PSYCHOMETRIKA—VOL. 62, NO. 1, 7–28
MARCH 1997

JOINT CONSISTENCY OF NONPARAMETRIC ITEM CHARACTERISTIC CURVE AND ABILITY ESTIMATION

JEFF DOUGLAS

DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF WISCONSIN

The simultaneous and nonparametric estimation of latent abilities and item characteristic curves is considered. The asymptotic properties of ordinal ability estimation and kernel smoothed nonparametric item characteristic curve estimation are investigated under very general assumptions on the underlying item response theory model as both the test length and the sample size increase. A large deviation probability inequality is stated for ordinal ability estimation. The mean squared error of kernel smoothed item characteristic curve estimates is studied and a strong consistency result is obtained showing that the worst case error in the item characteristic curve estimates over all items and ability levels converges to zero with probability equal to one.

Key words: item characteristic curve, kernel smoothing, large sample theory, nonparametric regression.

Introduction

Nonparametric regression has been a topic of much research in statistics over the past thirty years. This is due in part to the continuing increase in computing capability as well as the recognition that parametric modeling often lacks sufficient flexibility. More recently psychometricians have recognized that item characteristic curves (ICCs) cannot always be modeled well with parametric families such as the three-parameter logistic and the normal ogive models (Ramsay, 1991; Samejima, 1979, 1981, 1984, 1988, 1990). This paper concerns the asymptotic theory of a particular method of nonparametric ICC estimation based on a technique known in the statistical literature as *kernel smoothing*, together with the theory of obtaining the proper linear ordering of examinees with respect to their true latent abilities.

Kernel smoothing is introduced as a technique for nonparametrically estimating a regression function, followed by its formulation in item response theory (IRT). Let $(X_1, Y_1), \ldots, (X_J, Y_J)$ be an independent and identically distributed sequence of bivariate observations. Nadaraya (1964) and Watson (1964) proposed to estimate the regression function $E[Y|X = x] = m(x)$ by

$$\hat{m}(x) = \frac{\sum_{i=1}^{J} K\left(\frac{x - X_i}{h}\right) Y_i}{\sum_{i=1}^{J} K\left(\frac{x - X_i}{h}\right)}. \tag{1}$$

Here K is a nonnegative symmetric function that is nonincreasing as its argument gets further from zero, and h is a parameter called the bandwidth that is selected by the user to control the amount of smoothing. Thus, $\hat{m}(x)$ is just a smoothly weighted average, with

The author thanks James O. Ramsay and William F. Stout for helpful discussions of this material. Please send requests for reprints to Jeff Douglas, Department of Biostatistics, K6/438 Clinical Science Center, 600 Highland Avenue, Madison, WI 53792-4675.

0033-3123/97/0300-95259$00.75/0

the weights determined by the kernel function K and the bandwidth h. The bandwidth is chosen to obtain a desirable trade-off between the bias and the variance of estimation. For instance, if a small bandwidth is chosen only observations Y_i with X_i very close to x will influence the estimate $\hat{m}(x)$, resulting in a nearly unbiased but highly variable estimate. On the other hand, choosing a large bandwidth will allow Y_i with X_i far from x to contribute to the estimate of $m(x)$, resulting in low variance but potentially high bias stemming from the difference in $m(X_i)$ and $m(x)$ for X_i far from x. Ordinarily this trade-off is stated in terms of mean squared error of estimation, and the bandwidth is chosen in a way that comes close to minimizing this quantity.

In ICC estimation the dependent variable Y that we consider is a binary variable denoting success or failure to answer a given item correctly, and the independent variable is the latent ability variable Θ. The regression function previously denoted $m(x)$, can now be expressed as the item characteristic curve $P(\theta)$. Suppose that J examinees are randomly sampled. If each of the J latent ability variables could be measured without error, the Nadaraya-Watson estimator given in (1) could be written for $P(\theta)$ by

$$\hat{P}^*(\theta) = \frac{\sum_{k=1}^{J} K\left(\frac{\theta - \Theta_k}{h}\right) Y_k}{\sum_{K=1}^{J} K\left(\frac{\theta - \Theta_k}{h}\right)}. \tag{2}$$

However, the latent variables Θ_k cannot be directly observed. Kernel smoothing can still be used, but each Θ_k in (2) must be replaced with a reasonable estimator $\hat{\Theta}_k$, resulting in an estimate $\hat{P}(\theta)$ shown below:

$$\hat{P}(\theta) = \frac{\sum_{k=1}^{J} K\left(\frac{\theta - \hat{\Theta}_k}{h}\right) Y_k}{\sum_{k=1}^{J} K\left(\frac{\theta - \hat{\Theta}_k}{h}\right)}. \tag{3}$$

The asymptotic theory behind the Nadaraya-Watson estimator given in (1) has been thoroughly researched and some very strong and precise results on convergence and asymptotic normality have been obtained (see for example, Devroye, 1978; Schuster, 1972). The central problem in applying these results to ICC estimation is that in IRT the values of the independent variable cannot be measured without error. If each Θ_k in (2) could be observed, the asymptotic theory for kernel smoothing with random arguments would apply. However, the situation considered in this paper, requiring an estimator of the form given by (3), more nearly resembles an errors-in-variables regression problem. Fan and Truong (1993), considered the asymptotic theory of kernel smoothing when the independent variables X_k cannot be observed, but rather we observe $Z_k = X_k + \varepsilon_k$, where ε_k is thought of as an error of measurement. Under this model the estimate $\hat{m}(x)$ is obtained by replacing X_k in (1) with Z_k. This is similar to replacing Θ_k in (2) with $\hat{\Theta}_k$ to obtain the estimator in (3). Although the model that Fan and Truong consider is similar in many ways to IRT models, the assumptions that they make regarding the distribution of the error term ε, and the independence of ε with (X_k, Y_k), render their results inapplicable to the asymptotic theory of kernel smoothed ICC estimation. Thus, existing statistical theory will not suffice and the asymptotic theory of kernel smoothed ICC estimation requires its own special investigation.

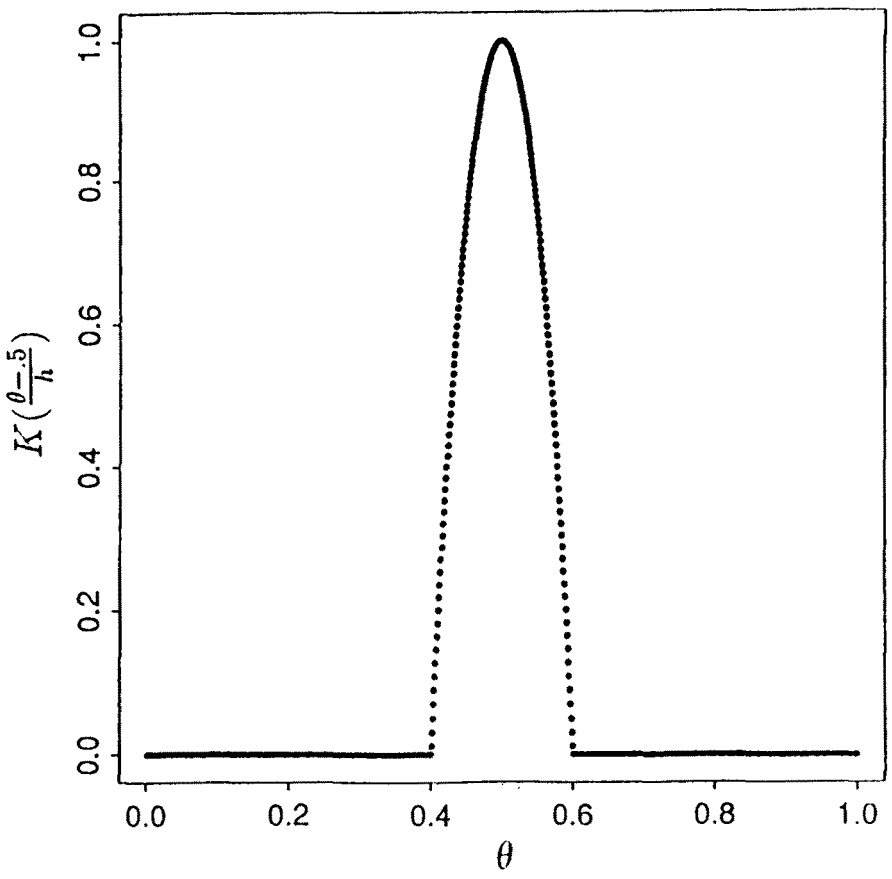

FIGURE 1.
The intended weights for ICC estimation at $\theta = 0.5$. The weights $K(\theta - 0.5/h)$, are plotted against θ. This assignment of weights requires that the true value of the latent ability of each examinee is known.

An example is given to further elucidate the difficulty involved when dealing with unobservable independent variables. Recall that $\hat{P}(\theta)$ in (3) is just a weighted average of the dichotomous item scores $Y_1, \ldots, Y_J$. Leaving the details of how to form the ability estimates $\hat{\Theta}_k$ in (3) to a later section, a 100 item exam with 1000 examinees was simulated to illustrate how the intended weights differ from the actual weights assigned to the Y_k for estimation at a particular θ value. In this example, the latent ability random variable Θ is assumed to have a uniform distribution on the open interval (0, 1), and a quadratic kernel function is used ($K(x) = 1 - x^2$, $-1 \leq x \leq 1$; $K(x) = 0$ otherwise), together with a bandwidth of $h = 0.1$. Both the intended and actual weights, $K(0.5 - \Theta_k/h)$ and $K(0.5 - \hat{\Theta}_k/h)$ respectively, were obtained to estimate $P(\theta)$ at $\theta = 0.5$.

In Figure 1 the intended weights for each Y_k are plotted against θ. In practice, as revealed by Figure 2, the weight assignment becomes somewhat imprecise but still tends to give high weight to Y_k for examinees with Θ_k near 0.5.

The software TESTGRAF (Ramsay, 1994) has recently become popular, and the asymptotic theory developed in this paper can be viewed as asymptotic theory for ICC estimates produced by this software. It should be pointed out that the method of estimation considered here slightly differs in some details from the method employed by TESTGRAF. However, the general approach is the same and the results presented here are suggestive of the asymptotic properties of ICC estimates produced by TESTGRAF.

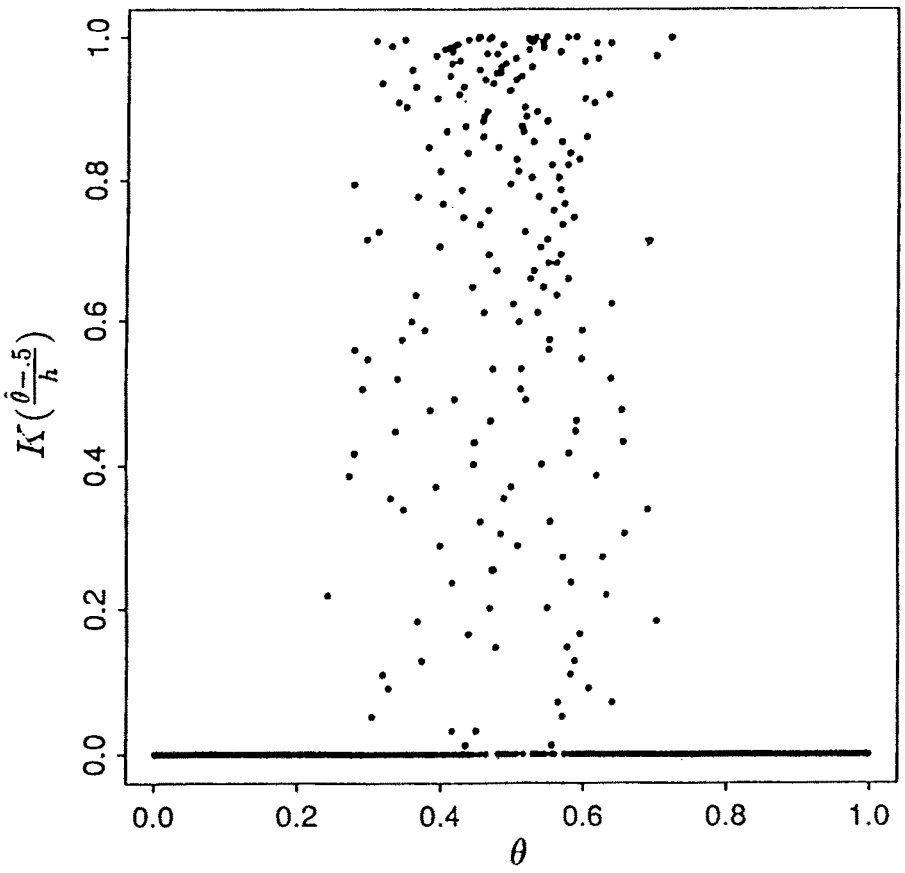

FIGURE 2.
The actual weights for ICC estimation at $\theta = 0.5$. The weights $K(\hat{\theta} - 0.5/h)$, are plotted against θ. This assignment of weights results from replacing the true values of θ with their estimated values.

In the next section a triangular sequence of nonparametrically specified IRT models is introduced, and a comprehensive list of notation is supplied for the convenience of the reader. The procedures for ICC and ability estimation are given in the third section. The fourth section includes statements of the main theoretical results as well as the assumptions and regularity conditions needed to obtain them. In the fifth section the practical implications of the theoretical results are discussed. An appendix is provided to outline the proofs of the theorems stated in the fourth section.

Triangular Sequence of IRT Models and Notation

Ordinarily, asymptotic theory only concerns the behavior of statistics as the sample size approaches infinity. However, in order to obtain jointly consistent estimates of ability parameters and ICCs, it is required that the test length and examinee sample size increase together. This "double asymptotics" in both sample size and test length was utilized by Haberman (1977) to show the joint consistency of maximum likelihood estimates of Rasch model item and examinee parameters. This paper considers a problem similar to Haberman's, but with less restrictive assumptions regarding the functional form of the ICCs.

A triangular sequence of IRT models is needed to conduct an asymptotic analysis in which both sample size and test length increase simultaneously. First consider an n item

test administered to J_n randomly chosen examinees. Using the methods described in the next section, the ICCs corresponding to the n items and the abilities of the J_n examinees are estimated. Then a test of $n + 1$ items, possibly containing no items in common with the length n exam, is considered along with a corresponding larger sample of J_{n+1} examinees. The idea is to carry out the estimation procedures for each row in the triangular arrays given below, and the asymptotic theory will concern the accuracy of the ICC and ability estimates as we progress down the rows of the triangular arrays of items and examinees.

Triangular Sequence of Tests:

$P_{n,1}, P_{n,2}, \ldots, P_{n,n}$
$P_{n+1,1}, P_{n+1,2}, \ldots, P_{n+1,n}, P_{n+1,n+1}$
$P_{n+2,1}, P_{n+2,2}, \ldots, P_{n+2,n+1}, P_{n+2,n+2}$
. . . . . .
. . . . . .

Triangular Sequence of Latent Ability Variables:

$\Theta_{n,1}, \Theta_{n,2}, \ldots, \Theta_{n,J_n}$
$\Theta_{n+1,1}, \Theta_{n+1,2}, \ldots, \Theta_{n+1,J_{n+1}}$
$\Theta_{n+2,1}, \Theta_{n+2,2}, \ldots, \Theta_{n+2,J_{n+2}}$
. . . . . .
. . . . . .

Here a test is identified by the set of ICCs associated with its items. Each row in the array of tests given above denotes a separate test, with $P_{n,i}$ denoting the ith item on the length n test. As has already been mentioned, the ICCs and their corresponding items in a given row may be entirely distinct from those in another row of this triangular array. In a test manufacturing process one might imagine that, due to test specifications among other considerations, the optimal choice of items for a length $n + 1$ test would very likely differ from merely combining the items chosen for a length n test with one additional item. For the test of length n a sample of J_n examinees, with latent abilities coming from a distribution F, is randomly selected from the test taking population. $\Theta_{n,k}$ denotes the ability of the kth examinee taking the length n test. For each $P_{n,i}$ we need to construct an estimate $\hat{P}_{n,i}$, and similarly we must construct estimates $\hat{\Theta}_{n,k}$ of the latent abilities $\Theta_{n,k}$.

The procedures for obtaining these estimates are detailed in the next section, but first a list of some important notation that will be used throughout the remaining sections and the appendix is given. Much of the notation listed below has not yet been introduced, and the list is provided for readers who may need to refer back to remind themselves about the meaning of some notation that is encountered in later sections.

Notation:

1. n = test length index.
2. J_n = the number of examinees taking the length n exam.
3. $P_{n,i}$ = the ICC associated with i-th item on the length n test.
4. $\bar{P}_n$ = the test characteristic curve.

$$\bar{P}_n = \frac{1}{n} \sum_{i=1}^{n} P_{n,i}.$$

5. $\bar{P}_n^{-1}$ = inverse function of $\bar{P}_n$ defined in the appendix.
6. $\hat{P}_{n,i}$ = estimate of $P_{n,i}$ obtained by kernel smoothing.
7. $\delta_{n,a,b}$ = index of worst case error in ICC estimation.

$$\delta_{n,a,b} = \sup_{\theta \in (a,b)} \max_{1 \le i \le n} |\hat{P}_{n,i}(\theta) - P_{n,i}(\theta)|.$$

8. $\Theta_{n,k}$ = the latent ability parameter of the kth examinee taking the length n exam.
9. $\hat{\Theta}_{n,k}$ = the estimate of $\Theta_{n,k}$ given in (6).
10. $\hat{\Theta}_{n,-i,k}$ = the leave-item-i-out estimate of $\Theta_{n,k}$ defined in (4).
11. Δ_n = the maximum of $|\hat{\Theta}_{n,j} - \Theta_{n,j}|$, $j = 1, \ldots, J_n$.
12. $\Delta_{n,-i}$ = the maximum of $|\hat{\Theta}_{n,-i,j} - \Theta_{n,j}|$, $j = 1, \ldots, J_n$.
13. $\bar{U}_{n,k}$ = the proportion correct score of examinee k taking the length n exam.
14. $\bar{U}_{n,-i,k}$ = the proportion correct score of examinee k excluding item i.
15. $p_{nk} = P[\bar{U}_n = k]$.
16. F = the chosen distribution function of the random ability variable Θ.
17. F_n = the population distribution function of $\bar{U}_n$.
18. $\hat{F}_{J_n}$ = the empirical distribution function of $\bar{U}_{n,k}$, $k = 1, \ldots, J_n$.
19. $F_{n,-i}$ = the population distribution function of $\bar{U}_{n,-i}$.
20. $\hat{F}_{J_n,-i}$ = the empirical distribution function $\bar{U}_{n,-i,k}$, $k = 1, \ldots, J_n$.
21. $W_{n,-i,k}$ = random variables used to break ties in the $\bar{U}_{n,-i,k}$.
22. $\bar{U}^*_{n,-i,k} = \bar{U}_{n,-i,k} + W_{n,-i,k}$.
23. $G_{J_n,-i}$ = the empirical distribution function of $\bar{U}^*_{n,-i,k}$, $k = 1, \ldots, J_n$.
24. m = constant regarding the slope of $\bar{P}_n$ discussed in assumption A3.
25. C = constant discussed in assumption A4.
26. M_L and M_U = constants discussed in assumption A5.
27. $M_{a,b}$ = constant discussed in assumption A7.
28. K = kernel function used for smoothing that follows conditions C1 through C3.
29. M_K = upper bound for the function K discussed in condition C1.
30. C_K = constant regarding the support of K discussed in condition C2.
31. L_K = maximum slope of function K discussed in condition C3.
32. h_n = the bandwidth used for smoothing for the test of length n.
33. L_α and U_α = constants related to the size of h_n discussed in condition C4.
34. O: A sequence a_n is said to be less than or equal to $O(b_n)$ if there is some constant D such that $a_n < Db_n$ for all n.

Kernel Smoothed ICC and Ordinal Ability Estimation

First, a five step procedure is given for kernel smoothed ICC estimation. An estimate of $P_{n,i}$, the ith ICC of the length n exam, is obtained by applying Steps 1 through 5.

Step 1. Observe the proportion correct on the remaining items $\bar{U}_{n,-i,1}, \ldots, \bar{U}_{n,-i,J_n}$. The ability estimates playing the role of the independent variable in the Nadaraya-Watson estimator are based on these scores. It is desirable to exclude the item score of the item whose ICC is currently being estimated from the set of data used to estimate abilities. Among other advantages, this avoids artificially anchoring the upper and lower asymptotes of the ICC estimate. Also, from a theoretical perspective, under an assumption of local independence the ability estimate will be independent of the score on the ith item after conditioning on the examinee's true latent ability. This property is convenient when studying the statistical properties of the ICC estimates.

Step 2. Add $W_{n,-i,k}$ to $\bar{U}_{n,-i,k}$ to obtain a sequence with no ties: $\bar{U}^*_{n,-i,1}, \ldots, \bar{U}^*_{n,-i,J_n}$. In practice this random breaking of ties or "jittering" has little influence on the ICC estimates but serves to simplify the problem both computationally and theoretically.

Step 3. Find the empirical percentile of tie-broken proportion correct (leaving out the item i score) for each examinee.

$$\hat{G}_{J_n,-i}(\bar{U}^*_{n,-i,1}), \ldots, \hat{G}_{J_n,-i}(\bar{U}^*_{n,-i,J_n})$$

This step is equivalent to ranking the adjusted total scores.

Step 4. Convert the empirical percentiles obtained in Step 3 to the appropriate quantile on ability scale determined by the distribution function F.

$$\hat{\Theta}_{n,-i,1}, \ldots, \hat{\Theta}_{n,-i,J_n} = F^{-1}(\hat{G}_{J_n,-i}(\bar{U}^*_{n,-i,1})), \ldots, F^{-1}(\hat{G}_{J_n,-i}(\bar{U}^*_{n,-i,J_n})). \quad (4)$$

In Step 4 a scale is chosen on which to represent the ICCs. The ordinal ability estimates of step 3 are assigned to the appropriate quantile on the chosen ability scale to arrive at the new ability estimates of Step 4. It should be pointed out that the choice of scale for Θ is arbitrary and is not imposing any assumptions or parametric constraints on the model. That is, once one has an IRT model for the *manifest distribution* of the test data, one can always monotonically transform the scale of the latent ability distribution as long as the corresponding transformation is made to the ICCs to retain the same manifest distribution. Because no parametric assumptions are made concerning the form of the ICCs, the choice of scale becomes perfectly arbitrary as long as certain qualitative assumptions, such as those given in the next section, remain satisfied.

Step 5. Calculate $\hat{P}_{n,i}(\theta)$.

$$\hat{P}_{n,i}(\theta) = \frac{\sum_{k=1}^{J_n} K\left(\frac{\theta - \hat{\Theta}_{n,-i,k}}{h_n}\right) Y_{n,i,k}}{\sum_{k=1}^{J_n} K\left(\frac{\theta - \hat{\Theta}_{n,-i,k}}{h_n}\right)} \quad (5)$$

Certainly there are many other details that one must address such as choosing the kernel function K and the bandwidth h_n. For these matters the reader is referred to Härdle (1990) and Ramsay (1991). Also see Ramsay for real data examples and a simulation study.

The next step is to estimate the latent ability of each examinee on the scale determined by the distribution function F. In modern parametric IRT, full-information ability estimation methods are preferred over limited-information methods that only make use of sample moments such as proportion correct score. However, in most practical classroom testing situations total score is used to order examinees. In this paper we consider the asymptotic theory for ability estimation obtained by ordering the observed total test scores and transforming this ordinal information to the corresponding quantile on the chosen latent ability scale. Because the latent ability estimates are based on ranked total scores, they are referred to as *ordinal ability estimates.*

The procedure for obtaining ordinal ability estimates is quite similar to Steps 3 and 4 given above to obtain proxies for the independent variables in the kernel smoothing formula of (5). However, no ties are broken and all of the item responses are used to obtain the total test scores. That is, proportion correct scores $\bar{U}_{n,k}$ are used rather than $\bar{U}^*_{n,-i,k}$. Thus, the ability estimates for the J_n examinees taking the length n test are given by

$$\hat{\Theta}_{n,1}, \ldots, \hat{\Theta}_{n,J_n} = F^{-1}(\hat{F}_{J_n}(\bar{U}_{n,1})), \ldots, F^{-1}(\hat{F}_{J_n}(\bar{U}_{n,J_n})). \quad (6)$$

Large Sample, Long Test Theory

Next, assumptions are stated that are needed to carry out the asymptotic analysis of the ability and ICC estimates that were presented in the previous section. It is worthwhile

to discuss each assumption in some depth in order to see how each is psychometrically plausible and relatively unrestrictive. The basic purpose of doing nonparametric statistics is to be able to perform accurate estimation and inference within a very general and unrestrictive model, and the assumptions used for an analysis of a nonparametric procedure should be consistent with this philosophy.

Assumptions:

A1. The latent ability Θ is a scalar valued random variable and item responses are mutually independent conditional on Θ. The assumption that Θ is scalar valued is usually referred to as *unidimensionality*, while standard terminology for the latter condition is *local independence.*

A2. Θ has a uniform distribution on the open interval (0, 1). Given the unidimensionality assumption above, this assumption merely amounts to a particular choice of scale. Of course, a choice of scale has a direct influence on the shape of the ICCs when represented on the chosen scale. Thus, the smoothness assumptions on the ICCs given in assumptions below need only make sense when the ICCs are represented on the uniform (0, 1) scale.

A3. $\bar{P}_n'$ exists everywhere in (0, 1), and there is a number m such that for each θ in (0, 1), $0 < m < \bar{P}_n'(\theta)$. Because ICCs are usually defined on the entire real line where each ICC is forced to asymptote, this assumption may seem restrictive. However, when the ability distribution is compressed into the finite length interval (0, 1), it is reasonable to assume that the slope of the test characteristic curve is uniformly bounded from below. The assumption is used to guarantee that the test is discriminating at each percentile of the ability distribution. To illustrate the validity of this assumption, the test characteristic curve for an ASVAB auto shop exam using fitted three-parameter logistic functions obtained by marginal maximum likelihood estimation (Bock & Aitkin, 1981) with a standard normal ability scale are shown in Figure 3, after a change-of-variables transformation to the uniform (0, 1) ability distribution. By insisting that the test characteristic curve has a positive derivative it is constrained to be strictly monotone increasing. However, this assumption still allows for the existence of some disfunctional items that are not monotone. It is only required that the number of such items is relatively small when compared to the test length. A benefit of nonparametric ICC estimation is that it can detect non-monotone ICCs if the test characteristic curve is increasing.

A somewhat less restrictive assumption on the derivatives of the test characteristic curves is to assume that they are uniformly bounded from below within some compact subinterval $[a, b]$ of (0, 1). This assumption would require only slight alterations to the statements of the theorems to follow, but makes the proofs much more tedious.

A4. Uniform boundedness of observed proportion correct histograms: This assumption is simply that there is a constant C that does not depend upon test length n, such that the discrete theoretical density or histogram for the proportion correct score $\bar{U}_n$ lies entirely below the line $y = C$. This is equivalent to the assumption that $np_{n,k} < C$ for all k.

For many testing situations, one would encounter score distributions that may be skewed or nonnormal in some way, but one would not expect severely sharp spikes in the score distribution that would invalidate this assumption. See Figure 4.

A5. Note that the sample size J_n is indexed by the test length n. A random sample of J_n examinees is drawn for the length n exam, and an independent sample of J_{n+1} examinees is randomly selected for the length $n + 1$ exam. Assume that there are positive constants M_L, M_U, and r, such that for all n,

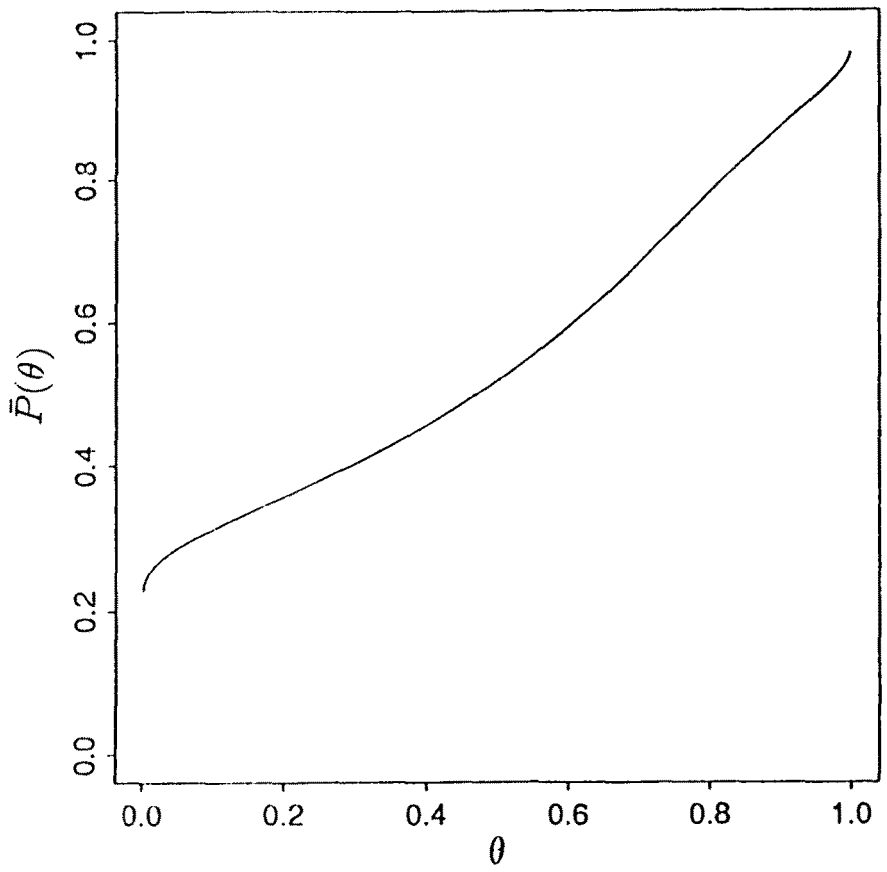

FIGURE 3.
The estimated test characteristic curve $\bar{P}(\theta)$, is plotted against θ on the Unif(0, 1) ability scale. The minimum value of the derivative $\bar{P}'(\theta)$ is equal to 0.43. In assumption A3, m may be set equal to 0.43.

$$M_L n^{3/2} < J_n < M_U n^r.$$

One would expect the sample size to considerably exceed the test length. The constant r in the upper bound can be taken arbitrarily large, and in that sense does not impose a serious constraint. The purpose of this assumption is to limit the number of examinees. The joint consistency of ICC and ability estimation cannot be obtained if too many examinees are present. Haberman (1977) used a similar, though somewhat less restrictive, assumption to show the joint consistency of maximum likelihood estimation of Rasch model parameters.

A6. For each pair (n, i) in the triangular array of items, $P'_{n,i}$ exists and is continuous in the open interval (0, 1).

A7. For each compact subset $[a, b]$ of (0, 1), there is a positive constant $M_{a,b}$ such that for all $\theta \in [a, b]$ and all pairs (n, i),

$$|P'_{n,i}(\theta)| < M_{a,b}.$$

This assumption could be strengthened by insisting on a uniform and global bound for the derivatives of the $P_{n,i}$. However, such a strengthening would be unrealistic and restrictive when working with an ability distribution with a finite range as in our case in which the distribution of Θ is supported on (0, 1), because the derivatives can become very steep at the boundaries.

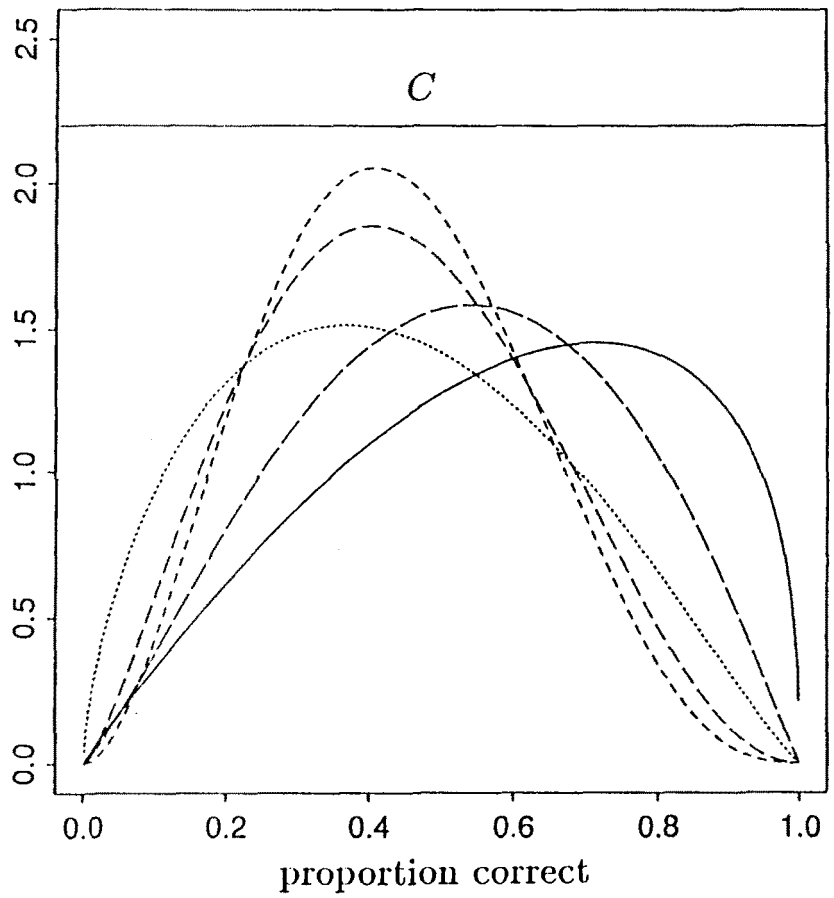

FIGURE 4.
Theoretical distributions for proportion correct scores are plotted. Assumption A4 merely states that these histograms are uniformly bounded over all test lengths.

Ordinal Ability Estimation

First, the probability of a large deviation of $\hat{\Theta}_n$ from Θ_n is studied. Consider the probability that $|\hat{\Theta}_n - \Theta_n|$ is greater than a multiple of $n^{-\alpha}$, where α is any positive number less than 1/2. It will be seen that the possibility of such a large deviation becomes exponentially small as the test length increases. Requiring only minor adjustments to the proofs, the results stated here are also true for the leave-one-out ability estimates of equation (4) that play a crucial role in the theory of the kernel smoothed ICC estimates. Later we will see a connection between α and the smoothing parameter h_n. The idea is to let h_n be of the order $n^{-\alpha}$, and from Theorem 1 that is stated below we will know that the chance of smoothing over examinees whose estimated abilities are more than a bandwidth from their true abilities is exponentially small. Clearly the $\hat{\Theta}_{n,j}$, $j = 1, \ldots, J_n$ are identically distributed, though stochastically dependent, and the following theorem is stated in terms of $\hat{\Theta}_{n,1}$.

Theorem 1. Let ε_n be a positive sequence and $0 < \alpha < 1/2$ such that for some positive constant C_α, $\varepsilon_n > C_\alpha n^{-\alpha}$. Given assumptions A1 through A7, there exist positive constants C_1 and C_2 (depending on α) such that for all n

$$P[|\hat{\Theta}_{n,1} - \Theta_{n,1}| > \varepsilon_n] \leq C_1 \exp[-C_2 n^{1-2\alpha}].$$

The proof of Theorem 1 is outlined in the appendix.

Recall from the list of notation that Δ_n denotes the maximum of $|\hat{\Theta}_{n,j} - \Theta_{n,j}|$ for $j = 1, \ldots, J_n$. It is the maximum error of ability estimation over all J_n examinees taking the length n exam. Thus

$$P[\underline{\Delta}_n > \varepsilon_n] = P[\cup_{j=1}^{J_n}\{|\hat{\Theta}_{n,j} - \Theta_{n,j}| > \varepsilon_n\}].$$

By the subadditivity of probability measures

$$P[\Delta_n > \varepsilon_n] \leq \sum_{j=1}^{J_n} P[|\hat{\Theta}_{n,j} - \Theta_{n,j}| > \varepsilon_n]. \tag{7}$$

Applying Theorem 1 and assumption A5 ($J_n \leq M_U n^r$), it is seen that the right side of (7) is less than or equal to

$$M_U n^r C_1 \exp[-C_2 n^{1-2\alpha}].$$

Thus,

$$P[\Delta_n > \varepsilon_n] < M_U n^r C_1 \exp[-C_2 n^{1-2\alpha}]. \tag{8}$$

The right side of inequality (8) is summable over the natural numbers, and an application of the Borel-Cantelli lemma (Billingsley, 1986) can be applied to prove Theorem 2 stated below:

Theorem 2. Given assumptions A1 through A7, $\Delta_n \to 0$ with probability equal to 1.

Thus, the maximum error in the ability estimates obtained by transforming the ranked proportion correct scores as described by (6), converges to 0 as the test length increases.

Convergence of Kernel Smoothed ICC Estimates

Next the large sample, long test behavior of the ICC estimates is considered. An upper bound on the mean squared error of estimation as a function of the test length is given in Theorem 3. Then the simultaneous accuracy of ICC estimation over all items within an interval of ability levels is considered in Theorem 4. The proofs of these theorems make use of some quite general conditions C1 through C4 given below, concerning the kernel function K, and the sequence of shrinking bandwidths h_n.

Conditions on the kernel function and bandwidth:

C1. The kernel function K is bounded by a constant M_K.
C2. K has bounded support. In particular, there is a constant C_K such that $K(x) > 0$ when $|x| < C_K$ and $K(x) = 0$ when $|x| \geq C_K$.
C3. There is a constant L_K such that for each pair (x, x'),

$$|K(x) - K(x')| \leq L_k|x - x'|.$$

Note that C2 and C3 imply C1. However, including this slight redundancy is harmless and the notation M_K of C1 is used in proofs of the theorems for the kernel smoothed ICC estimates.

C4. There is a number $\alpha \in (0, 1/2)$, and positive numbers L_α and U_α such that for all n, $L_\alpha n^{-\alpha} < h_n < U_\alpha n^{-\alpha}$. This condition is used to keep h_n from converging to 0 more quickly than the ability estimates in equation (4) converge to the true latent abilities.

By Theorem 1 we know that for $\alpha \in (0, 1/2)$, the chance that that any of the ability estimates is more than $n^{-\alpha}$ from their true values is exponentially small. There is no advantage to smoothing over an interval of ability levels that shrinks more quickly than the errors of measurement in the independent variables, and α is chosen to balance the error related to bandwidth size together with errors related to mismeasurement of the ability parameters.

A standard measure of the quality of estimation at a particular point θ is given by the mean squared error of estimation:

$$MSE(\hat{P}_{n,i}(\theta)) \equiv E[(\hat{P}_{n,i}(\theta) - P_{n,i}(\theta))^2].$$

Generally for parametric models (e.g., least squares regression) the mean squared error for estimating an unknown regression function $m(x)$ decreases at the rate of $1/J$, where J is the sample size. Of course this presumes that the parametric model is correct. If one opts for a more flexible approach using a linear smoother such as kernel smoothing, a good choice of bandwidth can at best ensure that the mean squared error of estimating $m(x)$, decreases at the rate of $1/J^{4/5}$. However, both of these cases pertain to the more usual regression problem in which the independent variable can be observed. Here mean squared error as a function of the test length is considered. Recall that test length is loosely related to the sample size through assumption A7.

Condition C4 states that h_n is of the order $n^{-\alpha}$, where $0 < \alpha < 1/2$. As was previously mentioned, this condition is used to regulate the size of the "window" for kernel smoothing, in relation to the accuracy of the stand-in values $\hat{\Theta}_{n,-i}$ for the independent variable Θ_n. Theorem 3 considers the highest mean squared error of estimation at a fixed point over all items on the exam, for any choice of α less than 1/2.

Theorem 3. Given assumptions A1 through A7 and conditions C1 through C4, for a fixed $\theta_0 \in (0, 1)$

$$\max \{MSE(\hat{P}_{n,i}(\theta_0)) : 1 \leq i \leq n\} \leq O(n^{-2\alpha})$$

where α is the constant in condition C4.

The proof of Theorem 3 is outlined in the appendix.

Of course α is arbitrary as long as $0 \leq \alpha \leq 1/2$. By choosing α close to 1/2, it is easy to see that the *MSE* can be made close to $O(1/n)$. If in replace of assumption A5 it is assumed that the sample size J_n is some multiple of $n^{3/2}$ it is clear that the *MSE* can be made close to $O(J_n^{-2/3})$.

While $MSE(\hat{P}_{n,i}(\theta))$ describes the accuracy of estimation at a single value of θ, it is perhaps more important to know that all of the items are accurately estimated at each point in a wide interval. In other words, one would like to know something about the simultaneous estimation of all the ICCs in some subinterval (a, b) of $(0, 1)$. An index of the worst case departure from the true ICCs within a fixed interval is given by

$$\delta_{n,a,b} = \sup_{\theta \in (a,b)} \max_{1 \leq i \leq n} |\hat{P}_{n,i}(\theta) - P_{n,i}(\theta)|.$$

It would be encouraging to know that for all ability levels in (a, b), the ICCs are all estimated within some small number ε of the true ICCs ($\delta_{n,a,b} \leq \varepsilon$). Theorem 4 asserts that $\delta_{n,a,b}$ converges to 0 with probability equal to 1 as the test length $n \to \infty$, ensuring that all of the estimates eventually become good at all points in (a, b). Note that a and b can

be taken arbitrarily close to 0 and 1, but for technical reasons cannot be set equal to these boundary points of the ability scale.

Theorem 4. Given assumptions A1 through A7 and conditions C1 through C4, $\delta_{n,a,b} \to 0$ with probability equal to 1 as test length $n \to \infty$.

The proof of Theorem 4 is outlined in the appendix.

The strong consistency results of Theorems 2 and 4 imply that triangular arrays of latent abilities and ICCs satisfying assumptions A1 through A7 can be jointly estimated to any degree of accuracy. A formal statement of this joint consistency is given below:

Theorem 5. $\max \{\Delta_n, \delta_{n,a,b}\} \to 0$ with probability equal to 1 as test length $n \to \infty$.

Theorem 5 follows directly from Theorems 2 and 4 by noting that the probability measure of the union of two sets, each of probability measure 0, also must have probability equal to 0. The two sets of probability 0 alluded to here are just those sets where the ICC and ability estimates fail to converge.

It is worth mentioning that although the ICC estimates considered here are on the uniform ability scale, they perform equally well on any other scale. One only needs an initial scale on which to construct the estimates, but neither the theory nor the performance of the ICC estimates are diminished by a change of scale. The only requirement is that assumptions A1 through A7 are satisfied when the ICCs are represented on the uniform (0, 1) scale. If one wishes to represent the ICCs on the ability scale determined by the distribution function F, and θ is on the uniform (0, 1) scale, $\eta = F^{-1}(\theta)$ has F as its distribution function. Thus, if estimation is accurate on some interval (a, b) on the uniform (0, 1) scale, the accuracy is unchanged by simply viewing the ICCs as a function of η on the interval $(F^{-1}(a), F^{-1}(b))$.

Discussion

An appealing feature of nonparametric curve estimation is that under very mild assumptions, the true curve can be consistently estimated. The large sample, long test theory presented in this paper shows that this desirable property holds in IRT for kernel smoothed ICC estimation. In addition, Theorem 5 states that all of the ICC estimates and ordinal ability estimates simultaneously converge to their true values. This joint consistency gives hope that to the degree to which some smooth unidimensional IRT model explains psychological reality, that model will be asymptotically recovered by kernel smoothing and ordinal ability estimation.

In contrast, this property is not necessarily shared by joint maximum likelihood estimation of parametric models, even when the correct family of models is chosen. Neyman and Scott (1948) showed that when the number of structural parameters (item parameters), increase with the number of incidental parameters (ability parameters), consistency is not always satisfied. However, Haberman (1977) was able to prove the joint consistency of maximum likelihood estimation of item and examinee parameters in the special case of the Rasch model. He obtained strong consistency results similar to Theorem 5, showing that the maximum error over all item and examinee parameters converges almost surely to 0 as both the test length and sample size go to infinity. Much like assumption A5 was used here to bound the number of examinees, Haberman had the somewhat less stringent requirements that $J_n > n$, and $\log (J_n)/n$ converges to 0.

Although kernel smoothed ICC estimation can be performed with little worry about

the functional form of the underlying ICCs, users must still consider the dimensionality of the latent ability space. Recall that assumption A1 requires the existence of a unidimensional latent trait under which local independence holds. The theorems presented in this paper all hinge upon this assumption being satisfied. Although nonparametric ICC estimation relieves the user of the need to correctly guess a parametric family of curves to fit the data, great care must be taken to ensure that the exam is measuring a well defined unidimensional trait.

A user of the nonparametric methods described in this paper may ask how long the test needs to be and how many examinees are required before the techniques described here can be effective. Ramsay (1991) reported on a simulation study in which a three-parameter logistic model and a standard normal ability distribution were used to randomly generate data. The parameter values are given in Ramsay and Winsberg (1991). By repeated simulation the root-integrated-mean-square-error (RIMSE) of kernel smoothed ICC estimation was compared to the RIMSE of ICCs obtained from marginal maximum likelihood estimation of three-parameter logistic item parameters.

$$RIMSE = \left[\int (\hat{P}(\theta) - P(\theta))^2 \phi(\theta)\, d\theta \right]^{1/2}$$

The function ϕ denotes the probability density function of a standard normal random variable. With 35 items and 500 examinees, the kernel smoothed ICC estimates were as efficient as estimates using marginal maximum likelihood. Each method resulted in $RIMSE = 0.037$, even though the data was generated from the very model assumed by the marginal maximum likelihood procedure. Although maximum likelihood estimation is *asymptotically efficient*, kernel smoothing was equally effective for the pre-asymptotic sample size and test length used in Ramsay's study. Thus, it appears that kernel smoothing can be as effective as standard parametric methods for realistic sample sizes and test lengths. Furthermore, the convergence of kernel smoothed ICC estimation is unaffected if the underlying ICCs do not belong to any particular well known parameterized family of curves.

TESTGRAF (Ramsay, 1993) produces kernel smoothed ICC estimates as well as option characteristic curve estimates for multiple choice items. The asymptotic results given here provide theoretical support for TESTGRAF produced ICCs estimates and can easily be extended to the more general option characteristic curves.

In conclusion, the flexibility of kernel smoothed ICC estimation, supported by the theoretical results of this paper, provide further evidence that TESTGRAF type ICC estimates can be relied upon in many testing situations when the validity of common parametric modeling techniques and their estimation procedures may be questionable.

Appendix

The proofs of Theorems 1, 3, and 4 are outlined in this appendix. Due to limited space, most propositions and lemmas needed to prove these theorems are stated without proof. A supplement to this appendix that provides complete proofs of these lemmas may be obtained from the author by request. Please refer to the list of notation provided in a previous section when unfamiliar notation is encountered.

Sums of item scores form the basis for obtaining the ordinal ability estimates of equation (6). By local independence, after conditioning on an examinees ability, such a sum is just a sum of independent Bernoulli variables. Also, in the context of kernel smoothing, weighted sums of scores on a single item over the set of examinees are of primary interest as seen in equation (5). Both of these sums can be thought of as sums of

independent and bounded random variables. Thus, to prove theorems for both ability estimates and kernel smoothed ICC estimates, it is useful to be able to obtain probability inequalities for such sums of random variables. A particular theorem concerning sums of independent and bounded random variables was used extensively to prove lemmas leading to the proofs of Theorems 1, 3, and 4. A statement of the theorem is given below. A good reference containing this theorem and other results that were useful to this study is Serfling (1980).

Theorem 6. (Hoeffding, 1986) Let $X_1, \ldots, X_n$ be independent random variables satisfying $P(a \leq X_i \leq b) = 1$ for each i, where $a < b$. Then, for $t > 0$

$$P\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} E[X_i] \geq nt\right) \leq \exp\left[\frac{-2nt^2}{(b-a)^2}\right].$$

Next some lemmas that are used in the proof of Theorem 1 are stated. The first concerns a probabilistic exponential bound for large discrepancies between observed and true proportion correct, as a function of test length n.

Prior to stating Lemma 1, an inverse function of the test characteristic curve $\bar{P}_n(\theta)$ must be defined. Denote the upper and lower limits of the monotone (A3) functions $\bar{P}_n$ as follows:

$$\alpha_n = \lim_{\theta \to 0} \bar{P}_n(\theta) \qquad \beta_n = \lim_{\theta \to 1} \bar{P}_n(\theta)$$

Define:

$\bar{P}_n^{-1}(\mu) = 0$ if $\mu \leq \alpha_n$
$\bar{P}_n^{-1}(\mu) =$ the unique $\theta \in (0, 1)$ for which $\bar{P}_n(\theta) = \mu$ if $\alpha_n < \mu < \beta_n$. The existence of such a unique θ is guaranteed by assumption A3.
$\bar{P}_n^{-1}(\mu) = 1$ if $\mu \geq \beta_n$.

Lemma 1. Let $\varepsilon > 0$, then

$$P[|\bar{P}_n^{-1}(\bar{U}_n) - \Theta| > \varepsilon] \leq 2 \exp[-2n\varepsilon^2 m^2].$$

Next a lemma is stated regarding the difference between the observed and true percentiles based on F_n, the distribution function of the proportion correct score $\bar{U}_n$. This lemma exploits assumption (A4) regarding the regularity of the score distributions.

Lemma 2. Given $\varepsilon > 0$,

$$P[|F_n(\bar{U}_n) - F_n(\bar{P}_n(\Theta))| > \varepsilon] \leq 2 \exp\left[-2n\left(\frac{\varepsilon}{C} - \frac{2}{n}\right)^2 m^2\right].$$

Lemma 3. Let α be a positive number less than 1/2 and r_n be a positive sequence such that as $n \to \infty$

$$r_n \to 0 \quad \text{and} \quad r_n \geq C_\alpha n^{-\alpha} \text{ for some positive constant } C_\alpha.$$

Then there is an integer n_0 such that for $n \geq n_0$, $|F_n(\bar{P}_n(\theta)) - \theta| < r_n$ for all $\theta \in (0, 1)$.

Next, the proof of Theorem 1 is outlined.

Proof of Theorem 1. The proof consists of expanding $P[|\hat{\Theta}_{n,1} - \Theta_{n,1}| > \varepsilon_n]$, and bounding each term in the expansion.

From (6)

$$P[|\hat{\Theta}_{n,1} - \Theta_{n,1}| > \varepsilon_n] \equiv P[|\hat{F}_{J_n}(\bar{U}_{n,1}) - \Theta_{n,1}| > \varepsilon_n].$$

Thus, by expanding the right side

$$P[|\hat{\Theta}_{n,1} - \Theta_{n,1}| > \varepsilon_n] \leq P\left[|\hat{F}_{J_n}(\bar{U}_{n,1}) - F_n(\bar{U}_{n,1})| > \frac{\varepsilon_n}{2}\right] + P\left[|F_n(\bar{U}_{n,1}) - \Theta_{n,1}| > \frac{\varepsilon_n}{2}\right]$$

$$\leq P\left[|\hat{F}_{J_n}(\bar{U}_{n,1}) - F_n(\bar{U}_{n,1})| > \frac{\varepsilon_n}{2}\right] + P\left[|F_n(\bar{U}_{n,1}) - F_n(\bar{P}_n(\Theta_{n,1}))| > \frac{\varepsilon_n}{4}\right]$$

$$+ P\left[|F_n(\bar{P}_n(\Theta_{n,1})) - \Theta_{n,1}| > \frac{\varepsilon_n}{4}\right]. \quad (9)$$

The first term of (9) can be bounded by a well known theorem regarding the Kolmogorov-Smirnov distance (Serfling, 1980).

$$P\left[|\hat{F}_{J_n}(\bar{U}_{n,1}) - F_n(\bar{U}_{n,1})| > \frac{\varepsilon_n}{2}\right] \leq B_\Delta \exp\left[\frac{-J_n\varepsilon_n^2}{2}\right], \quad (10)$$

where B_Δ is a universal constant that does not depend on n or the distribution function F_n. Lemma 2 can be applied to bound the second term of (7).

$$P\left[|F_n(\bar{U}_{n,1}) - F_n(\bar{P}_n(\Theta_{n,1}))| > \frac{\varepsilon_n}{4}\right] \leq 2 \exp\left[-2n\left(\frac{\varepsilon}{4C} - \frac{2}{n}\right)^2 m^2\right] \quad (11)$$

Finally, an application of Lemma 3 shows that for large enough n

$$P\left[|F_n(\bar{P}_n(\Theta)) - \Theta| > \frac{\varepsilon_n}{4}\right] = 0. \quad (12)$$

Thus, the third term of (7) eventually becomes equal to 0 and drops out.

After combining (8), (9), and (10), we see that for large enough n,

$$P[|\hat{\Theta}_{n,1} - \Theta_{n,1}| > \varepsilon_n] \leq B_\Delta \exp\left[\frac{-J_n\varepsilon_n^2}{2}\right] + 2 \exp\left[-2n\left(\frac{\varepsilon}{4C} - \frac{2}{n}\right)^2 m^2\right]. \quad (13)$$

By noting that $\varepsilon_n \geq C_\alpha n^{-\alpha}$ and $J_n \geq M_L n^{3/2}$ (A5), it is seen that there is a positive C_2 such that both the exponents on the right side of (11) are less than $-C_2 n^{1-2\alpha}$, and to account for n small enough that equation (10) does not yet hold, a constant C_1 can be chosen large enough so that for all n the right side of (13) is less than

$$C_1 \exp[-C_2 n^{1-2\alpha}].$$

This completes the outline of the proof of Theorem 1. □

Next some lemmas are stated that are needed for the proofs of Theorems 3 and 4 regarding the kernel smoothed ICC estimates. Recall that the leave-one-out ability estimates, $\hat{\Theta}_{n,-i,j}$ of (4), used in the estimation of the ith item's ICC requires adding a uniformly distributed random variable to each observed proportion $\bar{U}_{n,-i,j}$ (observed proportion for examinee j excluding item i) to obtain $\bar{U}^*_{n,-i,j}$, which are then ranked over $j = 1, \ldots, J_n$. It is important to consider how much randomly breaking ties influences ability estimation. A lemma is stated below (Lemma 4) to show that randomly breaking ties has a very small influence on ability estimation as the test length increases. This amounts to showing that with high probability the empirical distribution function of $\bar{U}_{n,-i}$, $\hat{F}_{J_n,-i}$, when

evaluated at $\bar{U}_{n,-i,j}$ will be very close to the empirical distribution function of $\bar{U}^*_{n,-i}$, $\hat{G}_{J_n,-i}$, when evaluated at $\bar{U}^*_{n,-i,j}$. The result is that the exact same statement of Theorem 1 holds when the ability estimates $\hat{\Theta}_{n,j}$ are replaced with the leave-one-out ability estimates $\hat{\Theta}_{n,-i,j}$. This is because the $\bar{U}_{n,-i,j}$ are almost identical to the $\bar{U}_{n,j}$, and breaking ties is not sufficiently consequential to perturb the asymptotic theory.

Lemma 4. Let ε_n be a positive sequence such that as $n \to \infty$, $\varepsilon_n/n \to 0$. Then for some constant D that is independent of the test length n and the item index i

$$P[|\hat{G}_{J_n,-i}(\bar{U}^*_{n,-i}) - \hat{F}_{J_n,-i}(\bar{U}_{n,-i})| > \varepsilon_n] \leq D \exp\left[\frac{-2J_n\varepsilon_n^2}{9}\right].$$

Next a particularly vital lemma regarding the probability that $\hat{P}_{n,i}(\theta)$ is more than a multiple of the bandwidth away from $P_{n,i}(\theta)$ is stated and proved.

Lemma 5. Let $0 < c < a < b < d < 1$. There is an integer n_0 such that for $n > n_0$, the following inequality holds for all i and $\theta \in (a, b)$:

$$P[|\hat{P}_{n,i}(\theta_0) - P_{n,i}(\theta_0)| > 3M_{c,d}C_K h_n] \leq 2 \exp[-D_\alpha n^{3/2-3\alpha}] + M_U n^r C_1 \exp[-C_2 n^{1-2\alpha}]$$

where M_U, $M_{c,d}$ and C_K are defined in A5, A7, and C4 respectively. C_1 and C_2 are positive constants that depend only on α while D_α depends on c and d as well as α.

Lemma 5 plays a key role in the proofs of both Theorems 3 and 4, and it is worthwhile to provide an outline of its proof. First some observations are made that are helpful in the proof of Lemma 5.

Recall that

$$\hat{P}_{n,i}(\theta) = \frac{\sum_{j=1}^{J_n} K\left(\frac{\theta - \hat{\Theta}_{n,-i,j}}{h_n}\right) Y_{n,i,j}}{\sum_{j=1}^{J_n} K\left(\frac{\theta - \hat{\Theta}_{n,-i,j}}{h_n}\right)} \equiv \frac{1}{D_n(\theta)} \sum_{j=1}^{J_n} K\left(\frac{\theta - \hat{\Theta}_{n,-i,j}}{h_n}\right) Y_{n,i,j}.$$

It is helpful to see that the denominator $D_n(\theta)$ is not random and does not depend on i, the index of the item whose ICC is currently being estimated. After sorting $(\hat{\Theta}_{n,1}, \ldots, \hat{\Theta}_{n,J_n})$ as defined in equation (4) we get $(1/J_n, 2/J_n, \ldots, 1)$. Thus, $D_n(\theta)$ can be simplified to

$$D_n(\theta) = \sum_{j=1}^{J_n} K\left(\frac{\theta - \frac{j}{J_n}}{h_n}\right). \tag{14}$$

By using conditions C1 through C3 it can be shown that there are positive constants m_l and m_u such that for sufficiently large n and all $\theta \in (0, 1)$

$$m_l J_n h_n < D_n(\theta) < m_u J_n h_n. \tag{15}$$

In fact, as $n \to \infty$

$$\frac{D_n(\theta)}{J_n h_n} \to \int K(x)\, dx.$$

Next we consider the number of examinees contributing to the estimate of $P_{n,i}$ at a particular value of θ. By condition C.1, $K(\theta - \hat{\Theta}_{n,-i,j}/h_n) > 0$ if and only if

$$-C_K < \frac{\theta - \hat{\Theta}_{n,-i,j}}{h_n} < C_K.$$

Let $A_{n,i,\theta} = \{j : |\theta - \hat{\Theta}_{n,-i,j}/h_n| < C_K\}$. Then we can rewrite $\hat{P}_{n,i}(\theta)$ by

$$\hat{P}_{n,i}(\theta) = \frac{1}{D_n(\theta)} \sum_{j \in A_{n,i,\theta}} K\left(\frac{\theta - \hat{\Theta}_{n,-i,j}}{h_n}\right) Y_{n,i,j}. \tag{16}$$

The number of terms in this sum is just the cardinality of $A_{n,i,\theta}$, denoted by $\#A_{n,i,\theta}$. Because the $\hat{\Theta}_{n,-i,j}$ partition (0, 1) into intervals of width $1/J_n$, as $n \to \infty$

$$\frac{\#A_{n,i,\theta}}{2C_K h_n J_n} \to 1.$$

In any open subinterval (a, b) of (0, 1) there is n_0 such that for $\theta \in (a, b)$ and $n > n_0$,

$$C_K h_n J_n < \#A_{n,i,\theta} < 3C_K h_n J_n . \tag{17}$$

Proof of Lemma 5. Let $\theta_0 \in (a, b)$ and $\Delta_{n,-i}$ denote the maximum of $|\hat{\Theta}_{n,-i,j} - \Theta_{n,j}|$ for $j = 1, \ldots, J_n$. Then

$$\begin{aligned} P[|\hat{P}_{n,i}(\theta_0) - P_{n,i}(\theta_0)| > 3M_{c,d}C_k h_n] &= P[|\hat{P}_{n,i}(\theta_0) - P_{n,i}(\theta_0)| > 3M_{c,d}C_k h_n | \Delta_{n,-i} \\ &< C_k h_n] P[\Delta_{n,-i} < C_k h_n] + P[|\hat{P}_{n,i}(\theta_0) - P_{n,i}(\theta_0)| \\ &> 3M_{c,d}C_k h_n | \Delta_{n,-i} > C_k h_n] P[\Delta_{n,-i} > C_k h_n]. \end{aligned} \tag{18}$$

Thus,

$$\begin{aligned} P[|\hat{P}_{n,i}(\theta_0) - P_{n,i}(\theta_0)| > 3M_{c,d}C_k h_n] \leq P[|\hat{P}_{n,i}(\theta_0) - P_{n,i}(\theta_0)| \\ > 3M_{c,d}C_k h_n | \Delta_{n,-i} < C_k h_n] + P[\Delta_{n,-i} > C_k h_n]. \end{aligned} \tag{19}$$

By applying (8) to the second term to the right of the inequality in (19) we see that

$$\begin{aligned} P[|\hat{P}_{n,i}(\theta_0) - P_{n,i}(\theta_0)| > 3M_{c,d}C_k h_n] \leq P[|\hat{P}_{n,i}(\theta_0) - P_{n,i}(\theta_0)| > 3M_{c,d}C_k h_n | \Delta_{n,-i} \\ < C_k h_n] + M_U n^r C_1 \exp[-C_2 n^{1-2\alpha}]. \end{aligned} \tag{20}$$

Let $\boldsymbol{\Theta}_\mathbf{n}$ and $\hat{\boldsymbol{\Theta}}_{\mathbf{n},-\mathbf{i}}$ denote the vectors $(\Theta_{n,1}, \ldots, \Theta_{n,J_n})$, and $(\hat{\Theta}_{n,-i,1}, \ldots, \hat{\Theta}_{n,-i,J_n})$ respectively, and let $\boldsymbol{\theta}_\mathbf{n}$ and $\hat{\boldsymbol{\theta}}_{\mathbf{n},-\mathbf{i}}$ denote realizations of these vectors. Also let F_Δ denote the joint distribution of these vectors of estimated and true abilities conditional on $\Delta_{n,-i} < C_k h_n$, and let α correspond to the α in condition C4. Then the right side of (20) can be expressed as

$$\begin{aligned} \int P[|\hat{P}_{n,i}(\theta_0) - P_{n,i}(\theta_0)| > 3M_{c,d}C_k h_n | \Delta_{n,-i} < C_k h_n, \hat{\boldsymbol{\theta}}_{\mathbf{n},-\mathbf{i}}, \boldsymbol{\theta}_\mathbf{n}] \, dF_\Delta(\hat{\boldsymbol{\theta}}_{\mathbf{n},-\mathbf{i}}, \boldsymbol{\theta}_\mathbf{n}) \\ + M_U n^r C_1 \exp[-C_2 n^{1-2\alpha}]. \end{aligned} \tag{21}$$

The strategy is to bound the integrand in the first term of (21) uniformly over all vectors $\hat{\boldsymbol{\theta}}_{\mathbf{n},-\mathbf{i}}$ and $\boldsymbol{\theta}_\mathbf{n}$ satisfying $\Delta_{n,-i} < C_k h_n$. Assume that $\Delta_{n,-i} < C_k h_n$, and condition on $\hat{\boldsymbol{\theta}}_{\mathbf{n},-\mathbf{i}}$, and $\boldsymbol{\theta}_\mathbf{n}$. Then by both local independence (A.1) and the representation given in (16)

$$\hat{P}_{n,i}(\theta_0) = \frac{1}{D_n(\theta)} \sum_{j \in A_{n,i,\theta_0}} K\left(\frac{\theta_0 - \hat{\theta}_{n,-i,j}}{h_n}\right) Y_{n,i,j}$$

is just a weighted sum of independent and bounded random variables. Thus, in order to show that $\hat{P}_{n,i}(\theta_0)$ is not likely to be too far from $P_{n,i}(\theta_0)$, we need to exploit the fact that $\Delta_{n,-i} < C_k h_n$ to show that $E[\hat{P}_{n,i}(\theta_0)]$ is close to $P_{n,i}(\theta_0)$, and use Theorem 6 to show that with high probability, $E[\hat{P}_{n,i}(\theta_0)]$ and $\hat{P}_{n,i}(\theta_0)$ are very near one another. Applying the triangle inequality in the event, it is seen that

$$P[|\hat{P}_{n,i}(\theta_0) - P_{n,i}(\theta_0)| > 3M_{c,d}C_k h_n | \Delta_{n,-i} < C_k h_n, \hat{\boldsymbol{\theta}}_{\mathbf{n},-\mathbf{i}}, \boldsymbol{\theta}_{\mathbf{n}}] \leq P[|\hat{P}_{n,i}(\theta_0) - E(\hat{P}_{n,i}(\theta_0))| + |E(\hat{P}_{n,i}(\theta_0)) - P_{n,i}(\theta_0)| > 3M_{c,d}C_k h_n | \Delta_{n,-i} < C_k h_n, \hat{\boldsymbol{\theta}}_{\mathbf{n},-\mathbf{i}}, \boldsymbol{\theta}_{\mathbf{n}}]. \quad (22)$$

Consider $|E(\hat{P}_{n,i}(\theta_0)) - P_{n,i}(\theta_0)|$, where the expectation is conditional on $\hat{\boldsymbol{\theta}}_{\mathbf{n},-\mathbf{i}}$, and $\boldsymbol{\theta}_{\mathbf{n}}$.

$$E(\hat{P}_{n,i}(\theta_0)) = \frac{1}{D_n(\theta_0)} \sum_{j \in A_{n,i,\theta_0}} K\left(\frac{\theta_0 - \hat{\theta}_{n,-i,j}}{h_n}\right) E[Y_{n,i,j} | \theta_{n,j}]$$

$E(\hat{P}_{n,i}(\theta_0))$ is just a weighted average of the $P_{n,i}(\theta_{n,j})$ for i such that $\hat{\theta}_{n,-i,j}$ satisfies $|\theta_0 - \hat{\theta}_{n,-i,j}| \leq C_k h_n$. By using the triangle inequality we see that for $j \in A_{n,i,\theta_0}$

$$|\theta_0 - \theta_{n,j}| \leq |\theta_0 - \hat{\theta}_{n,-i,j}| + |\hat{\theta}_{n,-i,j} - \theta_{n,j}| \leq C_k h_n + \Delta_{n,-i} \leq 2C_k h_n.$$

Choose n large enough so that

$$c < a - 2C_k h_n < b + 2C_k h_n < d.$$

Then $|\theta_0 - \theta_{n,j}| \leq 2C_k h_n$ implies that for $j \in A_{n,i,\theta_0}$, $\theta_{n,j}$ is in $[c, d]$. Consequently, by assumption A7,

$$|P_{n,i}(\theta_0) - P_{n,i}(\theta_{n,j})| \leq M_{c,d}|\theta_0 - \theta_{n,j}| \leq 2M_{c,d}C_k h_n.$$

This implies that $|E(\hat{P}_{n,i}(\theta_0)) - P_{n,i}(\theta_0))| \leq 2M_{c,d}C_k h_n$. Thus, from (22) it is seen that for big enough n

$$P[|\hat{P}_{n,i}(\theta_0) - P_{n,i}(\theta_0)| > 3M_{c,d}C_k h_n | \Delta_{n,-i} < C_k h_n, \hat{\boldsymbol{\theta}}_{\mathbf{n},-\mathbf{i}}, \boldsymbol{\theta}_{\mathbf{n}}] \leq P[|\hat{P}_{n,i}(\theta_0) - E(\hat{P}_{n,i}(\theta_0))| > M_{c,d}C_k h_n | \Delta_{n,-i} < C_k h_n, \hat{\boldsymbol{\theta}}_{\mathbf{n},-\mathbf{i}}, \boldsymbol{\theta}_{\mathbf{n}}]. \quad (23)$$

The problem has been reduced to obtaining a probability inequality for the difference between a sum of independent bounded random variables and its expected value, suggesting an application of Theorem 6. An alternative expression for $\hat{P}_{n,i}(\theta_0)$ is given by

$$\hat{P}_{n,i}(\theta_0) = \frac{1}{\#A_{n,i,\theta_0}} \sum_{j \in A_{n,i,\theta_0}} w_j Y_{n,i,j}$$

where $w_j = \#A_{n,i,\theta_0} K(\theta_0 - \hat{\theta}_{n,-i,j}/h_n)/D_n(\theta_0)$. By (15), (17), and condition C1 we see that for large enough n, $w_j \leq 3M_K C_K/m_l$, which is an upper bound for the random variables $w_j Y_{n,i,j}$.

Thus, by applying Theorem 6 for sums of bounded independent random variables, and (17) to bound the number of summands $\#A_{n,i,\theta_0}$, it is seen that

$$P[|\hat{P}_{n,i}(\theta_0) - E(\hat{P}_{n,i}(\theta_0))| > M_{c,d}C_k h_n | \Delta_{n,-i} < C_k h_n, \hat{\boldsymbol{\theta}}_{\mathbf{n},-\mathbf{i}}, \boldsymbol{\theta}_{\mathbf{n}}] \leq 2 \exp\left[\frac{-m_l^2 M_{c,d}^2 J_n h_n^3}{9M_k^2}\right],$$

which is less than or equal to

$$\leq 2 \exp\left[\frac{-M_L m_l^2 M_{c,d}^2 L_\alpha^{3/2} n^{3/2-3\alpha}}{9M_k^2}\right]. \tag{24}$$

By absorbing $M_L m_l^2 M_{c,d}^2 L_\alpha^{3/2}/9M_k^2$ into a constant D_α that depends only on α, conditions C1 through C4, and the endpoints c and d, (24) can be expressed as

$$2 \exp[-D_\alpha n^{3/2-3\alpha}]. \tag{25}$$

Thus, combining (25) with (21) we see that for large enough n and all $\theta_0 \in (a, b)$

$$P[|\hat{P}_{n,i}(\theta_0) - P_{n,i}(\theta_0)| > 3M_{c,d}C_k h_n] \leq 2 \exp[-D_\alpha n^{3/2-3\alpha}] + M_U n^r C_1 \exp[-C_2 n^{1-2\alpha}].$$

This completes the proof of Lemma 5. □

Proof of Theorem 3. Choose (a, b) and (c, d) so that $0 < c < a < \theta_0 < b < d < 1$.

$$MSE(\hat{P}_{n,i}(\theta_0)) = E[(\hat{P}_{n,i}(\theta_0) - P_{n,i}(\theta_0))^2] = \int_0^1 P[(\hat{P}_{n,i}(\theta_0) - P_{n,i}(\theta_0))^2 > t]\, dt$$

$$\leq (3M_{c,d}C_K h_n)^2 + P[(\hat{P}_{n,i}(\theta_0) - P_{n,i}(\theta_0))^2 > (3M_{c,d}C_K h_n)^2].$$

Thus,

$$MSE(\hat{P}_{n,i}(\theta_0)) \leq (3M_{c,d}C_K h_n)^2 + P[|\hat{P}_{n,i}(\theta_0) - P_{n,i}(\theta_0)| > 3M_{c,d}C_K h_n]. \tag{26}$$

By applying Lemma 5 to the second term in the right side of (26), it is seen that for large enough n and all item indices i,

$$MSE(\hat{P}_{n,i}(\theta_0)) \leq (3M_{c,d}C_K h_n)^2 + 2 \exp[-D_\alpha n^{3/2-3\alpha}] + M_U n^r C_1 \exp[-C_2 n^{1-2\alpha}]$$

$$\leq O(h_n^2) + O(\exp[-D_\alpha n^{3/2-3\alpha}]) + O(n^r \exp[-C_2 n^{1-2\alpha}]). \tag{27}$$

By C4 and the fact that $0 < \alpha < 1/2$, the right side of (27) is equal to $O(n^{-2\alpha})$, which completes the outline of the proof. □

Proof of Theorem 4. Let $\varepsilon > 0$. If it can be shown that

$$\sum_{n=1}^{\infty} P[\delta_{n,a,b} > \varepsilon] < \infty, \tag{28}$$

it will follow from the Borel-Cantelli lemma that $\delta_{n,a,b}$ converges to 0 with probability equal to 1.

Choose a sequence of sets of distinct points $\{\theta_1^n < \theta_2^n < \cdots < \theta_{2J_n^2}^n$ so that $a = \theta_1^n < \theta_2^n < \cdots < \theta_{2J_n^2}^n = b$, and the distance between any two neighbors, θ_i^n and θ_{i+1}^n, is less than or equal to $1/J_n^2$. Let θ be an arbitrarily chosen point in (a, b). Then out of the sequence above, for each n there is a k such that θ_k^n is no further than $1/J_n^2$ from θ.

Condition C3 and C4 ensure that for $\theta \in (a, b)$, and θ_k^n within $1/J_n^2$ of θ, there is a constant $C_{a,b}$ such that

$$|\hat{P}_{n,i}(\theta) - \hat{P}_{n,i}(\theta_k^n)| \leq \frac{C_{a,b}}{J_n}. \tag{29}$$

This means that regardless of the data, the smoothness of the kernel function and the size of the bandwidth impose some degree of smoothness on the ICC estimates. Using as-

sumption A3, it is clear that if numbers c and d are chosen so that $0 < c < a < b < d < 1$, then inequality (30) given below holds:

$$|P_{n,i}(\theta) - P_{n,i}(\theta_k^n)| \leq M_{c,d}|\theta - \theta_k^n| \leq M_{c,d}/J_n^2. \tag{30}$$

Next suppose that for all $k \in \{1, 2, \ldots, 2J_n^2\}$ and $i \in \{1, 2, \ldots, n\}$

$$|P_{n,i}(\theta_k^n) - \hat{P}_{n,i}(\theta_k^n)| \leq \frac{\varepsilon}{3}. \tag{31}$$

For all $\theta \in (a, b)$

$$|P_{n,i}(\theta) - \hat{P}_{n,i}(\theta)| \leq |P_{n,i}(\theta) - P_{n,i}(\theta_k^n)| + |P_{n,i}(\theta_k^n) - \hat{P}_{n,i}(\theta_k^n)| + |\hat{P}_{n,i}(\theta_k^n) - \hat{P}_{n,i}(\theta)|, \tag{32}$$

and by inequalities (29) and (30) for large enough n the first two terms of (32) are less than $\varepsilon/3$, and if (31) holds as well $|P_{n,i}(\theta) - \hat{P}_{n,i}(\theta)| \leq \varepsilon$. This would imply that $\delta_{n,a,b} \leq \varepsilon$.

Thus in order to show that (28) holds, it suffices to show that

$$\sum_{n=1}^{\infty} \sum_{i=1}^{n} \sum_{k=1}^{2J_n^2} P\left[|P_{n,i}(\theta_k^n) - \hat{P}_{n,i}(\theta_k^n)| > \frac{\varepsilon}{3}\right] < \infty, \tag{33}$$

because the inner double sum in (33) is greater than the summand in (28) for n large enough that the right side of (29) and (30) are less than $\varepsilon/3$.

Choose n_0 large enough so that for $n > n_0$, the right sides of (29) and (30) are less than $\varepsilon/3$, the hypothesis of Lemma 5 is satisfied, and $3M_{c,d}C_K h_n < \varepsilon/3$. Then the triple sum given in (33) is less than

$$2n_0^2 J_{n_0}^2 + \sum_{n>n_0}^{\infty} \sum_{i=1}^{n} \sum_{k=1}^{2J_n^2} P[|P_{n,i}(\theta_k^n) - \hat{P}_{n,i}(\theta_k^n)| > 3M_{c,d}C_K h_n]. \tag{34}$$

Lemma 5 can be applied to show that (34) is less than

$$2n_0^2 J_{n_0}^2 + \sum_{n>n_0}^{\infty} 2nJ_n^2(2 \exp[-D_\alpha n^{3/2-3\alpha}] + M_U n^r C_1 \exp[-C_2 n^{1-2\alpha}]). \tag{35}$$

By replacing J_n with its upper bound, $M_U n^r$ (A5), (35) can be bounded above by

$$2n_0^2 J_{n_0}^2 + \sum_{n>n_0}^{\infty} 2M_U n^{2r+1}(2 \exp[-D_\alpha n^{3/2-3\alpha}] + M_U n^r C_1 \exp[-C_2 n^{1-2\alpha}]). \tag{36}$$

Because $0 < \alpha < 1/2$ the sum in (36) is finite, completing the outline of the proof. □

References

Billingsley, P. (1986). *Probability and measure*. New York: Wiley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443–459.

Devroye, L. P. (1978). The uniform convergence of the Nadaraya-Watson regression function Estimate. *Canadian Journal of Statistics, 6*, 179–191.

Fan, J., & Truong, Y. K. (1993). Nonparametric regression with errors in variables. *The Annals of Statistics, 21*, 1900–1925.

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *Annals of Statistics, 5*, 815–841.

Härdle, W. (1990). *Applied nonparametric regression*. Cambridge: Cambridge University Press.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. London: Chapman Hall.

Hoeffding, W. (1986). Probability inequalities for sums of bounded random variables. *Encyclopedia of Statistical Sciences, 7*, 222–225.

Nadaraya, E. A. (1964). On estimating regression. *Probability Theory and Its Applications, 9*, 141–142.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrika, 16*, 1–32.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611–630.

Ramsay, J. O. (1994). *TESTGRAF: A program for the graphical analysis of multiple choice test and questionnaire data*. Unpublished user's guide to TESTGRAF.

Ramsay, J. O., & Winsberg, S. (1991). Maximum marginal likelihood estimation for semiparametric item analysis. *Psychometrika, 56*, 365–379.

Samejima, F. (1979). *A new family of models for the multiple-choice item* (Research report No. 79-4). Knoxville, TN: University of Tennessee, Department of Psychology.

Samejima, F. (1981). *Efficient methods of estimating the operating characteristic of item response categories and challenge to a new model for the multiple-choice item*. Unpublished manuscript, Knoxville, TN: University of Tennessee, Department of Psychology.

Samejima, F. (1984). *Plausibility functions of the Iowa Vocabulary Test items estimated by the simple sum procedure of the conditional* P.D.F. *approach* (Research report No. 84-1). Knoxville, TN: University of Tennessee, Department of Psychology.

Samejima, F. (1988). *Advancement of latent trait theory* (Research Report No. 79-4). Knoxville, TN: University of Tennessee, Department of Psychology.

Samejima, F. (1990). *Differential weight procedure of the conditional* p.d.f. *approach for estimating the operating characteristic of discrete item responses* (ONR research report No. 90-4). Knoxville, TN: University of Tennessee, Department of Psychology.

Schuster, E. F. (1972). Joint asymptotic distribution of the estimated regression function at a finite number of distinct points. *Annals of Mathematical Statistics, 43*, 84–88.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. New York: John Wiley.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya, Series A, 26*, 359–372.

Wingersky, M. S., Patrick, R., & Lord, F. M. (1988). *LOGIST User's Guide*. Princeton, New Jersey: Educational Testing Service.

Manuscript received 12/9/94
Final version received 7/20/95