8524078

**Warm, Thomas Albert**

WEIGHTED LIKELIHOOD ESTIMATION OF ABILITY IN ITEM RESPONSE THEORY WITH TESTS OF FINITE LENGTH

*The University of Oklahoma* PH.D. 1985

University
Microfilms
International 300 N. Zeeb Road, Ann Arbor, MI 48106

THE UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

# WEIGHTED LIKELIHOOD ESTIMATION
# OF ABILITY IN ITEM RESPONSE THEORY
# WITH TESTS OF FINITE LENGTH

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

DOCTOR OF PHILOSOPHY

By

THOMAS ALBERT WARM

Norman, Oklahoma

1985

# WEIGHTED LIKELIHOOD ESTIMATION OF ABILITY IN ITEM RESPONSE THEORY WITH TESTS OF FINITE LENGTH

A DISSERTATION

APPROVED FOR THE DEPARTMENT OF PSYCHOLOGY

By

Morris L Marx

Joseph Lee Rodgers III

Larry E. Toothaker

A.R. Schwarzkopf

# TABLE OF CONTENTS

Manuscript to be submitted for publication

Page

ABSTRACT . . . . . . . . . . . . . . . . . iv

LIST OF TABLES . . . . . . . . . . . . . . v

LIST OF FIGURES . . . . . . . . . . . . . vi

INTRODUCTION . . . . . . . . . . . 1

TWO EXPERIMENTS TO EVALUATE THE WEIGHTED LIKELIHOOD ESTIMATOR . 10

METHOD . . . . . . . . . . . . . 15

RESULTS . . . . . . . . . . . . . . 20

DISCUSSION . . . . . . . . . . . . 44

SUMMARY AND CONCLUSIONS . . . . . 53

REFERENCES . . . . . . . . . . . . . . . . 55

APPENDIX A ... PROOF OF THEOREM . . . . 62

APPENDIX B ... COMPUTER PROGRAM FOR CONVENTIONAL TESTS. . . . . . . . 76

APPENDIX C ... COMPUTER PROGRAM FOR TAILORED TESTS. . . . . . . . . . 82

APPENDIX D ... ADDITIONAL FIGURES OF RESULTS . . . . . . . . . . . . . . 93

# WEIGHTED LIKELIHOOD ESTIMATION OF ABILITY IN ITEM RESPONSE THEORY WITH TESTS OF FINITE LENGTH

## Abstract

Applications of Item Response Theory, which depend upon its parameter invariance property, require that parameter estimates be unbiased. All current estimation methods produce statistically biased estimates of both item and ability parameters. A new method, Weighted Likelihood Estimation (WLE), is derived, and proved to be less biased than Maximum Likelihood Estimation (MLE) with the same asymptotic variance and normal distribution. WLE removes the first order bias term from MLE. Two Monte Carlo studies compare WLE with MLE and Bayesian Modal Estimation (BME) of ability in conventional tests and tailored tests. The Monte Carlo studies favor WLE over MLE and BME on several criteria over a wide range of the ability scale.

Keywords: Maximum Likelihood Estimation, unbiased estimation, statistical bias, Bayesian Modal Estimation, Item Response Theory, tailored testing, adaptive testing.

# LIST OF TABLES

TABLE Page

1. Minimum, MLE($\theta^-$), and Maximum, MLE($\theta^+$), of "Sensible" MLE($\theta$), and of WLE($\theta$) and BME($\theta$), for Conventional Tests with n items, All a = 1 or 2, Normally Distributed b, and All c = 0.20 . - - -- - - - - 50

## LIST OF FIGURES

FIGURE Page

1. Test Information Curves of the 10 and 60 Item Conventional Tests with a = 1 and a = 2. . . . . . . . 21

2. Average Estimation Error of θ^ on Conventional Tests with 10 Items, All a = 1, Normally Distributed b, and All c = 0.20 . . . . . . . . . 22

3. Absolute Average Estimation Error of WLE(θ) with 10 Items, and of MLE(θ) with 10, 30, and 60 Items; a = 1, Normally Distributed b, and All c = 0.20 . . . . . . . . . . . 25

4. Absolute Average Estimation Error of WLE(θ) with 10 Items, and of MLE(θ) with 10, 30, and 60 Items; All a = 2, Normally Distributed b, and All c = 0.20 . . . . . . . . . 26

5. Absolute Average Estimation Error of WLE(θ) with 10 Items, and of BME(θ) with 10, 30, and 60 Items; All a = 1, Normally Distributed b, and All c = 0.20 . . . . . . . . . 27

6. Absolute Average Estimation Error of WLE(θ) with 10 Items, and of BME(θ) with 10, 30, and 60 Items; All a = 2, Normally Distributed b, and All c = 0.20 . . . . . . . . . 28

7. Standard Deviation of θ^ on Conventional Tests with 10 Items, All a = 1, Normally Distributed b, and All c = 0.20 . . . . . . . . . 29

8. Mean Squared Error of θ^ on Conventional Tests with 10 Items, All a = 1, Normally Distributed b, and All c = 0.20 . . . . . . . . . 30

9. Average Estimation Error of T^ on Conventional Tests with 10 Items, All a = 1, Normally Distributed b,

and All c = 0.20 . . . . . . . . 32

10. Standard Deviation of T^ on Conventional Tests with 10 Items, All a = 1, Normally Distributed b, and All c = 0.20 . . . . . . . . 33

11. Mean Squared Error of T^ on Conventional Tests with 10 Items, All a = 1, Normally Distributed b, and All c = 0.20 . . . . . . . . 34

12. Average Estimation Error of θ^ on Tailored Test with Declining a-parameter, Optimal b-parameter, and All c = 0.20 . . . . . . . . 36

13. Standard Deviation of θ^ on Tailored Test with Declining a-parameter, Optimal b-parameter, and All c = 0.20 . . . . . . . . . . 37

14. Standard Deviation of θ^ on Tailored Test with a = 2, Optimal b-parameter, and All c = 0.20 . 38

15. Mean Squared Error of θ^ on Tailored Test with Declining a-parameter, Optimal b-parameter, and All c = 0.20 . . . . . . . . . . 39

16. Mean Squared Error of θ^ on Tailored Test with a = 2, Optimal b-parameter, and All c = 0.20 . . 40

17. Average Number of Items Administered on Tailored Test with Declining a-parameter, Optimal b-parameter, and All c = 0.20 . . 41

18. Average Computation Time Between Items on Tailored Test with Declining a-parameter, Optimal b-parameter, and All c = 0.20 . 42

D.1. Average Estimation Error of θ^ on Conventional Tests with 20 Items, All a = 1, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 94

D.2. Average Estimation Error of θ^ on Conventional Tests with 30 Items, All a = 1, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 95

D.3. Average Estimation Error of θ^ on Conventional Tests with 40 Items, All a = 1, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 96

D.4. Average Estimation Error of θ^ on Conventional Tests with 50 Items, All a = 1, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 97

D.5. Average Estimation Error of θ^ on Conventional Tests with 60 Items, All a = 1, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 98

D.6. Average Estimation Error of θ^ on Conventional Tests with 10 Items, All a = 2, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 99

D.7. Average Estimation Error of θ^ on Conventional Tests with 20 Items, All a = 2, Normally Distributed b,

and All c = 0.20 . . . . . . . . . . 100

D.8. Average Estimation Error of θ^ on Conventional Tests with 30 Items, All a = 2, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 101

D.9. Average Estimation Error of θ^ on Conventional Tests with 40 Items, All a = 2, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 102

D.10. Average Estimation Error of θ^ on Conventional Tests with 50 Items, All a = 2, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 103

D.11. Average Estimation Error of θ^ on Conventional Tests with 60 Items, All a = 2, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 104

D.12. Standard Deviation of θ^ on Conventional Tests with 20 Items, All a = 1, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 105

D.13. Standard Deviation of θ^ on
Conventional Tests with 30 Items,
All a = 1, Normally Distributed b,
and All c = 0.20 . . . . . . . . . . 106

D.14. Standard Deviation of θ^ on
Conventional Tests with 40 Items,
All a = 1, Normally Distributed b,
and All c = 0.20 . . . . . . . . . . 107

D.15. Standard Deviation of θ^ on
Conventional Tests with 50 Items,
All a = 1, Normally Distributed b,
and All c = 0.20 . . . . . . . . . . 108

D.16. Standard Deviation of θ^ on
Conventional Tests with 60 Items,
All a = 1, Normally Distributed b,
and All c = 0.20 . . . . . . . . . . 109

D.17. Standard Deviation of θ^ on
Conventional Tests with 10 Items,
All a = 2, Normally Distributed b,
and All c = 0.20 . . . . . . . . . . 110

D.18. Standard Deviation of θ^ on Conventional Tests with 20 Items, All a = 2, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 111

D.19. Standard Deviation of θ^ on Conventional Tests with 30 Items, All a = 2, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 112

D.20. Standard Deviation of θ^ on Conventional Tests with 40 Items, All a = 2, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 113

D.21. Standard Deviation of θ^ on Conventional Tests with 50 Items, All a = 2, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 114

D.22. Standard Deviation of θ^ on Conventional Tests with 60 Items, All a = 2, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 115

D.23. Mean Squared Error of θ^ on

Conventional Tests with 20 Items, All a = 1, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 116

D.24. Mean Squared Error of θ^ on Conventional Tests with 30 Items, All a = 1, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 117

D.25. Mean Squared Error of θ^ on Conventional Tests with 40 Items, All a = 1, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 118

D.26. Mean Squared Error of θ^ on Conventional Tests with 50 Items, All a = 1, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 119

D.27. Mean Squared Error of θ^ on Conventional Tests with 60 Items, All a = 1, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 120

D.28. Mean Squared Error of θ^ on Conventional Tests with 10 Items,

All a = 2, Normally Distributed b,
and All c = 0.20 . . . . . . . . . . 121

D.29. Mean Squared Error of θ^ on
Conventional Tests with 20 Items,
All a = 2, Normally Distributed b,
and All c = 0.20 . . . . . . . . . . 122

D.30. Mean Squared Error of θ^ on
Conventional Tests with 30 Items,
All a = 2, Normally Distributed b,
and All c = 0.20 . . . . . . . . . . 123

D.31. Mean Squared Error of θ^ on
Conventional Tests with 40 Items,
All a = 2, Normally Distributed b,
and All c = 0.20 . . . . . . . . . . 124

D.32. Mean Squared Error of θ^ on
Conventional Tests with 50 Items,
All a = 2, Normally Distributed b,
and All c = 0.20 . . . . . . . . . . 125

D.33. Mean Squared Error of θ^ on
Conventional Tests with 60 Items,
All a = 2, Normally Distributed b,

and All c = 0.20 . . . . . . . . . . 126

D.34. Average Estimation Error of T^ on Conventional Tests with 60 Items, All a = 2, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 127

D.35. Standard Deviation of T^ on Conventional Tests with 60 Items, All a = 2, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 128

D.36. Mean Squared Error of T^ on Conventional Tests with 60 Items, All a = 2, Normally Distributed b, and All c = 0.20 . . . . . . . . . . 129

D.37. Average Estimation Error of θ^ on Tailored Test with a = 2, Optimal b-parameter, and All c = 0.20 . . . 130

D.38. Average Number of Items Administered on Tailored Test with All a = 2, Optimal b-parameter, and All c = 0.20 . . . . . . . . . . . 131

D.39. Average Computation Time Between Items on Tailored Teat with All a = 2, Optimal b-parameter, and All c = 0.20 . . . . . . . . . . . . . . 132

# INTRODUCTION

Item Response Theory (IRT) is an elegant model of examinee behavior on multiple-choice tests in terms of item and person parameters that are invariant within a linear transformation. The mathematical form of IRT that is almost exclusively used is the three parameter logistic model, which gives the probability, P, that a scored item response, $u_i$, to item i is correct ($u_i = 1$) is a function of the ability parameter, θ, and three item parameters, $a_i$, $b_i$, and $c_i$.

$$P(u_i=1|\theta;a_i,b_i,c_i) = c_i+(1-c_i)/(1+\exp(-1.7a_i(\theta-b_i))) \quad (1)$$

The left hand side (LHS) of (1) is often abbreviated $P_i(\theta)$, $P(\theta)$, or just P, when the context excludes ambiguity.

The true parameters, of course, are never known, and must be estimated. The estimates are unidentifiable unless an origin and unit of the θ scale are given. Usually, the mean and standard deviation of the ability estimates, θ^, of some reference group of examinees is chosen for the origin and unit, respectively. The estimates of the parameters then are on their own specific scale, and may not be directly comparable with other estimates.

In principle all parameters are invariant within a linear transformation. Given the true parameters, scaled with respect to any two tests with at least two common items, the linear transformation that links scores and parameters of the two tests is easy to solve for. Since the true parameters are never known, any application of IRT which makes use of the parameter invariance property depends upon an assumption of parameter estimate invariance. Parameter estimates are not invariant because of estimation error. As a result the linear transformation for linking tests is also an estimate. The greater the error of the parameter estimates, the greater will be the error of the linear transformation. Hence, in practice, the invariance principle must be phrased in terms of the expectation of the transformation. In order to minimize the error of transformation, averages of parameter estimates are used in place of the parameters themselves when solving for the linking transformation.

Averages have reduced variance, a valuable property which should reduce the variability of the linking transformation. However, if the estimates are statistically biased, then the averages will also be biased, and so will the linking transformation. Among the many strengths of IRT, the invariance property distinguishes it most clearly from other approaches. Thus, unbiased estimation is critical to applications of IRT that make use of the invariance property, and is a fundamental requirement of IRT as a statistical theory. This paper derives and tests by Monte Carlo methods a new procedure, called Weighted Likelihood Estimation (WLE), for estimating θ, the ability parameter. The new estimator is relatively unbiased and is computationally efficient.

Estimation Methods and Bias. There are five basic estimation methods that are used in IRT for parameter estimation: Maximum Likelihood Estimation (MLE) (Lord, 1980), Bayesian Modal Estimation (BME) (Samejima, 1980) [also called Modal A-Posteriori (MAP), (Bock, 1983)], Owen's Sequential Bayesian (OSB) (Owen, 1975), Expected A-Posteriori (EAP) (Bock, 1983), and Marginal Maximum Likelihood (MML) (Bock and Aitkin, 1981). In addition to these, there are variations such as the Robustified Jackknife (Wainer & Wright, 1980), h-estimators (Jones, 1982), and biweight estimates (Bock & Mislevy, 1981).

All of these estimation methods produce estimates that are biased to some degree. MLE (Lord,1983a), and BME (Lord, 1983b, 1984) were shown to be biased to order $n^{-1}$; that is to say the bias is inversely proportional to n, the number of items in the test, other things being equal. OSB, EAP, and MML are all Bayesian procedures (in spite of MML's title), and, therefore, also are biased to order $n^{-1}$. The biweight and h-estimators are robust M-estimators [modified MLE, (See Andrews et al, 1972)] designed to reduce the influence of outliers rather than to reduce bias.

The bias of jackknifed estimators, which were introduced by Quenouille (1956), is of one order less than the order of bias of the estimator jackknifed (Kendall & Stuart, 1973). Therefore, the bias of the Robustified Jackknife, which jackknifes MLE, should be of order $n^{-2}$ except for any bias caused by the robustification. Unfortunately, the reduced bias of jackknifing is achieved by increasing the required computations by an order of magnitude. That is to say, for a test of n-items, the computational time for the jackknifed MLE is n times the computational time for the MLE itself. Even on large computers this computational intensity increases CPU time from minutes to hours; this increase is unacceptable in most settings.

Other methods of bias reduction have been used with some success. For MLE, Cox & Hinkley (1974) suggest evaluating the bias at the value of the estimate, and then subtracting the estimated bias from the estimate to produce an improved estimate. Anderson and Richardson (1979) and Schaefer (1983) used this technique successfully on discrimination and location parameters, respectively, of logistic models. One difficulty with this approach is that for models in which bias is a monotonic function of the parameter being estimated (as is often the case in IRT), error can actually be increased rather than reduced.

Lord (1983c, Personal Communication) has suggested using as an estimate of θ that value of the parameter, which when added to the bias evaluated at the value, is equal to the maximum likelihood estimate. There are two difficulties with this estimator: 1) it is not necessarily unique, and 2) if the maximum likelihood estimate is infinite, so is this estimator. It is unknown whether this approach, used with other estimators, would overcome these difficulties. A similar, untried estimator is the value of the parameter, which when added to its bias, maximizes the likelihood function.

Weighted Likelihood Estimation. For a test of n items the MLE of θ, MLE(θ), is the value of θ that maximizes the likelihood function, $L(\underline{u}|\theta)$, where

$$L(\underline{u}|\theta) = \prod_{i=1}^{n} P(\theta)^{u_i} \cdot Q(\theta)^{1-u_i} , \qquad (2)$$

and $\underline{u}$ is the vector of n scored item responses ($u_i$ = 1 if item i is correctly answered, and $u_i$ = 0 if item i is incorrectly answered; i = 1, 2, ---, n), and $Q(\theta) = 1 - P(\theta)$. Hereafter, the subscript i will be dropped for convenience, unless it is needed for clarity. MLE(θ) is found at the zero of the likelihood equation,

$$l_1 = \delta/\delta\theta \ln L(\underline{u}|\theta) = \sum_{i=1}^{n} (u-P)P'/PQ = 0 \quad , \qquad (3)$$

where P′ = δ/δθ P.

A class of M-estimators, z = M(Z), of the parameter Z may be defined as the value of Z that maximizes,

$$f(Z)\cdot L(\underline{u}|\theta) = f(Z)\cdot\pi_{i=1}^{n} P(\theta)^{u_i} \cdot Q(\theta)^{1-u_i} \quad , \qquad (4)$$

where Z is a function of θ. The M-estimate of Z is found at the zero of the M-estimate equation,

$$\sum_{i=1}^{n} (u-P)P'/PQ + \delta/\delta\theta \ln f(Z) = 0 \quad . \qquad (5)$$

If f(Z) is a constant, z is a maximum likelihood estimate of Z, MLE(Z), and (5) reduces to (3). If f(Z) is an assumed prior density function of Z, then (4) is the posterior density function, and z is a Bayesian Modal Estimate of Z, BME(Z).

Lord (1983a) gives the following asymptotic expression for the bias of MLE(θ), BIAS(MLE(θ)), which is of order $n^{-1}$:

$$\mathrm{BIAS(MLE(\theta))} = -J/(2I^2) \qquad (6)$$

where I is test information,

$$I = \Sigma P'^2/PQ \quad ,$$

$$J = \Sigma P'P''/PQ \quad ,$$

and $P'' = \delta^2/\delta\theta^2 \; P$ . Equation (6) is equivalent to the general expression of BIAS(MLE(θ)) for a multinomially distributed variable, given by Cox & Hinkley (1974, p. 310). Lord (1984) gives the bias of BME(θ) with a standard normal prior.

$$\mathrm{BIAS(BME(\theta))} = \mathrm{BIAS(MLE(\theta))} - \theta/I \qquad (7)$$

BIAS(BME(θ)) is also of order $n^{-1}$. The last term on the right hand side (RHS) of (7) is the derivative, with respect to θ, of the log of the standard normal density divided by test information. From this observation, we can conjecture that the bias of the M-estimator defined by (5) is

$$\mathrm{BIAS(M(Z))} = \mathrm{BIAS(MLE(\theta))} + (\delta/\delta\theta \ln f(Z))/I \quad . \qquad (8)$$

Thus, in order to find the M-estimator that is unbiased, we only need set the RHS of (8) equal to zero, and solve for f(Z). Substituting (6) into (8), we obtain

$$\delta/\delta\theta \ln f(Z) = -J/(2I) \quad . \tag{9}$$

Replacing f(.) with w(.) in (5) to emphasize that the function is now specifically defined, and letting Z = θ yields

$$\sum_{i=1}^{n} (u-P)P'/PQ + \delta/\delta\theta \ln w(\theta) = 0 \quad , \tag{10}$$

and substituting (9) into (5), gives

$$\sum_{i=1}^{n} (u-P)P'/PQ + J/(2I) = 0 \quad , \tag{11}$$

where I and J are as defined in (6). An estimate satisfying (11) is called a Weighted Likelihood Estimate (WLE). If the conjecture above is correct, then BIAS(WLE(θ)) should be only of order $n^{-2}$, one order less than MLE(θ) and BME(θ).

THEOREM: WLE(θ) is unbiased to order $n^{-1}$, i.e.

$$\text{BIAS(WLE}(\theta)) = 0 + o(n^{-1}) \; . \tag{12}$$

Appendix A gives the mathematical proof of the theorem for rather restrictive conditions. Apparently, however, this

"... method of removing the first bias term will work in complete generality, and can be extended to any form of consistent estimating equation where the mathematical form of bias is computable." (Hinkley, 1985, Personal Communication). WLE($\theta$) is asymptotically normally distributed with variance equal (asymptotically) to the variance of MLE($\theta$), i.e.,

$$\mathrm{VAR(WLE}(\theta)) = \mathrm{VAR(MLE}(\theta)) = I^{-1} .$$

Warm (1985) affirmed (12) in a Monte Carlo study of the asymptotic properties of WLE($\theta$), and found that VAR(WLE($\theta$)) converges to the asymptotic value ($I^{-1}$) more rapidly than VAR(MLE($\theta$)).

In general there is no closed-form solution for the indefinite integral of $\delta/\delta\theta \ln w(\theta)$ in order to solve for $w(\theta)$. However, if the c-parameters are equal to zero for all items (as they are in the one and two parameter models of IRT), then

$$w(\theta) = I^{½} \quad (c_i = 0, \quad \text{all } i) .$$

# TWO EXPERIMENTS TO EVALUATE WEIGHTED LIKELIHOOD ESTIMATION

The proof in Appendix A and the results of Warm (1985) speak well of the asymptotic properties of WLE. While these properties are crucial to the mathematical statistician, they are of less interest to users of IRT in practical testing applications. In real testing situations, the number of items in a test seldom exceeds 100, which may be too few for asymptotic results to hold.

To be especially useful, an estimation method must be shown to be practical for conventional tests with fewer than 50 items (the fewer the better). By "conventional test" is meant a paper and pencil test on which all examinees take the same items. A limitation of Monte Carlo studies is that the results are specific to the design of the study. The usual solution for the choice of the design is to select either an idealized case, which may not generalize to more than a few realistic situations, or to select a "typical" case. The second alternative was chosen for this study. The conventional tests, simulated here, were designed to have a Test Information function roughly "normal" in shape. Lord & Novick (1967) point out that this is a common design. This shape was accomplished by choosing normally distributed b-parameters for the items. Once the gross shape was decided upon, other details were resolved; namely the values of the a- and c-parameters, and the range of the number of items in the simulated tests. Both the a- and c-parameters affect the amount of test information (but in different ways). Variability among the a- and c-parameters in a given test has "local" effects, but not important overall effects. On the other hand, the amount of test information is important to the behavior of an estimator. To keep the number of possibile permutations of the variables in the design manageable, the c-parameter for all items in all tests was held constant at 0.20 - a typical value. Two values were chosen for the a-parameters, a "moderate" value (1.0), and a "high" value (2.0). To avoid the local effects mentioned above the a-parameters for all items in a given test were set to the same value. A wide range of test lengths was chosen (10 to 60). The six test lengths by two values for the a-parameter give 12 different conventional tests.

With the pending adoption of tailored testing by the U.S. Armed Forces (Green et al, 1984) for the ASVAB (Armed Services Vocational Aptitude Battery), the behavior of the θ-estimator in tailored tests (Lord, 1980) is of current interest. The same conflict between "idealized" versus "typical" occurs in this case for the design of the hypothetical item pool. Various solutions have been chosen. McBride (1977) used both an ideal, finite item pool and an infinite item pool. Maurelli (1978) and McKinley & Rekase (1981) used a finite item pool with normally distributed b-parameters. Gorman (1980) used finite item pools with both rectangularly and normally distributed b-parameters. Jensema (1977) showed that the design of the item pool strongly influences the results, and recommended rectangularly distributed b-parameters.

For the tailored testing experiment in the present inquiry two item pools with rectangularly distributed b-parameters were used -- differing only in the a-parameter. The a-parameters of one pool were held constant to a high value (2.0) in order to simulate an ideal, infinite item bank. The a-parameters of the other item bank declined as the number of items administered increased. The purpose of the declining a-parameter was to simulate the depletion of items with high item information as the tailored test proceeds; this always occurs with a finite item bank. The c-parameters of all items in both item banks were held constant (0.20) as with the conventional test.

The evaluation of a new estimator such as WLE requires a standard of comparison. Both MLE and BME were chosen as "benchmark" methods, representing opposite extremes in several ways.

The criteria of comparison are also issues. Since the major motivation for the development of WLE is unbiasedness, conditional on $\theta$, bias is an important criterion. The variance of an estimator is also a common criterion for evaluating estimation methods. However, it is trivial to produce small variance in an estimator -- simply setting every estimate to some constant value produces zero variance (although it wreaks havoc with bias). The mean squared error combines both criteria, since it is the sum of the variance and the squared bias by a common analysis of variance identity.

In tailored testing, the number of items administered is important, reflecting both total testing time required and the exposure of items to potential compromise. In addition, the speed of estimating ability between items may be important. If computation time is too slow, total testing time increases, and "dead time" may occur, i.e. the examinee may have to sit and wait for the next item to be presented. Dead time can induce boredom and cause underestimation of ability.

The measurement of ability on the $\theta$-scale is not the only scale of interest. True score, $T \equiv T(\theta) = \Sigma P(\theta)$, the expected value of the raw, number-right score on a conventional test, is often desired. A frequent criticism of Bayesian estimators is an internal contradiction; the expected value of the Bayesian estimator of a non-linear transformation of $\theta$ is not equal to the non-linear transformation of the Bayesian estimator of $\theta$. On the other hand MLE is invariant to non-linear transformation. True score is a non-linear transformation of $\theta$, and $E(T(BME(\theta)|\theta))$ =/= $T(\theta)$, where E is the expectation operator. But MLE(T) is also biased. Thus, since WLE is ostensibly less biased than MLE and BME, it would be interesting to know whether $E(T(WLE(\theta)|\theta)) = T(\theta)$.

# METHOD

Design of the Conventional Tests: For the comparison of the three estimators under a wide range of test lengths, 12 conventional tests were constructed. There were two tests for each of six test lengths, n=10, 20, 30, 40, 50, and 60. For each test length, the a-parameters of one test were set to $a_i = 1.0$, for all i=1,2,---n, and for the other test $a_i = 2.0$, all i. The b-parameters of all 12 tests were distributed "normally", using the inverse normal transformation, $\Phi^{-1}(.)$. That is, for a test of length n, $b_i = \Phi^{-1}((i-.5)/n)$. The c-parameters of all items in all tests were set to 0.20 . These item parameters produce a test information curve that is roughly "normally" shaped, and is a commonly used conventional test design.

At each of 17 values of θ (= -4, -3.5, ---, 4), 1000 simulated examinees were administered all 12 tests, and WLE(θ), MLE(θ), and BME(θ) was computed for each examinee and for each test. The same item responses were used for all three estimators. The mean, standard deviation, and mean squared error of the 1000 estimates of θ was computed at each of the 17 values of θ for each test and each estimator. In addition, the mean, standard deviation, and mean squared error of estimated True Score, $T^{\wedge} = T(\theta^{\wedge})$, was computed for each of the 1000 estimates of $\theta$ at each of the 17 values of $\theta$ for each estimator on the two tests with most extreme Test Information functions. Appendix B contains the computer program used for the Monte Carlo study of the conventional test. It is written in the matrix algebra language (PROC MATRIX) of the Statistical Analysis System (SAS, 1980), and was executed on an IBM 3081 mainframe computer.

Design of Tailored Tests: Six tailored tests, two for each estimator, were administered to 100 simulated examinees at each of the 17 values of $\theta$. For all tailored tests all $c_i = .20$ . For each of the three estimators one tailored test had all $a_i = 2.0$ . The other tailored test for each estimator had declining a-parameters to simulate the declining item information available from a finite item pool, specifically $a_i = (71-i)/35$ . Following Weiss & McBride (1984), the values of the b-parameters for all tailored tests were chosen so that the maximum of the item information for the item a- and c-parameter was at the current estimate of $\theta$. That is, the b-parameter of the $(i + 1)^{th}$ item was

$$b_{i+1} = \theta^{\wedge} - \ln(\tfrac{1}{2}(1 + (1 + 8c)^{\frac{1}{2}}))/(1.7a) \quad ,$$

where $\theta$^ is the current estimate of $\theta$ after the $i^{th}$ item.

The stopping rules for administering items were: 1) stop if test information exceeds 20 at the current estimate of $\theta$, or 2) stop if the number of items administered (n) = 50, whichever occurred first. The mean, standard deviation, and mean squared error of the 100 estimates was computed for each of the 17 values of $\theta$ for each test and each estimator. In addition, for each tailored test and estimator the average number of items administered, and the average iteration computation time per item were computed. Appendix C contains the computer program of the tailored tests, which was written in TURBO PASCAL 3.0 (Borland, 1985), and was executed on a Corona PC portable computer with an Intel 8087 high speed arithmetic processor chip.

Estimating $\theta$: For the conventional tests, $\theta$ was iteratively estimated by the interval bisection method with r = 15 iterations.

$$\theta^\wedge_{m,r} = \theta^\wedge_{m,r-1} + \delta_{m,r} \quad , \; r = 1,2,---15,$$

$$m = \text{WLE, MLE, or BME, and } \theta^\wedge_{m,0} = \theta \quad .$$

For the first four iterations, r = 1, 2, 3, 4, $|\delta_{m,r}| = 1$ with the sign the same as the objective function. In the remaining iterations, $r = 5, 6, ---, 15$, $|\delta_{m,r}| = |\delta_{m,r-1}|/2$. This iteration method has several advantages: 1) $|\theta\hat{}-\theta| < 5$, 2) it will find the local maximum closest to true $\theta$, 3) the magnitude of the difference between the true maximum and the final estimate < .001, and 4) convergence is guaranteed.

Iterative estimation of MLE($\theta$) for the tailored tests was accomplished by a modification of Newton-Raphson, called "Fisher Scoring" (Lord, 1980). In "Fisher Scoring" the second derivative of the log likelihood in the denominator of the Newton-Raphson "delta" is replaced with its expected value (which is equivalent to the negative of test information). The respective analogies of Fisher Scoring were used for WLE and BME.

$$\theta\hat{}_{m,r} = \theta\hat{}_{m,r-1} + \delta_{m,r} \quad , r = 1,2,---,$$

$$m = \text{WLE, MLE, BME, and } \theta\hat{}_{m,0} = 0 \quad ,$$

with the magnitude of $\delta_{m,r}$ limited to 2.0 .

For $m$ = MLE [with $l_1$ as defined in (3)],

$$\delta_{m,r} = (l_1/I), \text{ evaluated at } \theta\hat{}_{m,r-1} \quad .$$

For $m$ = BME,

$$\delta_{m,r} = (l_1 - \theta)/(I+1),\ \text{evaluated at}\ \hat{\theta}_{m,r-1}\ .$$

For $m$ = WLE,

$$\delta_{m,r} = (l_1 + J/(2I))/(I - (IJ'-I'J)/(2I^2)),$$

evaluated at $\hat{\theta}_{m,r-1}$ . Iterations were continued until $|\delta_{m,r}| < .001$ , $r = 21$, or $|\hat{\theta}_{m,r+1}| > 5$, whichever occurred first.

<u>Generating the scored item responses, $u_i$</u>: For each item response $P(\theta)$ was calculated, and a pseudo-random number, $y$, uniformly distributed over the interval $(0,1)$, was generated. Then $u_i = 0$, if $y > P(\theta)$; else $u_i = 1$. The seeds for the computer program random number generators were arbitrarily taken from the real-time clock of the computer.

# RESULTS

Conventional Tests: The results for the 12 conventional tests are remarkably similar, and virtually identical with respect to the relative results among the three estimators. Therefore, complete results of conventional tests will be presented here only for the test with 10 items and a = 1. Some results for the 30 and 60 item tests will also be shown to provide a range of values. Figures showing the results of the other tests are in Appendix D, labeled D.1, D.2, etc.

To show the range of the designs of the 12 conventional tests, the Test Information functions of the 10 and 60 item tests with a = 1 and a = 2 are presented in Figure 1. Figure 1 shows that the Test Information functions become higher and broader as the number of items increases, but more peaked as the a-parameters increase .

---

Insert Figures 1 and 2 about here.

---

Figure 2 (and D.1 through D.11) shows the average error of each estimator at each of the 17 values of $\theta$. As predicted by equations (6) and (7) (and demonstrated by

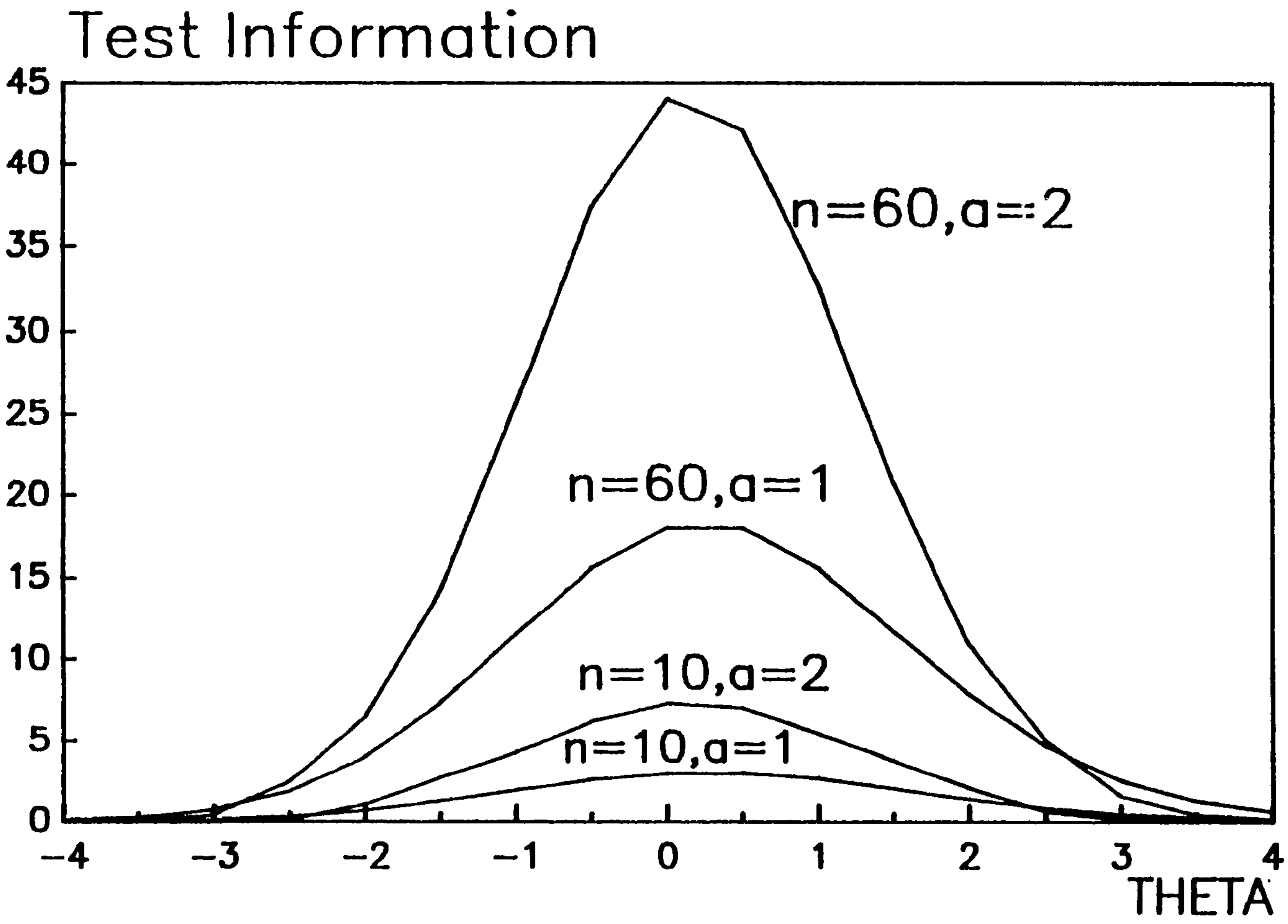

Figure 1
Test Information Curves of the 10 and 60 Item Conventional tests with a = 1 and a = 2.

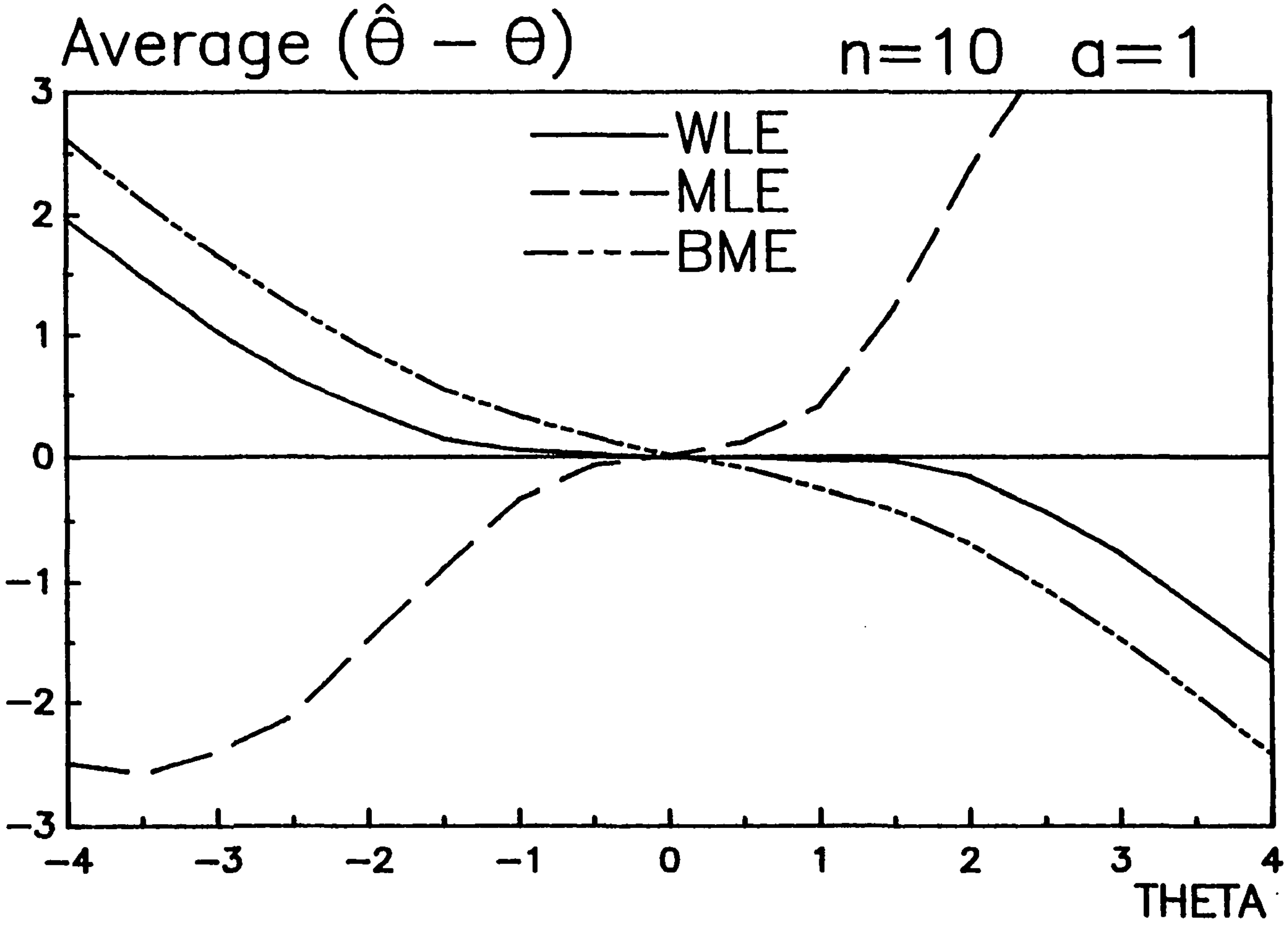

Figure 2
Average Estimation Error of θ^ on Conventional Test with 10 Items, All a = 1, Normally Distributed b, and All c = 0.20 .

Lord, 1984), the bias of MLE(θ) is positively correlated with θ, while the bias of BME(θ) is negatively correlated. The bias of WLE(θ) is also negatively correlated with θ. It is very clear from these figures that WLE(θ) is considerably less biased than both MLE(θ) and BME(θ) over the entire range of θ, for all test lengths and both values of the a-parameter. Since biases of the three estimators are zero at slightly different points on the θ-scale, inevitably there will exist very small intervals when the bias of WLE(θ) will exceed the biases of MLE(θ) and BME(θ). However, these instances will occur only when the bias of WLE(θ) is itself negligible. Note that the range of θ over which the bias of WLE(θ) is apparently negligible (i.e. indistinguishable from the zero reference line) is relatively broad, whereas the other two estimators have small bias essentially at a point.

The relative magnitudes of the biases are difficult to compare across tests from the figures. Figures 3 and 4, which display results for a = 1 and a = 2 respectively, show the absolute bias of WLE(θ) for the 10 item test and the absolute bias of MLE(θ) for tests with 10, 30 and 60 items. Figures 3 and 4 indicate that WLE(θ) with 10 items is less biased than MLE(θ) with two to six or more times as many items, depending upon the value of θ. Figures 5 and 6 give the same comparisons between WLE(θ) and BME(θ) in terms of absolute average error. WLE(θ) is also less biased than

BME(θ) in tests with two or three times as many items.

---

Insert Figures 3 through 8 about here.

---

Comparison of the standard deviations of estimated θ in Figure 7 (and D.12 through D.22) reveals a different picture than the comparisons on bias. BME(θ) has less variability than both WLE(θ) and MLE(θ) at all values of θ. (The decline of the standard deviations of MLE(θ) at high and low values of θ were caused by the artificial boundary of |MLE(θ)-θ| < 5. In some cases all 1000 values of MLE(θ) were equal to the artificial boundary.) As discussed above, it is not difficult to reduce variability at the expense of bias. This trade-off accounts for the low standard deviation of BME(θ), and will be expanded upon below. On the other hand WLE(θ) also has small standard deviation, and is considerably less biased than BME(θ).

The mean squared error (MSE) of WLE(θ) (See Figures 8 and D.23 through D.33) is smaller than that of MLE(θ) at all values of θ for all 12 tests. BME(θ) has smaller MSE still at the more central values of θ. However, as the value of the a-parameter or the number of items increases, the advantage of BME(θ) over WLE(θ) shrinks rapidly. At all other values of θ, MSE(WLE(θ)) is considerably smaller than

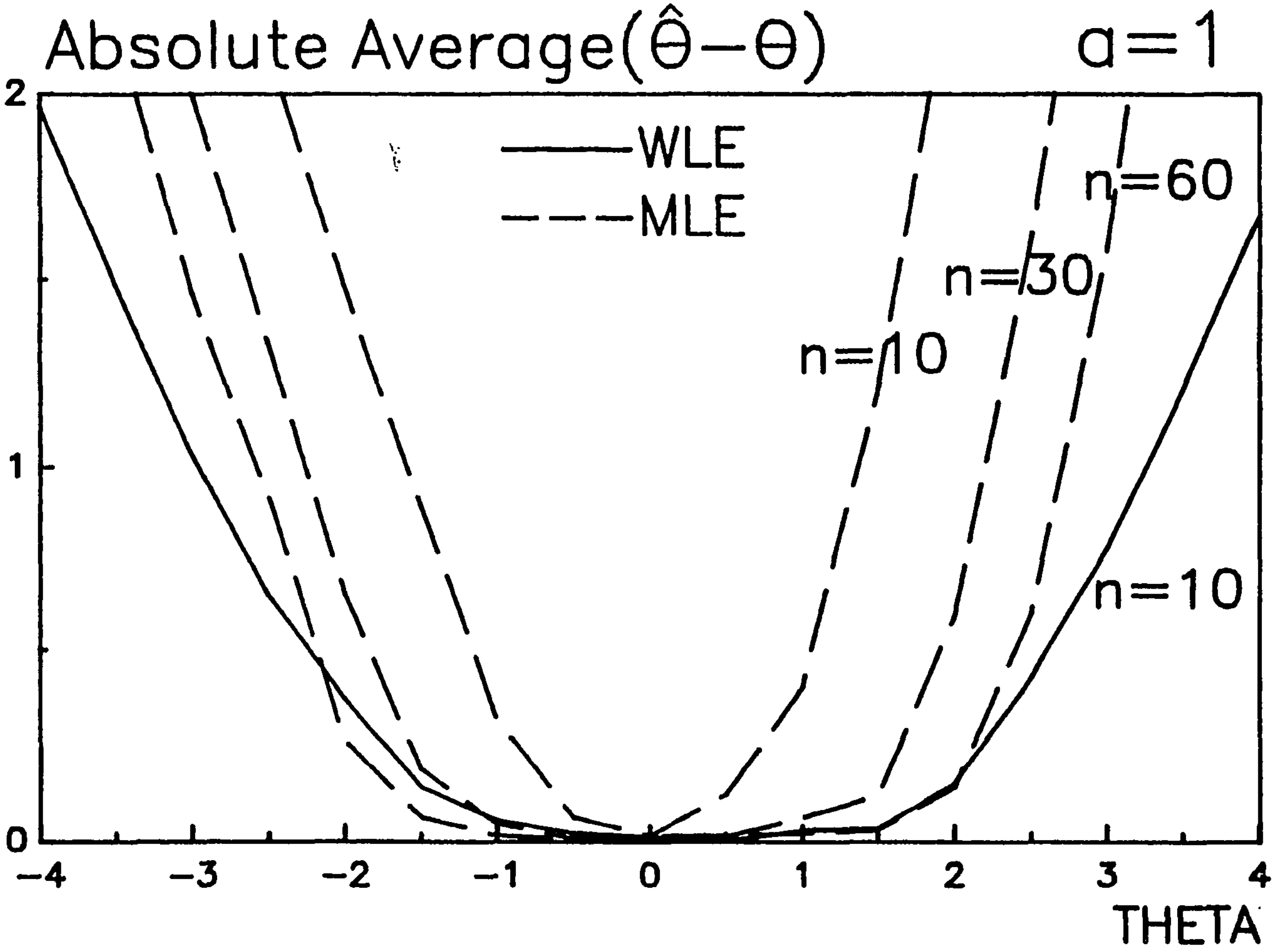

Figure 3
Absolute Average Estimation Error of WLE(θ) with 10 Items, and of MLE(θ) with 10, 30, and 60 Items; All a = 1, Normally Distributed b, and All c = 0.20 .

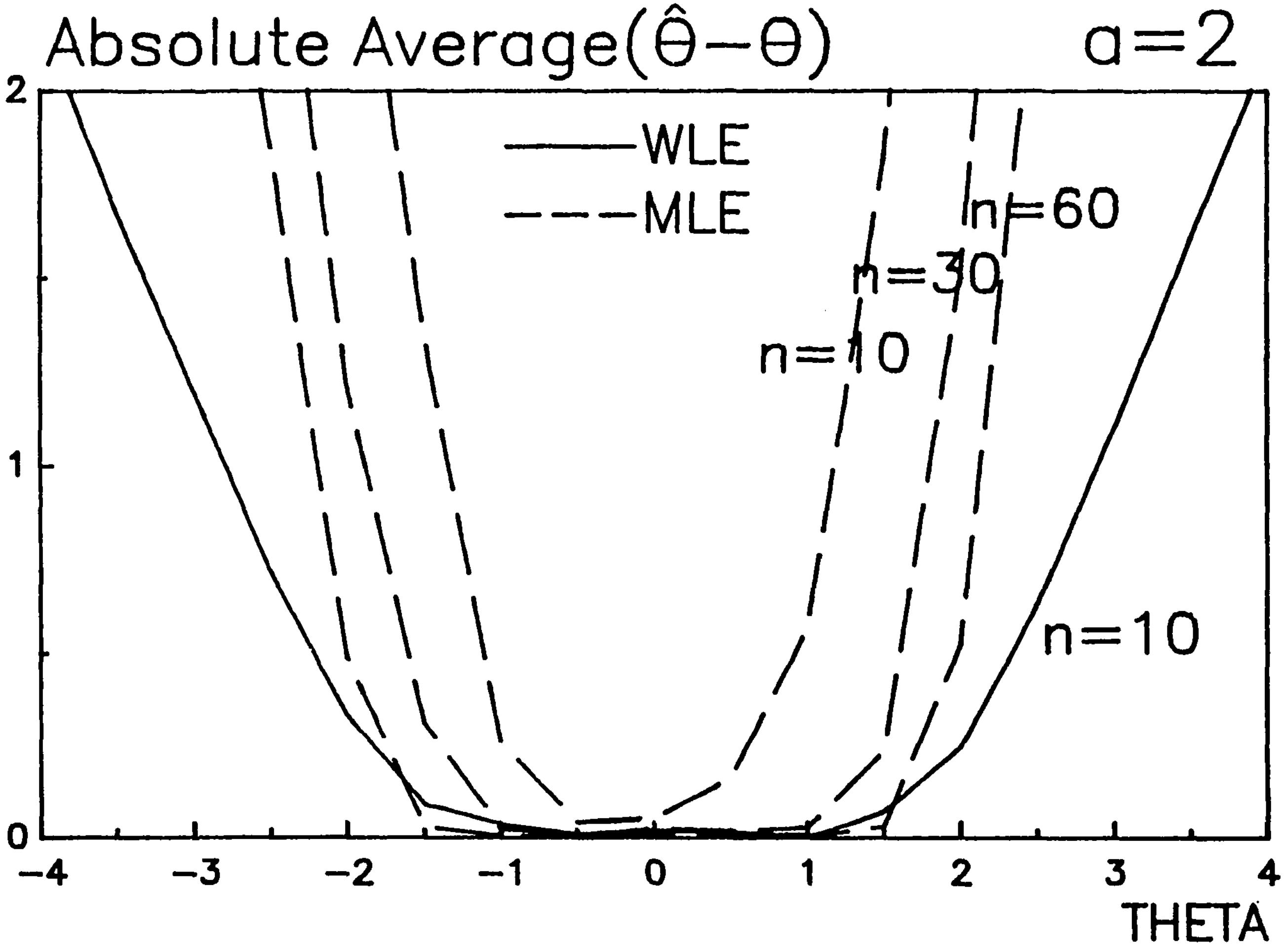

Figure 4
Absolute Average Estimation Error of WLE(θ) with 10 Items, and of MLE(θ) with 10, 30, and 60 Items; All a = 2, Normally Distributed b, and All c = 0.20 .

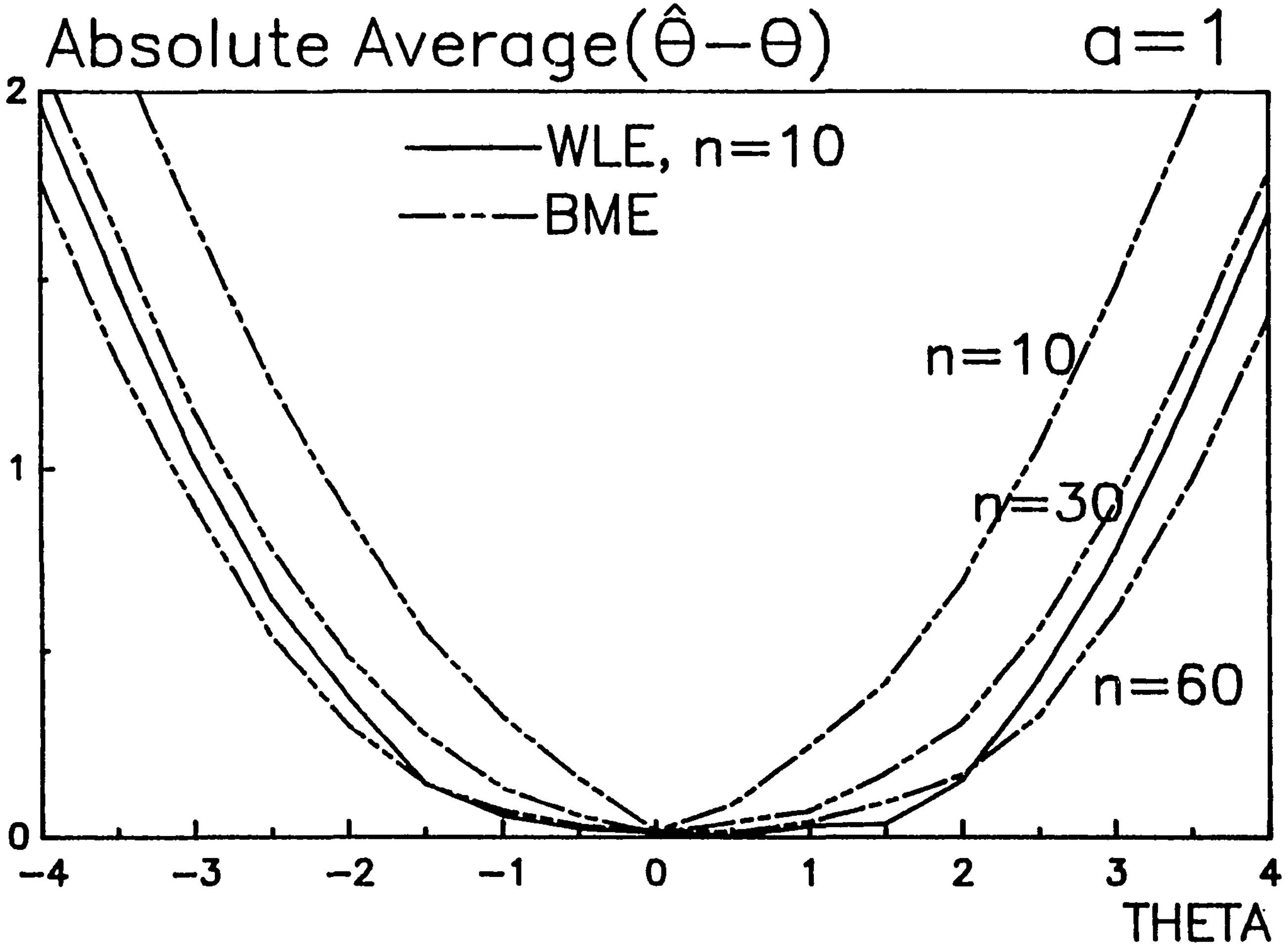

Figure 5
Absolute Average Estimation Error of WLE(θ) with 10 Items, and of BME(θ) with 10, 30, and 60 Items; All a = 1, Normally Distributed b, and All c = 0.20 .

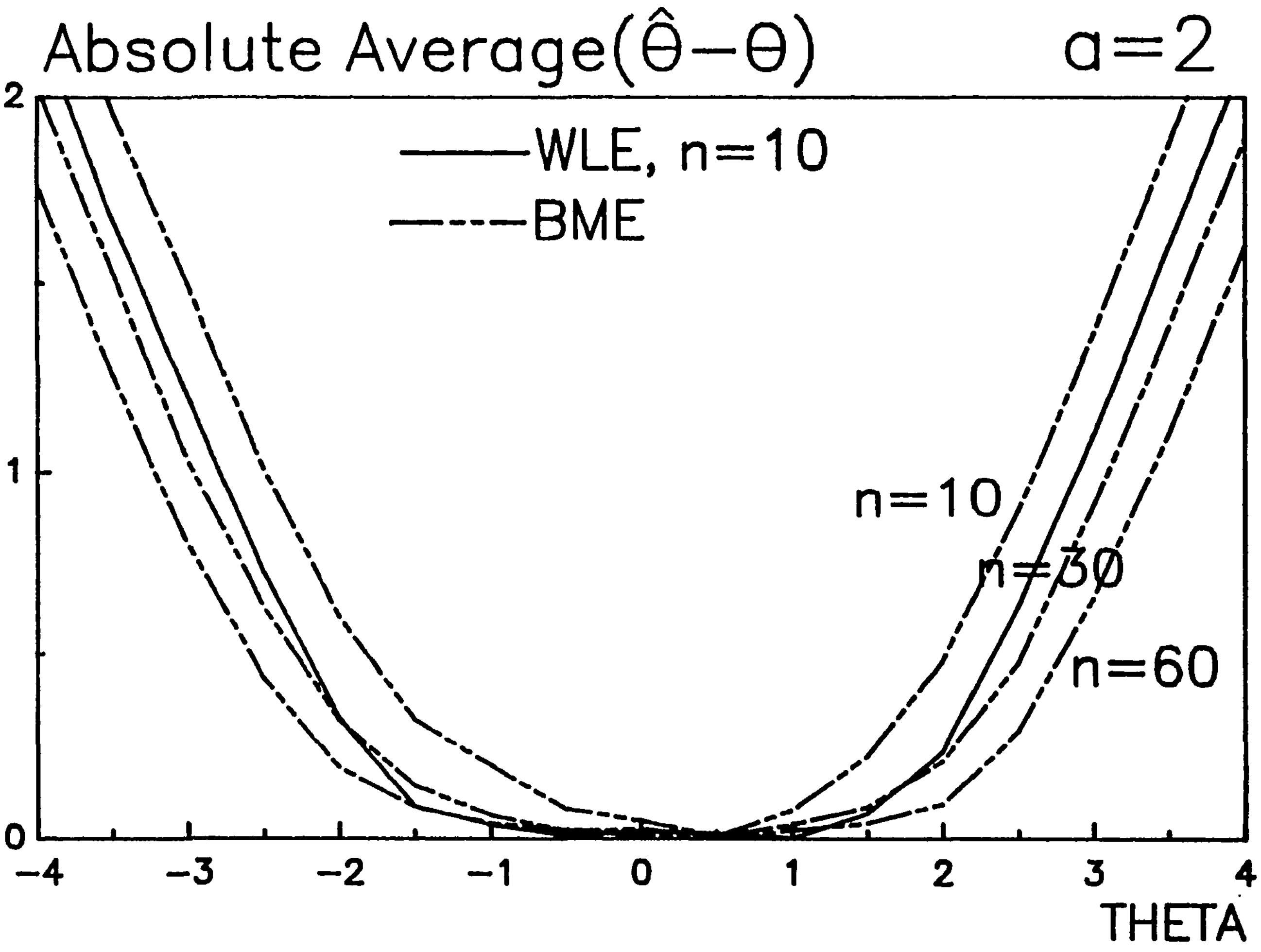

Figure 6
Absolute Average Estimation Error of WLE(θ) with 10 Items, and of BME(θ) with 10, 30, and 60 Items; All a = 2, Normally Distributed b, and All c = 0.20 .

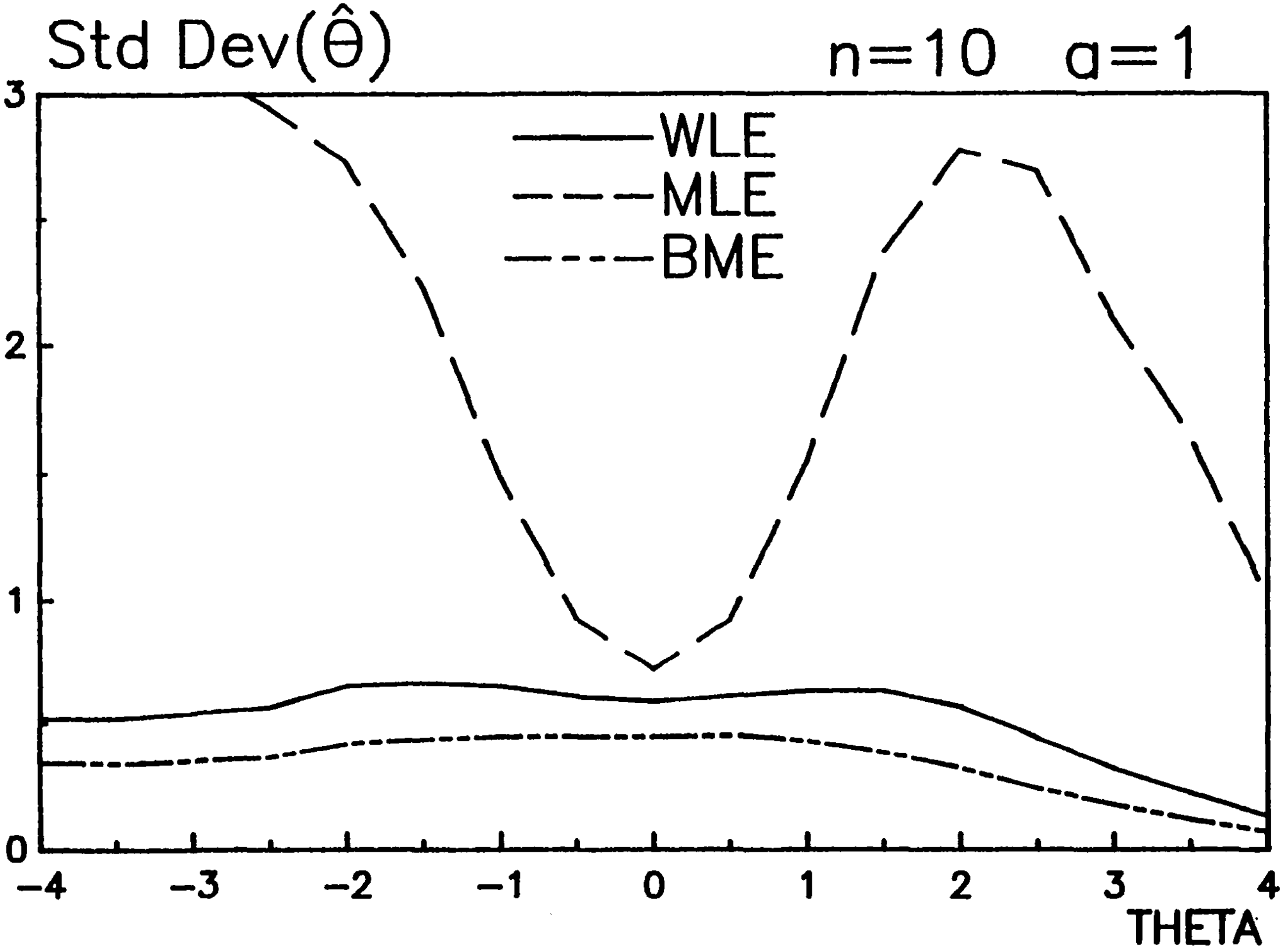

Figure 7
Standard Deviation of θ^ on Conventional Test with 10 Items, All a = 1, Normally Distributed b, and All c = 0.20 .

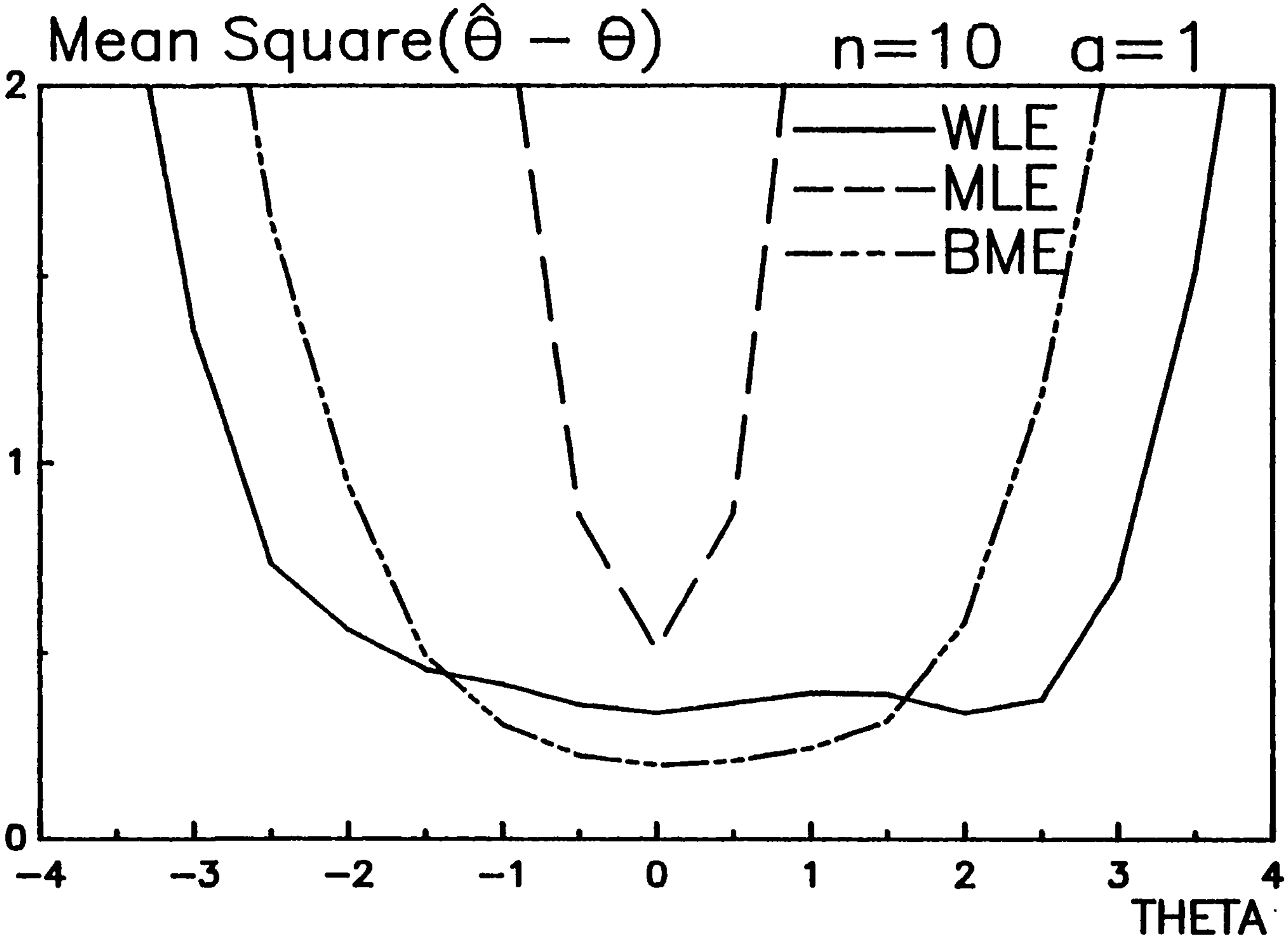

Figure 8
Mean Squared Error of $\theta$^ on Conventional Test with 10 Items, All a = 1, Normally Distributed b, and All c = 0.20 .

MSE(BME(θ)).

---

Insert Figures 9 through 11 about here.

---

True Scores (T) were estimated by evaluating T(θ) at each estimate of θ, T(θ^). T(θ) = Σ P(θ), and T(θ^) = Σ P(θ^). Figures 9 through 11 (and D.34 through D.36) give the average error (T^ - T), standard deviation of T^, SD(T^), and MSE(T^) of the tests with the lowest and highest Test Information functions. In general T(MLE(θ))≡MLE(T) is less biased than T(WLE(θ))≡WLE(T) and T(BME(θ))≡BME(T) at all values of θ for all tests. BME(T)) is more biased than the other two estimators. The bias of WLE(T) falls roughly half way between the biases of MLE(T) and BME(T). SD(BME(T)) is smallest on all tests at most values of θ. Except at extreme θ, where MLE(T) is severely affected by the artificial boundary on MLE(θ), SD(MLE(T)) is largest. Using the standard deviation criterion, WLE(T) is usually intermediate falling between BME(T) and MLE(T). MSE(WLE(T)) is also usually intermediate in value with respect to MSE(MLE(T)) and MSE(BME(T)). However, MSE(BME(T)) is smallest at central values of θ, and largest at extreme values of θ. MSE(MLE(T)), contrary to MSE(BME(T)), is largest at central θ, and smallest at extreme θ.

Tailored Tests: Figures 12 through 16 (and D.37) give

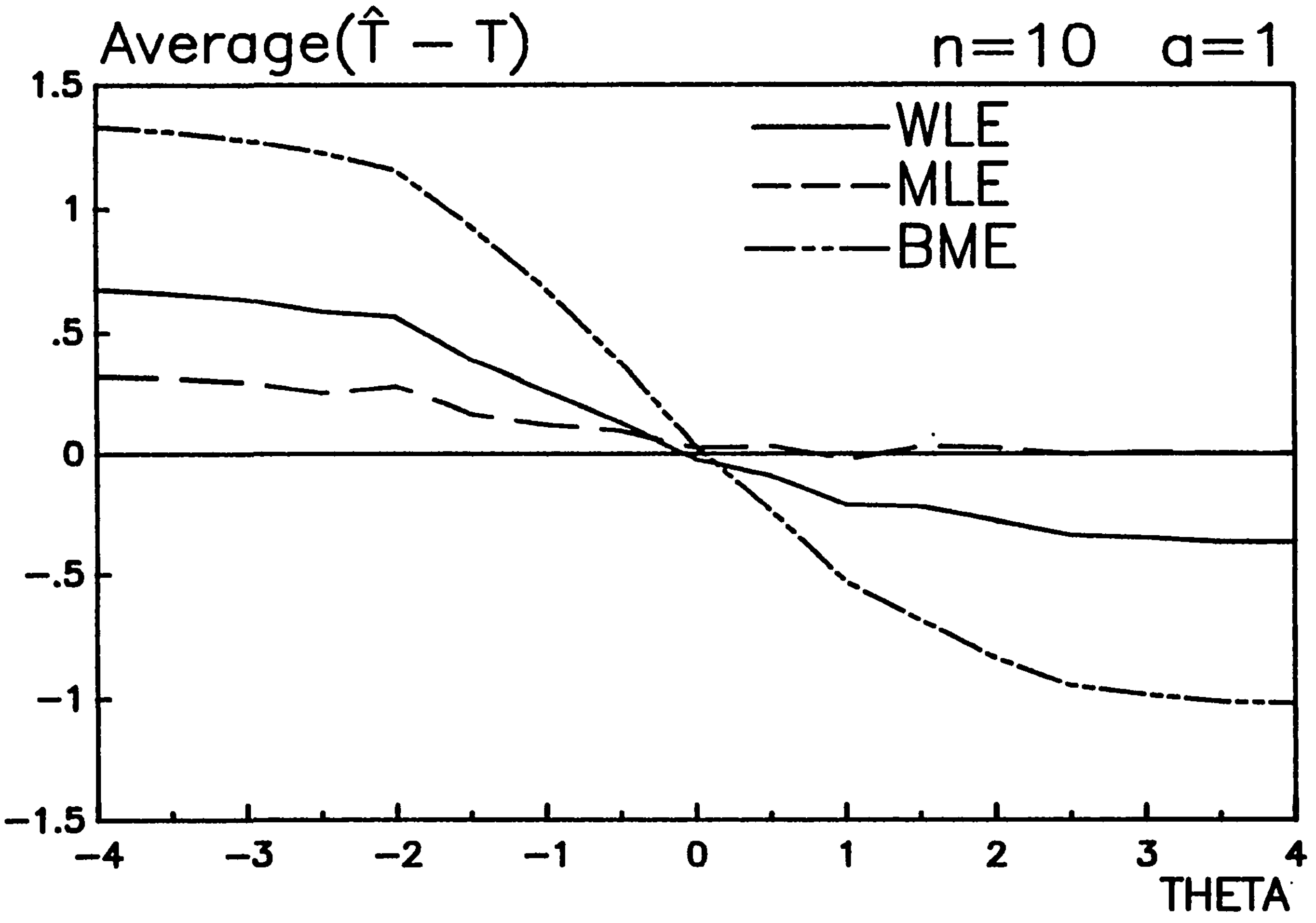

Figure 9
Average Estimation Error of T^ on Conventional Test with 10 Items, All a = 1, Normally Distributed b, and All c = 0.20 .

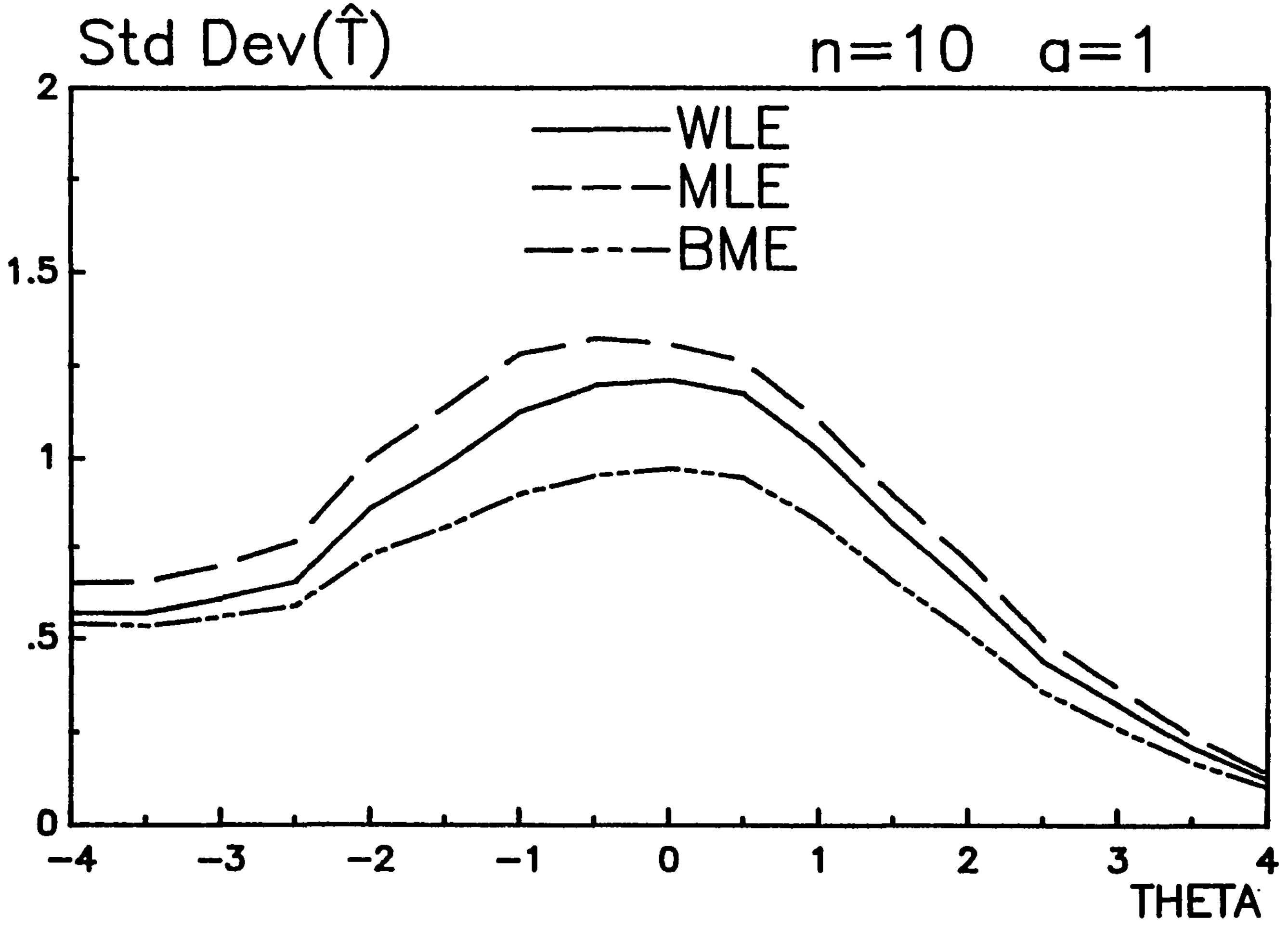

Figure 10
Standard Deviation of T^ on Conventional Test with 10 Items, All a = 1, Normally Distributed b, and All c = 0.20 .

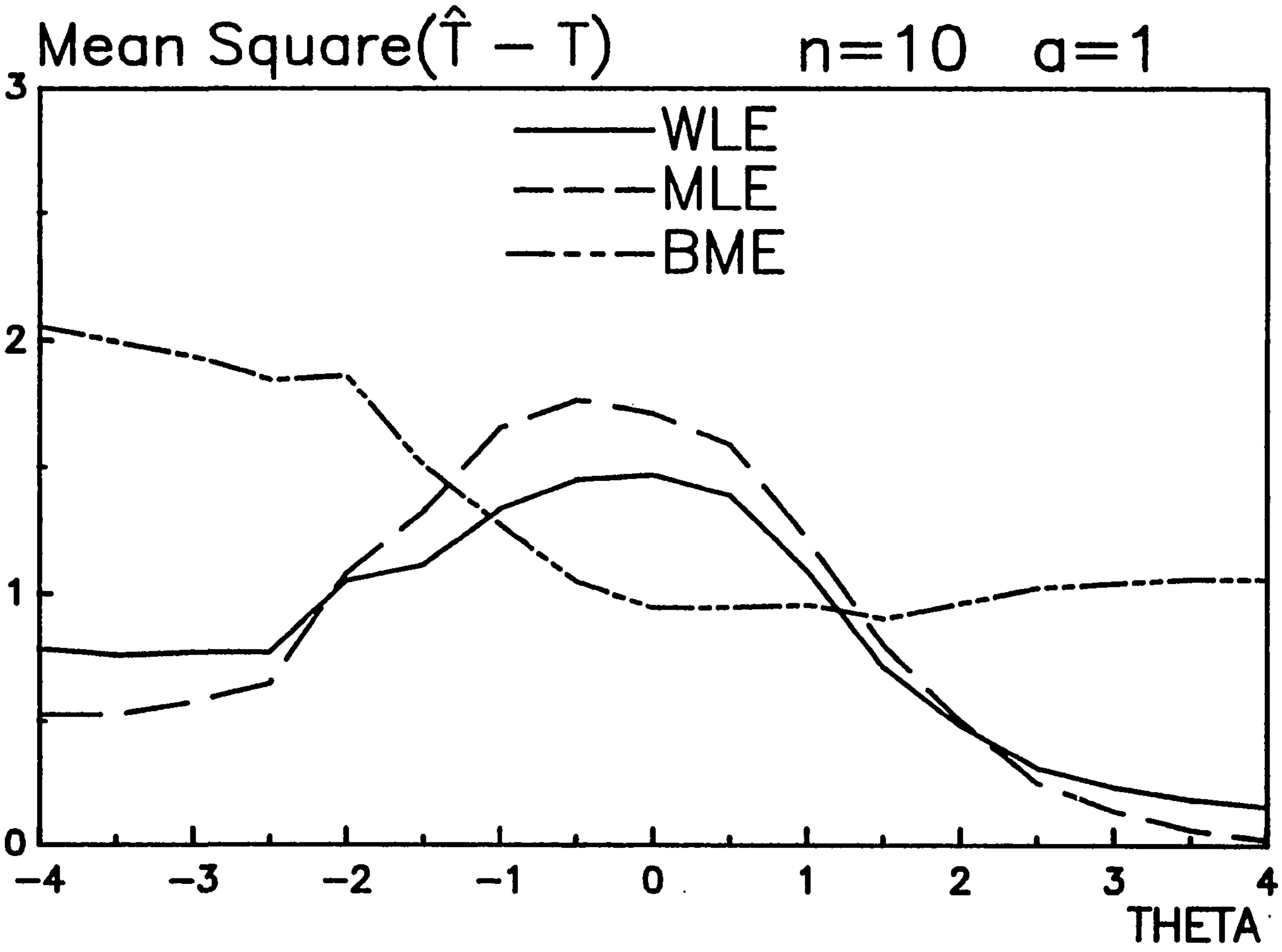

Figure 11
Mean Squared Error of T^ on Conventional Test with 10 Items, All a = 1, Normally Distributed b, and All c = 0.20 .

the average error of θ^, SD(θ^), and MSE(θ^) for the tailored tests. BME(θ) is very biased at all non-zero values of θ on both tests. MLE(θ) is slightly positively biased at most values of θ on both tests. WLE(θ) is the least biased estimator at positive values of θ. At low values of θ , the biases for WLE(θ) and MLE(θ) are about equal and small. SD(WLE(θ)) and SD(BME(θ)) are about equal on both tailored tests, as are MSE(WLE(θ)) and MSE(BME(θ)) at most values of θ. SD(MLE(θ)) on the test with declining a-parameters is also about the same as SD(MLE(θ)) and SD(BME(θ)). However, SD(MLE(θ)) and MSE(MLE(θ)) are very large at central values of θ on the tailored test with a = 2.

---

Insert Figures 12 through 18 about here.

---

The relative average number of items administered to reach the stopping rule under the three estimators are virtually identical for the two tailored tests. (See Figures 17 and D.38.) At central values of θ, WLE(θ) and BME(θ) used about the same number of items, and as few as half as many items as MLE(θ). Both WLE(θ) and BME(θ) required considerably fewer items at central values of θ than at more extreme values. The number of items required by MLE(θ) declined slightly on both tests as θ increased.

The average computation times for estimating θ between

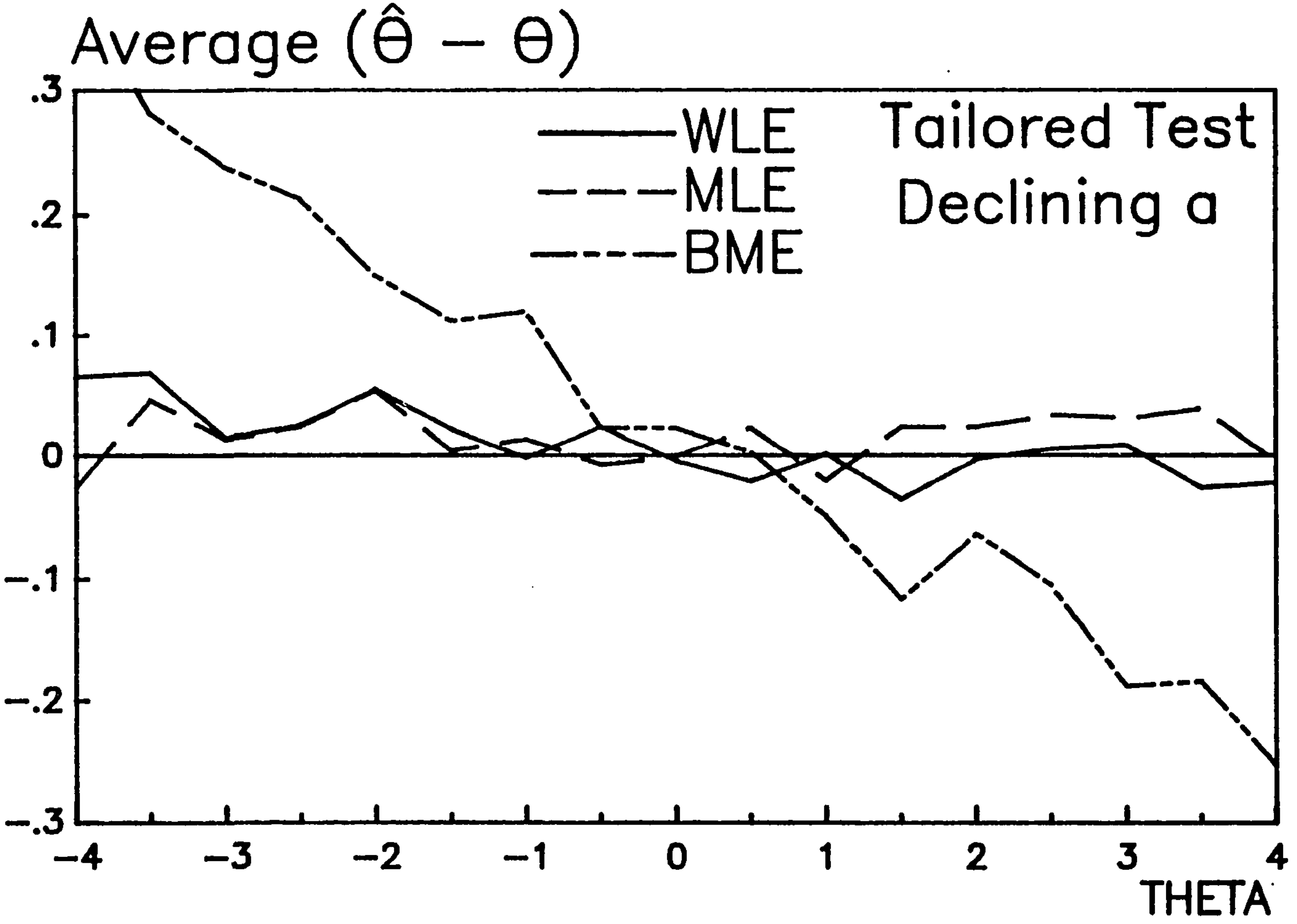

Figure 12
Average Estimation Error of θ^ on Tailored Test with Declining a-parameter, Optimal b-parameter, and All c = 0.20 .

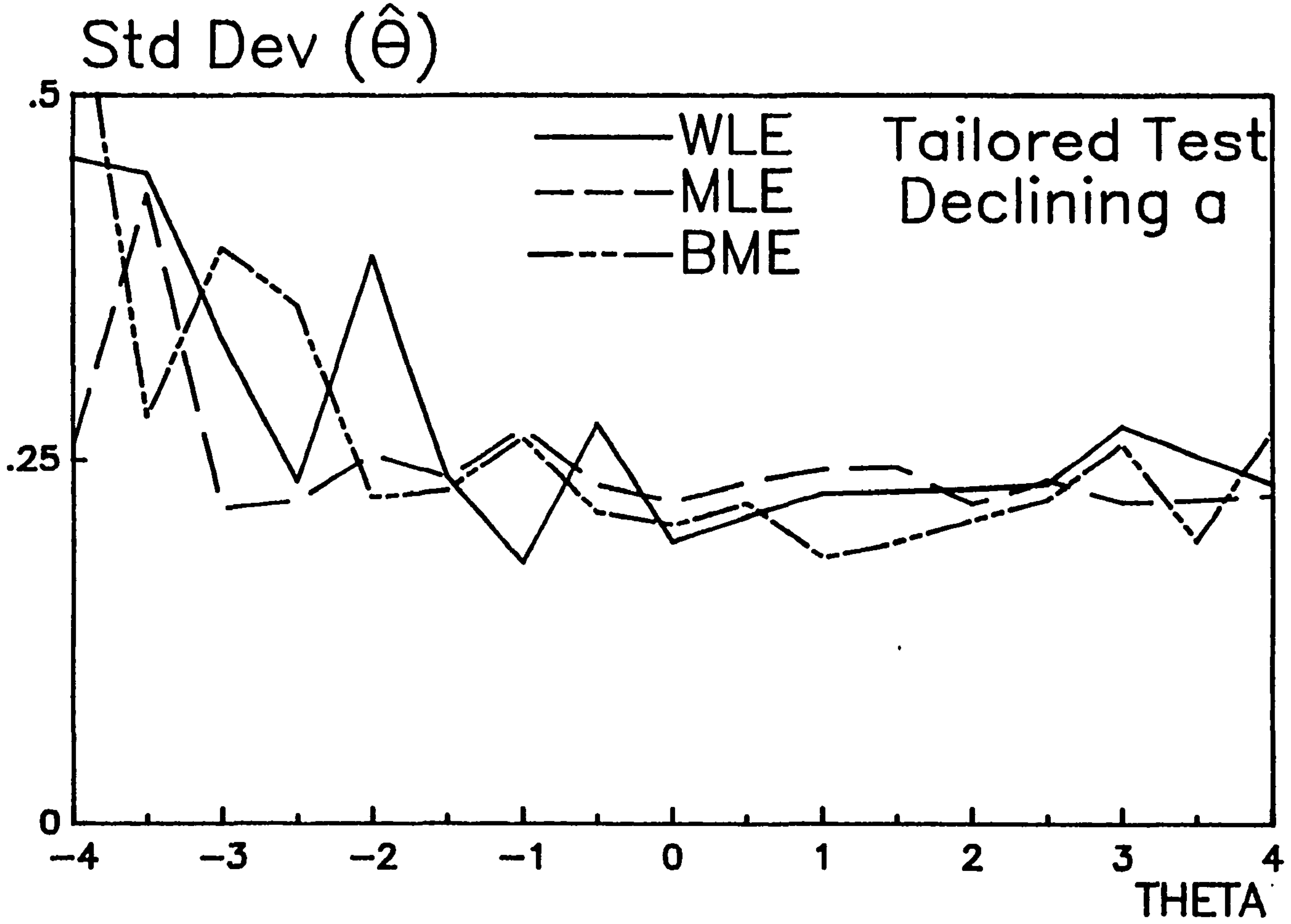

Figure 13
Standard Deviation of θ^ on Tailored Test with Declining a-parameter, Optimal b-parameter, and All c = 0.20 .

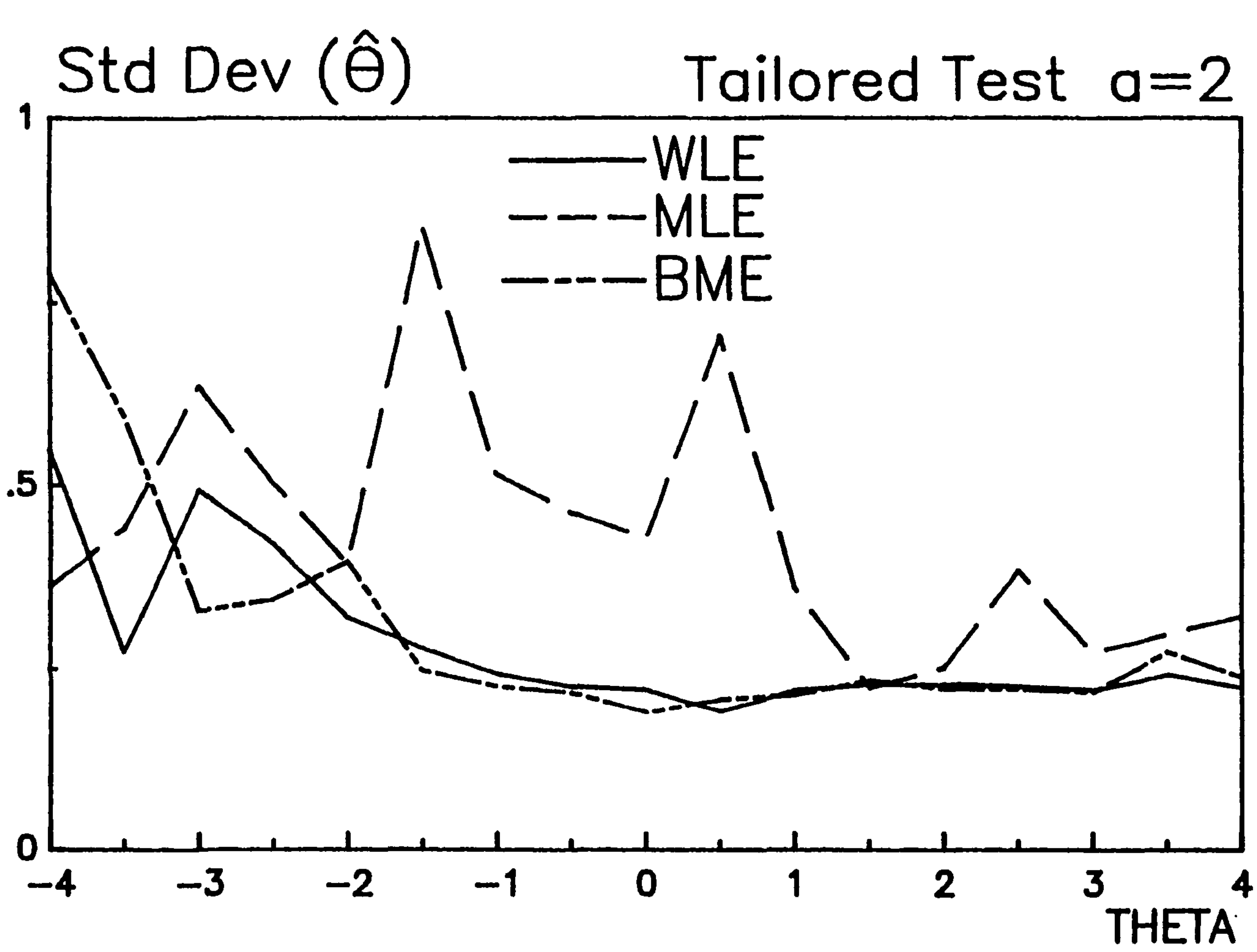

Figure 14
Standard Deviation of θ^ on Tailored Test with a = 2, Optimal b-parameter, and All c = 0.20 .

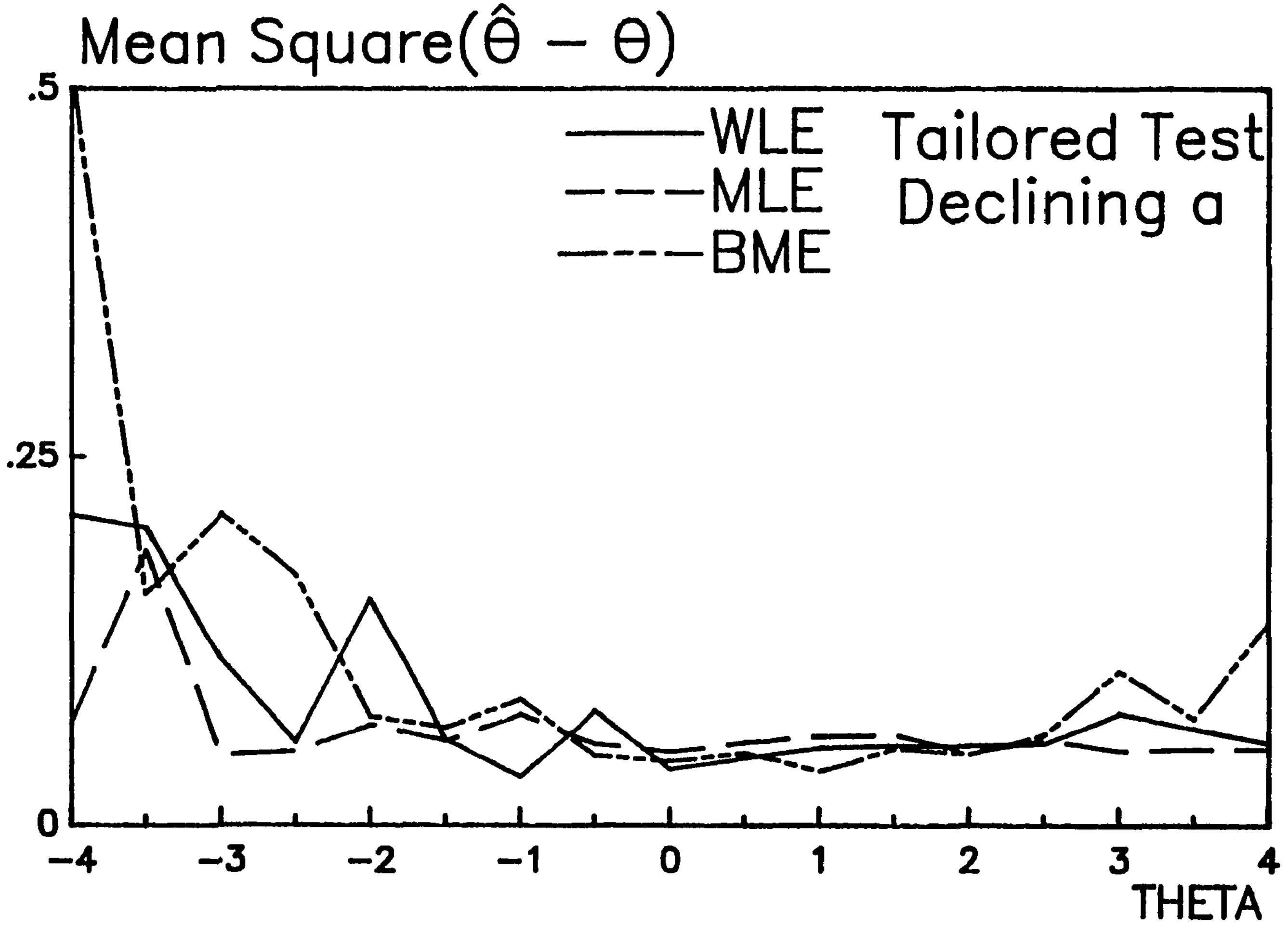

Figure 15
Mean Squared Error of θ^ on Tailored Test with Declining a-parameter, Optimal b-parameter, and All c = 0.20 .

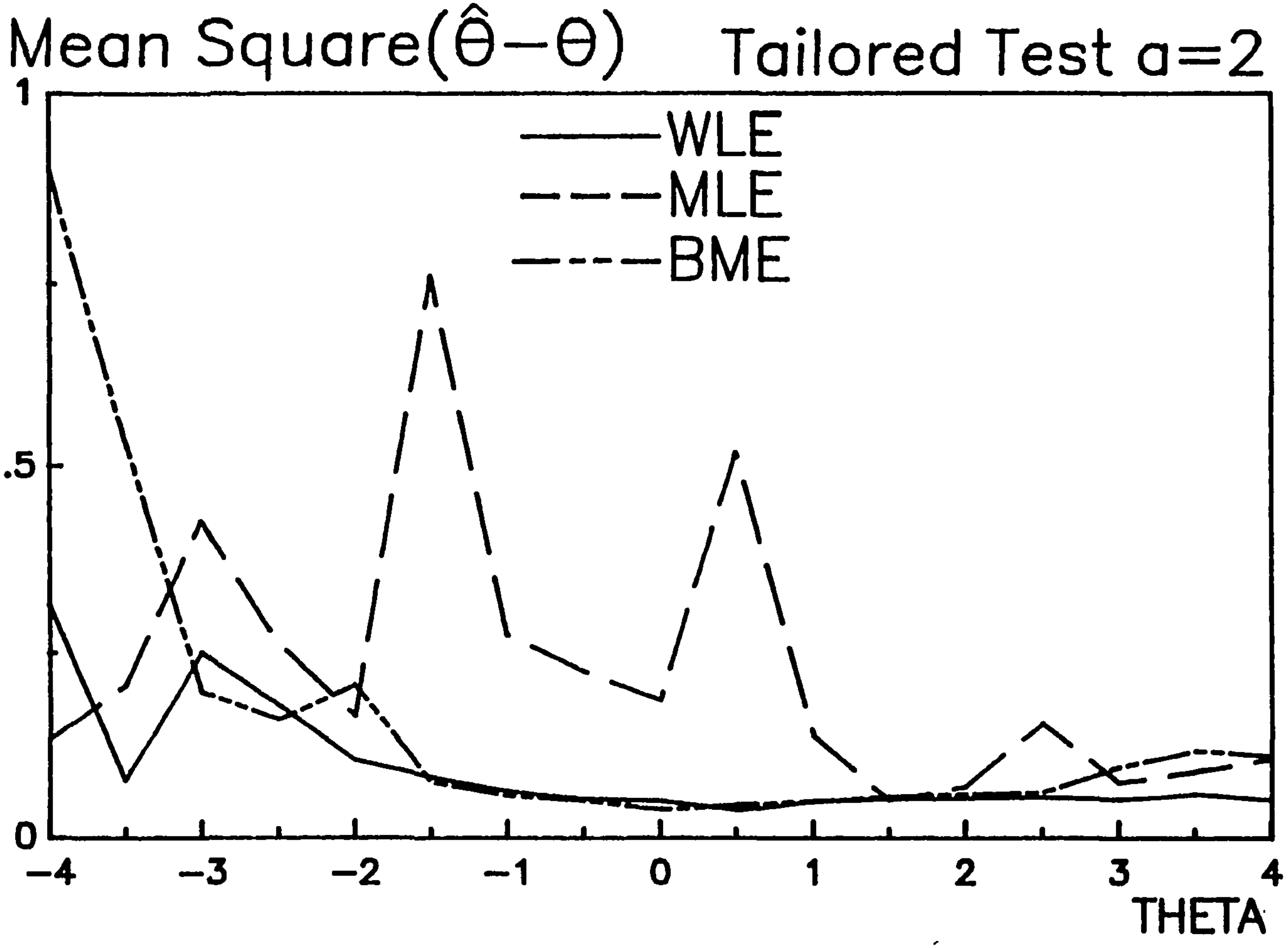

Figure 16
Mean Squared Error of θ^ on Tailored Test with a = 2, Optimal b-parameter, and All c = 0.20 .

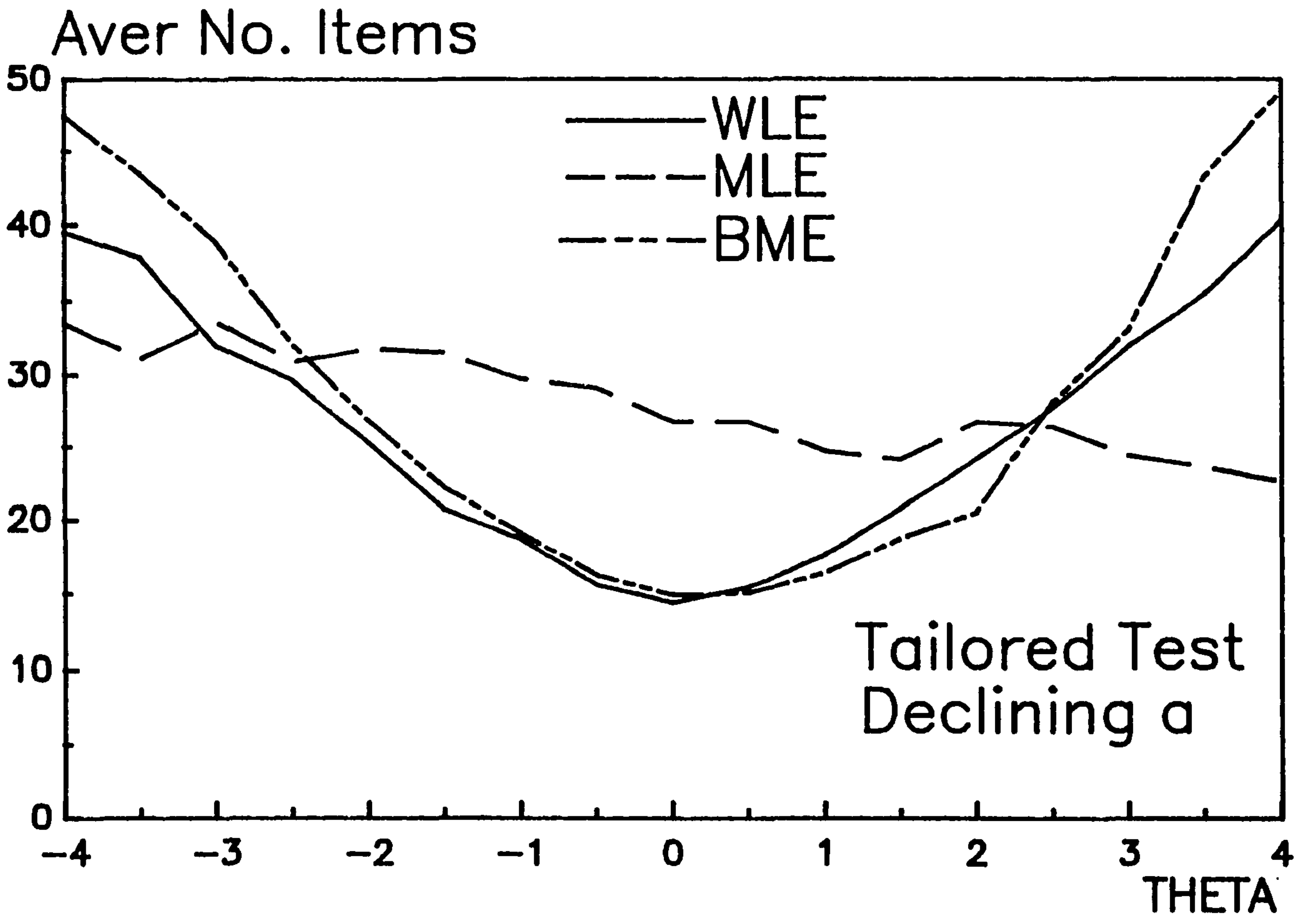

Figure 17
Average Number of Items Administered on Tailored Test with Declining a-parameter, Optimal b-parameter, and All c = 0.20 .

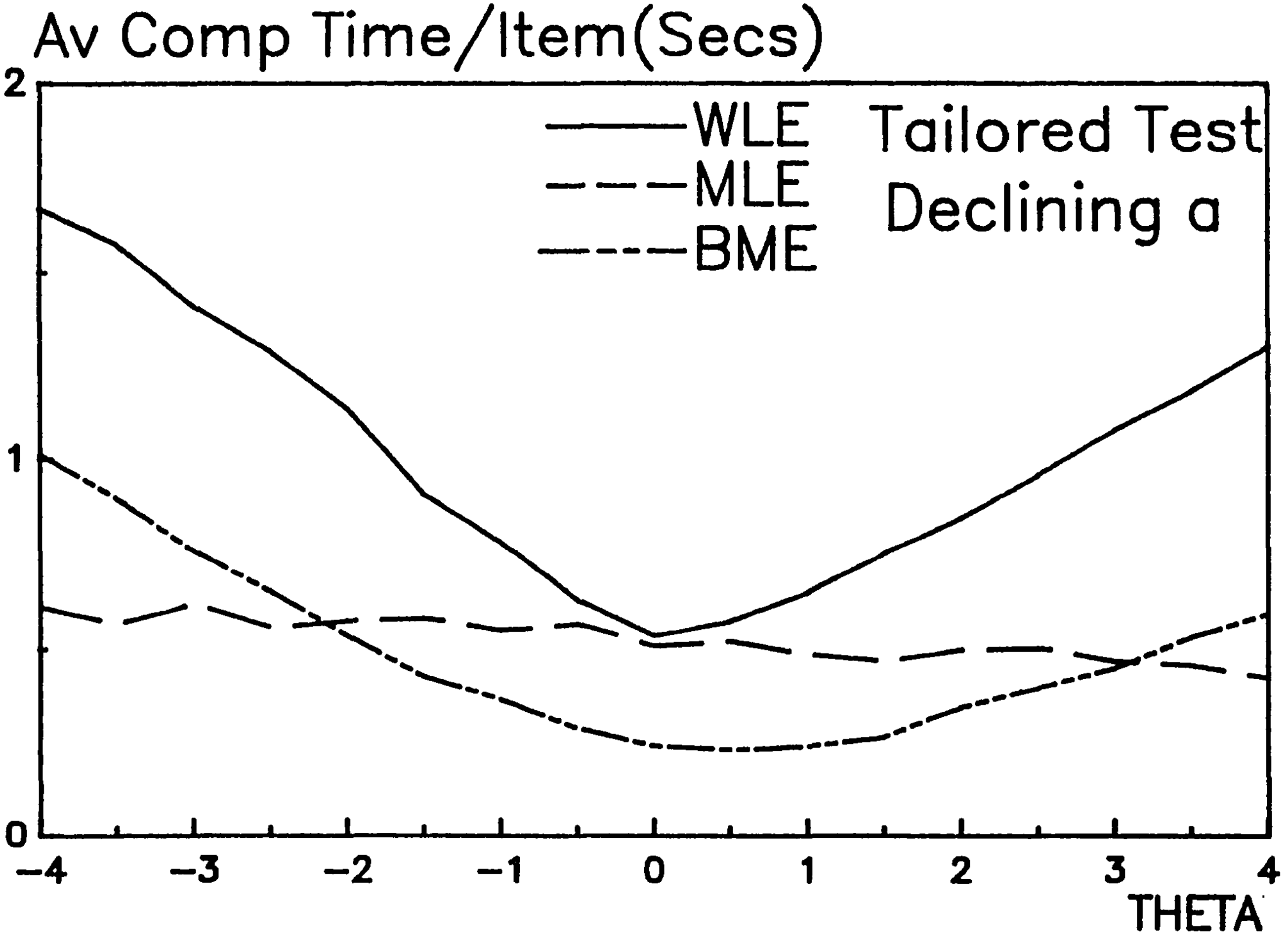

Figure 18
Average Computation Time Between Items on Tailored Test with Declining a-parameter, Optimal b-parameter, and All c = 0.20 .

items is displayed in Figure 18 (and D.39). For both tailored tests, WLE(θ) required more computation time than either of the other two estimators at all values of θ. Computation time for MLE was nearly constant on both tests and at all values of θ. BME used the least computation time of the three estimators on both tests at all except extreme values of θ.

# DISCUSSION

WLE(θ) is clearly a less biased estimator of θ than either MLE(θ) or BME(θ). In light of the proof in Appendix A and the results found by Warm (1985), this outcome should not be surprising in some circumstances. What is surprising is that the superiority of WLE(θ) was demonstrated in every condition investigated in this study: i.e., for test lengths ranging from 10 to 60 items; for a-parameters of $a = 1$ and $a = 2$; for all values of θ, and for conventional tests as well as for tailored tests (utilizing both infinite item banks and simulated finite item banks). Moreover, WLE(θ) has small and roughly constant variance over a wide range of the θ-scale, as well as small MSE over a much wider interval than either MLE(θ) or BME(θ).

For the conventional tests BME(θ) also fared better than MLE(θ) on all three criteria -- bias, SD, and MSE. However, BME(θ) benefitted greatly by the design of the simulated tests. The tests were designed to have roughly "normally" distributed test information. The general effect of this design was to bias estimates outwardly (digress estimates away from the peak of test information). BME(θ) tends to regress estimates toward the peak of the prior probability distribution. Since the peaks of test information for these tests were located near $\theta = 0$ (which was the peak of the standard normal prior), the digression of test information and the regression of the prior tended to cancel out each other. For differently designed tests or a different prior, where the peak of the prior is not nearly coincident with the peak of test information, BME($\theta$) would be much worse.

The width of the interval over which an estimator performs well is an important consideration. It is trivially simple to create an estimator that has small bias, variance, and MSE over a small interval of the $\theta$-scale. For example, consider an estimator, $\Omega$, which is defined as $\Omega = k$, where k is a constant. Then, used as an estimate of $\theta$, $\mathrm{BIAS}(\Omega) = (k - \theta)$, $\mathrm{VAR}(\Omega) = 0$, and $\mathrm{MSE}(\Omega) = (k - \theta)^2$. In the interval on the $\theta$-scale where $|k - \theta|$ is sufficiently small, $\Omega$ may be superior to any other estimator by all three criteria. Note that a nonlinear transformation of $\Omega$ is equally superior over a small interval of the nonlinearly transformed $\theta$-scale. The estimator $\Omega$ can be considered to be a Bayesian modal estimator for which the prior has non-zero density only where $\theta = k$. $\Omega$ compresses all estimates to the constant k, which has no functional relationship to the test. Other Bayesian modal estimators have the same characteristic -- they "regress" estimates toward the peak of a prior distribution, which also has no functional relationship to the test.

This analysis of the estimator $\Omega$ makes it clear why Bayesian modal estimators sometimes work well -- by coincidence, they sometimes "regress" estimates in a direction opposite to the bias of MLE($\theta$). Bayesian estimators may also, by coincidence, compress the estimates into an interval of the $\theta$-scale that is of interest; this is exactly the reason that BME($\theta$) had the smallest MSE for certain relatively small intervals. On the other hand, if the coincidences do not hold, then the "regression" of the Bayesian estimators will worsen estimation, rather than improve it. In general, BME($\theta$) will work well if the prior approximates w($\theta$), the weighting function for WLE($\theta$). For the tests in this study a normal prior with standard deviation of two would have approximated w($\theta$) much better than the standard normal prior. Since the tests in this study have more peaked test information curves than many actual tests, in practice the priors for BME($\theta$) should have standard deviations greater than two in order to approximate w($\theta$).

Apparently, Bayesians have discovered this fact by trial and error. Lord (1984) states that leading Bayesian testing practitioners prefer to use priors more diffuse than the standard normal prior due to practical considerations. Thus, it can be argued that <u>the appeal of BME($\theta$) in IRT is</u>

not due to the properties of the posterior distribution, but rather is due to the coincidence that in IRT, $w(\theta)$ can often be approximated by a diffuse normal curve.

WLE($\theta$) also "regresses" estimates. However, the regressing function, $w(\theta)$, is a function of the test, and is in no way arbitrary. Its action is always to regress estimates in the direction opposite of the bias of MLE($\theta$), and into an interval for which the test has sufficient information to reduce bias, variance, and MSE.

Pure MLE is everybody's whipping boy, (See, for example, Thissen & Wainer, 1983) because (even though it has very attractive asymptotic properties) its performance with tests of finite length is miserable by most criteria. The unbounded nature of MLE($\theta$) is one source of its difficulties with finite tests. It is well known that if the response vector $\underline{u} = \underline{1}$, (all items answered correctly), then MLE($\theta$) = $+\infty$, and if $\underline{u} = \underline{0}$ (all items answered incorrectly), then MLE($\theta$) = $-\infty$. The problem is basically that of mapping the finite set of ($2^n$) possible response vectors onto the (infinite) set of real numbers. Thus, in practical applications pure MLE($\theta$) is never used. Upper and lower bounds on MLE($\theta$) are always set. Even in theoretical work these bounds are necessary; e.g. in Lord's (1983a) derivation of the bias of MLE($\theta$), $\theta$ is assumed to be bounded. The practical bounds are set arbitrarily at values that are not expected to have a significant impact. It has not been recognized that there exist rational, natural upper and lower bounds for MLE(θ) in IRT. The rational, natural bounds occur at the points where the slope of BIAS(MLE(θ)) is equal to one. That is, MLE(θ) should be defined only in the interval(s) where

$$\theta^{-} <= \text{MLE}(\theta) <= \theta^{+} \;;$$

where $\theta^{-}$ is defined such that

$$\delta/\delta\theta \text{ BIAS}(\text{MLE}(\theta^{-})) = 1 \;,$$

and

$$\delta^{2}/\delta\theta^{2} \text{ BIAS}(\text{MLE}(\theta^{-})) < 0 \;;$$

and, furthermore, where $\theta^{+}$ is defined such that

$$\delta/\delta\theta \text{ BIAS}(\text{MLE}(\theta^{+})) = 1 \;,$$

and

$$\delta^{2}/\delta\theta^{2} \text{ BIAS}(\text{MLE}(\theta^{+})) > 0 \;.$$

Outside these bounds the slope of BIAS(MLE(θ)) is always greater than one, i.e. the magnitude of BIAS(MLE(θ)) increases faster than $\theta$ changes. Therefore, the expected value of MLE($\theta$) is always farther from true $\theta$ than is the closer of the two bounds, and the closer bound is a less biased estimate of $\theta$ than is MLE($\theta$). Moreover, the replacement of the closer bound for MLE($\theta$) as the estimate of $\theta$ reduces the variance of the estimates, and, consequently, also the MSE of the estimates. Thus, for all three of the criteria used in this study (bias, variance, and MSE), MLE($\theta$) is always degraded by permitting it to fall outside the closed interval $[\theta^-, \theta^+]$. These bounds are proposed as reasonable or "sensible" limits on MLE($\theta$). Therefore, MLE($\theta$) with these bounds may be termed "Sensible MLE($\theta$)", [SMLE($\theta$)].

---

Insert Table 1 about here.

---

Table 1 lists the minimum and maximum values of SMLE($\theta$), and the actual minimum and maximum values of WLE($\theta$) and BME($\theta$) for the 12 conventional tests. Note that the minima of WLE($\theta$) is always more extreme (more negative) than the minima of the other two estimators. In contrast, the maxima of SMLE($\theta$) are by far the most extreme. This extreme upper bound on SMLE($\theta$) would do little to reduce bias, but at least would prevent infinite estimates of $\theta$. The extraordinary asymmetry of the bounds of SMLE($\theta$) is apparently due to the loss of test information at low $\theta$, caused by the non-zero c-parameters.

TABLE 1

Minimum, MLE($\theta^-$), and Maximum, MLE($\theta^+$), of "Sensible" MLE($\theta$), and of WLE($\theta$) and BME($\theta$), for Conventional Tests with n items, All a = 1 or 2, Normally Distributed b, and All c = 0.20 .

| | a = 1 | | | a = 2 | | |
|---|---|---|---|---|---|---|
| | | Min | Min | | Min | Min |
| n | MLE($\theta^-$) | WLE($\theta$) | BME($\theta$) | MLE($\theta^-$) | WLE($\theta$) | BME($\theta$) |
| 10 | -1.907 | -2.543 | -1.769 | -1.685 | -2.002 | -1.763 |
| 20 | -2.284 | -2.989 | -2.175 | -2.028 | -2.337 | -2.068 |
| 30 | -2.486 | -3.242 | -2.410 | -2.208 | -2.519 | -2.237 |
| 40 | -2.623 | -3.418 | -2.574 | -2.329 | -2.642 | -2.354 |
| 50 | -2.725 | -3.553 | -2.701 | -2.419 | -2.735 | -2.444 |
| 60 | -2.807 | -3.664 | -2.803 | -2.491 | -2.809 | -2.515 |
| | | Max | Max | | Max | Max |
| n | MLE($\theta^+$) | WLE($\theta$) | BME($\theta$) | MLE($\theta^+$) | WLE($\theta$) | BME($\theta$) |
| 10 | 22.847 | 2.347 | 1.587 | 12.246 | 1.903 | 1.630 |
| 20 | 23.163 | 2.800 | 2.009 | 12.561 | 2.229 | 1.941 |
| 30 | 23.331 | 3.058 | 2.252 | 12.729 | 2.408 | 2.116 |
| 40 | 23.444 | 3.238 | 2.422 | 12.842 | 2.530 | 2.238 |
| 50 | 23.529 | 3.376 | 2.553 | 12.927 | 2.622 | 2.330 |
| 60 | 23.596 | 3.489 | 2.659 | 12.995 | 2.696 | 2.404 |

For both tailored tests, WLE($\theta$) was only slightly less biased than MLE($\theta$) (and at high $\theta$ only). For the tailored test with a simulated finite item bank, WLE($\theta$) and MLE($\theta$) were also about the same on the standard deviation and MSE criteria, although MLE($\theta$) may have some small advantage at low $\theta$. However, for the tailored test with an infinite item bank ($a = 2$), MLE($\theta$) was considerably worse than WLE($\theta$) over a wide central region of $\theta$ (using the standard deviation and MSE criteria). This effect is apparently due to the high values of the a-parameters, and would seem to be a conditional (on $\theta$) analogy to the "attenuation paradox" (Lord & Novick, 1967); i.e. the conditional variance of MLE($\theta$) increased (for central $\theta$) when the a-parameters increased from one to two. If so, then this result suggests that the selection of items for item banks so as to maximize item information may not be optimal for MLE($\theta$).

Since WLE($\theta$), over a wide range of $\theta$, used many fewer items than MLE($\theta$) in order to achieve the stopping criteria, tailored tests using WLE($\theta$) would, in general, be much shorter than if MLE($\theta$) were used to estimate $\theta$. This advantage translates into savings in terms of testing time and the exposure of items to potential compromise.

Computation time for estimating $\theta$ (in the interval between items in tailored testing) is an important consideration for tailored tests. The time required for WLE($\theta$) is always more than for MLE($\theta$) at all values of $\theta$ --

ranging from a slight increase in time to more than three times as long. This delay may or may not be significant, however. The times found in this study are only relative. The absolute times required for applications will depend upon several factors, such as the speed of the computer, the programming language, and the cleverness of the programmer. If the actual computation times can be decreased to one second or less, then no harm is caused by the extra calculations required for WLE(θ).

# SUMMARY AND CONCLUSIONS

A new method of estimation , Weighted Likelihood Estimation (WLE), was derived, and proved to yield asymptotically normally distributed estimates, with finite variance, and unbiased estimates to order $n^{-1}$. The unbiasedness of WLE(θ) is in contrast to Maximum Likelihood Estimation (MLE) and Bayesian Modal Estimation (BME), both of which are biased to order $n^{-1}$. The new estimator was applied to ability estimation in IRT. Using Monte Carlo methods, WLE(θ) was compared to MLE(θ) and BME(θ) on 12 conventional tests with 10 to 60 items, and a-parameters of 1 or 2. The three estimators were also compared on two tailored tests. One tailored test had an infinite item bank and all a = 2 . The other tailored test simulated a finite item bank with declining a-parameters.

In all tests WLE(θ) was less biased than both of the other estimators. In addition WLE(θ) had small variance over the entire range of the θ-scale, as well as small mean squared error even at non-central θ. The relative unbiasedness of WLE(θ) makes this estimator particularly appropriate in applications of Item Response Theory (IRT) for which the parameter invariance property is important.

Two new insights for MLE($\theta$) were discovered: 1) natural, rational bounds, and 2) a conditional analogy to the attenuation paradox in tailored tests with high a-parameters.

The heart of WLE is a weighting function, w($\theta$), which is multiplied times the likelihood function, and the product maximized. This weighting function, which removes the bias and uncontrolled variance of MLE($\theta$), is a function of $\theta$ and the item parameters, and is specific to each test. It was shown to be equal to the square root of test information for the one- and two-parameter models of IRT, and equal to a closely related function for the three-parameter model.

# REFERENCES

Anderson, J.A. & Richardson, S.C.(1979) Logistic Discrimination and Bias Correction in Maximum Likelihood Estimation. Technometrics, 21(1), 71-78.

Baker, F.B.(1984) Ability Metric Transformation Involved in Vertical Equating Under Item Response Theory. Applied Psychological Measurement, 8(3), 261-271.

Birnbaum, A.(1967) Some Latent Trait Models and Their Use in Inferring an Examinee's Ability(Part 5). In Lord, F.M. & Novick, M. Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA, 1967.

Bock, R.D.(1983) The Discrete Bayesian. In Wainer H., & Messick, S.(Eds.) Principals of Modern Psychological Measurement. A Festschrift for Frederick M. Lord. Lawrence Erlbaum Associates, Publishers. Hillsdale, NJ., 103-115.

Bock, R.D. & Aitkin, M.(1981) Marginal Maximum Likelihood

Estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 443-459.

Bock, R.D. & Mislevy, R.J.(1981) Biweight Estimates of Latent Ability. Manuscript.

Borland(1985) Turbo Pascal, Version 3.0 Reference Manual. Borland International Inc., Scotts Valley, CA 95066

Bradley, R.A. & Gart, J.J.(1962) The Asymptotic Properties of ML Estimators When Sampling From Associated Populations. Biometrika, 49, 205-214.

Cox, D.R. & Hinkley, D.V.(1974) Theoretical Statistics. Chapman & Hall (distributed by Halsted Press, New York).

Gorman, S.(1980) A Comparative Evaluation of Two Bayesian Adaptive Estimation Procedures with a Conventional Test Strategy. Unpublished Doctoral Dissertation. Catholic University of America. Washington D.C.

Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.L. & Rekase, M.D.(1984) Evaluation Plan for the Computerized Adaptive Vocational Aptitude Battery. MPL TN 85-1, Manpower and Personnel Laboratory, Navy Personnel

Research and Development Center, San Diego, CA 92152.

Gulliksen, H.(1950) Theory of Mental Tests. John Wiley & Sons, Inc., New York.

Hinkley, D.V.(1985) Professor, Department of Mathematics, The University of Texas at Austin, Austin TX 78712. Personal Communication.

Jensema, C.J.(1977) Bayesian Tailored Testing and the Influence of Item Bank Characteristics. Applied Psychological Measurement, 1(1), 111-120.

Jones, D.H.(1982) Redescending M-type Estimators of Latent Ability. Program Statistics Technical Report, 82-30. Educational Testing Service, Princeton, NJ.

Kendall, M.G. & Stuart,A.(1973) The Advanced Theory of Statistics, vol 2. Hafner Publishing Company, New York.

Kendall, M.G. & Stuart,A.(1977) The Advanced Theory of Statistics, vol 1. Charles Griffin & Co., London.

Kullback, S.(1978) Information Theory and Statistics. Peter Smith, Gloucester, MA.

Larson, R.J. & Marx, M.L.(1981) An Introduction to Mathematical Statistics and Its Applications. Prentice Hall, Inc., Englewood Cliffs, NJ 07632.

LeCam, L.(1953) On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes' Estimates. University of California Publications in Statistics, 1(11), 277-330.

Lord, F.M.(1980) Applications of Item Response Theory to Practical Testing Problems. Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ 07642.

Lord, F.M.(1983a) Unbiased Estimators of Ability Parameters, of Their Variance, and of Their Parallel-Forms Reliability. Psychometrika, 48(2), 233-245.

Lord, F.M.(1983b) Memorandum for: Ms. Stocking, Ms. M. Wang, Ms. Wingersky. Subject: Sampling Variance and Bias for MLE and Bayesian Estimation of θ. August 26, 1983. Internal Memorandum, Educational Testing Service, Princeton, NJ 08541.

Lord, F.M.(1983c) Distinguished Research Scientist, Educational Testing Service, Princeton, NJ 08541. Personal Communication.

Lord, F.M.(1984) Maximum Likelihood and Bayesian Parameter Estimation in Item Response Theory. Research Report RR-84-30-ONR, Educational Testing Service, Princeton, NJ 08541.

Lord, F.M. & Novick, M.(1967) Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA.

Maurelli, V.A.(1978) A Comparison of Bayesian and Maximum Likelihood Scoring in a Simulated Stradaptive Test. Unpublished Master's Thesis. St. Mary's University. San Antonio, TX.

McBride, J.R.(1977) Some Properties of a Bayesian Adaptive Ability Testing Strategy. Applied Psychological Measurement, 1(1), 121-140.

McKinley, R.L., & Reckase, M.D.(1981) A Comparison of a Bayesian and a Maximum Likelihood Tailored Testing Procedure. Research Report 81-2. Educational Psychology Department. University of Missouri. Columbia, MO 65211.

Quenouille, M.H.(1956) Notes on Bias in Estimation. Biometrika, 43, 353-360.

Samejima, F.(1980) Is Bayesian Estimation Proper For Estimating the Individual's Ability. Research Report 80-3, Dept. of Psychology, University of Tennessee, Knoxville, TN 37916.

Schaefer, R.L.(1983) Bias Correction in Maximum Likelihood Logistic Regression. Statistics in Medicine, 2,71-78.

Shenton, L.R. & Bowman, K.(1963) Higher Moments of a Maximum-likelihood Estimate. Journal of the Royal Statistical Society, Series B, 25, 305-317.

Thissen, D. & Steinberg L.(1984) A Response Model for Multiple Choice Items. Psychometrika, 49(4), 501-519.

Thissen, D. & Wainer, H.(1982) Some Standard Errors in Item Response Theory. Psychometrika, 47(4), 397-412.

Vale, C.D., Maurelli,V.A., Gialluca,K.A., Weiss, D.J., & Ree, M.J.(1981) Methods For Linking Item Parameters. Technical Report AFHRL-TR-81-10. Air Force Human Resources Laboratory, Brooks Air Force Base, TX 78235.

Wainer, H. & Wright, B.D.(1980) Robust Estimation of Ability in the Rasch Model. Psychometrika, 45(3).

Warm, T.A.(1979) A Primer of Item Response Theory. Technical Report 940279. National Technical Information Service, AD-A063072. Springfield, VA 22161.

Warm, T.A.(1985) Weighted Likelihood Estimation of Ability in Item Response Theory. Technical Report CGI-TR-85-01. U.S. Coast Guard Institute. Oklahoma City OK 73169. Submitted for publication.

Weiss, D.J. & McBride, J.R.(1984) Bias and Information of Bayesian Adaptive Testing. Applied Psychological Measurement, 8(3), 273-285.

# APPENDIX A

## PROOF THAT THE WEIGHTED LIKELIHOOD ESTIMATE IS UNBIASED TO ORDER $n^{-1}$

The approach and techniques of this derivation were taken from, and parallel closely, the derivations in Lord(1983a, 1983b, & 1984) of the first order biases of the Maximum Likelihood and Bayesian Modal Estimates in Item Response Theory (See Lord,1980), both of which biases are of order $n^{-1}$. The Weighted Likelihood Estimator removes the first order bias term from the Maximum Likelihood estimate. The derivation is limited to a single parameter for a multinomially distributed variable and a regular, "smooth" mathematical model with rather restrictive assumptions. Apparently, however, this "... method of removing the first bias term will work in complete generality, and can be extended to any form of consistent estimating equation where the mathematical form of bias is computable." (Hinkley, 1985, Personal Communication)

<u>Preliminaries</u>: For a set of n independent experiments,

$H_i(i = 1,2,---,n)$ with binary outcomes, $u_i$, (success or failure), let $P = P_i(\theta)$ denote the probability of a success($u_i = 1$), and let $Q = Q_i(\theta) = 1 - P_i(\theta)$ denote the probability of failure($u_i = 0$), where $P_i(\theta)$ is a strictly increasing function of the common parameter $\theta$ for all n experiments. $P_i(\theta)$ is not necessarily equal to $P_h(\theta)$, $h$ =/= $i$. Let $\underline{u} = \{u_i\}$ denote the multinomially distributed, n x 1 vector of outcomes of the n experiments.

Assumptions:

(a) $\theta$ is a bounded variable on a continuous scale.

(b) $P_i(\theta)$ is continuous and bounded away from 0 and 1 at all values of $\theta$, $i=1,2,---,n$.

(c) At least the first five derivatives with respect to $\theta$ of $P_i(\theta)$ exist at all values of $\theta$, and are bounded.

(d) For asymptotic considerations n is considered to be incremented with replications of all of the original n experiments.

From these assumptions and theorems 1(i) and 1(iv) of Bradley & Gart(1962) it follows that the Maximum Likelihood Estimate of $\theta$, MLE($\theta$) = $\theta$^, is a consistent estimator of $\theta$, and that $n^{1/2}\cdot(\hat{\theta} - \theta)$ is asymptotically normally distributed with zero mean and with variance given by

$$\lim_{n \to \infty} 1/(nI) \quad ,$$

where I is Fisher's Information. Assumption (d) guarantees the existence of this limit (Lord, 1983a).

## Maximum Likelihood Estimation

The likelihood function, $L(\underline{u}|\theta)$, is given by (A.1).

$$L(\underline{u}|\theta) = \prod_{i=1}^{n} P_i(\theta)^{u_i} \cdot Q_i(\theta)^{1-u_i} \qquad (A.1)$$

Let

$$l_s = \delta^s/\delta\theta^s \ln L(\underline{u}|\theta),$$

where $\delta^s/\delta\theta^s$ indicates the $s^{th}$ partial derivative with respect to $\theta$, and ln indicates the natural logarithm.

The Maximum Likelihood Estimate of $\theta$ is defined as the value of $\theta$ that maximizes (A.1). Usually $\theta$^ is found by setting $l_1$ equal to zero, and solving for $\theta$, as in (A.2).

$$\delta/\delta\theta \ln L(\underline{u}|\theta) \equiv l_1 = \Sigma(u - P)P'/PQ = 0 \quad , \qquad (A.2)$$

evaluated at $\theta$^. In (A.2) and hereafter the argument ($\theta$) and index i are usually dropped for convenience.

The asymptotic variance of MLE($\theta$) is the reciprocal of Fisher's Information (Kendall & Stuart, 1973, p. 10).

$$I = E(l_1^2) = -E(l_2) = \Sigma P'^2/PQ \quad , \qquad (A.3)$$

where E is the expectation operator, and $P' = \delta/\delta\theta\ P$.

The bias of MLE(θ), BIAS(MLE(θ)), from Cox & Hinkley (1974) is given by (A.4).

$$\mathrm{BIAS}(\mathrm{MLE}(\theta)) \equiv E(\theta\hat{} - \theta) = -J/2I^2 \quad , \qquad (A.4)$$

where

$$J = -2E(l_1 l_2) - E(l_3) = \Sigma P'P''/PQ \quad , \qquad (A.5)$$

and $P'' = \delta^2/\delta\theta^2\ P$ .

Equation (A.4) is equivalent to Lord's (1983a) equation (28) for BIAS(MLE(θ)). Note that I and J are of order n, and that since neither are a function of $\underline{u}$, $J/2I^2$ is of order $n^{-1}$.

## Weighted Likelihood Estimation

The Weighted Likelihood Estimate of θ, WLE(θ) $\equiv \theta^*$, is defined as the value of θ, such that the Weighted

Likelihood Function, given in (A.6) is maximized,

$$w(\theta) \cdot L(\underline{u}|\theta) \quad , \qquad \text{(A.6)}$$

where $\delta/\delta\theta \ \ln w(\theta) = J/2I$. WLE(θ) is found by solving the weighted likelihood equation as in (A.7),

$$l_1 + J/2I = 0 \quad , \qquad \text{(A.7)}$$

evaluated at θ*, or, letting $d_s = \delta^s/\delta\theta^s \ \ln w(\theta)$, as in (A.8).

$$l_1 + d_1 = 0 \qquad \text{(A.8)}$$

Note that

$$d_1 = J/2I = -I \cdot \text{BIAS}(\text{MLE}(\theta)) \quad .$$

Rather than finding WLE(θ) by maximizing (A.6), it will be useful to maximize the $n^{th}$ root of (A.6), which will always yield the same estimate for any given set of data since n is always positive. The reason for doing so is to help keep track of the order of the terms. Letting $T_s$ = the $s^{th}$ derivative of the log of $n^{th}$ root of (A.6), and $T_s^* = T_s$, evaluated at θ*,

$$T_s = \delta^s/\delta\theta^s \ \ln [\ w(\theta) \cdot L(\underline{u}|\theta)\ ]^{1/n}$$

$$= l_1/n + d_1/n \qquad (A.9)$$

<u>THEOREM</u>: WLE($\theta$) is unbiased to order $n^{-1}$, i.e.

$$\text{BIAS}(\text{WLE}(\theta)) = 0 + o(n^{-1}) \quad ,$$

where $o(n^{-r})$ represents terms such that

$$\lim_{n\to\infty} n^r \cdot o(n^{-r}) = 0 \quad .$$

These are of order higher than $n^{-r}$ (i.e. $n^{-r-1}$, $n^{-r-3/2}$, $n^{-r-2}$ etc.).

<u>PROOF</u>:

It is sufficient to solve the $n^{th}$ root of (A.6):

$$T_1^* = l_1/n + d_1/n = 0 \quad , \qquad (A.10)$$

evaluated at $\theta^*$. Letting $x = (\theta^* - \theta)$, expand (A.10) in terms of x and θ.

$$T_1^* = 0$$

$$= T_1 + xT_2 + ½x^2T_3 + x^3T_4/6 + \tau x^4V_5/24, \qquad (A.11)$$

where $V_5 = \mathrm{Max}(T_5)$ over θ, and $|\tau| < 1$. This closed form of the expansion is always valid, making the proof of the convergence of the Taylor series unnecessary. Letting

$$g_s = El_s/n,$$

and

$$e_s = l_s/n - g_s,$$

then

$$l_s/n = g_s + e_s \quad ,$$

and

$$T_s = g_s + e_s + d_s/n . \qquad (A.12)$$

The purpose of (A.12) is to separate $T_s$ into a sum of terms not containing $(u - P)$, $g_s$, a sum of terms containing $(u - P)$, $e_s$, and a term that is a ratio of sums not containing $(u - P)$, $d_s/n$.

Substituting (A.12) into (A.11), $s = 1, 2, 3, 4$, expresses the expansion of the weighted likelihood equation in terms of $g_s$, $e_s$, $d_s/n$, $V_5$, and powers of $x$ in (A.13).

$$
\begin{aligned}
-(e_1 + d_1/n) = \; & g_1 \\
& + x(g_2 + e_2 + d_2/n) \\
& + ½x^2(g_3 + e_3 + d_3/n) \\
& + x^3(g_4 + e_4 + d_4/n)/6 \\
& + \tau x^4 V_5/24 \quad . \qquad \text{(A.13)}
\end{aligned}
$$

We now need to evaluate some of the terms in (A.13), and their expected values.

$$g_1 = 0 \qquad \text{(A.14)}$$

$$g_2 = -n^{-1} \cdot \Sigma P'^2/PQ = -I/n \qquad \text{(A.15)}$$

$$g_3 = (-3J + 2K)/n \quad , \qquad \text{(A.16)}$$

where

$$K = \Sigma(1-2P)P'^3/(PQ)^2 = -3E(l_1l_2) - E(l_3) \quad .$$

$$d_1 = J/2I \tag{A.17}$$

$$d_2 = (IJ' - I'J)/2I^2 \tag{A.18}$$

$$d_3 = (I^2J'' - II''J - 2II'J' + 2I'^2J)/2I^3 \tag{A.19}$$

where (') and (") indicate first and second derivatives with respect to θ, respectively.

Since $g_s$ and $d_s$ do not contain (u - P),

$$Eg_s = g_s \quad , \tag{A.20}$$

and

$$Ed_s = d_s \quad . \tag{A.21}$$

$$e_1 = n^{-1}\cdot\Sigma(u-P)P'/PQ \tag{A.22}$$

$$e_2 = n^{-1}\cdot\Sigma \{(u-P)\cdot[(P''/PQ)-(1-2P)P'^2/(PQ)^2]\} \tag{A.23}$$

$$Ee_s = 0 \qquad (A.24)$$

Since $e_s$ is $n^{-1}$ times the sum of the terms of $l_s$ that contain $(u - P)$, $e_s$ may be expressed as

$$e_s = n^{-1} \cdot \Sigma \, (u-P) \cdot R_{si} \quad , s = 1, 2, 3, 4, 5,$$

where $R_{si}$ is the $(s - 1)^{th}$ derivative of $P'/PQ$, and does not depend on n nor on $u_i$. Since by assumption (b) P and Q are bounded, and by assumption (c) the required derivatives of P are bounded, the $R_{si}$ and, thus $e_s$, are bounded. By assumption (d) the bound does not depend on n. The same conclusion is true of $g_s$.

Since by assumption (d) $l_1/n$ in (A.10) is of order $n^0$, and $d_1/n$ is of order $n^{-1}$, (A.2) and (A.7) are asymptotically equivalent, and, asymptotically, $n^{½}(\theta^*-\theta) = n^{½}(\theta\hat{}-\theta)$. Because $n^{½}(\theta\hat{}-\theta)$ is asymptotically normally distributed with zero mean and finite variance, so is $n^{½}(\theta^*-\theta)$. Therefore, $Ex^r$ $(r=1,2,\cdots)$ is of order $n^{-r/2}$. By similar logic $e_s{}^r$ is of the same order. By the Cauchy-Schwartz inequality $Ex^r \cdot e_s{}^t <= (Ex^{2r} \cdot e_s{}^{2t})^{½}$, and therefore $Ex^r e_s{}^t$ is of order $n^{-(r+t)/2}$ $(r,t = 1,2,\cdots)$.

The Variance of WLE(θ).

To get the variance of θ*, VAR(θ*), square (A.13) and take expectations.

$$Ee_1^2 + 2d_1Ee_1/n + d_1^2/n^2 = g_2^2Ex^2$$

$$+ \quad g_2Ex^2e_2 + \tfrac{1}{2}g_2g_3Ex^3$$

$$+ Ex^2e_2^2 + \tfrac{1}{2}g_2Ex^3e_3 + \tfrac{1}{2}g_2g_3Ex^3 + \tfrac{1}{4}g_3Ex^4 + \cdots \qquad (A.25)$$

The terms in the first line of (A.25) are of order $n^{-1}$, in the second line of order $n^{-3/2}$, and in the third line of order $n^{-2}$ with the remaining terms $o(n^{-2})$. Dropping all terms of $o(n^{-1})$, we can rewrite (A.25) as

$$Ee_1^2 + 2d_1Ee_1/n + d_1^2/n^2 = g_2^2Ex^2 + o(n^{-1}) \quad .$$

Since $Ee_1^2 = I/n^2$, $Ee_1 = 0$, $g_2^2 = (-I)^2/n^2$, and $Ex^2 = \text{VAR}(\theta^*)$, (A.25) evaluates as (A.26).

$$(I + d_1^2)/n^2 = (I^2/n^2)\text{VAR}(\theta^*) + o(n^{-1}). \qquad (A.26)$$

Solving for VAR(θ*) gives (A.27),

$$VAR(\theta^*) = (I + d_1^2)/I^2 + o(n^{-1})$$

$$= I^{-1} + o(n^{-1}) \quad , \qquad (A.27)$$

which proves that the asymptotic variance of WLE($\theta$) is equal to the asymptotic variance of MLE($\theta$).

<u>The Bias of WLE($\theta$).</u>

Let $E_1$ indicate the expectation operator in which only terms of order $n^{-1}$ are retained. To get the first order statistical bias of $\theta^*$ take the first order expectation of (A.13) to obtain (A.28).

$$-d_1/n = g_2E_1x + E_1xe_2 + \tfrac{1}{2}g_3E_1x^2 \qquad (A.28)$$

To evaluate $E_1xe_2$ multiply (A.13) by $e_2$, and take first order expectations.

$$-E_1e_1e_2 = g_2E_1xe_2 \qquad (A.29)$$

To evaluate the LHS of (A.29) substitute (A.22) and (A.23), and take the expectation. Both (A.22) and (A.23) are sums of n terms indexed with i, each containing the factor $(u_i - P_i)$. The product is a double sum of $n^2$ terms, each, the product of a term in (A.22), indexed with i, and a term in (A.23), indexed with i', say. Because the n experiments, $H_i$, are independent, the expected value of all terms are equal to zero, except the n terms where i = i'. Noting from (A.16) that

$$K = \Sigma\ (1 - 2P)P'^3/(PQ)^2,$$

then

$$-E_1e_1e_2 = (-J + K)/n^2 \qquad (A.30)$$

Substituting (A.15) and (A.30) into (A.29), and solving for $E_1xe_2$ gives

$$E_1xe_2 = (J - K)/nI \quad . \qquad (A.31)$$

Substituting (A.15), (A.16), (A.17), (A.27), and (A.31) into (A.28) obtains

$$-J/2nI = (-I/n)Ex + (J-K)/nI + \tfrac{1}{2}(-3J + 2K)/nI \quad . \quad (A.32)$$

Finally, solving (A.32) for Ex,

$$Ex = E(\theta^* - \theta) = 0 + o(n^{-1}) \quad , \qquad (A.33)$$

which completes the proof.

It is interesting to note that, if the mathematical model is such that

$$P'' = P'' \cdot \delta/\delta\theta \,(\ln PQ)$$

as in Item Response Theory when $c_i = 0$, all i, then

$$E(l_1 l_2) = 0 \quad ,$$

$$J = K = \delta/\delta\theta \; I = -E(l_3) \quad ,$$

and

$$w(\theta) = I^{\frac{1}{2}} \quad .$$

Otherwise, $w(\theta) = I^{\frac{1}{2}} \cdot \exp(\tfrac{1}{2}\int K/I \;\delta\theta)$ for which there is no closed form solution for the indefinite integral.

## APPENDIX B

## SAS PROGRAM FOR CONVENTIONAL TEST MONTE CARLO STUDY

```
//    EXEC SAS824,REGION=3000K
 /*********************************************************/
 /*  THIS PROGRAM PERFORMS A MONTE CARLO COMPARISON OF    */
 /*  MAXIMUM LIKELIHOOD, WEIGHTED LIKELIHOOD, AND         */
 /*  BAYESIAN MODAL ESTIMATES OF THETA IN ITEM RESPONSE   */
 /*  THEORY.  THE NUMBER OF ITEMS ARE SET IN THE          */
 /*  "%LET N = " LINE, AND THE COMMON A- AND C-PARAMETERS */
 /*  IN THE 2 LINES BELOW IT.  THE B-PARAMETERS ARE       */
 /*  NORMALLY DISTRIBUTED.  1000 THETA ESTIMATES OF EACH  */
 /*  ESTIMATOR ARE MADE AT 17 VALUES OF THETA, USING THE  */
 /*  SAME ITEM RESPONSES FOR EACH ESTIMATOR.              */
 /*                  THOMAS A. WARM                       */
 /*              UNIVERSITY OF OKLAHOMA                   */
 /*                   JULY 30,1985                        */
 /*********************************************************/
%GLOBAL N  NZ  NR TA TC    ;
%LET N = 40 ;
%LET TA = 2.0 ;
```

```
%LET TC = .20 ;


PROC MATRIX ;    N = &N ;
                    /* CREATE VECTORS OF ITEM PARAMETERS */
A = J(1,N, &TA)  ;
B = PROBIT(((1:&N) - .5)#/&N)  ;
C = J(1,N, &TC) ;
CNP = 'A' 'B' 'C' ; PAR=A'||B'||C' ;
NN = 1000 ;                   /* 1000 SIMULATED EXAMINEES */
REPL = 1 ;
IF (N GT 30 ) THEN DO; NN = 500 ;  REPL = 2 ; END;
J1N = J(1,N) ;   JNN1 = J(NN,1) ;
CN1 = 'Z' 'LZ' 'WZ' 'BZ' 'ERLZ' 'ERWZ' 'ERBZ' 'ERLZ2'
     'ERWZ2' 'ERBZ2' 'INFZ' 'BIASMLEZ' 'T' 'ERLT'
     'ERWT' 'ERBT';                          /* Z MEANS THETA */
DO IZ = 1 TO 17;
TZ = J(NN,1,(IZ-9)#/2);                   /* TZ = TRUE THETA  */
                /* COMPUTE TEST INF AND BIAS(MLE(THETA)) */
P = 1 + EXP( (TZ*J1N - JNN1*B)#(JNN1*(-1.7#A ) ) ) ;
P = JNN1*C + (JNN1*(1-C))#/P ;
P1 = P(1,); T = J(NN,1, (P1(,+)) )  ;
DP1 = 1.7 # &TA # (P1 - &TC) # (1 - P1) #/ ((1) - (&TC)) ;
INFZ = ( DP1 # DP1 #/ (P1 # (1 - P1)))(,+) ;
DP2 = 1.7 # &TA # ((1) + (&TC) - (2#P1)) # DP1 #/
      ((1) - (&TC)) ;
BIASMLEZ = -((DP1#DP2#/(P1#(1-P1)))(,+))#/(2#INFZ#INFZ) ;
```

```
INFZ = J(NN,1,INFZ) ;

BIASMLEZ = J(NN,1,BIASMLEZ) ;

                                        /* BEGIN MONTE CARLO STUDY */


DO I2 = 1 TO REPL ;

          /* STARTING VALUES OF ESTIMATES ARE TRUE THETA */

LZ = TZ ;  WZ = TZ;  BZ = TZ;

              /* MAKE MATRIX OF SCORED ITEM RESPONSES, U */

U = (UNIFORM(J(NN,N,O)) LE P)  ;

                                   /* MAXIMUM LIKELIHOOD ESTIMATES */

LINK MLEZ ;   ERLZ = LZ - TZ ;  ERLZ2 = ERLZ##2 ;

PH = 1 + EXP( (LZ*J1N - JNN1*B)#(JNN1*(-1.7#A ) ) ) ;

ERLT = (JNN1*C + (JNN1*(1-C))#/PH)(,+) - T ;

                                  /* WEIGHTED LIKELIHOOD ESTIMATES */

LINK WLEZ ;   ERWZ = WZ - TZ ;  ERWZ2 = ERWZ##2 ;

PH = 1 + EXP( (WZ*J1N - JNN1*B)#(JNN1*(-1.7#A ) ) ) ;

ERWT = (JNN1*C + (JNN1*(1-C))#/PH)(,+) - T ;

                                       /* BAYESIAN MODAL ESTIMATES */

LINK BMEZ ;   ERBZ = BZ - TZ ;  ERBZ2 = ERBZ##2 ;

PH = 1 + EXP( (BZ*J1N - JNN1*B)#(JNN1*(-1.7#A ) ) ) ;

ERBT = (JNN1*C + (JNN1*(1-C))#/PH)(,+) - T  ;

W =TZ¦¦LZ¦¦WZ¦¦BZ¦¦ERLZ¦¦ERWZ¦¦ERBZ¦¦ERLZ2¦¦ERWZ2¦¦ERBZ2 ;

W = W¦¦INFZ¦¦BIASMLEZ¦¦T¦¦ERLT¦¦ERWT¦¦ERBT ;

OUTPUT W OUT=W COLNAME=CN1 ;

END ;  END ;
```

```
            /* MAXIMUM LIKELIHOOD ESTIMATON SUBROUTINE */
STOP ;   MLEZ:
DO I1 = 1 TO 15 ;
PH = 1 + EXP( (LZ*J1N - JNN1*B)#(JNN1*(-1.7#A) ) ) ;
PH = JNN1*C + (JNN1*(1-C))#/PH ;
DPH1 = (PH - JNN1*C)#(1 - PH)#(JNN1*(1.7#A#/(1-C)) );
SL = ((U-PH)#DPH1#/(PH#(1-PH)))(,+) ;
IF (I1 LE 4) THEN DELTA = 1 ; ELSE DELTA = DELTA #/ 2;
LZ = LZ + DELTA#SIGN(SL) ;
END ;
RETURN ;


           /* WEIGHTED LIKELIHOOD ESTIMATION SUBROUTINE */
STOP ;   WLEZ:
DO I1 = 1 TO 15 ;
PH = 1 + EXP( (WZ*J1N - JNN1*B)#(JNN1*(-1.7#A) ) ) ;
PH = JNN1*C + (JNN1*(1-C))#/PH  ;
DPH1 = (PH - JNN1*C)#(1 - PH)#(JNN1*(1.7#A#/(1-C)) );
DPH2 = (JNN1*(1 + C) - 2#PH)#DPH1#(JNN1*(1.7#A#/(1-C))) ;
INFW =  (DPH1#DPH1#/(PH#(1-PH)))(,+) ;
JNFW =  (DPH1#DPH2#/(PH#(1-PH)))(,+) ;
SW = ((U-PH)#DPH1#/(PH#(1-PH)))(,+) + JNFW#/(2#INFW) ;
IF (I1 LE 4) THEN DELTA = 1 ; ELSE DELTA = DELTA #/ 2;
WZ = WZ + DELTA#SIGN(SW) ;
END ;
RETURN ;
```

```
                 /* BAYESIAN MODAL ESTIMATION SUBROUTINE */
STOP ;   BMEZ:
DO I1 = 1 TO 15 ;
PH = 1 + EXP( (BZ*J1N - JNN1*B)#(JNN1*(-1.7#A) ) ) ;
PH = JNN1*C + (JNN1*(1-C))#/PH  ;
DPH1 = (PH - JNN1*C)#(1 - PH)#(JNN1*(1.7#A#/(1-C)) );
SB = ((U-PH)#DPH1#/(PH#(1-PH)))(,+) - BZ ;
IF (I1 LE 4) THEN DELTA = 1 ; ELSE DELTA = DELTA #/ 2;
BZ = BZ + DELTA#SIGN(SB) ;
END ;
RETURN ;

                                /* PRINT AND PLOT OUTPUT */
PROC SORT DATA=W  ;  BY  Z T INFZ BIASMLEZ;
TITLE Z=TRUE THETA, LZ=MLE(O), WZ=WLE(O), BZ=BME(O),
      INFZ=INF(O), N=&N, A=&TA, C=&TC, B= -2, 2;

PROC UNIVARIATE DATA=W PLOT;   BY  Z T INFZ BIASMLEZ;
VAR WZ LZ BZ;
OUTPUT OUT=W3 MEAN=AVWZ AVLZ AVBZ  STD=SDWZ SDLZ SDBZ
        MIN=MINWZ MINLZ MINBZ MAX=MAXWZ MAXLZ MAXBZ
               KURTOSIS=KURTWZ KURTLZ KURTBZ
               SKEWNESS=SKEWWZ SKEWLZ SKEWBZ    N=NN;

PROC UNIVARIATE DATA=W NOPRINT;   BY  Z INFZ BIASMLEZ;
```

```
     VAR ERLZ ERWZ ERBZ ERWT ERLT ERBT ERWZ2 ERLZ2 ERBZ2;
OUTPUT OUT=W4
  MEAN=AVERWZ AVERLZ AVERBZ AVERWT AVERLT AVERBT MSEWZ
       MSELZ MSEBZ
  STD=SDERWZ SDERLZ SDERBZ SDERWT SDERLT SDERBT
  MIN=ERMINWZ ERMINLZ ERMINBZ  MAX=ERMAXWZ ERMAXLZ ERMAXBZ
        N=NN ;

PROC PRINT DATA=W3 ;  PROC PRINT DATA=W4 ;

PROC PLOT DATA = W4 ;
     PLOT AVERWZ*Z='W' AVERLZ*Z='L' AVERBZ*Z='B'/OVERLAY ;
     PLOT AVERWT*Z='W' AVERLT*Z='L' AVERBT*Z='B'/OVERLAY ;
     PLOT SDERWZ*Z='W' SDERLZ*Z='L' SDERBZ*Z='B'/OVERLAY ;
     PLOT SDERWT*Z='W' SDERLT*Z='L' SDERBT*Z='B'/OVERLAY ;
     PLOT MSEWZ*Z='W'  MSELZ*Z='L'  MSEBZ*Z='B' /OVERLAY ;
  PLOT ERMINWZ*Z='W' ERMINLZ*Z='L' ERMINBZ*Z='B' /OVERLAY ;
  PLOT ERMAXWZ*Z='W' ERMAXLZ*Z='L' ERMAXBZ*Z='B' /OVERLAY ;

TITLE Z=TRUE(O), T=TRUE SCORE, LZ=MLE(O), WZ=WLE(O),
      BZ=BME(O), INFZ=INF(O), N=&N, A=&TA, C=&TC, B= -2, 2;

PROC PLOT DATA=W3 ;
     PLOT SKEWWZ*KURTWZ='W' SKEWLZ*KURTLZ='L'
          SKEWBZ*KURTBZ='B'/OVERLAY;
```

## APPENDIX C

## PASCAL PROGRAM FOR TAILORED TEST MONTE CARLO STUDY

```
PROGRAM W1tt1(INPUT,OUTPUT);
{$C-,U+}
{This program performs a Monte Carlo comparison  of  Maximum
Likelihood,   Weighted   Likelihood,   and   Bayesian  Modal
estimates of theta in Tailored Tests. 100 estimates of theta
are made at each of 17 values  of  theta.  The  c-parameters
= .2,  a-parameters  decline  from 2.0 in increments of 1/35
with each item administered,  and b-parameters are chosen to
maximize item information for the current estimate of theta,
given the a- and c- parameters.
                    Thomas A. Warm
                University of Oklahoma
                     July 30, 1985                        }
type
     TimeString = string[8];
VAR fileA,fileB  :  TEXT ;
    itavzh,avnit,a,b : ARRAY[1..101] OF REAL;
    u,nnit  :  ARRAY[1..101] OF INTEGER  ;
   sumn,d3p,avzh,avzh2,sdzh,djnf,dinf,d2lnw,sumzh3,sumzh4,
       skewzh,kurtzh: REAL ;
   msezh,time,z,zh,c,p,pq,dp,d2p,dlnL,dlnw,TestInf,sumzh,
       sumzh2 :  REAL;
   n,nn,maxn,i1,i2,i3,i4,method,iz  INTEGER;
   meth : STRING[3] ;

PROCEDURE Initialize1; BEGIN
      ASSIGN(fileA,'wltt1a.prt') ;
      REWRITE(fileA) ;             (* APPEND(fileA)  ;  *)
      ASSIGN(fileB,'b:wltt1b.prt') ;
      REWRITE(fileB) ;             (* APPEND(fileB)  ;  *)
      RANDOMIZE;
  END;

PROCEDURE Initialize2; BEGIN
      nn :=  100    ;
      maxn := 50 ;
      c := 0.2 ;
      sumn := 0.0  ;
      sumzh := 0.0 ;
      sumzh2 := 0.0 ;
      sumzh3 := 0.0 ;
      sumzh4 := 0.0 ;
      FOR i2 := 1 TO 101 DO BEGIN
         avnit[i2] := 0 ;
         nnit[i2]  := 0 ;
         itavzh[i2] := 0 ;
      END  ;
   END ;



FUNCTION realtime: REAL  ;
   TYPE
      regpack = RECORD
           ax,bx,cx,dx,bp,si,di,ds,es,flags: INTEGER;
   END;
   VAR
      recpack:          regpack;            {assign record}
      ah,al,ch,cl,dh:   BYTE;
      hour,min,sec:     STRING[2];
      hour2,min2,sec2 : REAL ;
      code :            INTEGER ;
   BEGIN ah := $2c;          {initialize correct registers}
      WITH recpack DO BEGIN
         ax := ah SHL 8 + al;
      END;
      INTR($21,recpack);                   {call interrupt}
      WITH recpack DO BEGIN
         STR(cx SHR 8,hour);            {convert to string}
         STR(cx MOD 256,min);              { " }
         STR(dx SHR 8,sec);                { " }
      END;
      VAL(hour,hour2,code);
      VAL(min,min2,code) ;
      VAL(sec,sec2,code) ;
      realtime := 3600*hour2 + 60*min2 + sec2 ;
   END ;

FUNCTION Pof(t:REAL;i5:INTEGER) : REAL ;{Compute P(Theta) }
   BEGIN
      Pof := c + (1.0 - c)/(1.0 + EXP(-1.7 * a[i5] * (t -
             b[i5] ) ) ) ;
   END;

FUNCTION dPdz(i6:INTEGER): REAL ;
       {1st deriv of P with respect to theta}
   BEGIN
      dPdz := 1.7 * a[i6] * (p - c) * (1.0 - p)/ (1.0 - c);
   END;

PROCEDURE ComputeTestInf(t:REAL)  ;
   VAR i7 : INTEGER ;
   BEGIN
      TestInf := 0 ;
      FOR i7 := 1 TO n DO BEGIN
         p := Pof(t,i7) ; dp := dpdz(i7) ;
         TestInf := TestInf + dp*dp/(p*(1.0 - p)) ;
      END; END;
```

```
FUNCTION d2Pdz2(i8:INTEGER) : REAL ;
    {2nd deriv of P with respect to theta}
   BEGIN
      d2Pdz2 :=  1.7 * a[i8] * (1.0 + c - 2.0 * p) * dp /
                (1.0 - c) ;
   END;

FUNCTION d3Pdz3(i9:INTEGER) : REAL ;
     {3rd deriv of P with respect to theta}
   BEGIN
     d3Pdz3 :=  sqr(1.7*a[i9]*(1.0 + c - 2.0 * p)/(1-c))*dp
                 - 2*1.7*a[i9]*sqr(dp)/(1-c);
   END;

FUNCTION  nextb : REAL ;
{Get b with max item info at theta^}
   VAR nextp : REAL ;
   BEGIN
      nextp := (1.0 + SQRT(1.0 + 8.0 * c)) / 4.0 ;
      nextb :=
        zh - LN((nextp - c)/(1.0 - nextp)) / (1.7 * a[n]) ;
   END;

FUNCTION nextu : INTEGER ;
     {Get u for next item}
   BEGIN
      p := pof(z,n) ;
      nextu := 0 ;
      IF (RANDOM < p) THEN nextu := 1 ;
   END;

FUNCTION sign(t:REAL):REAL ;             {SIGN function  }
   BEGIN
      sign := 0.0 ;
      IF (t > 0.0) THEN sign := 1.0 ;
      IF (t < 0.0) THEN sign := -1.0 ;
   END;

PROCEDURE EstimateTheta12or3  ;   {1=MLE, 2=WLE, and 3=BME}
   VAR
      jnf,delta: REAL ; nit : INTEGER ;
   BEGIN
      nit := 0 ; REPEAT
         nit := nit + 1 ;
         dlnL := 0.0 ;
         jnf := 0.0 ;
         djnf := 0.0 ;
         dinf := 0.0 ;
         TestInf := 0.0 ;
         FOR i1 := 1 to n DO BEGIN
            p := Pof(zh,i1) ;
      pq := p*(1 - p) ;
      dp := dPdz(i1) ;
      dlnL := dlnL + (u[i1] - p)*dp/pq ;
      TestInf := TestInf + dp*dp/pq ;
      CASE method OF
      2 : BEGIN
         d2p :=  d2pdz2(i1) ;
         d3p :=  d3PdZ3(i1) ;
         jnf := jnf + dp*d2p/pq ;
         djnf := djnf + ((d2p*d2p+dp*d3p)/pq)
                 - ((dp*dp*d2p*(1-2*p))/(pq*pq)) ;
         dinf := dinf  + (2*dp*d2p/pq)
                 - (dp*dp*dp*(1-2*p)/(pq*pq)) ;
      END;
   END ;
END;
CASE method OF
      1 : delta := (dlnL )/(TestInf )  ;
                                {Maximum Likelihood}
      2 : BEGIN                 {Weighted Likelihood}
             dlnw := jnf/(2*TestInf) ;
             d2lnw := (TestInf*djnf - dinf*jnf)/
                       (2*sqr(TestInf)) ;
             delta := (dlnL + dlnw)/(TestInf - d2lnw);
          END;
      3 : delta := (dlnL - zh)/(TestInf + 1) ;
```

```
                                        {Bayesian Modal}
        END ;
        IF (ABS(delta) > 2 ) THEN delta := 2*sign(delta) ;
        zh := zh + delta  ;
     UNTIL ((ABS(delta) < 0.001) OR (ABS(zh) > 5.0)
             OR (nit > 20)) ;
     avnit[n] := avnit[n] + nit ;
     itavzh[n] := itavzh[n] + zh ;
     nnit[n] := nnit[n] + 1 ;
  END;               {End of EstimateTheta12or3 Procedure}

PROCEDURE WriteHeading;
  BEGIN
     WRITELN(fileA) ;
     WRITELN(fileA,'G','A','2',
     '       True    Bias              Skewness  Kurtosis
             AvTime  Aver.', '        MSE' ) ;
     WRITELN(fileA,'Method ', 'θ    (θ^)
      SD(θ^)    (θ^)    (θ^)   ','  (secs)
     items N  (θ^) ') ;
  END;

PROCEDURE Summarize ;
  BEGIN
     avzh := sumzh/nn ;
     avzh2 := sumzh2/nn ;
```

```
sdzh := SQRT(avzh2 - avzh*avzh) ;
sumzh3 := sumzh3/nn ;
skewzh := (sumzh3 - 3*avzh*avzh2 + 2*avzh*sqr(avzh))/
          (sdzh*sqr(sdzh)) ;
sumzh4 := sumzh4/nn ;
kurtzh :=( sumzh4 - 4*avzh*sumzh3 + 6*avzh2*sqr(avzh)
     - 3*sqr(avzh)*sqr(avzh) )/(sqr(sdzh)*sqr(sdzh));
sumn := sumn/nn ;
avzh := avzh - z ;
msezh := sqr(avzh) + sqr(sdzh) ;
CASE method OF
   1 : meth := 'MLE' ;
   2 : meth := 'WLE' ;
   3 : meth := 'BME' ;
END;
Writeln(fileA,meth:6,z:5:1,avzh:10:4,sdzh:10:4,
           skewzh:10:4,kurtzh:10:4,
           time:7:1,sumn:8:1,'  ',nn:3,msezh:8:4) ;
FOR i1 := 1 TO 101 DO BEGIN
   IF (nnit[i1] > 0) THEN  BEGIN
      avnit[i1] := avnit[i1]/nnit[i1] ;
      itavzh[i1] := itavzh[i1]/nnit[i1] ;
      WRITELN(fileB,'Meth=',meth:3,'   θ=',z:4:1,'  Avθ^=',
              itavzh[i1]:8:3,'   n=',i1:3,
              '   avNits=', avnit[i1]:5:1,
              '  N=',nnit[i1]:2)  ;
```

```
         END ;
      END;        Writeln(fileB);
   END ;

BEGIN
 Initialize1;
 FOR method := 1 TO 3 DO BEGIN
 WriteHeading ;
  FOR iz := 1 TO 17 DO BEGIN
   z := (iz - 9)/2.0 ;
   Initialize2 ;
   time := realtime ;
   FOR  i2 := 1 TO nn DO BEGIN
      n := 0 ;
      zh := 0.0 ;
      TestInf := 0.0  ;
      REPEAT
         n := n + 1 ;
         a[n] := 2.0  ;  { a[n] :=  (71.0 - n)/35.0 ;   }
         b[n] := nextb ;
         u[n] := nextu ;
         EstimateTheta12or3 ;            {MLE, WLE, or BME}
      UNTIL ((TestInf > 20) OR (n >= maxn) );
      sumn  :=  ((sumn)  + ( n))  ;
      ComputeTestInf(zh)  ;
      Writeln(method:3,i2:5,'  z=',z:3:4,'  zh=', zh:3:4,
              '  I=',TestInf:3:4,'  n=',n);
      sumzh := sumzh + zh   ;
      sumzh2 := sumzh2 +  zh*zh ;
      sumzh3 := sumzh3 + zh*sqr(zh) ;
      sumzh4 := sumzh4 + sqr(zh) * sqr(zh) ;
   END;
   time := (realtime - time)/nn  ;
   Summarize ;
  END;
 END;
 CLOSE(fileA) ;
 CLOSE(fileB) ;
END.
```

# APPENDIX D

## ADDITIONAL FIGURES OF RESULTS

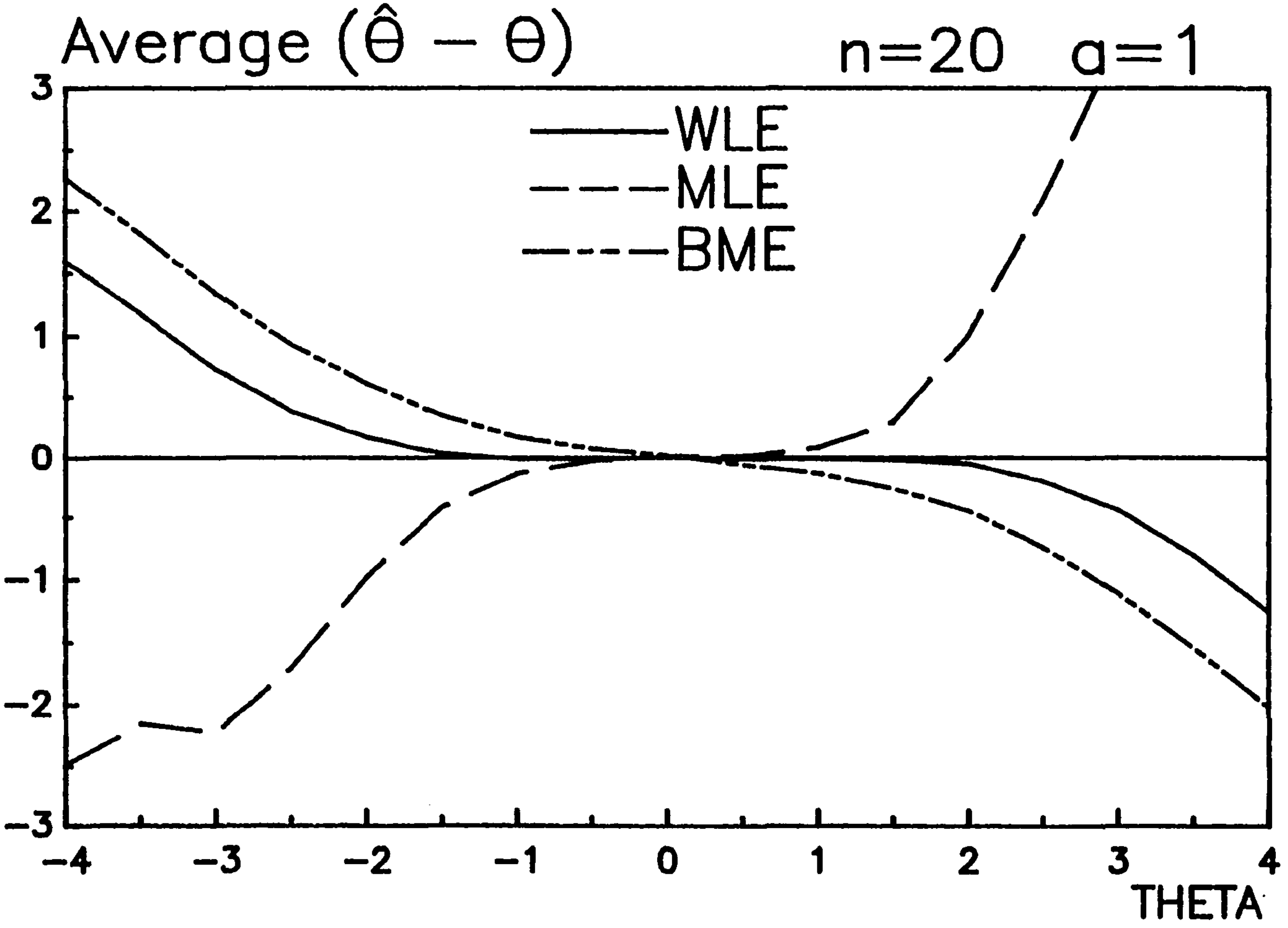

Figure D.1
Average Estimation Error of θ^ on Conventional Test with 20 Items, All a = 1,
Normally Distributed b, and All c = 0.20 .

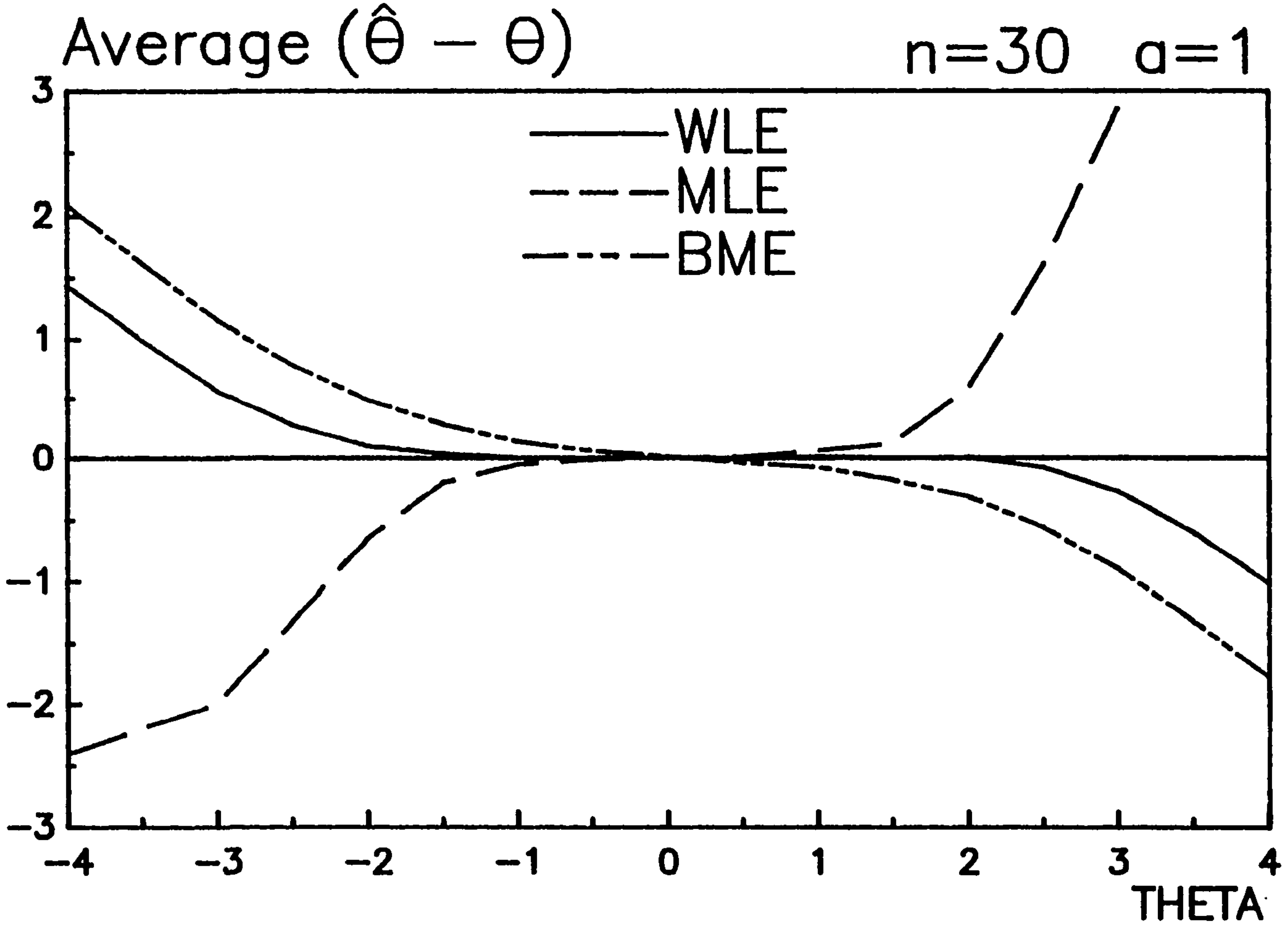

Figure D.2
Average Estimation Error of θ^ on Conventional Test with 30 Items, All a = 1, Normally Distributed b, and All c = 0.20 .

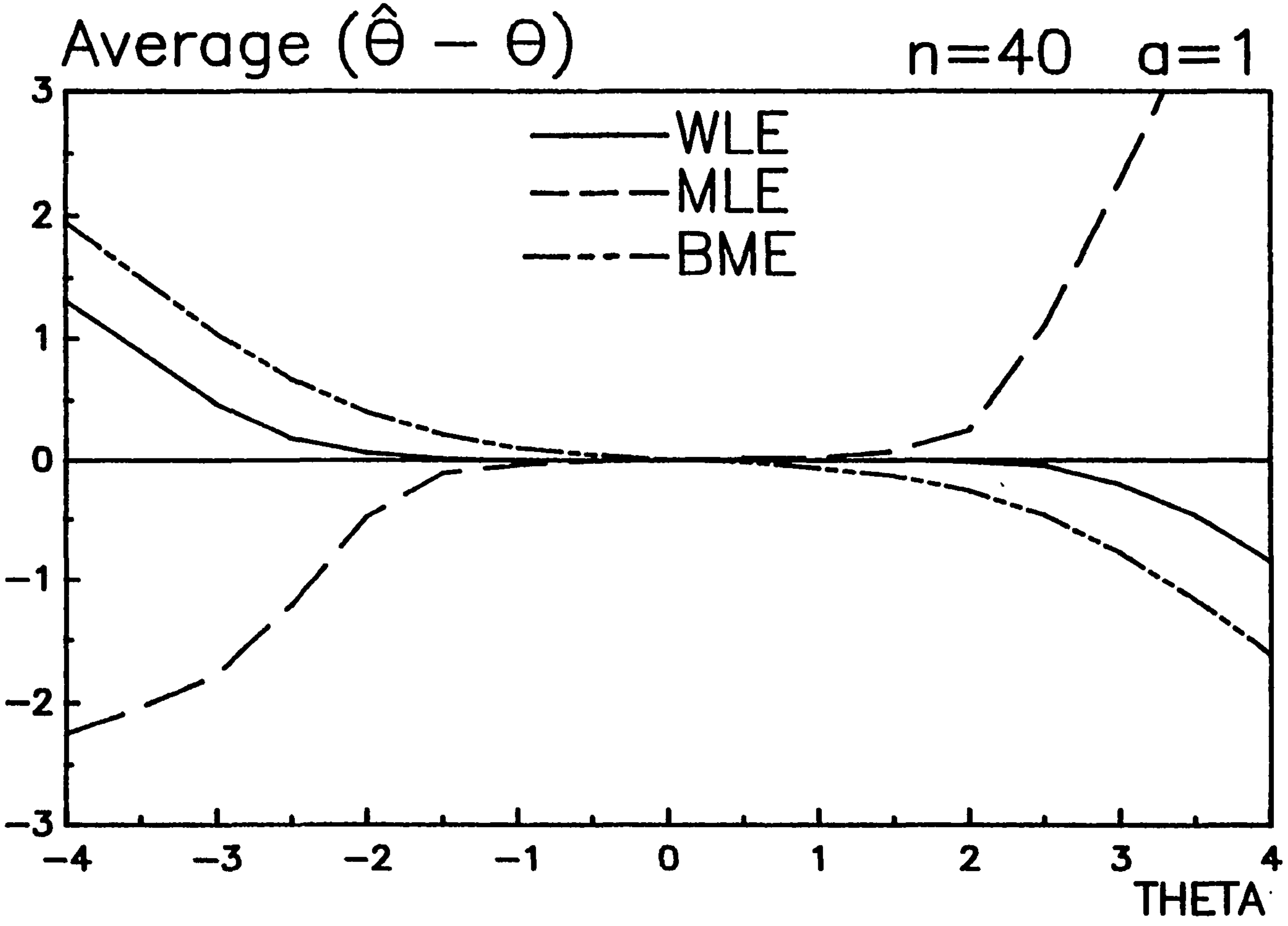

Figure D.3
Average Estimation Error of θ^ on Conventional Test with 40 Items, All a = 1,
Normally Distributed b, and All c = 0.20 .

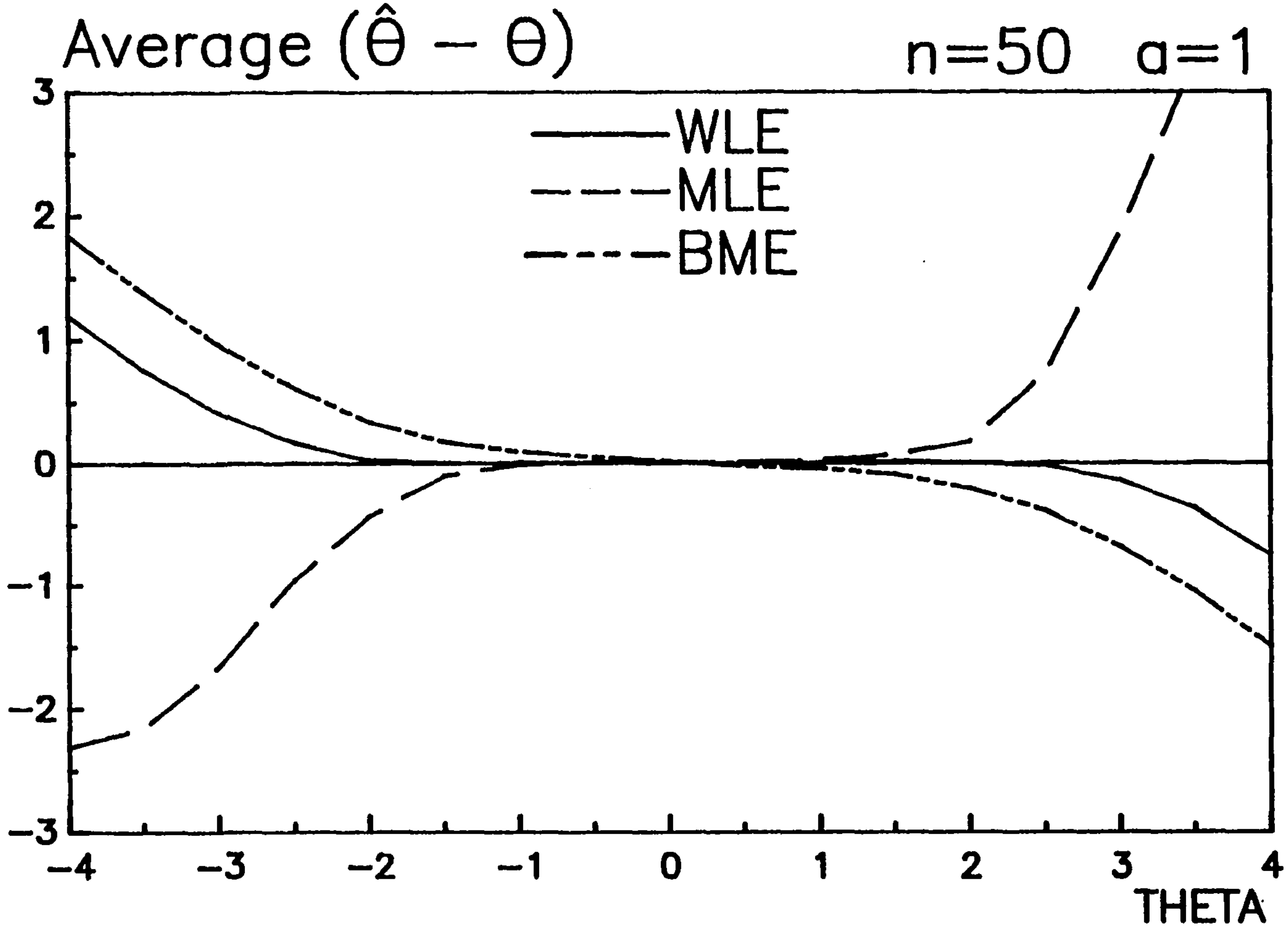

Figure D.4
Average Estimation Error of θ^ on Conventional Test with 50 Items, All a = 1, Normally Distributed b, and All c = 0.20 .

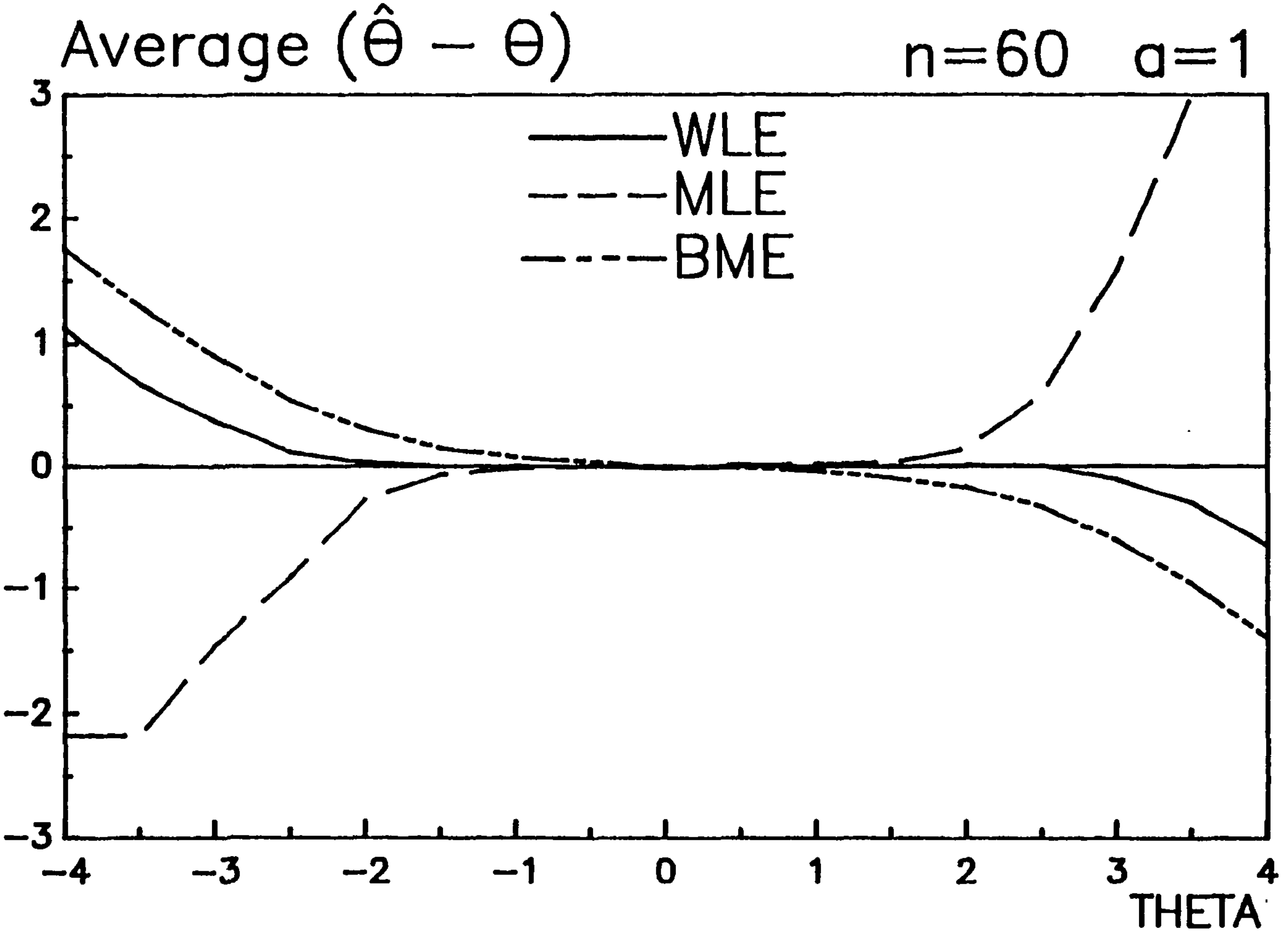

Figure D.5
Average Estimation Error of θ^ on Conventional Test with 60 Items, All a = 1, Normally Distributed b, and All c = 0.20 .

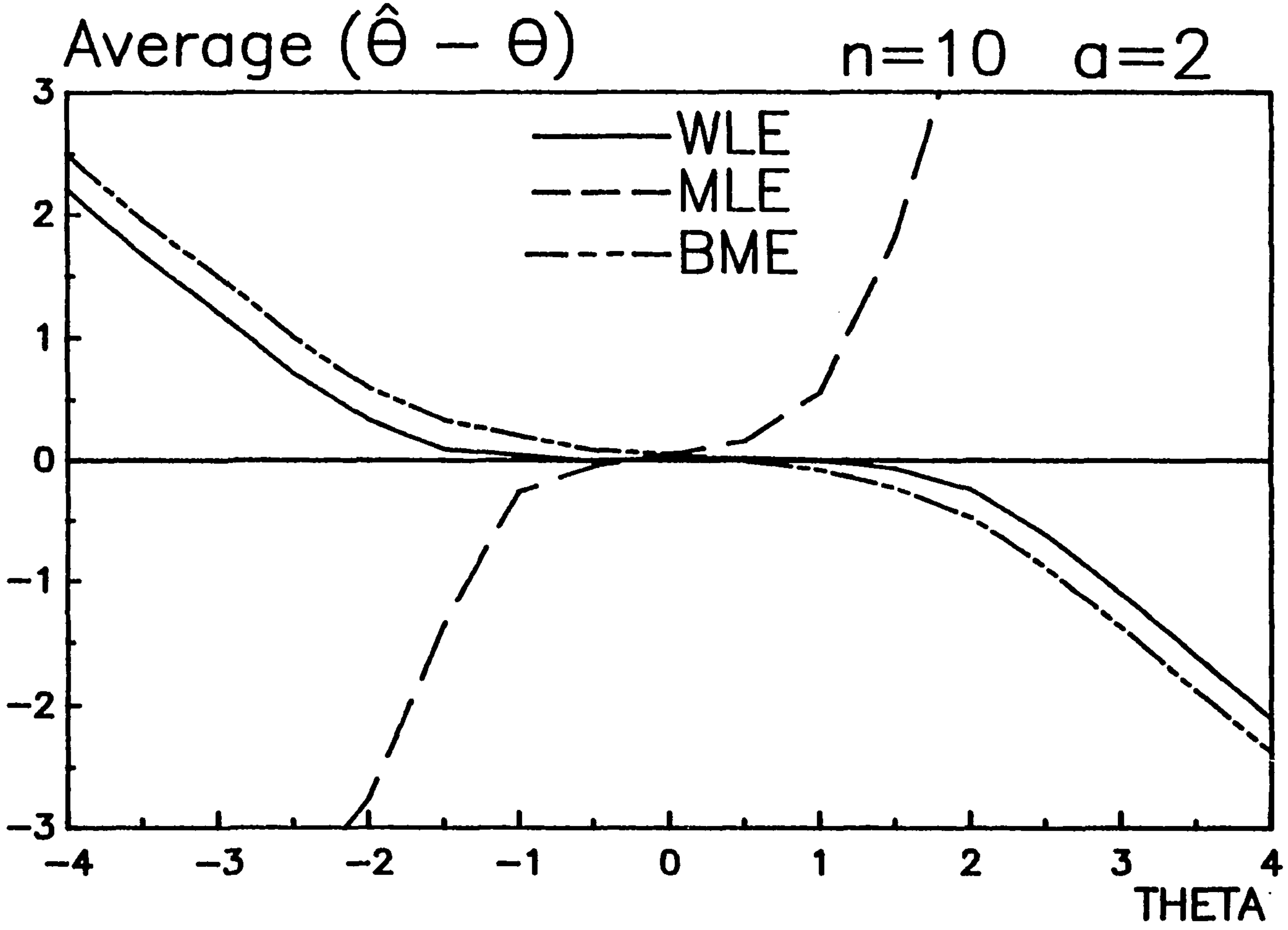

Figure D.6
Average Estimation Error of θ^ on Conventional Test with 10 Items, All a = 2, Normally Distributed b, and All c = 0.20 .

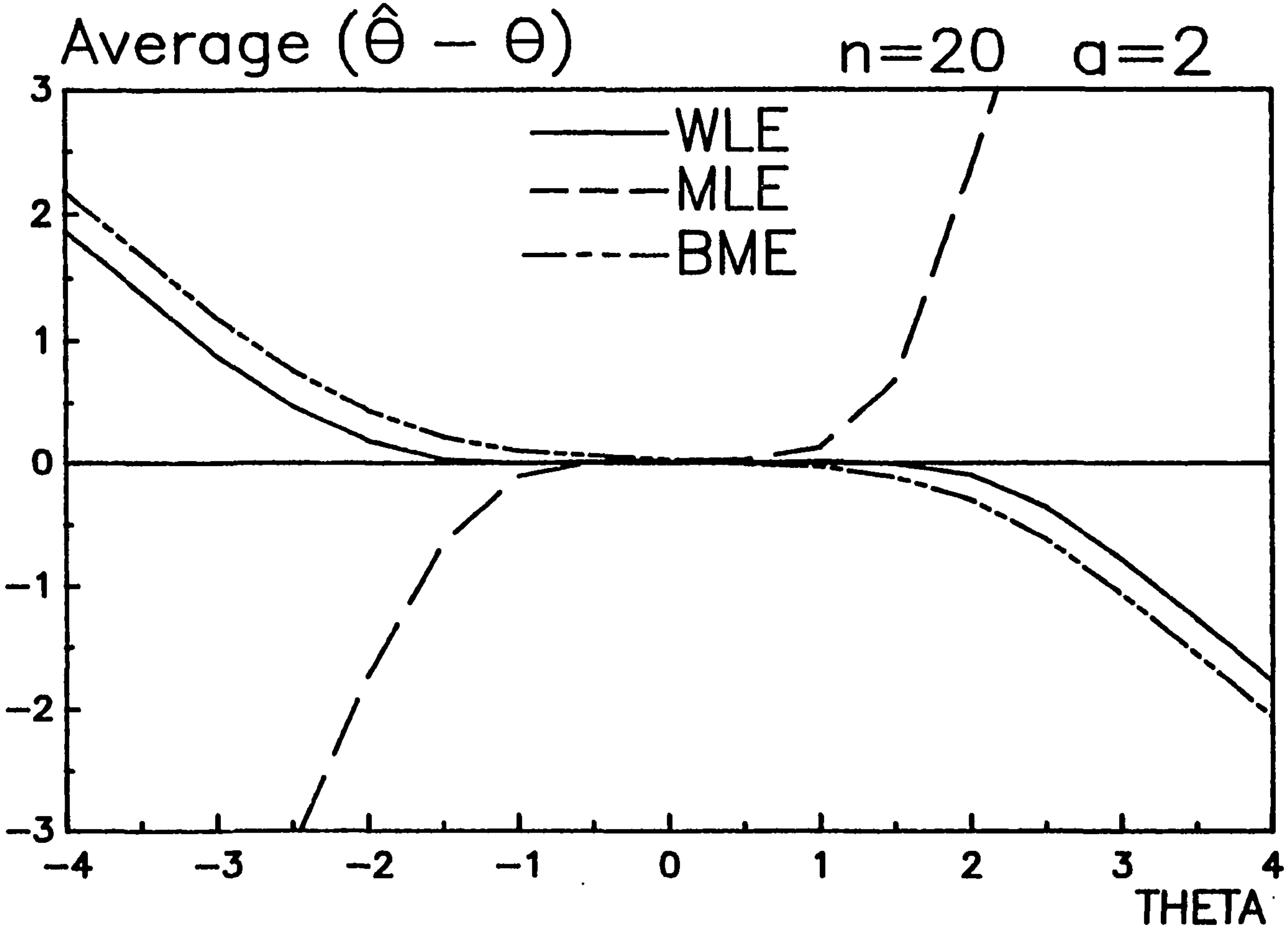

Figure D.7
Average Estimation Error of θ^ on Conventional Test with 20 Items, All a = 2, Normally Distributed b, and All c = 0.20 .

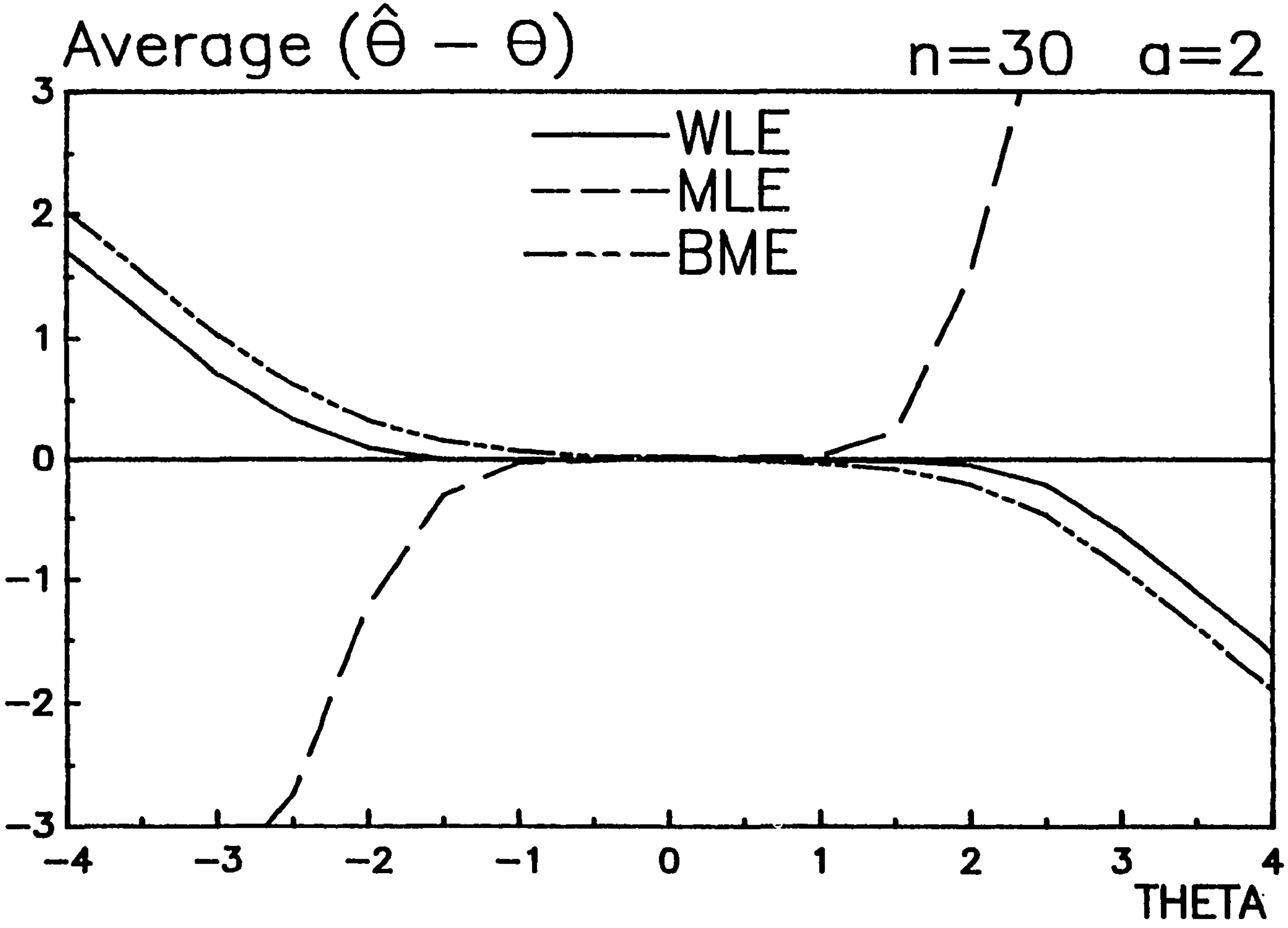

Figure D.8
Average Estimation Error of θ^ on Conventional Test with 30 Items, All a = 2, Normally Distributed b, and All c = 0.20 .

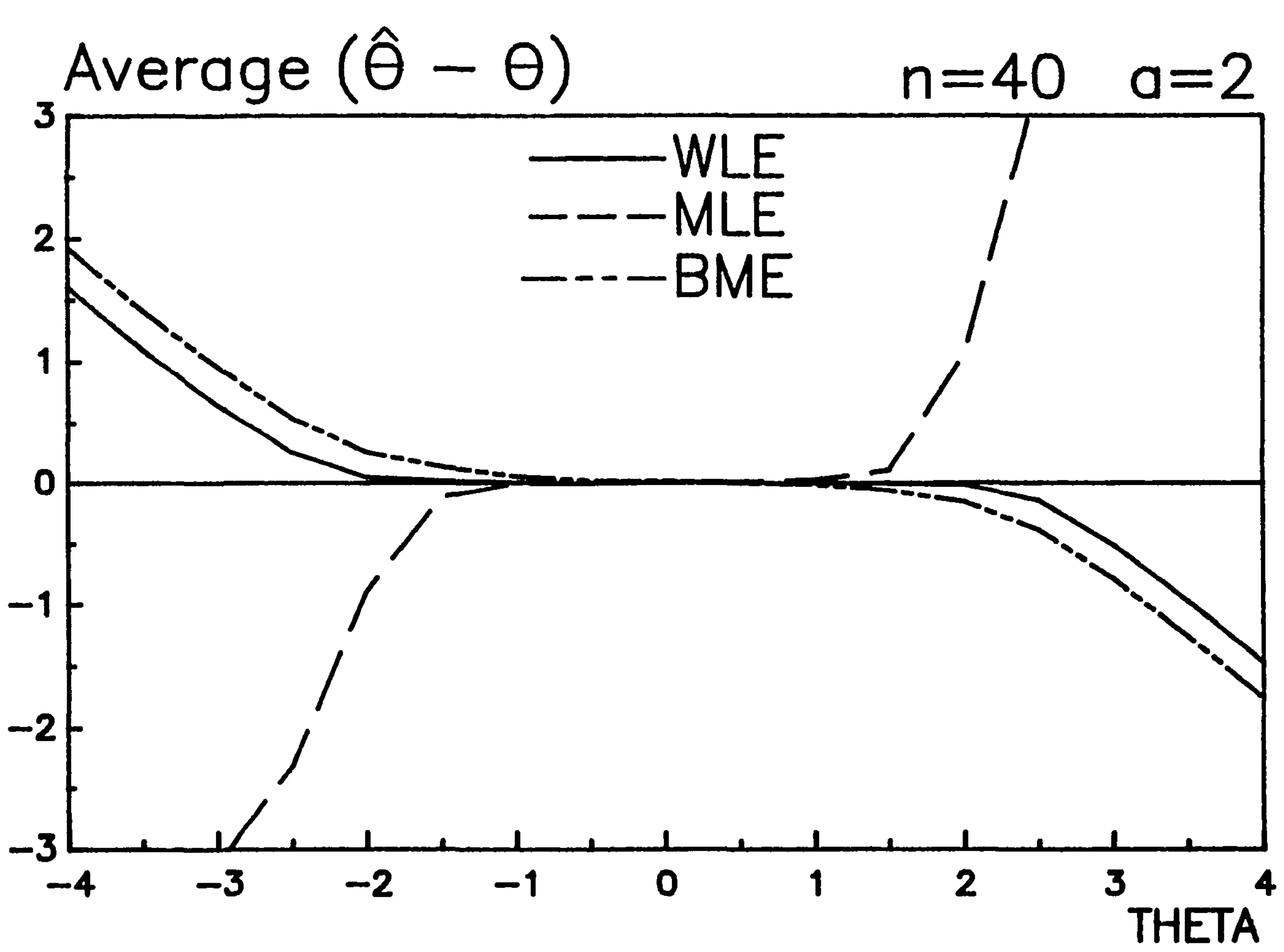

Figure D.9
Average Estimation Error of θ^ on Conventional Test with 40 Items, All a = 2, Normally Distributed b, and All c = 0.20 .

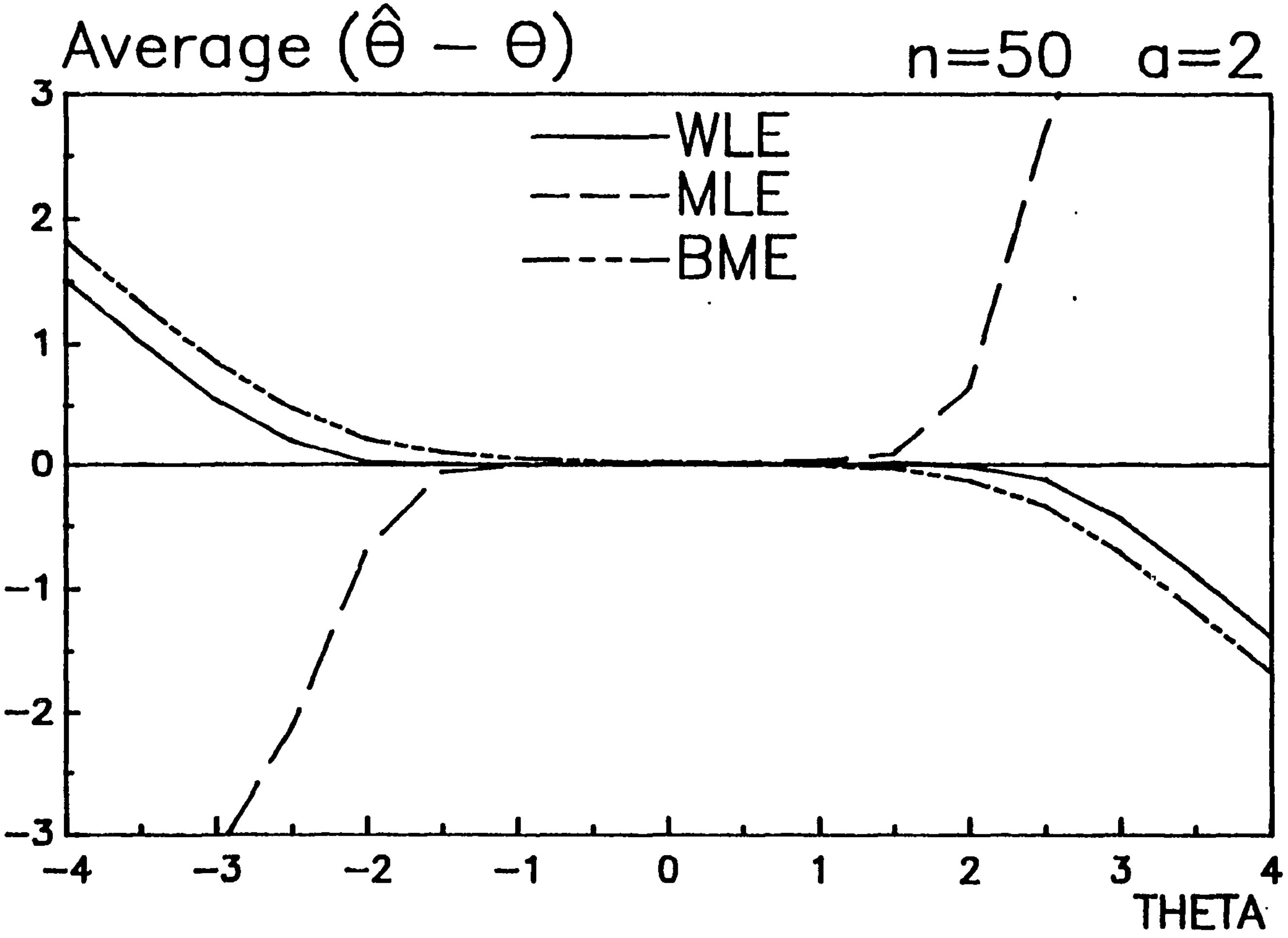

Figure D.10
Average Estimation Error of θ^ on Conventional Test with 50 Items, All a = 2, Normally Distributed b, and All c = 0.20 .

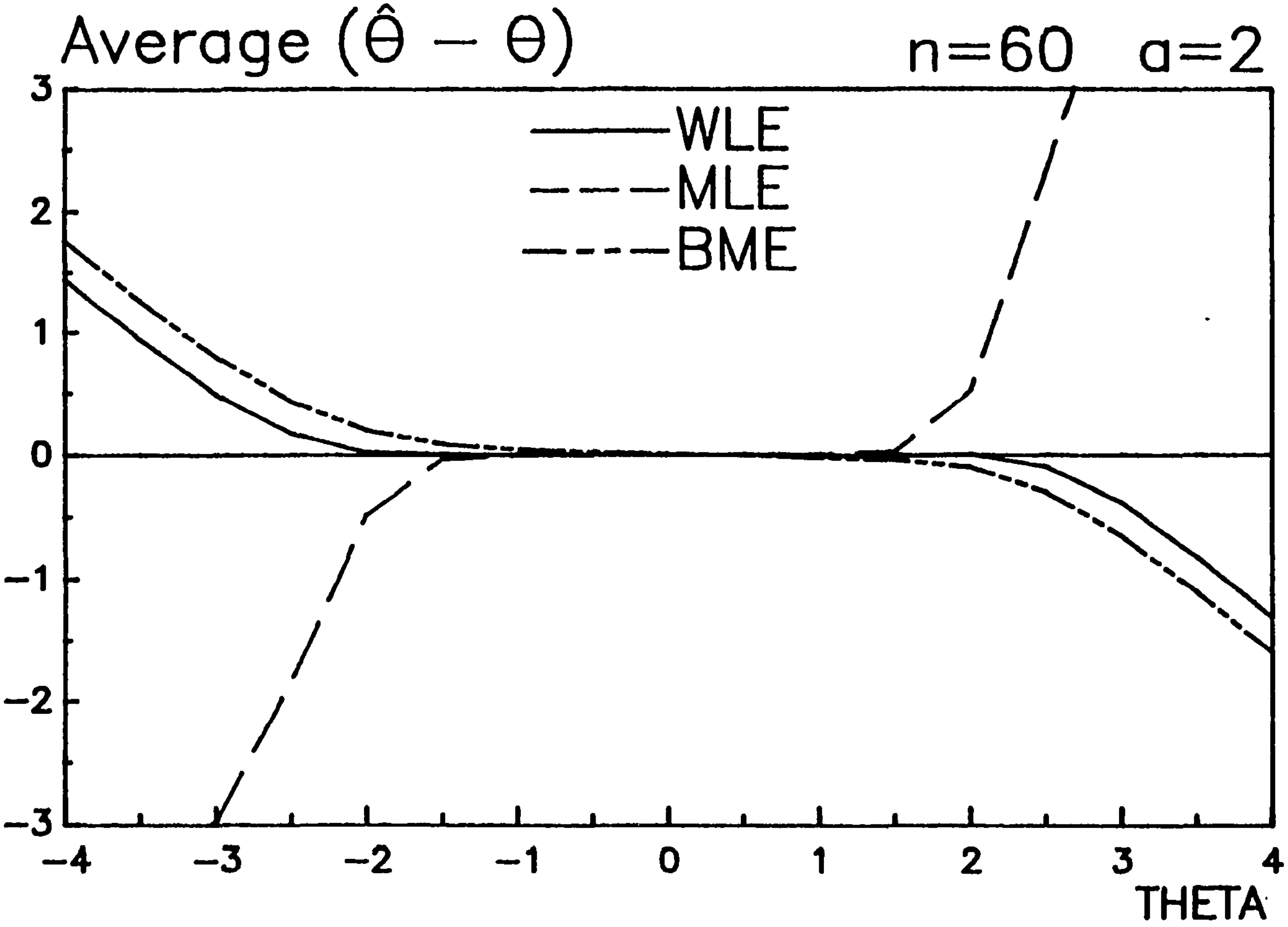

Figure D.11
Average Estimation Error of θ^ on Conventional Test with 60 Items, All a = 2, Normally Distributed b, and All c = 0.20 .

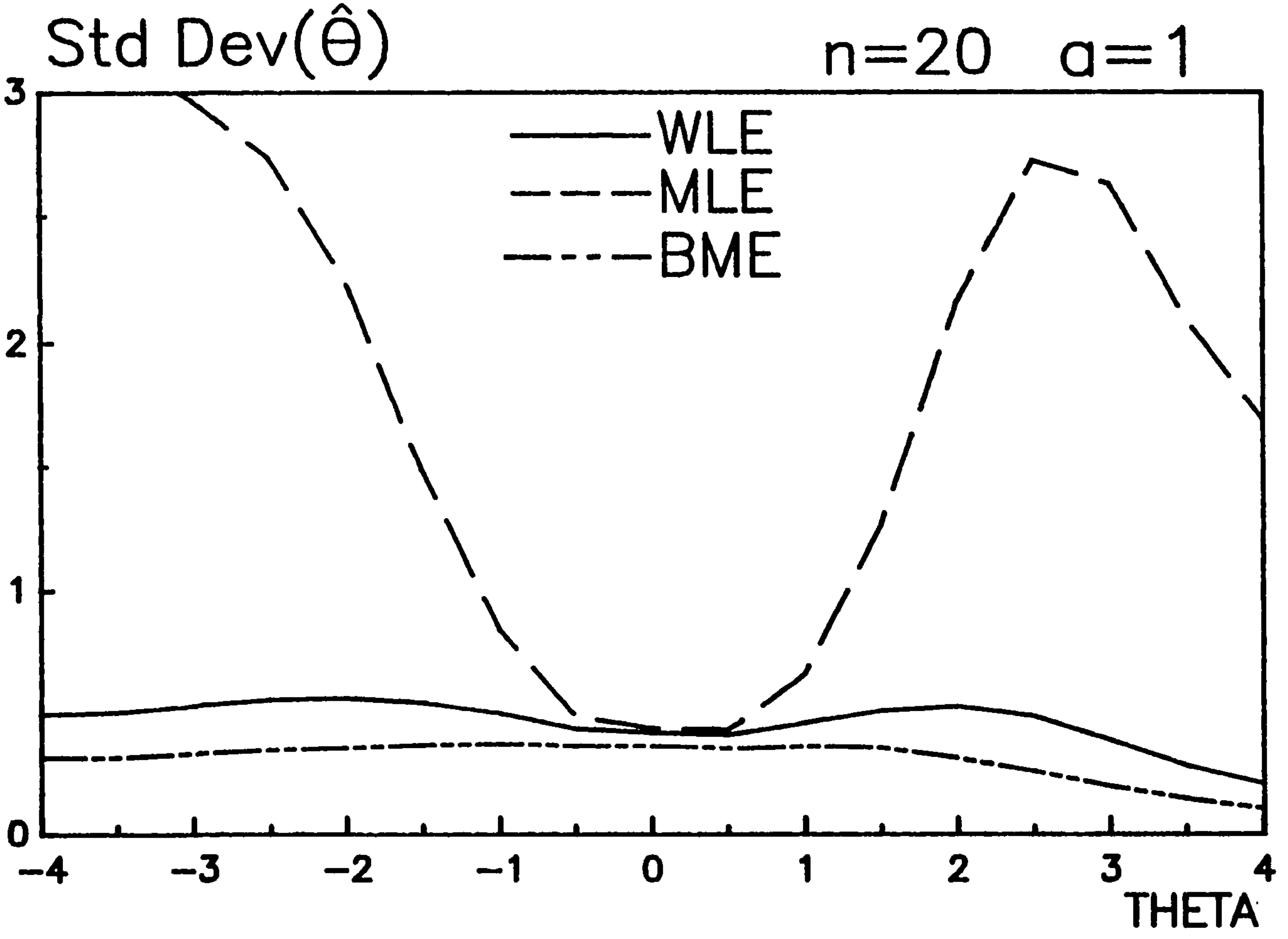

Figure D.12
Standard Deviation of θ^ on Conventional Test with 20 Items, All a = 1, Normally Distributed b, and All c = 0.20 .

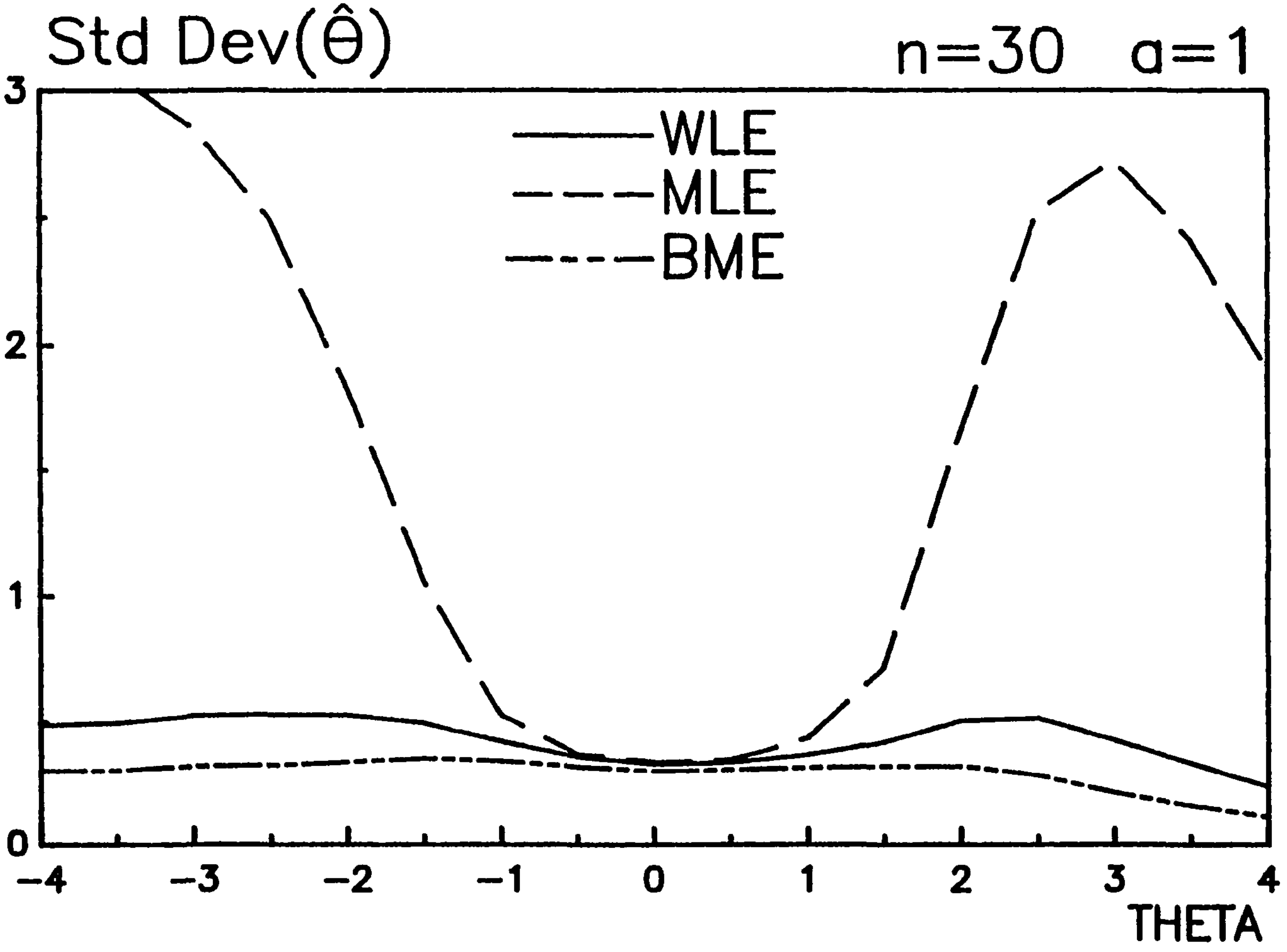

Figure D.13
Standard Deviation of θ^ on Conventional Test with 30 Items, All a = 1, Normally Distributed b, and All c = 0.20 .

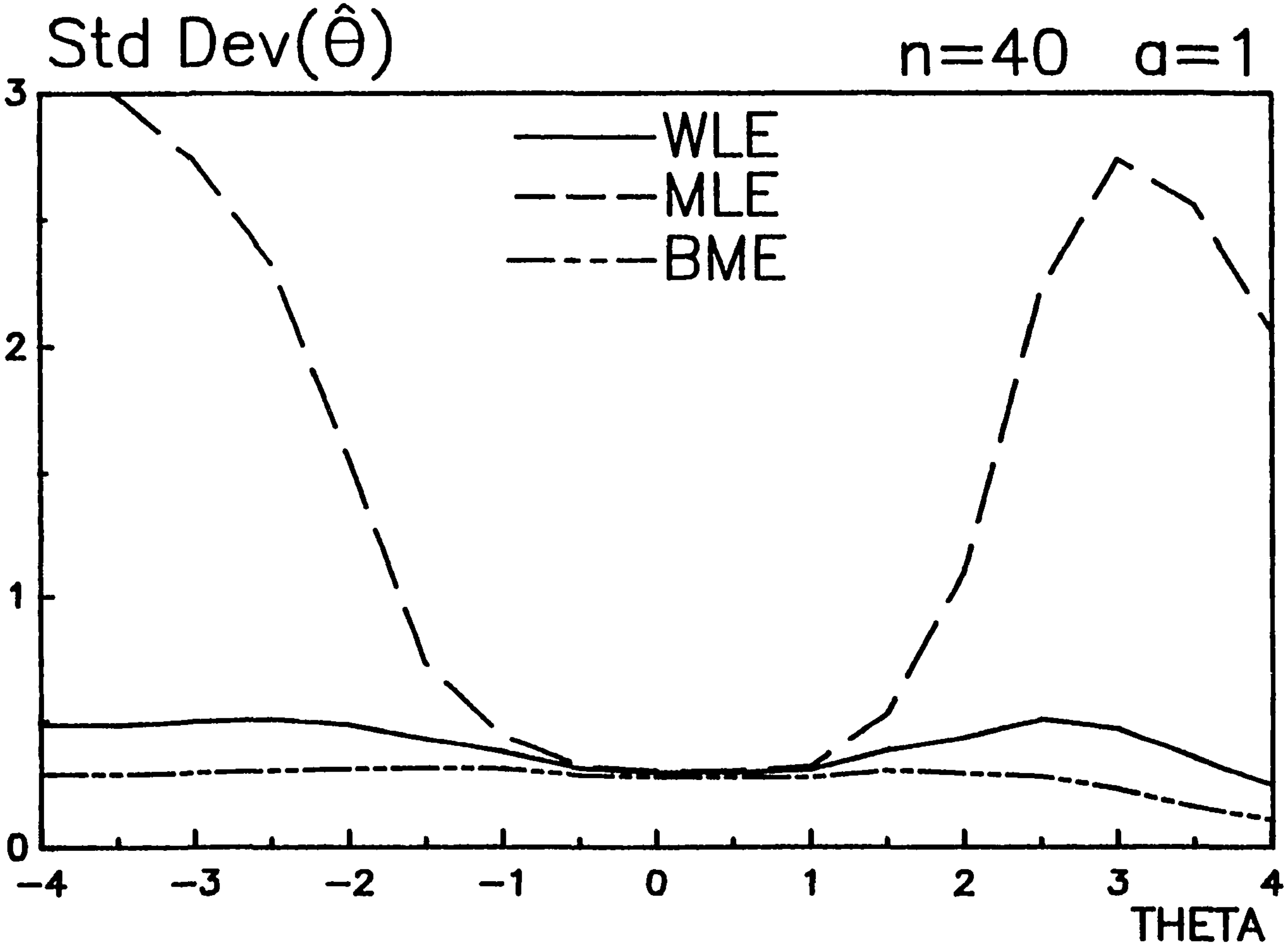

Figure D.14
Standard Deviation of θ^ on Conventional Test with 40 Items, All a = 1, Normally Distributed b, and All c = 0.20 .

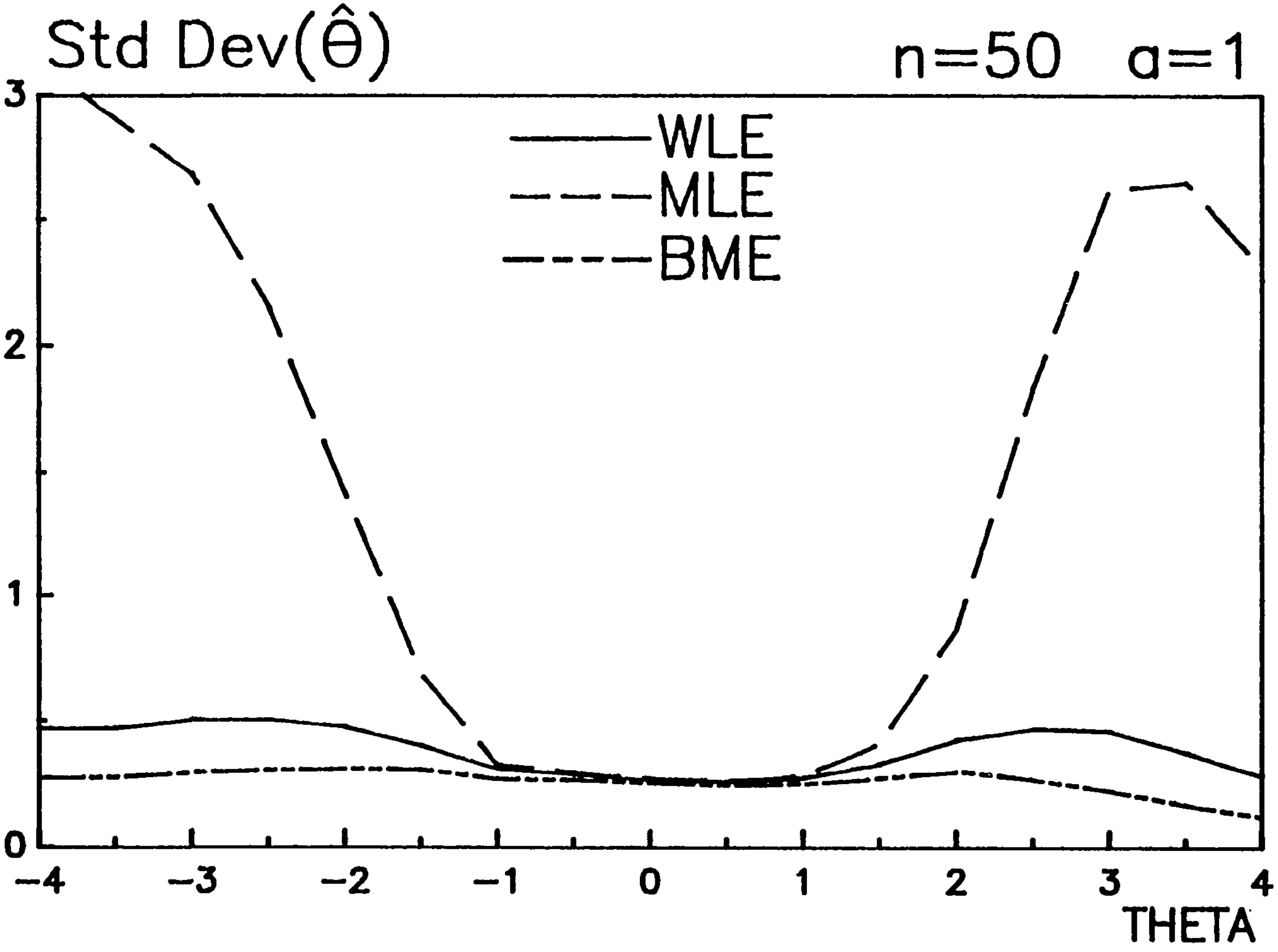

Figure D.15
Standard Deviation of θ^ on Conventional Test with 50 Items, All a = 1, Normally Distributed b, and All c = 0.20 .

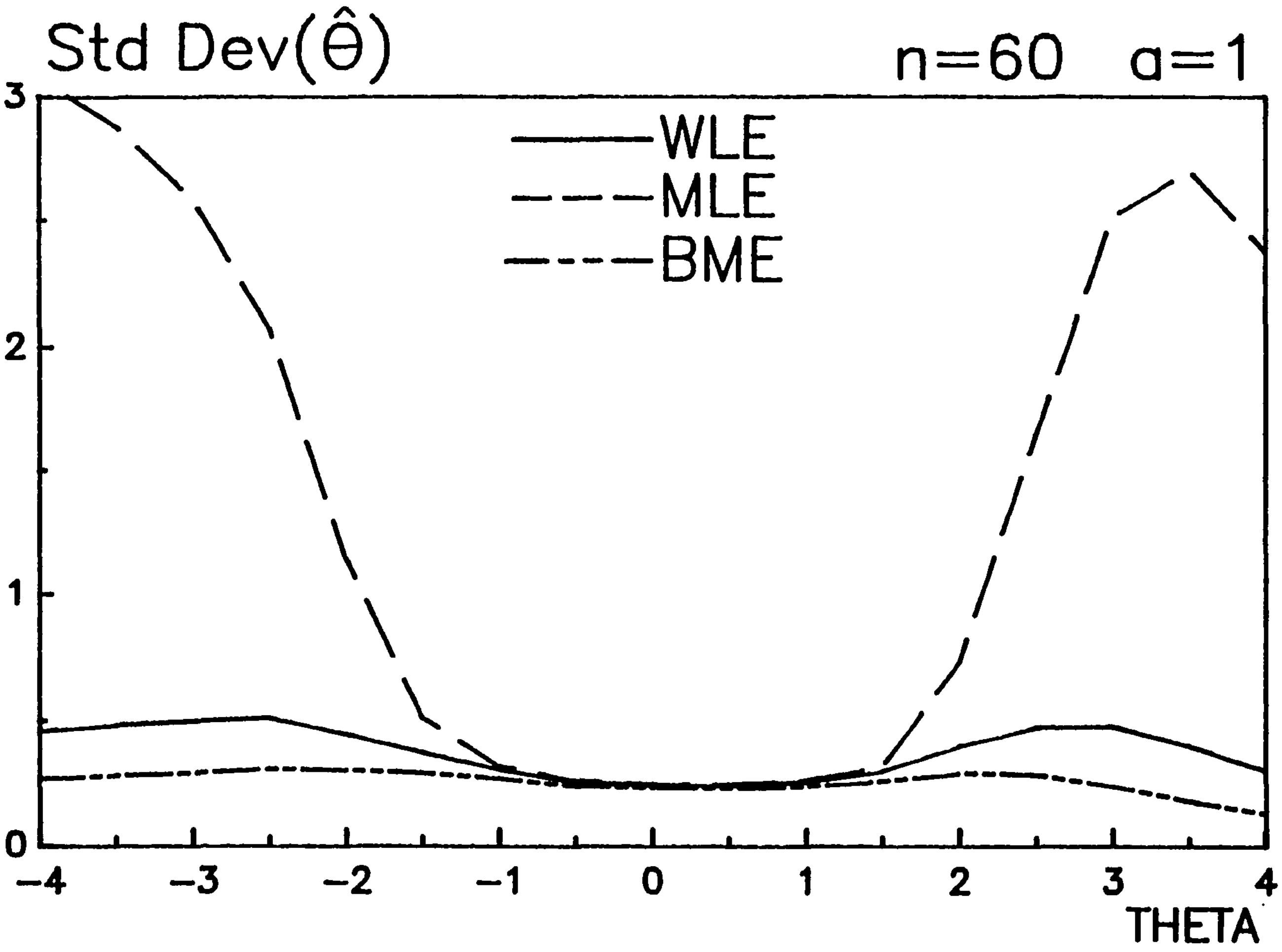

Figure D.16
Standard Deviation of θ^ on Conventional Test with 60 Items, All a = 1, Normally Distributed b, and All c = 0.20 .

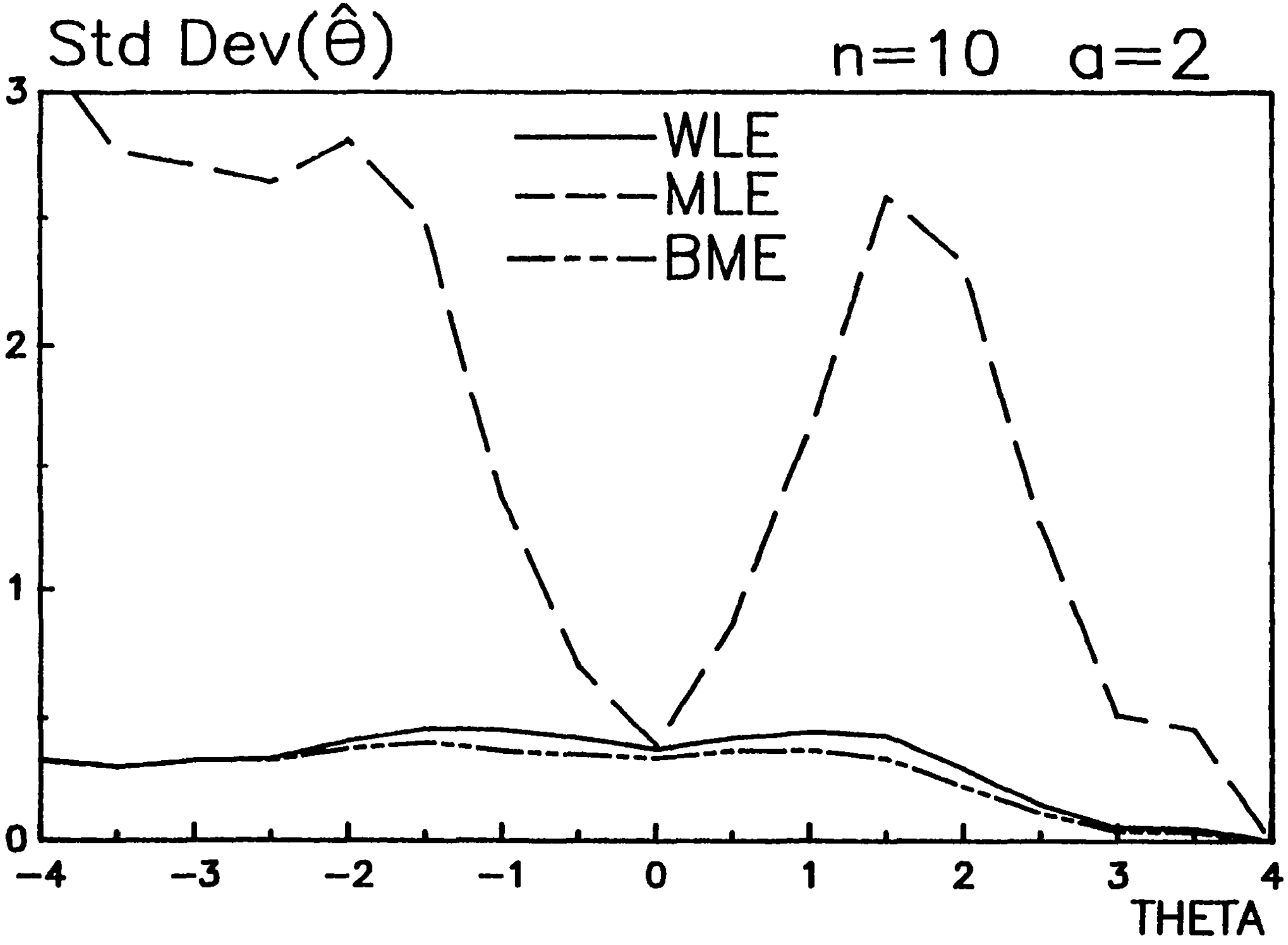

Figure D.17
Standard Deviation of θ^ on Conventional Test with 10 Items, All a = 2, Normally Distributed b, and All c = 0.20 .

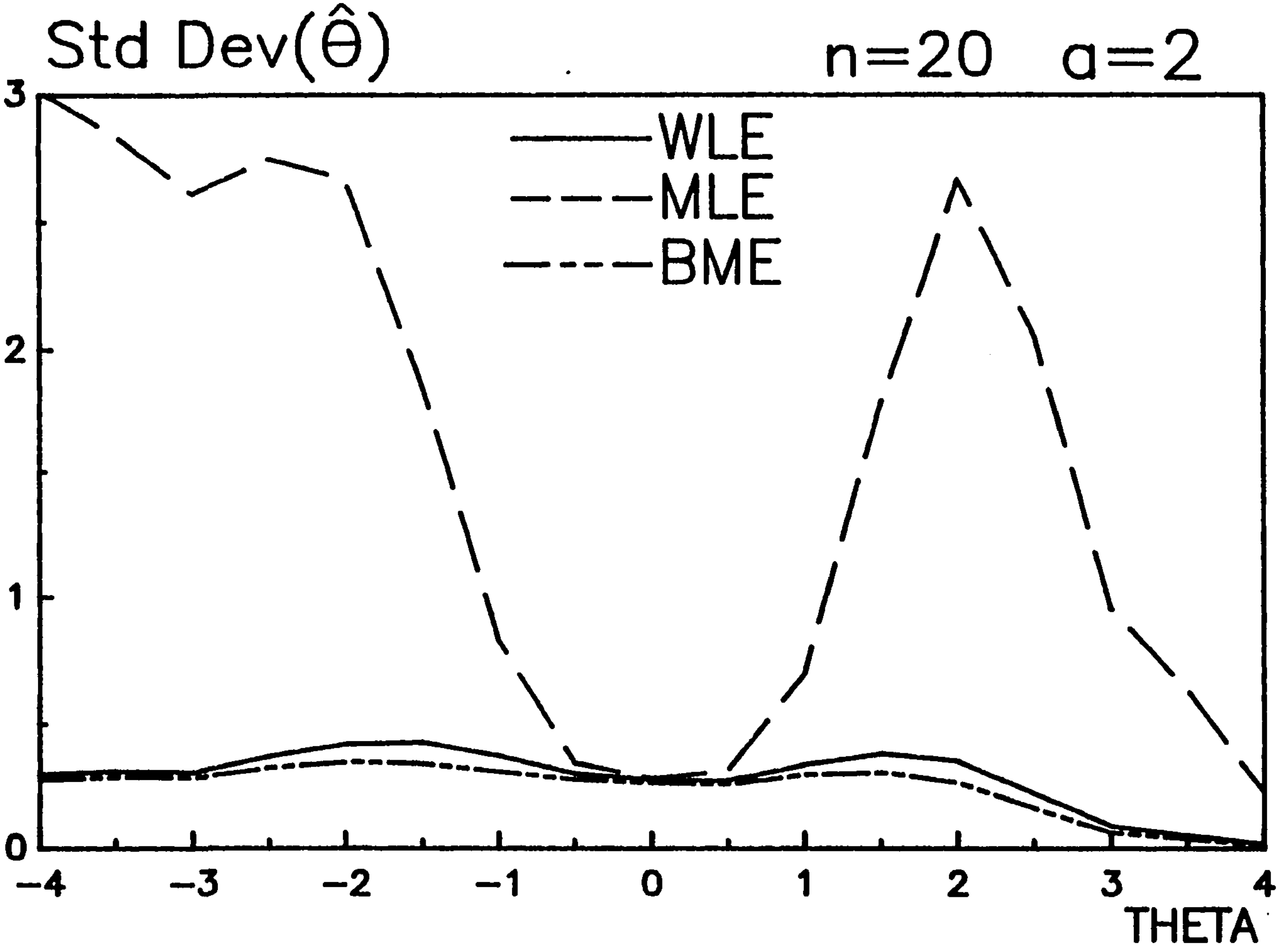

Figure D.18
Standard Deviation of θ^ on Conventional Test with 20 Items, All a = 2, Normally Distributed b, and All c = 0.20 .

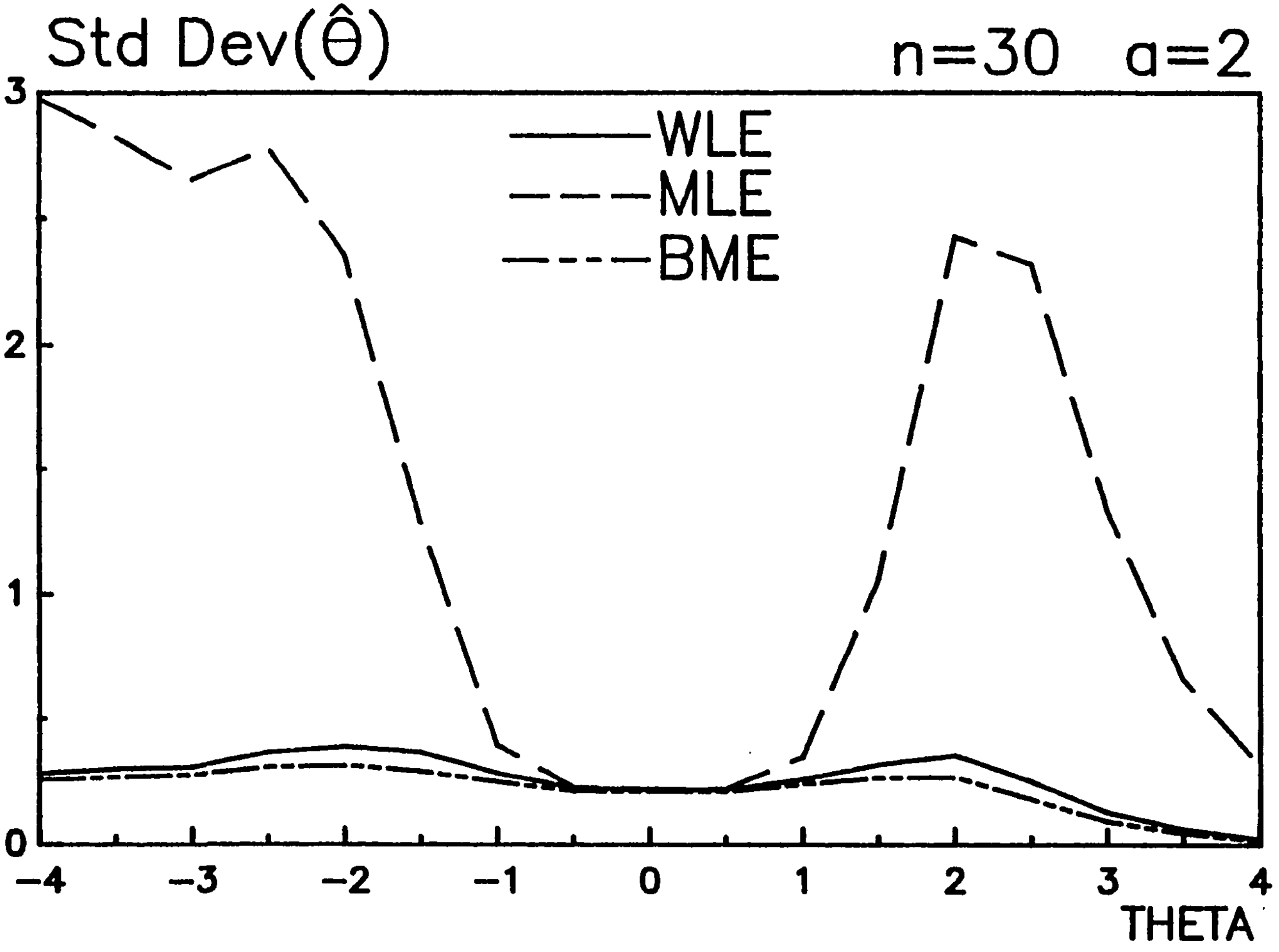

Figure D.19
Standard Deviation of θ^ on Conventional Test with 30 Items, All a = 2, Normally Distributed b, and All c = 0.20 .

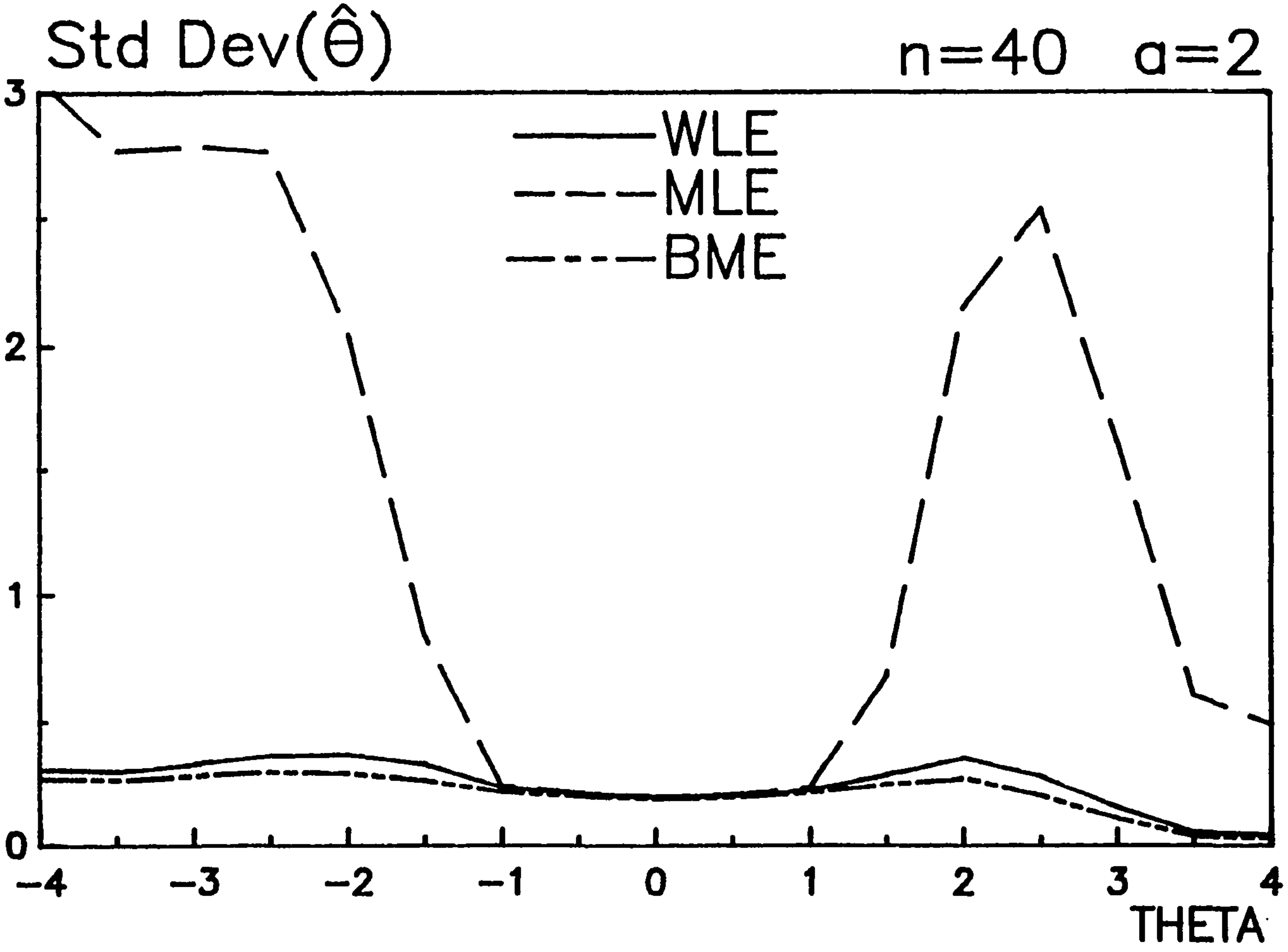

Figure D.20
Standard Deviation of θ^ on Conventional Test with 40 Items, All a = 2, Normally Distributed b, and All c = 0.20 .

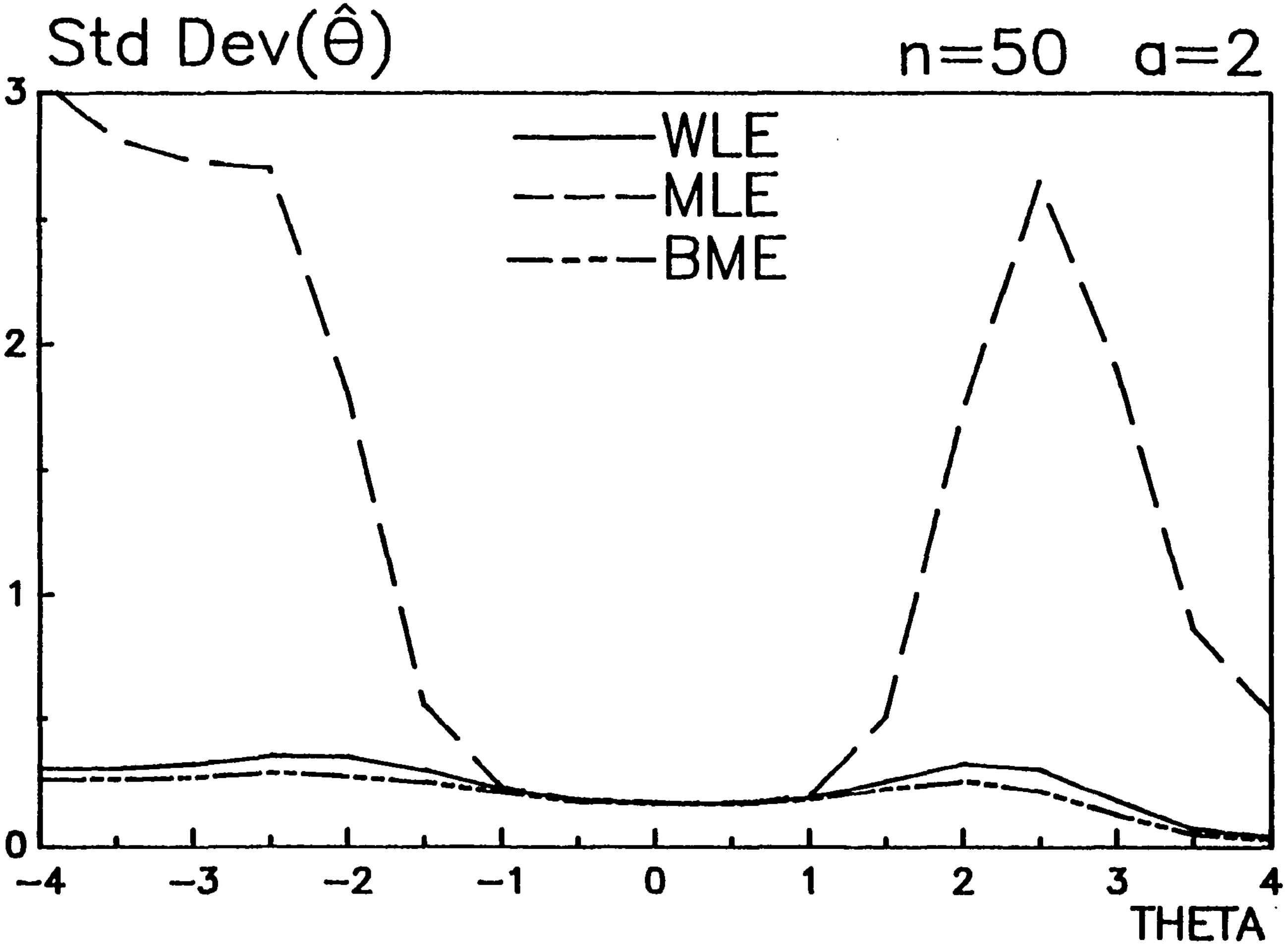

Figure D.21
Standard Deviation of θ^ on Conventional Test with 50 Items, All a = 2, Normally Distributed b, and All c = 0.20 .

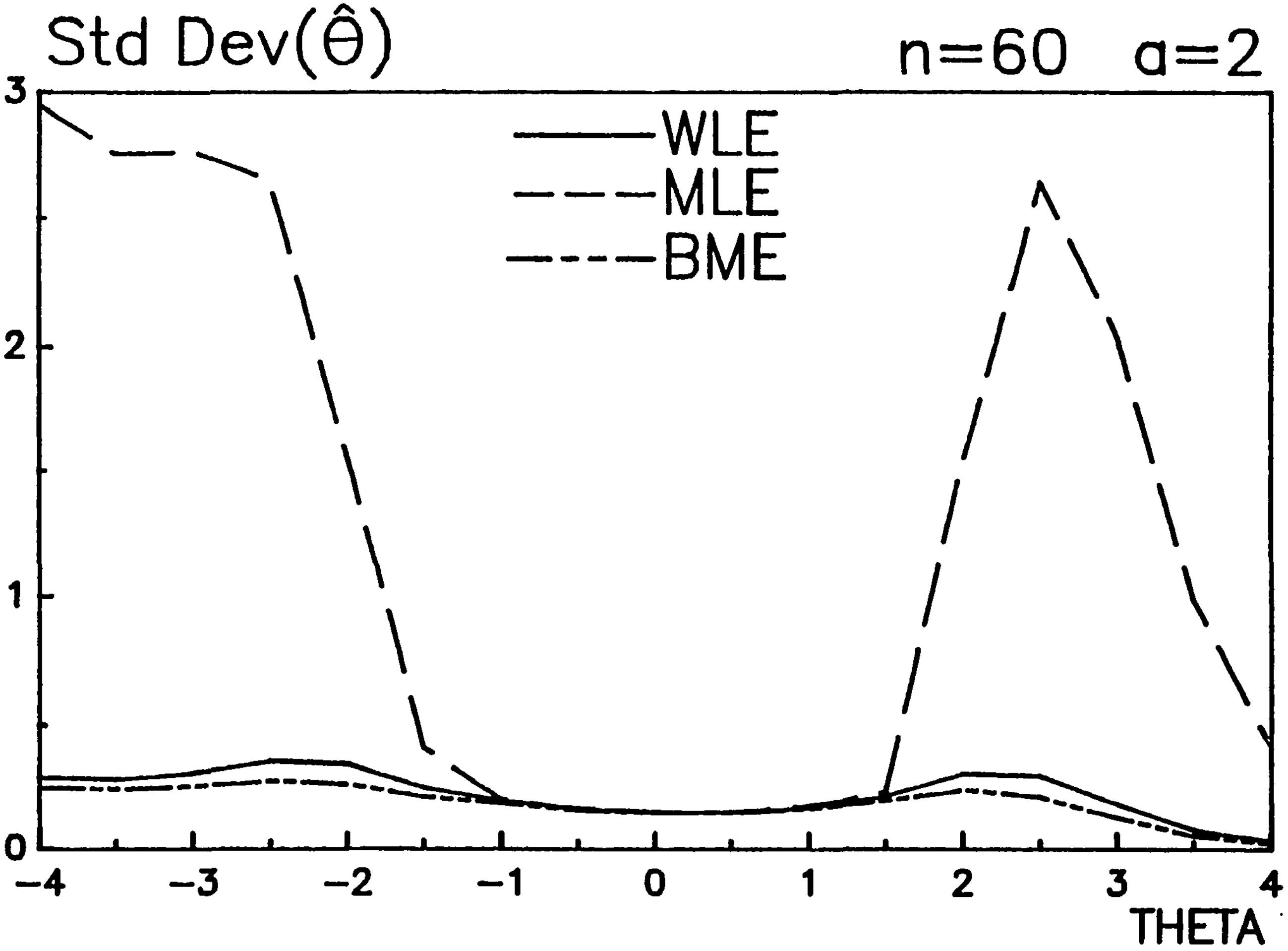

Figure D.22
Standard Deviation of θ^ on Conventional Test with 60 Items, All a = 2, Normally Distributed b, and All c = 0.20 .

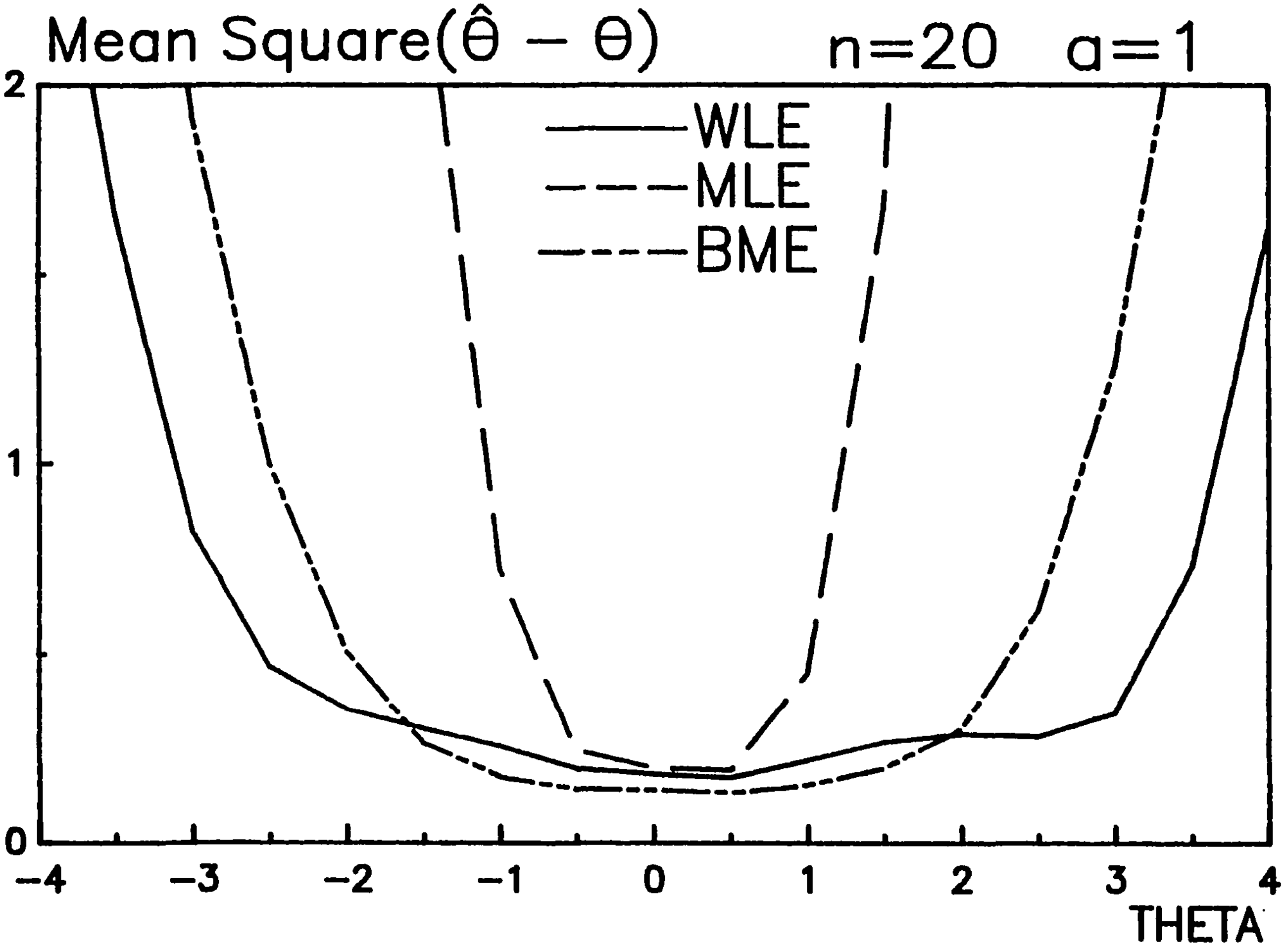

Figure D.23
Mean Squared Error of θ^ on Conventional Test with 20 Items, All a = 1, Normally Distributed b, and All c = 0.20 .

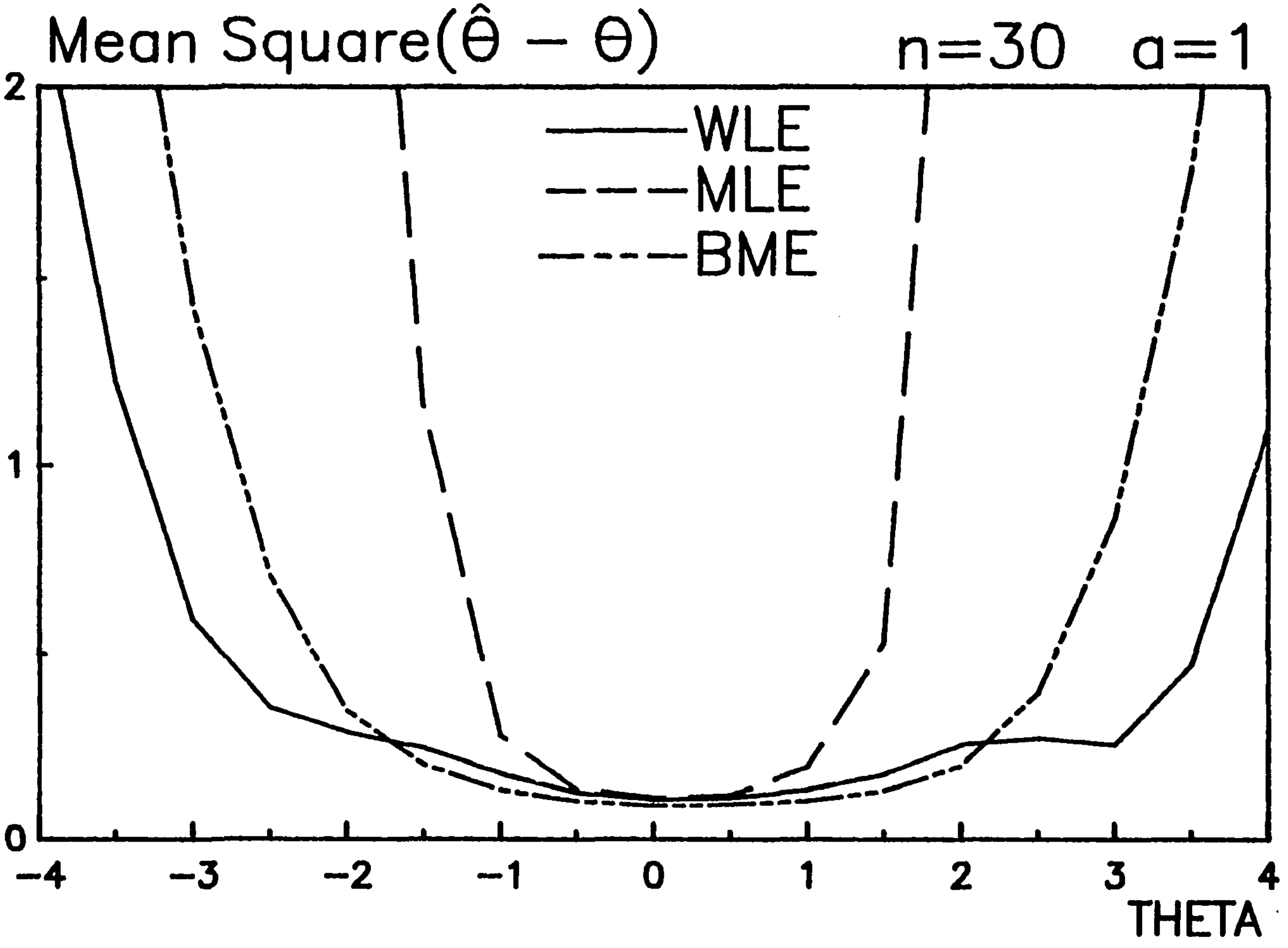

Figure D.24
Mean Squared Error of θ^ on Conventional Test with 30 Items, All a = 1, Normally Distributed b, and All c = 0.20 .

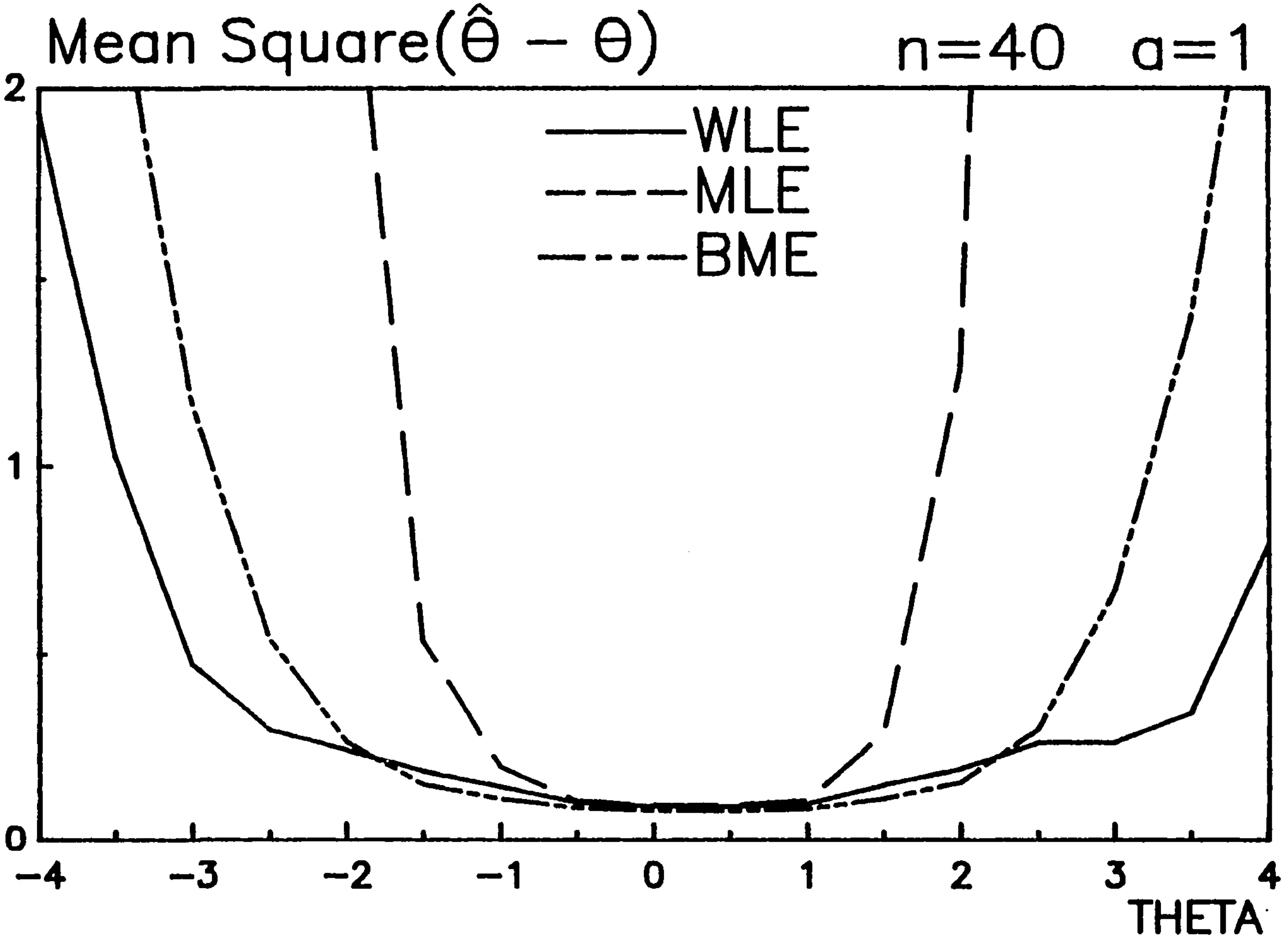

Figure D.25
Mean Squared Error of θ^ on Conventional Test with 40 Items, All a = 1, Normally Distributed b, and All c = 0.20 .

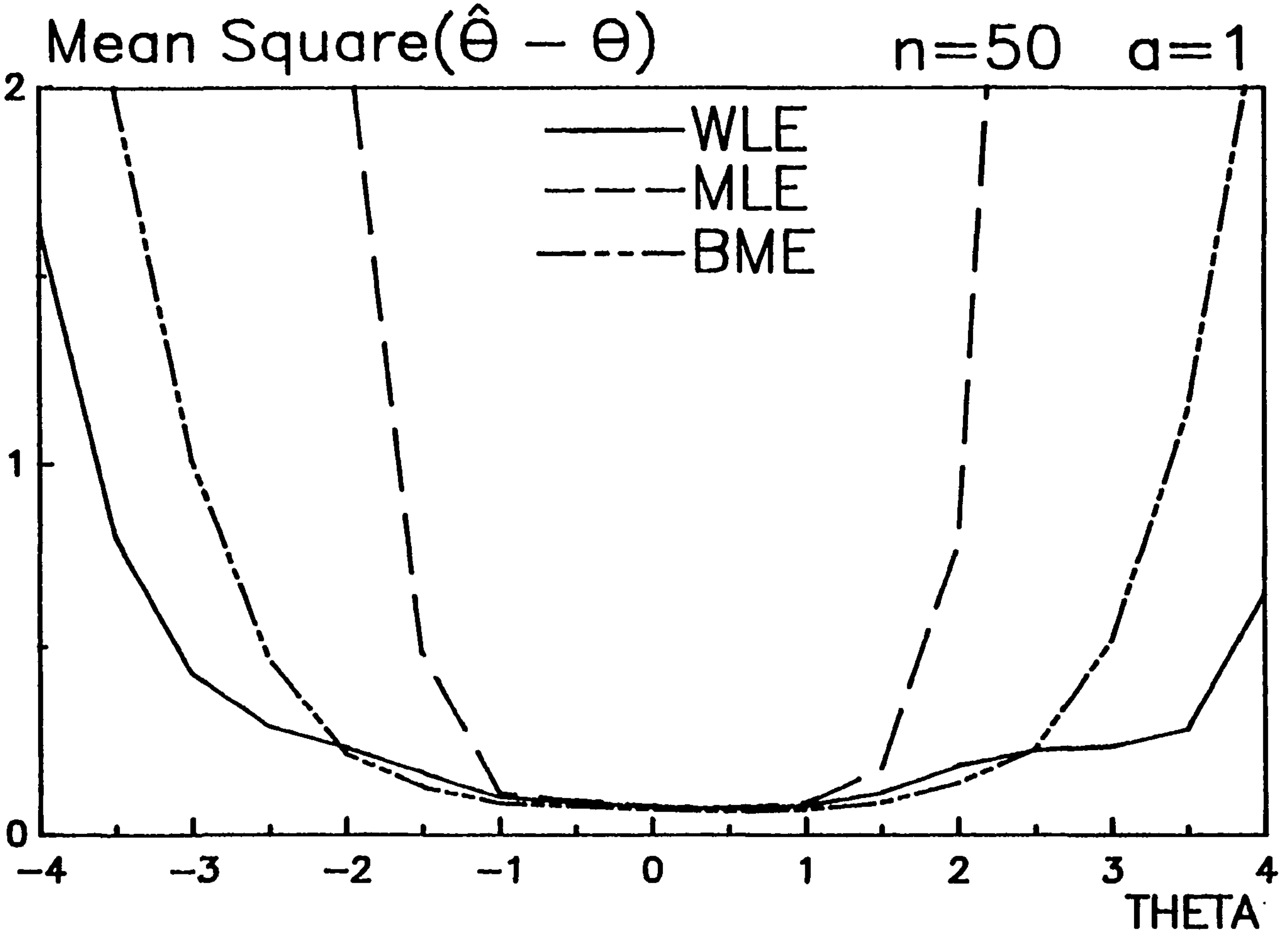

Figure D.26
Mean Squared Error of θ^ on Conventional Test with 50 Items, All a = 1, Normally Distributed b, and All c = 0.20 .

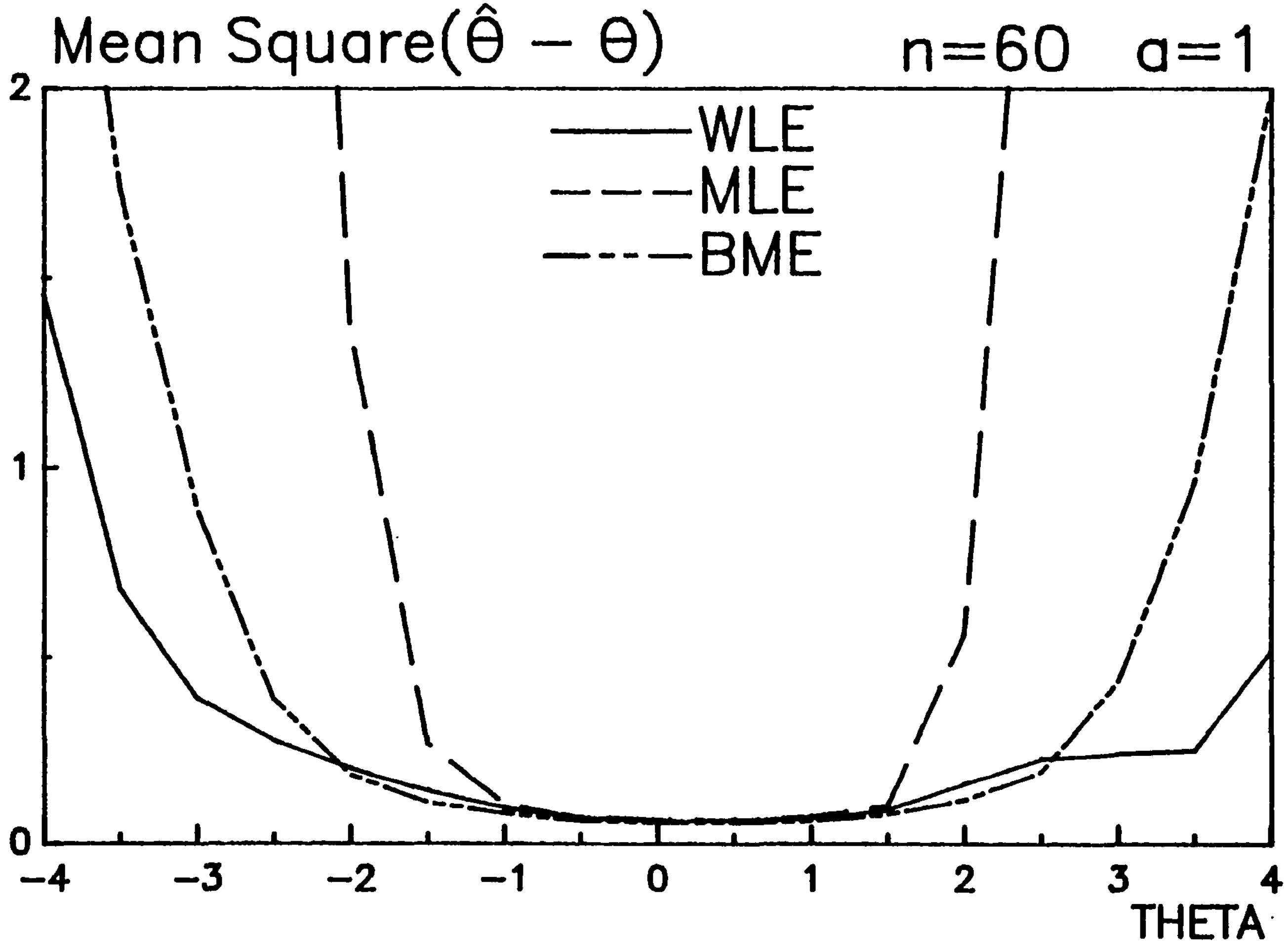

Figure D.27
Mean Squared Error of θ^ on Conventional Test with 60 Items, All a = 1, Normally Distributed b, and All c = 0.20 .

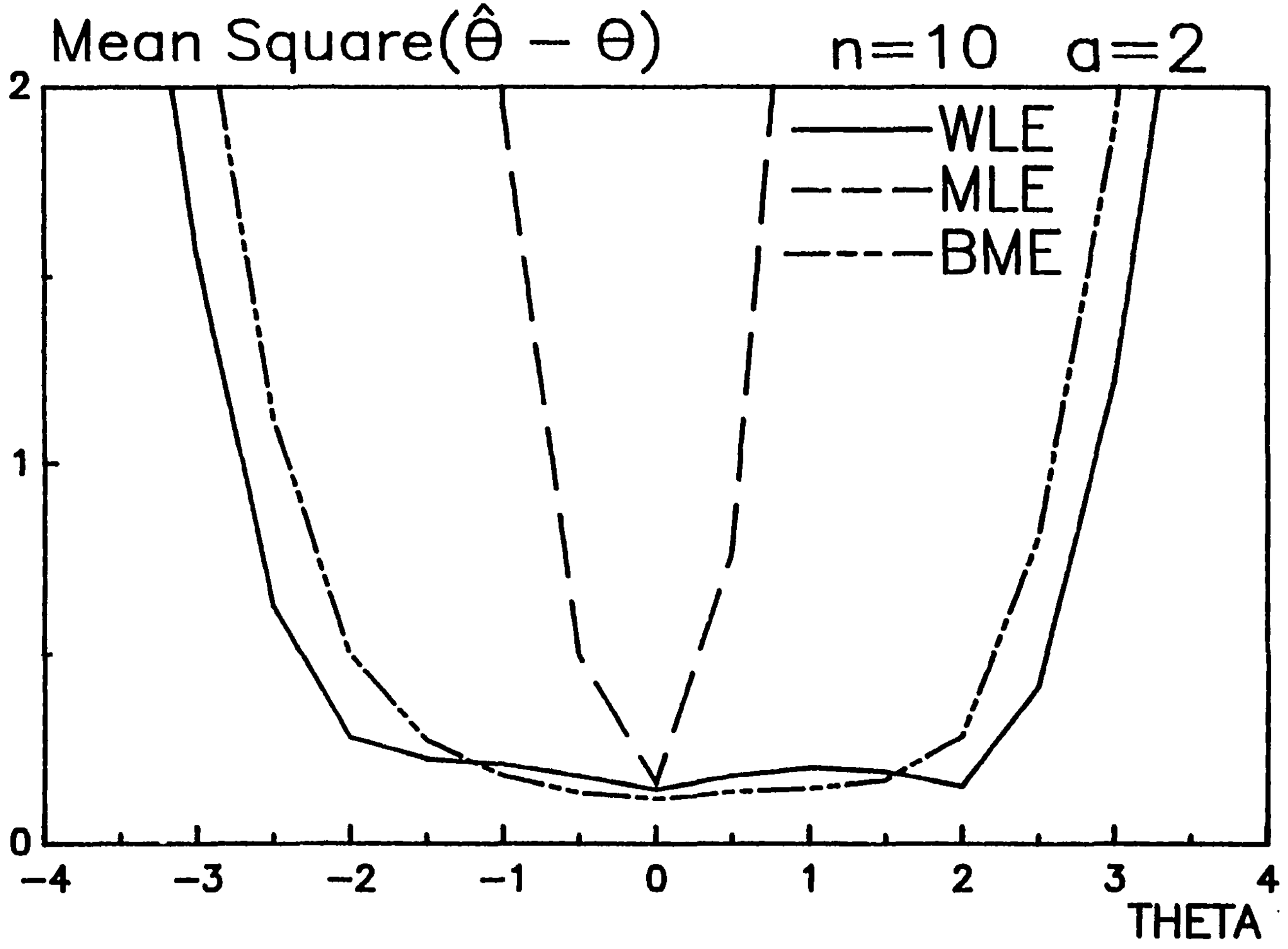

Figure D.28
Mean Squared Error of θ^ on Conventional Test with 10 Items, All a = 2,
Normally Distributed b, and All c = 0.20 .

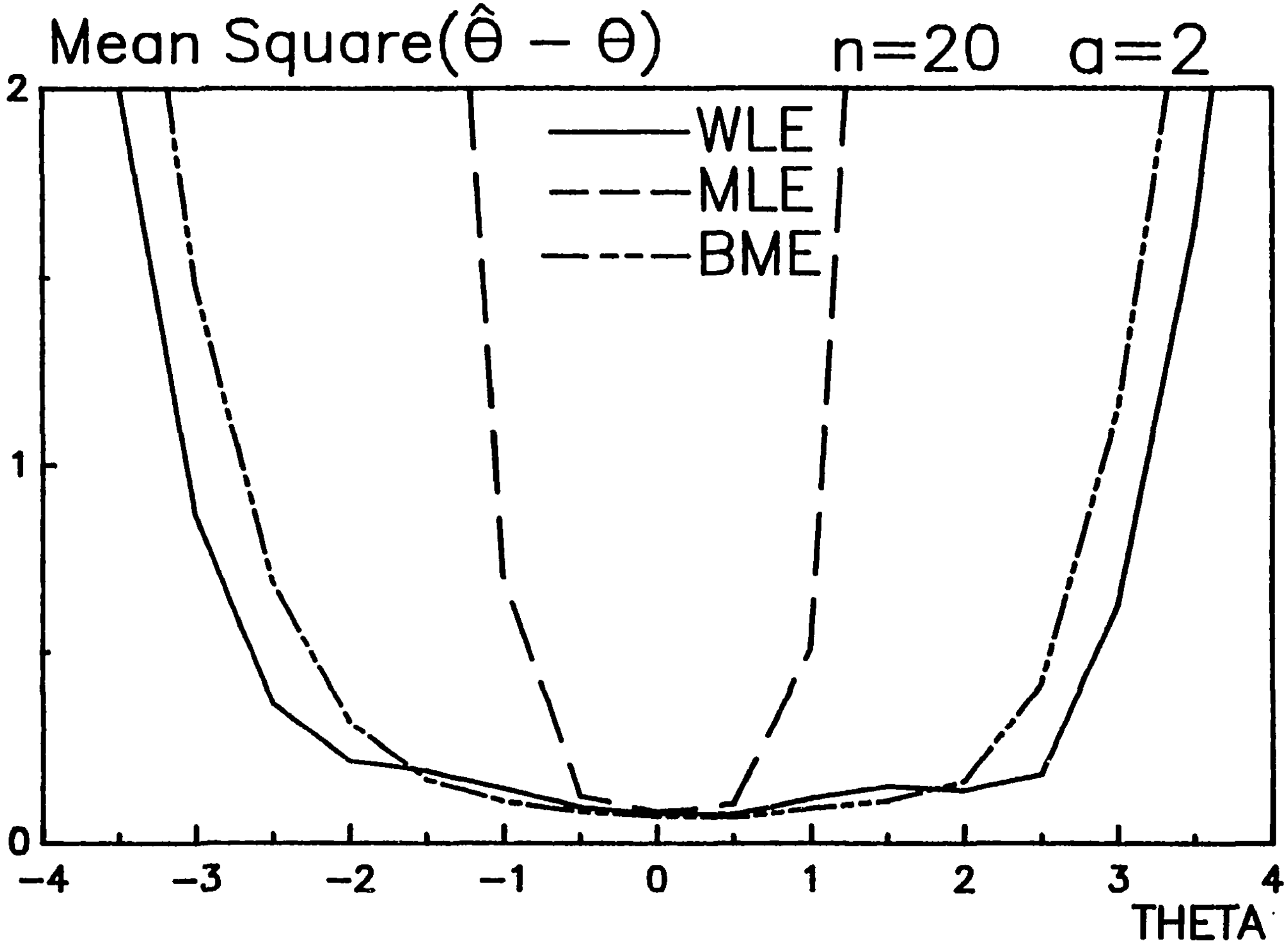

Figure D.29
Mean Squared Error of θ^ on Conventional Test with 20 Items, All a = 2, Normally Distributed b, and All c = 0.20 .

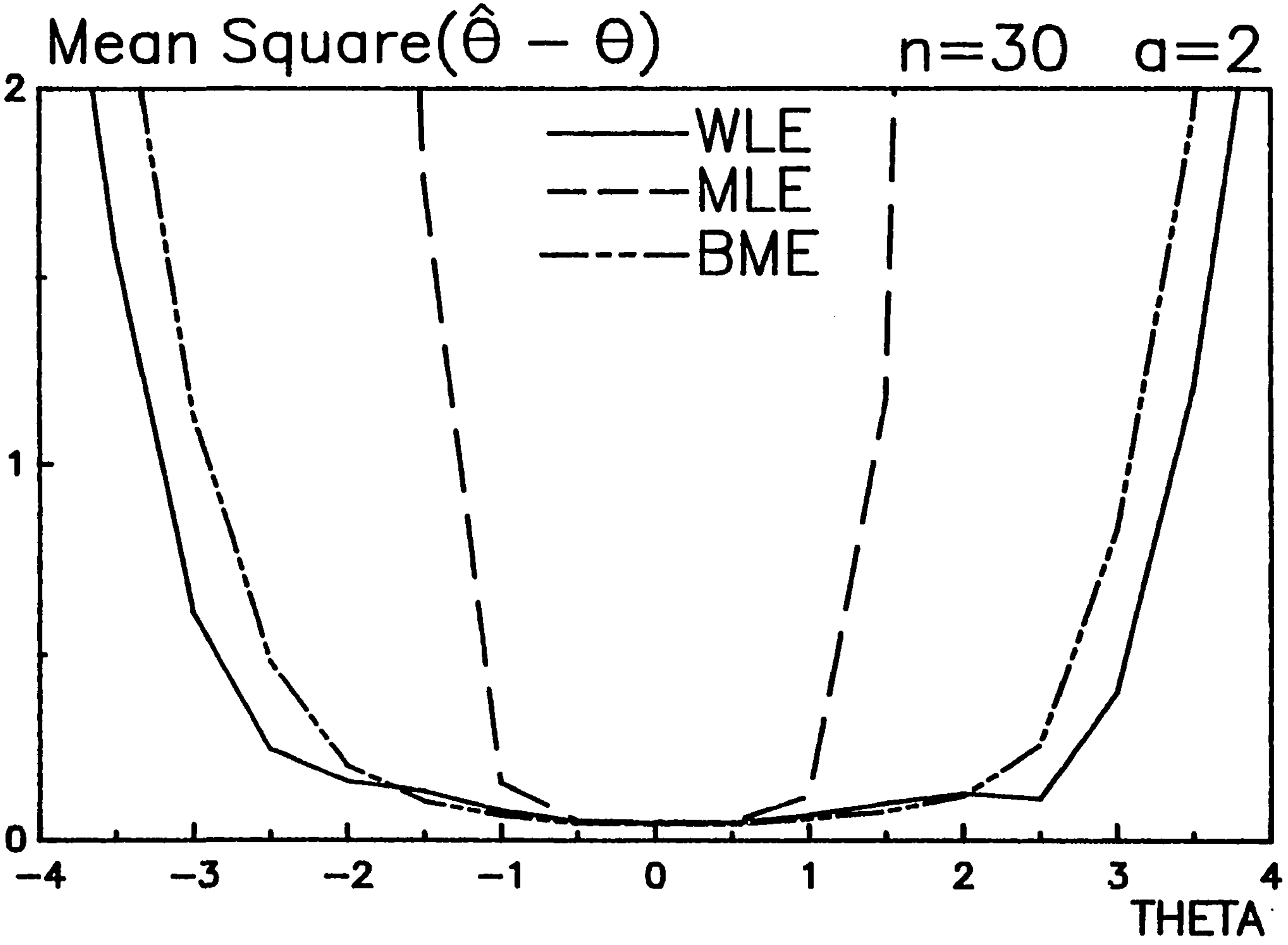

Figure D.30
Mean Squared Error of θ^ on Conventional Test with 30 Items, All a = 2, Normally Distributed b, and All c = 0.20 .

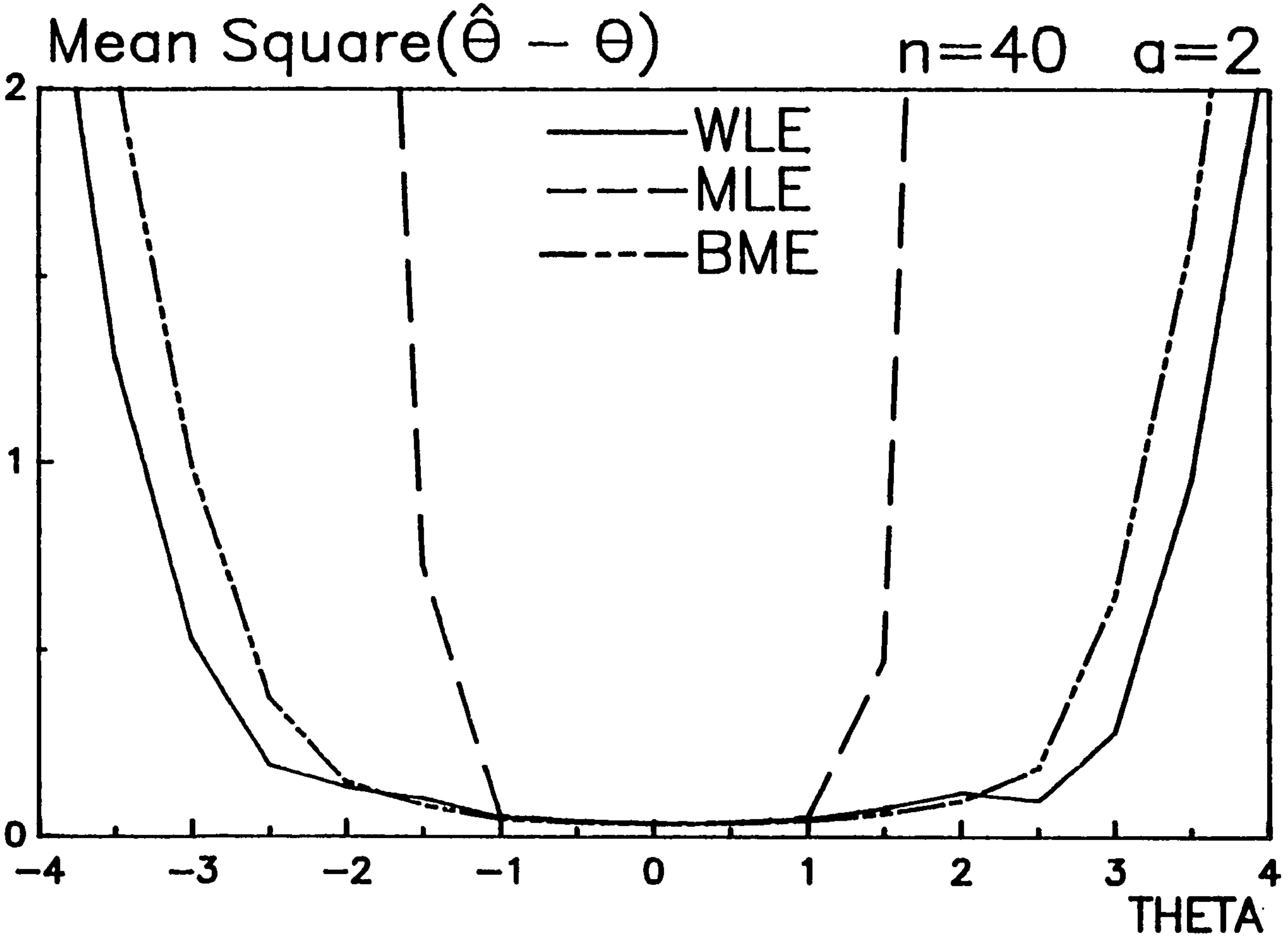

Figure D.31
Mean Squared Error of θ^ on Conventional Test with 40 Items, All a = 2, Normally Distributed b, and All c = 0.20 .

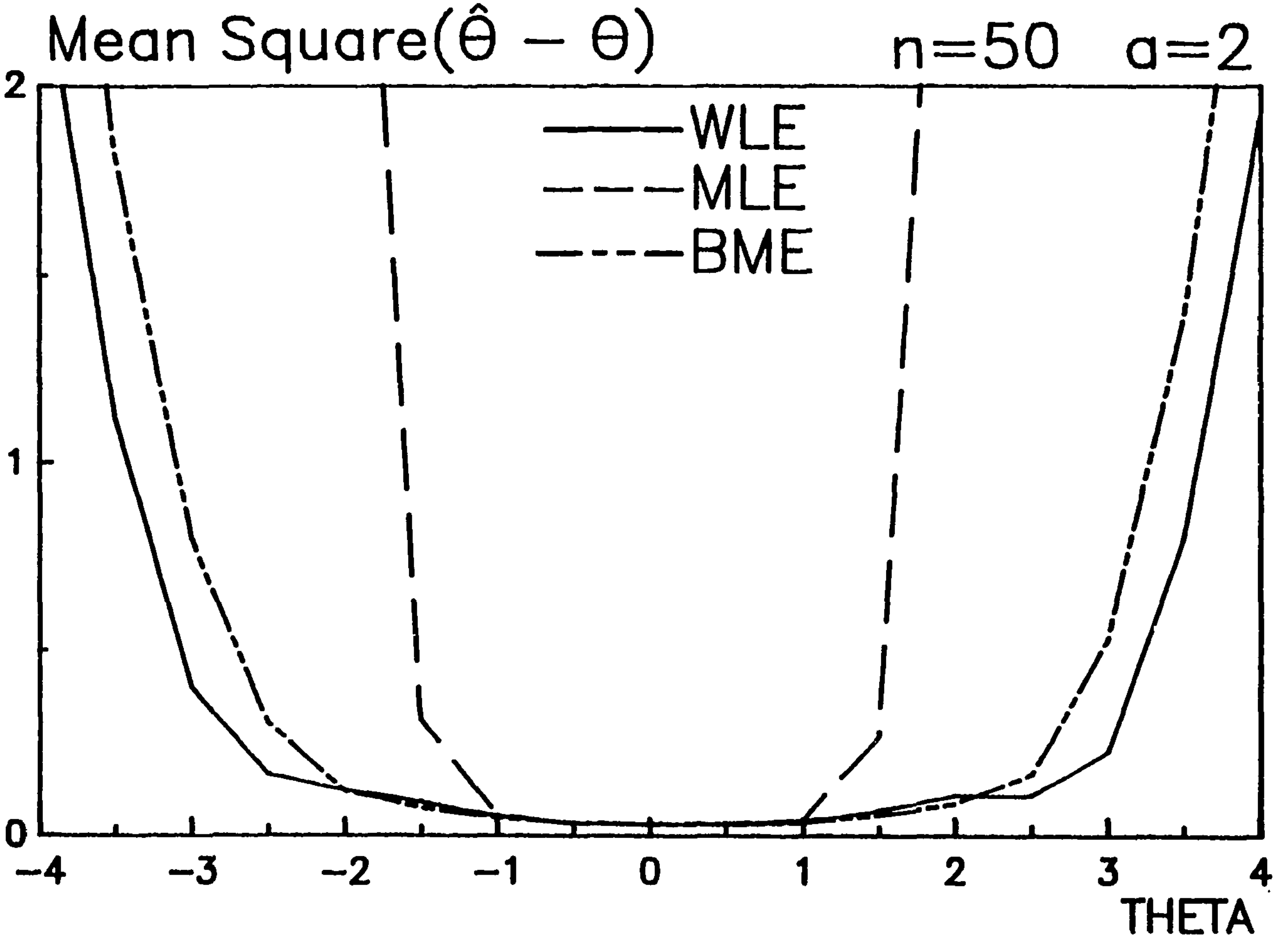

Figure D.32
Mean Squared Error of θ^ on Conventional Test with 50 Items, All a = 2, Normally Distributed b, and All c = 0.20 .

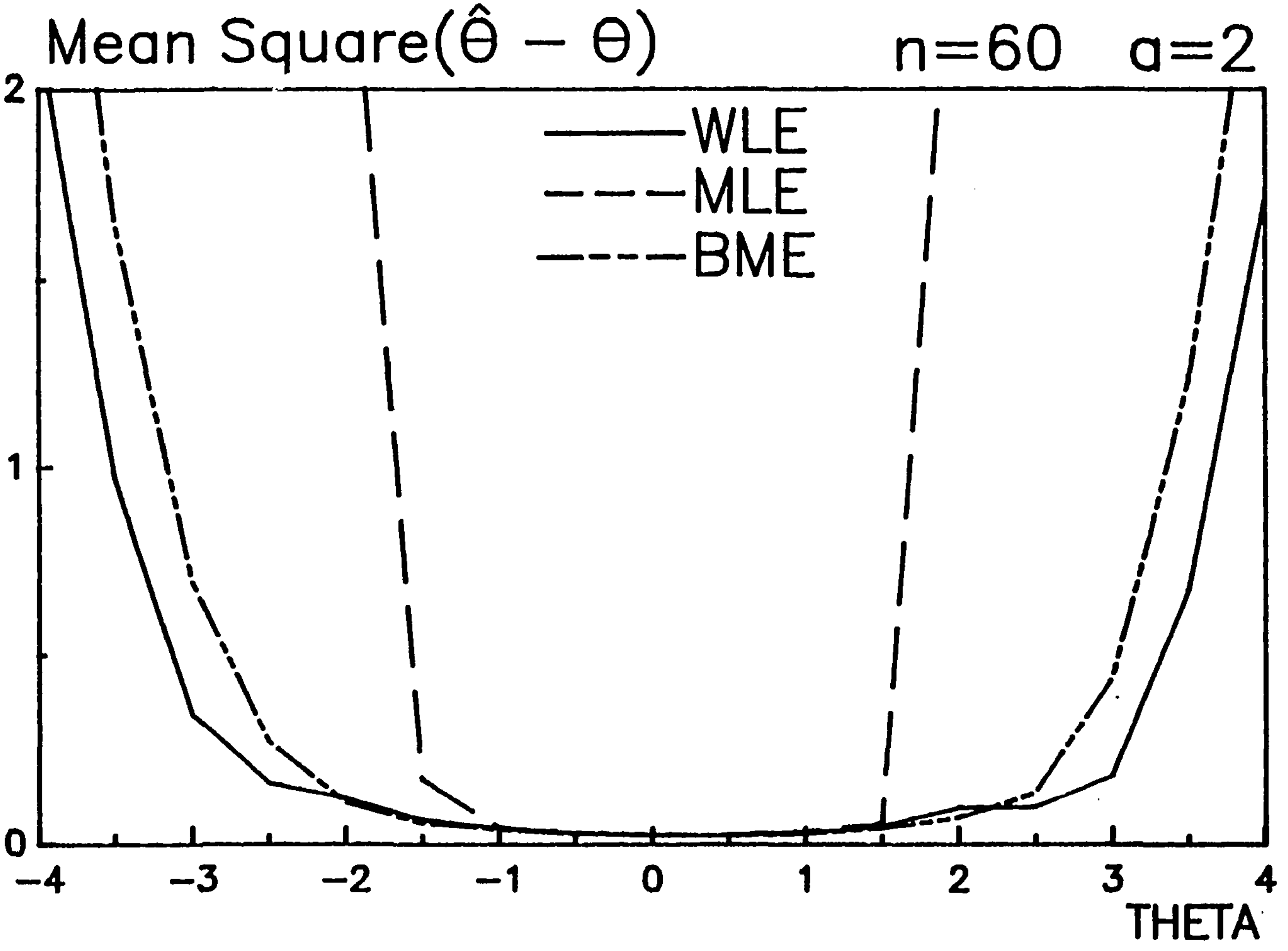

Figure D.33
Mean Squared Error of θ^ on Conventional Test with 60 Items, All a = 2, Normally Distributed b, and All c = 0.20 .

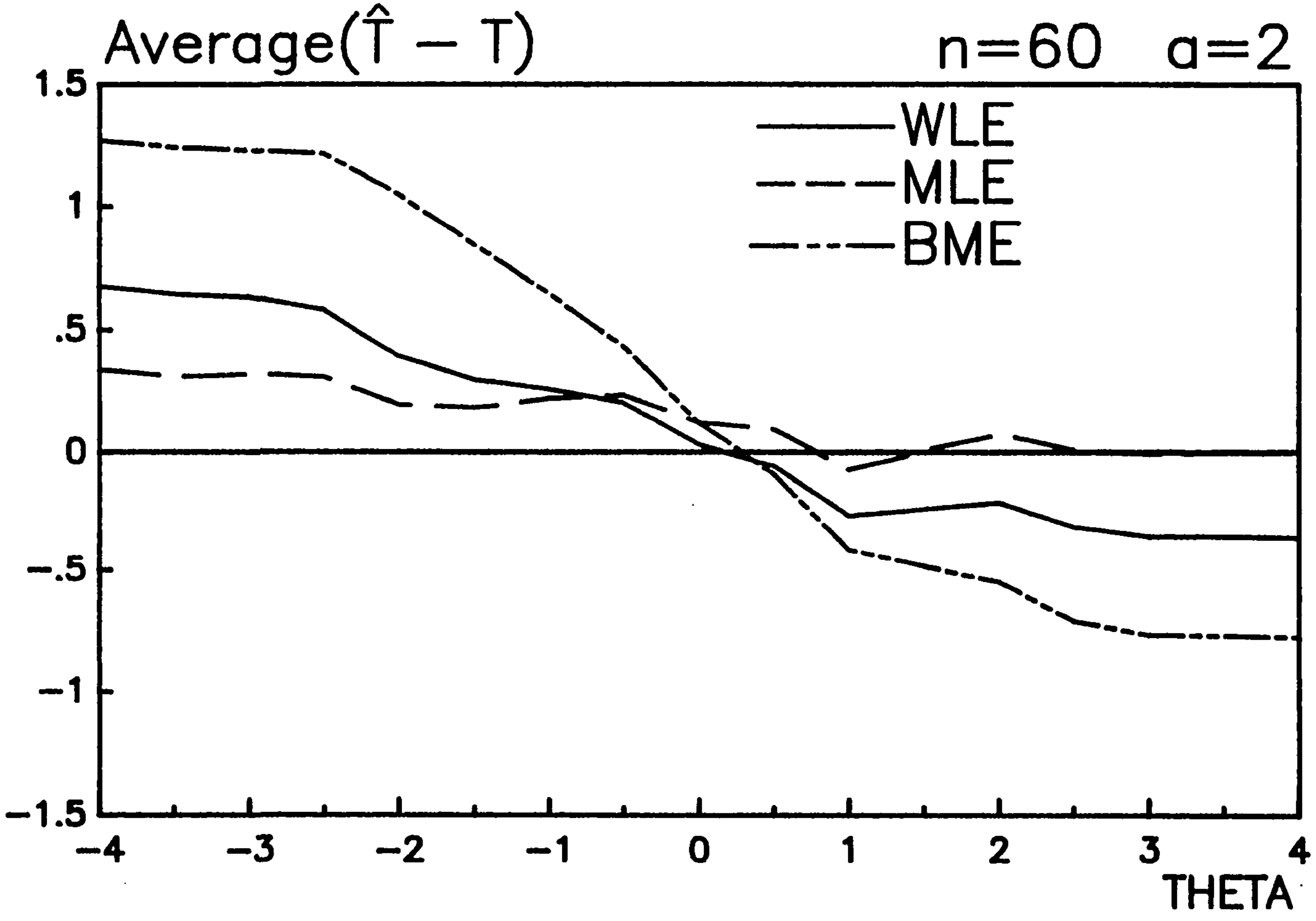

Figure D.34
Average Estimation Error of T^ on Conventional Test with 60 Items, All a = 2, Normally Distributed b, and All c = 0.20 .

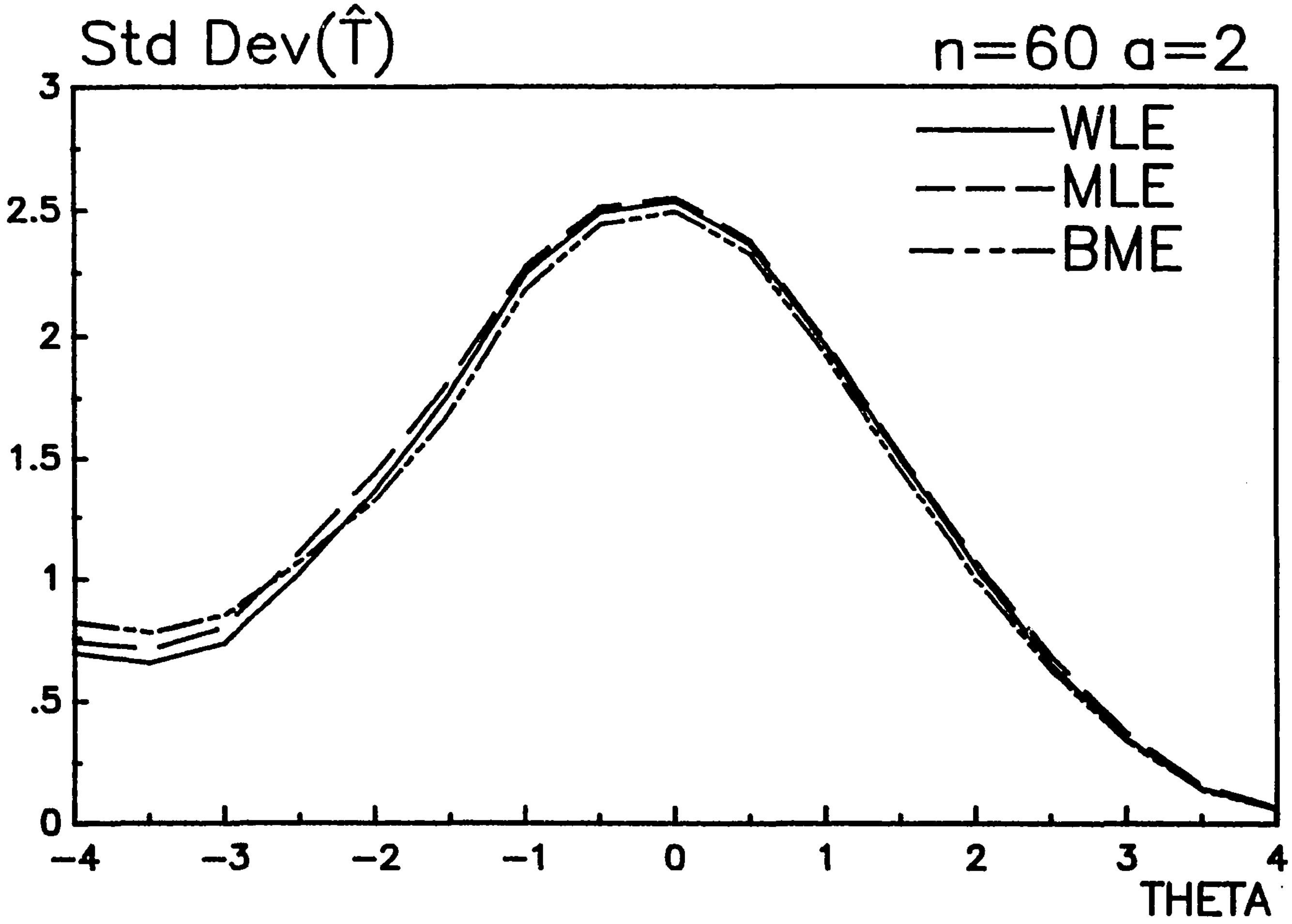

Figure D.35
Standard Deviation of T^ on Conventional Test with 60 Items, All a = 2, Normally Distributed b, and All c = 0.20 .

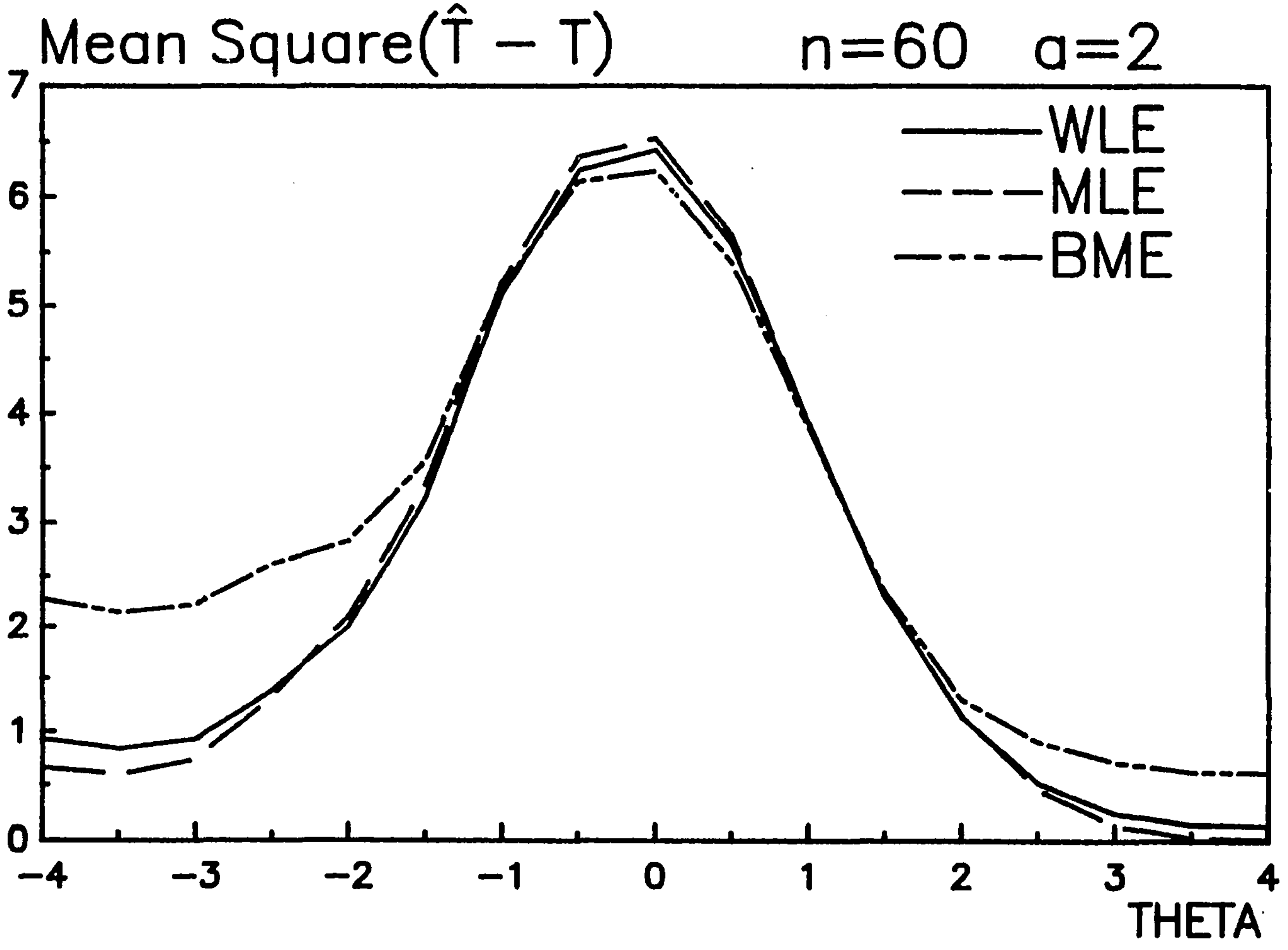

Figure D.36
Mean Squared Error of T^ on Conventional Test with 60 Items, All a = 2, Normally Distributed b, and All c = 0.20 .

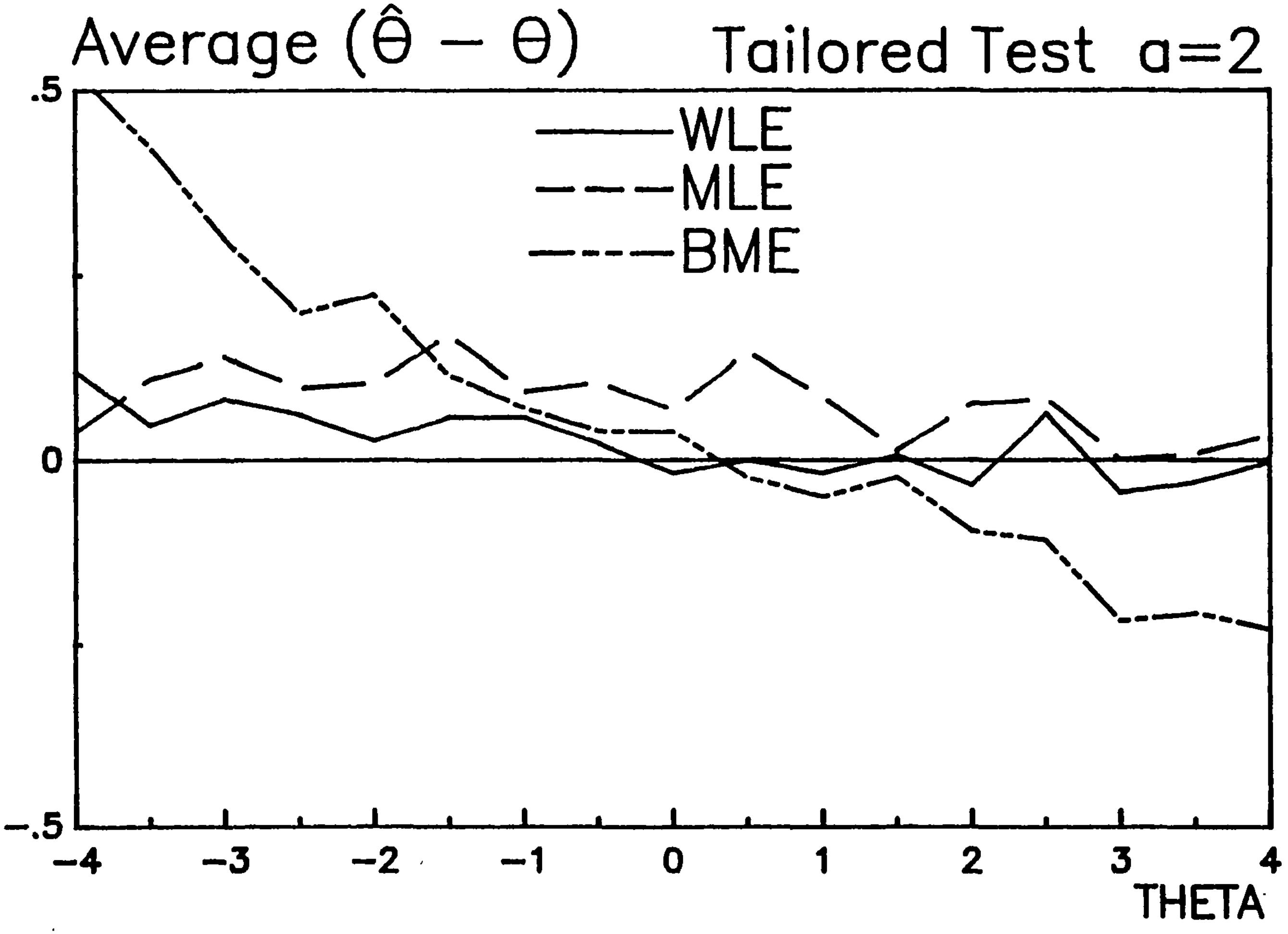

Figure D.37
Average Estimation Error of θ^ on Tailored Test with a = 2, Optimal b-parameter, and All c = 0.20 .

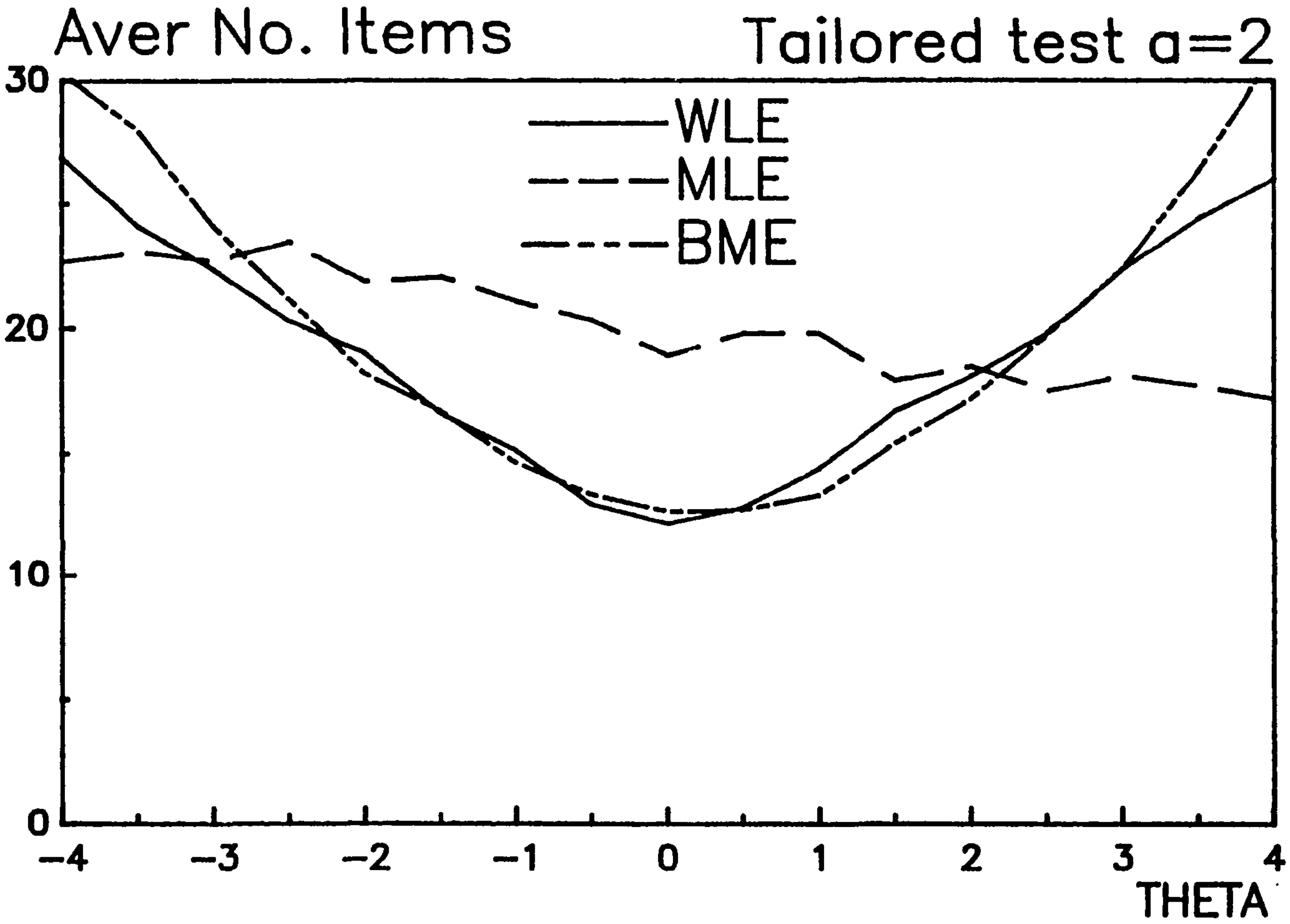

Figure D.38
Average Number of Items Administered on Tailored Test with All a = 2, Optimal b-parameter, and All c = 0.20 .

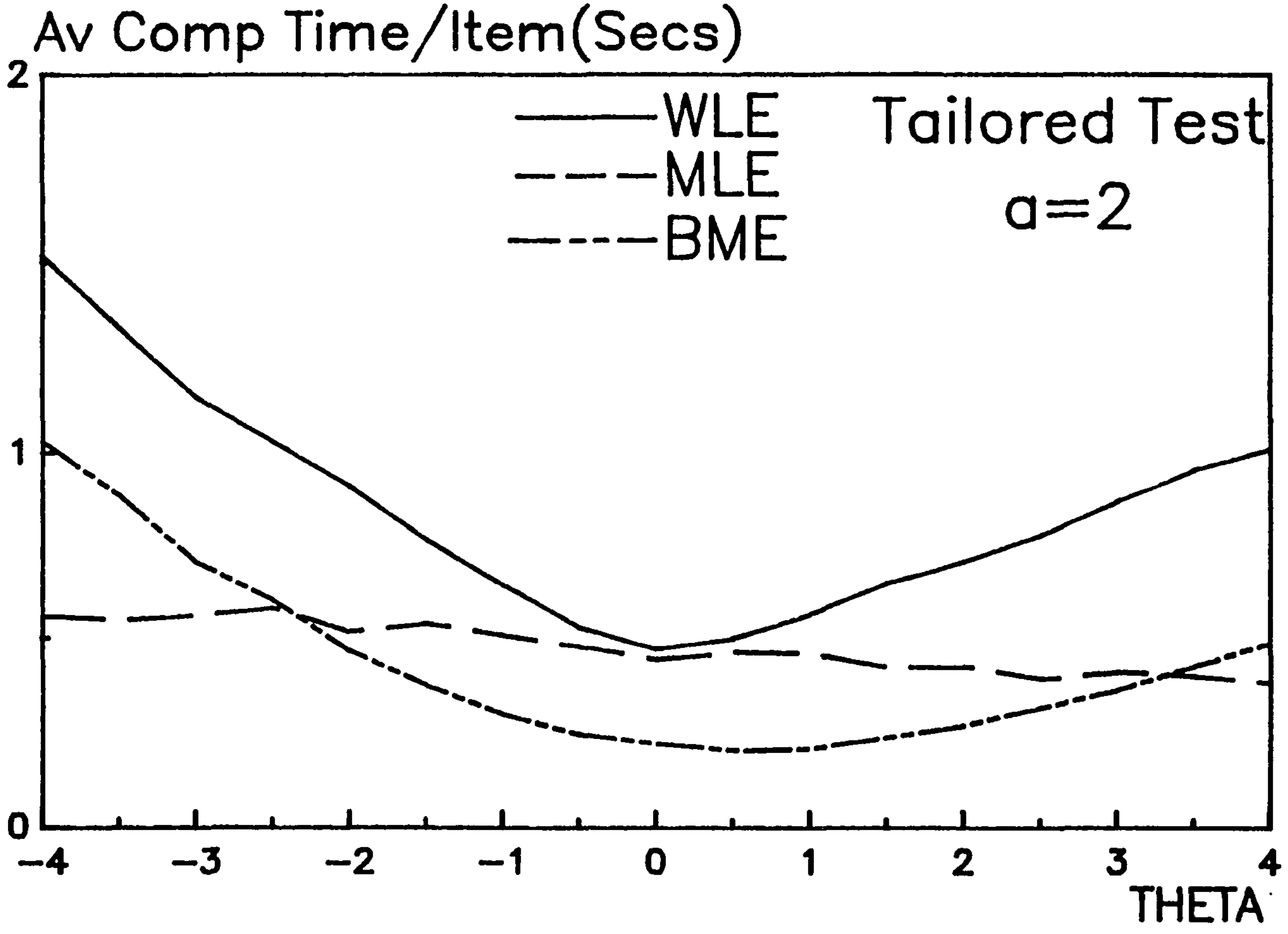

Figure D.39
Average Computation Time Between Items on Tailored Test with All a = 2, Optimal b-parameter, and All c = 0.20 .