. & Brennan, R. L. (1989). Reliability. In R. L. Linn (Eds.), *Educational measurement* (3rd ed.) (pp. 105-146). New York: American Council on Education and McMillan.

Hanson, B. A. (1991). *Method of moments estimates for the four-parameter beta compound binomial model and the calculation of classification consistency indexes.* ACT Research Report 91-5. Iowa City, IA: ACT, Inc.

Hanson, B. A. & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement*, *27*(4), 345-359.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13*, 253-264.

Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Lee, W. C., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indexes for multiple classifications. *Applied Psychological Measurement, 26*(4), 412-432.

Livingston, S. A. & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 9*, 13-26.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. New York: Addison-Wesley Publishing Company.

Lord, F. M. & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score 'equatings'. *Applied Psychological Measurement, 8*, 453-461.

Mislevy, R. J. & Bock, R. D. (1990). BILOG-3: *Item and test scoring with binary logistic models* (2nd ed.). Mooresville, IN: Scientific Software.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-references test. *Journal of Educational Measurement, 13*, 265-276.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1975). A reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement, 11*, 263-267.

Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessments using polytomous IRT. *Journal of Educational Measurement, 37*, 141-162.

Yi, H. S., Kim, S., & Brennan, R. L. (2004, April). *Methods for estimating classification consistency indices for two equated forms*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Yi, H. S., Kim, S., & Brennan, R. L. (2007). A Method for estimating classification consistency indices for two equated forms. *Applied Psychological Measurement, 31*(4), 275-291.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multi-group IRT analysis and test maintenance for binary items*. Chicago: IL: Scientific Software International.

• *논문 접수 : 2007년 3월 15일 / 수정본 접수 : 2007년 5월 15일 / 게재 승인 : 2007년 5월 25일*

초 록

반복측정의 관점에서 본 분류일치도 계수의 유형

이 현 숙
(한국교육과정평가원 부연구위원)

분류일치도는 피험자의 특성을 위계적 범주로 구분하고자 하는 검사의 신뢰도로서 종종 간주된다. 분류일치도를 측정하기 위해서는 적어도 두 개 이상의 반복측정 자료가 필요한데, 연구의 목적에 부합하는 반복측정 자료를 얻기 위해서 검사개발자는 검사과정의 반복을 통해 변화시키고자 하는 요인과 변화시키지 않고자 하는 요인을 정의해야 하며, 이에 따라 자료 수집의 설계와 측정방법이 결정되어야 한다. 검사개발자가 이러한 요인들을 어떻게 개념화하는가에 따라 분류일치도의 유형이 달라진다. 본 연구에서는 분류일치도의 유형을 검사-재검사 분류일치도, 동형검사 분류일치도, 검사내적 분류일치도 등의 세 가지로 구분하고, 문헌에서 제시된 분류일치도 측정방법 중 세 가지를 선택하여 반복측정의 관점에서 재조명하였다. 또한 분류일치도를 어떻게 개념화하는가에 따라 분류일치도 계수의 측정치가 얼마나 달라지는지를 보기 위해 본 연구에서 소개된 세 가지 방법을 실제 자료에 적용한 결과를 예시하였다.

주제어 : 분류일치도, 신뢰도, 반복측정