Does Adaptive Testing Violate Local Independence?

CSE Technical Report 476

Robert J. Mislevy and Hua-Hua Chang
Educational Testing Service

April 1998

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Publication of this report was supported under the Educational Research and Development Center Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

DOES ADAPTIVE TESTING VIOLATE LOCAL INDEPENDENCE?*

Robert J. Mislevy and Hua-Hua Chang
Educational Testing Service

Abstract

Item response theory posits "local independence," or conditional independence of item responses given item parameters and examinee proficiency parameters. The usual definition of local independence, however, addresses the context of fixed tests and appears to yield incorrect response-pattern probabilities in the context of adaptive testing. The paradox is resolved by introducing additional notation to deal with the item selection mechanism. Implications for estimation of examinee proficiency are noted.

Key words: Adaptive testing; conditional independence; item response theory; local independence

A Question

The cornerstone of item response theory (IRT) is the assumption of *local independence* (LI), which posits that an examinee's response to a given test item depends on an unobservable examinee parameter θ but not on the identity of or responses to other test items the examinee may have been presented (Lord, 1980, p. 19). More formally, responses to test items are conditionally independent, given item parameters and θ; or, equivalently, the joint distribution of item responses is equal to the product of the marginal distributions (Lord & Novick, 1968, p. 361). An IRT model satisfies LI in a domain of n dichotomous items if

* We are grateful to Charlie Lewis, Ming-Mei Wang, and Pao-Kuei Wu for discussions on this topic. The first author's work was supported in part by the National Center for Research on Evaluation, Standards, Student Testing (CRESST), Educational Research and Development Program, cooperative agreement number R117G10027 and CFDA catalog number 84.117G, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed in this report do not reflect the policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

$$\text{Prob}\left(U_1 = u_1,\ldots,U_n = u_n|\theta,\beta_1,\ldots,\beta_n\right) = \prod_{j=1}^{n}\text{Prob}\left(U_j = u_j|\theta,\beta_j\right), \quad (1)$$

where U_j is the response variable for Item *j*; u_j represents a value thereof, either 1 if correct or 0 if incorrect; and β_j is a possibly vector-valued parameter characterizing the dependence of response probabilities to Item *j* on θ. Assuming the β_js are known, denote the item response function $\text{Prob}\left(U_j = u_j|\theta,\beta_j\right)$ by $f_j(u;\theta)$.

As an example, consider a test consisting of two items, Item *a* and Item *b*. Examinees' responses follow the Rasch model, or

$$f_j(u;\theta) = \exp\left[u\left(\theta-\beta_j\right)\right]\Big/\left[1+\exp\left(\theta-\beta_j\right)\right],$$

with $\beta_a = 0$ and $\beta_b = 1$. Under LI,

$$\text{Prob}\left(U_a = 1, U_b = 0|\theta = 1\right) = f_a(1;\theta)f_b(0;\theta) = .731\times.500 = .3655. \quad (2)$$

Similarly,

$$\text{Prob}\left(U_a = 0, U_b = 0|\theta = 1\right) = f_a(0;\theta)f_b(0;\theta) = .269\times.500 = .1345,$$

$$\text{Prob}\left(U_a = 0, U_b = 1|\theta = 1\right) = f_a(0;\theta)f_b(1;\theta) = .269\times.500 = .1345,$$

and

$$\text{Prob}\left(U_a = 1, U_b = 1|\theta = 1\right) = f_a(1;\theta)f_b(1;\theta) = .731\times.500 = .3655.$$

Equation 1, the usual definition of LI, does not address the order or the mechanism by which items come to be administered to the examinee. It is typically used with fixed test forms, in which the identity and order of items is predetermined. As in the example, (1) specifies the probabilities of observing the 2^n possible response patterns given a particular value of θ. As such it serves as the basis for both Bayesian inference about θ and maximum likelihood estimation of θ with reference to repeated samples of $\left(U_1,\ldots,U_n\right)$ for fixed values of θ (e.g., Section 20.3 of Birnbaum, 1968).

Computerized adaptive testing (CAT) is a more recent development in IRT (Wainer et al., 1990). Items are selected sequentially in CAT in light of an examinee's previous responses, in order to provide more efficient estimation of

θ. Under the Rasch model, for example, an examinee answering items correctly would be administered successively more difficult items, while an examinee answering incorrectly would be administered successively easier items. Let us add to our example Item c, with β_c=-1, and define the following simple adaptive testing scheme. Two items are presented to an examinee, with the identity of the second item dependent on the response to the first:

1. Administer Item a and observe a value for U_a.

2a. If u_a=1, then with probability .75 administer Item b and observe a value for U_b, or with probability .25 administer Item c and observe a value for U_c; or else,

2b. if u_a=0, then with probability .75 administer Item c and observe a value for U_c, or with probability .25 administer Item b and observe a value for U_b.

3. Stop testing.

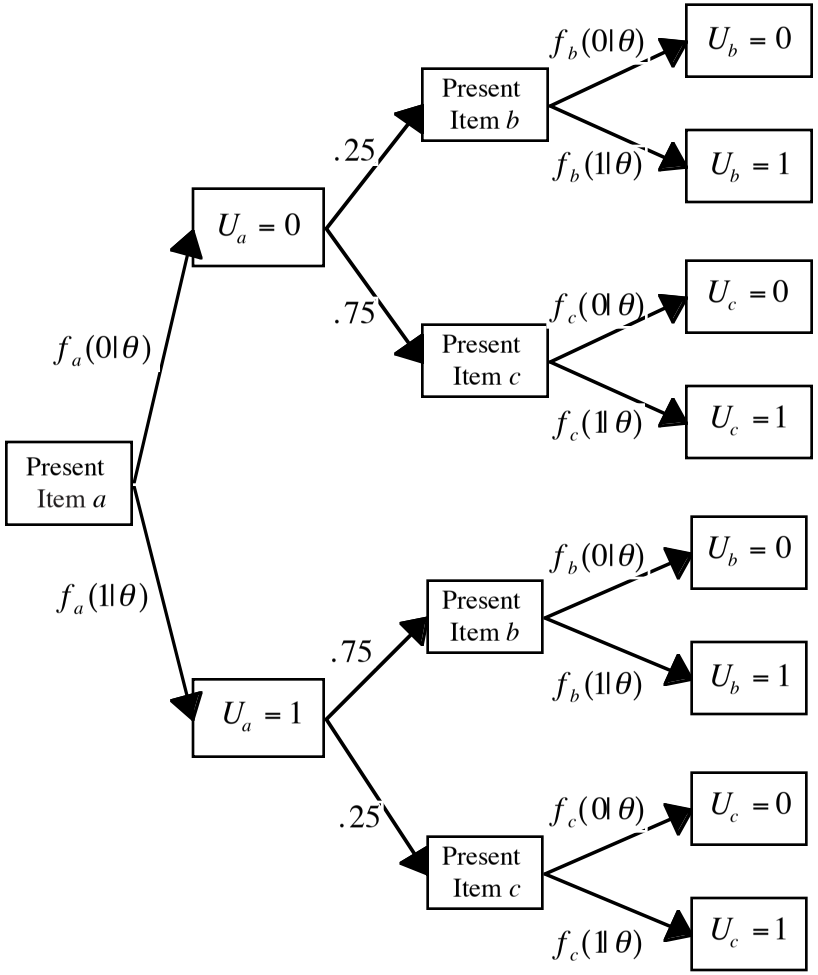
The tree in Figure 1 shows the eight possible response patterns that can occur and their respective probabilities when $\theta = 1$. We see in particular that

$$\text{Prob}\left(U_a = 1, U_b = 0 | \theta = 1\right) = .2741, \tag{3}$$

which contradicts the value of .3655 calculated as $f_a(1;\theta) f_b(0;\theta)$ using the definitional equation of LI with $\theta = 1$. The probability of observing the response pattern, it would appear, is *not* equal to the product of the marginal probabilities of the individual responses. Is this not a failure of local independence?

Notation for CAT

The trouble is ambiguous notation: The expression "$\text{Prob}\left(U_a = 1, U_b = 0 | \theta = 1\right)$" refers to different events in (2) and (3). In (2), only the values of responses to specified items in a specified order need be addressed. In (3), the process by which item identities and orders are determined is also at issue. We must extend the notation to first distinguish, then model, these two situations. We adapt the route taken by Mislevy and Wu (1996). We assume that, in Fred Lord's words, "the probability of success on an item depends on ... item parameters, on examinee ability θ, and nothing else" (Lord, 1980, p. 19).



Item Responses	Path probability when θ=1
$U_a = 0,\ U_b = 0,\ U_c = *$	.0336
$U_a = 0,\ U_b = 1,\ U_c = *$	.0336
$U_a = 0,\ U_b = *,\ U_c = 0$	.0240
$U_a = 0,\ U_b = *,\ U_c = 1$	.1777
$U_a = 1,\ U_b = 0,\ U_c = *$	.2741
$U_a = 1,\ U_b = 1,\ U_c = *$	.2741
$U_a = 1,\ U_b = *,\ U_c = 0$	.0217
$U_a = 1,\ U_b = *,\ U_c = 1$	.1610

* indicates response not observed, because item was not administered

Figure 1. Probability tree for adaptive testing example.

The datum observed in adaptive testing is actually a sequence of $N \leq n$ ordered pairs, $S = \left(\left(I_1, U_{I_1}\right), \ldots, \left(I_N, U_{I_N}\right)\right)$, where I_k identifies the k^{th} item administered and U_{i_k} is the response to that item. Define the partial response sequence S_k as the first k ordered pairs in S, with the null sequence s_0 representing the status as the test begins. Testing continues until, say, a desired level of precision is reached, or a predetermined number of items has been administered. We may augment the collection of items with the fictitious Item 0, the selection of which corresponds to a decision to terminate testing. It can be written as the N+1st item in the adaptive test, but no response is associated with it.

A test administrator defines an adaptive test design by specifying, for all items j and all realizable partial response sequences s_k, the probabilities $\phi(j, s_k)$ that Item j will be selected as the k+1st test item, after the partial response sequence s_k has been observed from an examinee. Under Bayesian minimum variance item selection, for example, the as-yet-unadministered item that minimizes the expected posterior variance of θ with respect to the current distribution $p(\theta|s_k)$ is chosen as the k+1st item with probability one (Owen, 1975). For the two tests in our running example, the item selection probabilities are as given in Figure 2.

The probability of S for an examinee with ability θ can be constructed sequentially. The probability of selection for the first item is $\phi(i_1, s_0)$. The probability of response u_{i_1} to Item i_1 is given by the IRT model as $f_{i_1}\left(u_{i_1}; \theta\right)$, which does not depend on the fact that Item i_1 happened to have been presented first. The probability of selection for the second item *given* s_1 is $\phi(i_2, s_1)$, which depends on the value of u_{i_1} but not on θ given u_{i_1}. The probability of the corresponding response is $f_{i_2}\left(u_{i_2}; \theta\right)$, independent of the identification of, and the response to, the first item. Continuing in this manner through the decision to stop testing (i.e., the selection of Item 0 as the N+1st item) and grouping similar terms yields

$$\phi_{\text{fixed}}(j, s_k) = \begin{cases} 1 \text{ if } & j = a \text{ and } k = 0 \\ 1 \text{ if } & j = b \text{ and } k = 1 \\ 1 \text{ if } & j = 0 \text{ and } k = 2 \\ 0 & \text{otherwise} \end{cases}$$

Fixed Test

Note: The interpretation of this item selection function is as follows: Item *a* is selected with probability 1 as the first item to administer, Item *b* is selected with probability 1 as the second item to administer, and Item 0, or the determination to stop testing, is selected with probability 1 as "the third item."

$$\phi_{\text{CAT}}(j, s_k) = \begin{cases} 1 & \text{if } \quad j = a \text{ and } k = 0 \\ .25 & \text{if } \quad j = b \text{ and } s_k = ((a,0)) \\ .75 & \text{if } \quad j = c \text{ and } s_k = ((a,0)) \\ .75 & \text{if } \quad j = b \text{ and } s_k = ((a,1)) \\ .25 & \text{if } \quad j = c \text{ and } s_k = ((a,1)) \\ 1 & \text{if } \quad j = 0 \text{ and } k = 2 \\ 0 & \text{otherwise.} \end{cases}$$

Adaptive Test

Figure 2. Item selection functions for fixed and adaptive test examples.

$$\text{Prob}\left[S=\left(\left(i_1,u_{i_1}\right),\ldots,\left(i_N,u_{i_N}\right)\right)\middle|\theta\right]=\prod_{k=1}^{N}f_{i_k}\left(u_{i_k};\theta\right)\prod_{k=1}^{N+1}\phi(i_k,s_{k-1}). \quad (4)$$

Rather than the ambiguous expression "$\text{Prob}\left(U_a=1,U_b=0\middle|\theta=1\right)$," we can now write for our fixed test

$$\begin{aligned}
&\text{Prob}\left[S=((a,1),(b,0))\middle|\theta,\phi_{\text{fixed}}\right]\\
&\quad=\prod_{k=1}^{N}f_{i_k}\left(u_{i_k};\theta\right)\prod_{k=1}^{N+1}\phi_{fixed}\left(i_k,s_{k-1}\right)\\
&\quad=f_a(1;\theta)\,f_b(0;\theta)\text{Prob}\left(I_1=a\right)\text{Prob}\left(I_2=b\middle|s_1=(a,1)\right)\text{Prob}\left(I_3=0\middle|s_2=((a,1),(b,0))\right)\\
&\quad=.731\times.500\times1\times1\times1\\
&\quad=.3655,
\end{aligned}$$

which agrees with (2). For our CAT, we can write

$$\begin{aligned}
&\text{Prob}\left[S=((a,1),(b,0))\middle|\theta,\phi_{\text{CAT}}\right]\\
&\quad=\prod_{k=1}^{N}f_{i_k}\left(u_{i_k};\theta\right)\prod_{k=1}^{N+1}\phi_{CAT}\left(i_k,s_{k-1}\right)\\
&\quad=f_a(1;\theta)\,f_b(0;\theta)\text{Prob}\left(I_1=a\right)\text{Prob}\left(I_2=b\middle|s_1=(a,1)\right)\text{Prob}\left(I_3=0\middle|s_2=((a,1),(b,0))\right)\\
&\quad=.731\times.500\times1\times.75\times1\\
&\quad=.2741,
\end{aligned}$$

which agrees with (3).

We note in passing some implications for estimation. Equation 4, the probability for a sequence of responses in CAT, factors into two terms. Only the first depends on θ, and it is just the product of marginal item-response probabilities that appears in (1). Inferences about θ that accord with the Likelihood Principle, therefore, need only address the first term. This includes Bayesian and direct likelihood inference about θ—but not sampling interpretations of the MLE $\hat{\theta}$. The correct point estimate is identified but no claims about its distribution in repeated samples for fixed θ necessarily follow (Mislevy & Wu, 1988). The correct sampling distribution for $\hat{\theta}$ must be verified with respect to repeated administrations of the entire adaptive test. Chang and Ying (in press) consider the sampling variance of $\hat{\theta}$ with respect to the second order derivative of $\prod f_j\left(u_j;\theta\right)$, and offer some large-sample conditions under

which the latter is a reasonable large-sample approximation of the former in CAT.

Conclusion

So, does CAT violate local independence? Since the standard notation and terminology for defining "local independence" is not rich enough to describe CAT, one must choose how to apply the term in the extension. We can write expressions with the previously missing item-selection random variables on either side of the conditioning bar, and see the answers they suggest. Again we assume that item responses depend on item parameters and θ only.

In CAT, we could, on the one hand, take $\text{Prob}\left(U_a = u_a,\ldots,U_q = u_q|\theta\right)$ to mean the probability of observing the CAT response vector with the implied identity of items and the indicated responses; that is, $\text{Prob}\left(\left\{U_a = u_a,\ldots,U_q = u_q\right\} \text{ and } \left\{I_1 = a,\ldots,I_N = q\right\}|\theta\right)$. But from (4),

$$\begin{aligned}
&\text{Prob}\left(\left\{U_a = u_a,\ldots,U_q = u_q\right\} \text{ and } \left\{I_1 = a,\ldots,I_N = q\right\}|\theta\right) \\
&\qquad = \text{Prob}\left[S = \left(\left(i_1,u_{i_1}\right),\ldots,\left(i_N,u_{i_N}\right)\right)|\theta\right] \\
&\qquad = \prod_{k=1}^{N} f_{i_k}\left(u_{i_k};\theta\right)\prod_{k=1}^{N+1}\phi\left(i_k,s_{k-1}\right) \\
&\qquad \neq \prod_{k=1}^{N} f_{i_k}\left(u_{i_k};\theta\right).
\end{aligned}$$

The answer is "yes, local independence is violated by CAT"—*if* "local independence" is taken to mean that the product of the item-by-item probabilities conditional on θ yields the probability of observing a response vector with those items and those responses, given θ.[1] To those interested in frequentist inference, such as sampling interpretations of the MLE, the other terms in the probability of observing the response vector can materially affect the distribution of these estimators.

[1] In this sense, LI is also violated by intentional omissions and examinee choice of items, even when student responses to items they have responded to depends only on θ (Mislevy & Wu, 1996). Unlike CAT, the missing responses in these cases can depend on θ even after conditioning on the identity of and responses to items for which responses are observed. The missingness process cannot be ignored even under Bayesian and direct likelihood estimation.

On the other hand, we could take $\text{Prob}\left(U_a = u_a,\ldots,U_q = u_q|\theta\right)$ to mean the probability of observing the CAT response vector with the indicated responses *given* the implied identity of items; that is, $\text{Prob}\left(\left\{U_a = u_a,\ldots,U_q = u_q\right\}|\theta,\left\{I_1 = a,\ldots,I_N = q\right\}\right)$. And

$$\begin{aligned}
&\text{Prob}\left(\left\{U_a = u_a,\ldots,U_q = u_q\right\}|\theta,\left\{I_1 = a,\ldots,I_N = q\right\}\right)\\
&\qquad = \prod_{k=1}^{N}\text{Prob}\left(U_{i_k} = u_{i_k}\middle|\theta,\left\{U_a = u_a,\ldots,U_{i_{k-1}} = u_{i_{k-1}}\right\},\left\{I_1 = a,\ldots,I_N = q\right\}\right)\\
&\qquad = \prod_{k=1}^{N}\text{Prob}\left(U_{i_k} = u_{i_k}|\theta,I_k = i_k\right)\\
&\qquad = \prod_{k=1}^{N} f_{i_k}\left(u_{i_k};\theta\right)\\
&\qquad = \text{Prob}\left(U_a = u_a|\theta\right)\times\ldots\times\text{Prob}\left(U_q = u_q|\theta\right).
\end{aligned}$$

The answer is "no, local independence is *not* violated by CAT"—if "local independence" is taken to mean that the product of the item-by-item probabilities conditional on θ yields the only term in the probability of observing a CAT response vector that depends on θ. And to those interested in inference that accords with the Likelihood Principle, this is all that matters.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Chang, H., & Ying, Z. (in press). Nonlinear sequential designs for logistic item response theory models, with applications to computerized adaptive tests. *Annals of Statistics.*

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Mislevy, R. J., & Wu, P-K. (1988). *Inferring examinee ability when some item responses are missing.* Princeton, NJ: Educational Testing Service. [ERIC Document No. ED 395 017]

Mislevy, R. J., & Wu, P-K. (1996). *Missing responses and Bayesian IRT ability estimation: Omits, choice, time limits, and adaptive testing*. Research Report RR-96-30-ONR. Princeton, NJ: Educational Testing Service.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*, 351-356.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum Associates.