An Examination of Item Review on a CAT Using the Specific Information Item Selection Algorithm

Ryan Bowles

University of Virginia

Mary Pommerich

ACT, Inc.

Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA, April 2001.

Abstract

Many arguments have been made against allowing examinees to review and change their answers after completing a computer adaptive test (CAT). These arguments include increased bias, decreased precision, and susceptibility to test-taking strategies. Results of simulations suggest that the strength of these arguments are reduced or eliminated by using specific information item selection (SIIS), under which items are selected to meet information targets, instead of the more common maximum information item selection, under which items are selected to maximize information. Because allowing item review offers several benefits, by using SIIS and an appropriate ability estimator, allowing item review on a CAT is probably warranted.

Introduction

An important controversy in computer adaptive testing is whether, after completing a test, examinees should be allowed to review their answers with the option of changing them. Examinees clearly want item review to be allowed, make use of review when it is allowed, and achieve more accurate estimates of ability with review (Benjamin, Cavell, & Shallenberger, 1984; Bowles, 2001; Vispoel, Hendrickson, & Bleiler, 2000; Waddell & Blankenship, 1994; Wise, 1996). Nonetheless, many preeminent researchers have concluded that item review should not be allowed on computer adaptive tests (CAT). For example, Wainer (1993) felt that review should not be allowed because of the loss of efficiency and the vulnerability to test-taking strategies. Green, Bock, Humphreys, Linn, and Reckase (1984, p. 356) argued that "the possible confusion and certain delay inherent in retracing the test to earlier items argues against allowing [review]." These studies argue against allowing item review on a CAT because of specific problems with item review. Because of its benefits, item review should be allowed, but only if the problems with item review can be sufficiently reduced or eliminated (Bowles, 2001; Wise, 1996).

One way to reduce the strength of the arguments against allowing review is through the choice of an appropriate item selection algorithm. Most of the research on item review is based on CATs using maximum information item selection (MIIS), under which items are selected to maximize the statistical information available for an examinee. This paper examines the effects of using a different item selection algorithm, specific information item selection (SIIS; Davey & Fan, 2000), on four of the problems with allowing review. These four problems are bias in estimation, decreased precision, and susceptibility to the Wainer and the Kingsbury strategies, two test-taking strategies available only with review on which examinees attempt to artificially increase their ability estimates.

Specific information item selection

Under MIIS, it is impossible to achieve a specified level of precision without varying across examinees the number of items administered, a procedure which yields variable speededness effects (Davey, Pommerich, & Thompson, 1999). SIIS was developed to avoid this problem through a fixed length test with a specified level of precision (Davey & Fan, 2000). Under SIIS, items are selected to meet a specified level of information, instead of a maximal level. Associated with each item administration is an information target chosen by the test-administrator, with the item selected to meet as closely as possible the information target given the current ability estimate. Figure 1 shows an example set of 35 information targets, for a 35 item test. The top line describes the information target for the entire test.

Given the information target and a posterior distribution of estimated ability, items are selected as follows:[1]

1. An *appropriate set* of items is identified. The appropriate set consists of all not yet administered items whose difficulty levels are within some distance ε of the current ability estimate. ε is chosen by the test administrator so that, even if little information is needed to meet the information target, the administered item is of an appropriate difficulty, that is, undiscriminating, rather than too easy or too difficult.

2. For each item *i* in the appropriate set, the item information function is integrated over the posterior distribution of estimated ability to yield an expected information gain, o_i.

3. At each ability level, the current information is subtracted from the information target for the next item. This difference is then integrated over the posterior distribution of estimated ability to yield an information need, T.

4. For each item *i*, o_i is compared to T, using the objective function $(T - o_i)^2$. This function can be adjusted to account for the contribution of other factors, such as content balancing.

5. The item that minimizes the objective function is selected, and submitted to whatever exposure control procedure is being used. If the exposure control procedure rejects the item, then select the item with the next smallest value of the objective function.

After each item is administered, the posterior distribution of estimated ability is updated, and the item selection algorithm is repeated with a new information target. The test ends when the specified number of items has been administered.

[1] The description of SIIS in this section is adapted from Davey and Fan (2000)

For a more detailed description of SIIS and its advantages over other item selection algorithms, please see Davey and Fan (2000).

Arguments against allowing review

Bias in estimation

The goal under MIIS is to achieve perfect targeting, where the test provides its maximal information at the true ability. When this occurs, the maximum likelihood estimator (MLE) and weighted likelihood estimator are unbiased (i.e. the estimated ability is not systematically different from the true ability; Lord, 1983; Wang & Vispoel, 1998; Warm, 1989). Answer changes under review results in a different ability estimate than that on which item selection was based. Therefore, with perfect targeting, review leads to a loss of information, and because the magnitude of bias is inversely related to the amount of information, an increase in the magnitude of bias (Lord, 1983).

With other commonly used estimators, the effect of review on bias is more complicated. For the two most common Bayesian estimators, expected *a posteriori* (EAP; Bock & Mislevy, 1982) and maximum *a posteriori* (MAP; Lord, 1986), the bias depends both on information and the distance from the mean of the prior distribution of ability. Thus, even for perfect targeting, there is bias toward the prior mean (Wang & Vispoel, 1998). When estimated ability changes because of answer changes, not only is there an increase in the magnitude of bias because of loss of information, but also a change in the bias from a move toward or away from the prior mean. Across the continuum of ability, though, the movement toward and away from the prior mean should even out, so that on average there will be an increase in the magnitude of bias.

In practice, of course, perfect targeting is impossible. In reality, for most ability levels the bias is not zero regardless of the estimator, and the magnitude depends at least in part on the quality of the item pool (Wang, Hanson, & Lau, 1999). This complicates the analysis of the change in bias resulting from item review, which may help explain why little research has addressed changes in bias outside of extreme situations (see Wainer strategy below; Vispoel, Rocklin, Wang, & Bleiler, 1999). However, an overall increase in bias can be expected, which may overwhelm any gains in accuracy resulting from review (Bowles, 2001; Wise, 1996)

Under SIIS, perfect targeting is not the goal of item selection. It is unclear *a priori* what effect item review will have on bias with SIIS. In fact, a test may provide more information after review than before, and therefore, the magnitude of bias may decrease after review. This study will look at the change in bias from allowing review on a CAT using SIIS, for each of the four estimators described above (MLE, WLE, EAP, MAP).

Decreased precision

Precision is defined as the inverse of the standard error, and is mathematically equivalent to the test information. When a test is perfectly targeted (that is, information is maximized), precision is maximized. Therefore, the answer changes during review lead unavoidably to a reduction in precision. While perfect targeting is not possible in reality, on average a loss in precision can be expected.

Several studies have addressed this problem with item review under MIIS. The results are consistent, in that there is a slight reduction, less than 2%, in precision under normal circumstances (Lunz & Bergstrom, 1995; Lunz, Bergstrom, & Wright, 1992; Stone & Lunz, 1994; Vispoel, 1998; Vispoel et al., 2000). Although the reductions in measurement precision may be small on average, it is important to note that results vary across examinees. For example, one examinee in Lunz et al. (1992) changed 10 items from wrong to right and only two from right to wrong, yielding a large increase in ability estimate and a corresponding large increase in standard error.

As Vispoel et al. (1999) point out, on a test with a cut score, losses of precision can be especially detrimental. They looked at an extreme case of precision loss (the Wainer strategy; see below) and found that the percentage of examinees with true ability below the cut score whose ability estimates were above the cut score was substantially higher than that for examinees with normal levels of precision. This percentage of false positives, however, will be much lower with precision losses in the 1% to 2% range typically observed, and probably negligible.

Under SIIS, perfect targeting for maximal information is not the goal of the item selection. In fact, when the current amount of information is above the SIIS goal, even if a perfectly targeted item is available, it will not be selected. There may then be no change in precision as a result of item review, and there may even be an increase in precision. A goal of SIIS is to measure examinees to a specified level of precision on a fixed-length test (Davey & Fan, 2000). Therefore, neither decreases nor increases in precision are desired, although it can be argued that increases in precision are less a problem than decreases. This study will look at the effects of item review on precision for a CAT implementing SIIS.

Wainer strategy

The Wainer strategy (Wise, 1996) is a slight variation on a strategy proposed by Wainer (1993).[2] Under the Wainer strategy, an examinee attempts to get all items incorrect during initial presentation, resulting in a test with items selected for a very low ability examinee, which under MIIS is an extremely easy test. During item review, the examinee attempts to answer all of the items correctly.

As Wise (1996) points out, the Wainer strategy is an attempt to violate the invariance principle of IRT, which says that the expected ability estimates should be the same no matter which items were administered. Systematic gains occur only when the bias of the ability estimator increases as a result of the Wainer strategy. In a simulation of the Wainer strategy, Vispoel et al. (1999) found that, with EAP, systematic gains occurred only for low to middle ability candidates, and the gains were very small. However, MLE was highly biased positively for high ability examinees, and was therefore more susceptible to the Wainer strategy.

Successful implementation of the Wainer strategy is very difficult. Gershon and Bergstrom (1995) simulated results from the Wainer strategy using MLE with a Bayesian approach to extreme scores as implemented in the computer program Winsteps (Linacre & Wright, 2000). They found that the Wainer strategy was highly risky, in that positively biased ability estimates were achieved only with a perfect score following review, and answering just one item incorrectly led to a large underestimate of ability.

Vispoel et al. (1999) found that, even on a CAT based on items designed for high school students, only 73% of a sample of college students were able to get all items

[2] Wainer (1993) developed the strategy for a test on which not only is review allowed, but also omitting items. In this case, the examinee should omit any items for which he or she is unsure of the correct answer. The examinee should answer incorrectly any items for which he or she is sure of the correct answer, then answer those items correctly during review.

incorrect before review, and fewer than 25% of the examinees were able to get all items correct after item review. Only 32% of the examinees received higher ability estimates with MLE under the Wainer strategy than they received on a non-adaptive fixed-item test (FIT) using items from the same source and administered in the same session. However, mean ability estimates on the CAT for the sample implementing the Wainer strategy were higher than a control group not implementing the Wainer strategy, although the difference was almost entirely a result of ceiling values for examinees with perfect scores. With EAP estimation, only 11% of examinees did better under the Wainer strategy than on the FIT. The mean ability estimate for those employing the Wainer strategy was substantially lower than for those who did not, even though subjects were randomly assigned to the two groups and received the same score on average on the FIT.

It is important to note that the Wainer strategy leads to an extreme loss in precision because, after review, the tests are highly mistargeted. Thus, even though the Wainer strategy generally does not yield a gain in estimated ability, some examinees will benefit from implementing the strategy. However, the greatest loss in precision will be for examinees with the highest abilities, for which a very easy test is most mistargeted (Vispoel et al., 1999). These examinees also presumably least need to use a risky strategy for a chance to artificially inflate their ability estimates.

The Wainer strategy is thus not a very good way to achieve an increased ability estimate. The benefits of the Wainer strategy depend strongly on getting a perfect score after review (Gershon & Bergstrom, 1995; Vispoel et al., 1999). Under MIIS, it is likely that a person attempting to implement the Wainer strategy will be administered the easiest items, maximizing the chance of getting a perfect score. Under SIIS, the items selected may not be the easiest ones available in the item pool, making a perfect score less likely, so it is even less likely that the Wainer strategy can be implemented successfully. This study will examine the effects of the Wainer strategy on a CAT using SIIS.

Kingsbury strategy

The Kingsbury strategy was operationally defined by Kingsbury (1996) but in essence was first proposed by Green et al. (1984). They noted that “to the extent that the applicant knows or concludes that item difficulty depends on previous responses, the perceived difficulty of the present item may be taken as a clue to the correctness of earlier responses" (p. 356). In other words, if the examinee recognizes that the next item is easier, he or she knows that the previous answer was probably incorrect and should change it during item review. The Kingsbury strategy refers to the changing of the answers that were initially guessed, based on perceived changes in difficulty.

Kingsbury (1996), in a simulation using a CAT calibrated with the Rasch model, assumed that an examinee guessed the answer to an item that was at least 1 logit more difficult than his or her ability. The guess was changed if the next item was at least 0.5 logits easier. Examinees had on average a small but possibly important increase of 0.11 in estimated ability. This result, however, is contingent on perfect recognition of decreases in item difficulty of at least 0.5 logits.

Wise, Finney, Enders, Freeman, and Severance (1999) looked at how well examinees were able to recognize changes in item difficulty. They also developed the generalized Kingsbury (GK) strategy, which is the Kingsbury strategy applied to all item responses, not just to those that were initially guessed. Subjects took a CAT calibrated with a standard normal prior for ability, and after each item, were asked if the item was harder or easier than the preceding item. Not surprisingly, item pairs that differed more in difficulty were easier to distinguish than pairs that were close in difficulty. For items less than 0.5 units apart, subjects were able to differentiate items at a rate essentially equal to chance (51.5% correct). For items more than 0.5 units apart, subjects were able to choose the harder item successfully about 73% of the time.

Wise et al. (1999) then used the responses to a different CAT to simulate results from implementing the Kingsbury and GK strategies. It was assumed that an examinee guessed on any item more than 1 unit above his or her final ability estimate. If the next item was at least 0.5 units easier than the guessed item (indicating a wrong answer), with probability 0.73 the response was changed randomly to one of the remaining three options (successful difficulty change recognition). If the next item was at least 0.5 units harder than the guessed item, the response was changed with probability 0.27 (unsuccessful difficulty change recognition). For the Kingsbury and the GK strategies, there was a mean increase in ability estimate from before to after implementation of the strategy of 0.01 and -0.03 respectively. The variation for the GK strategy was relatively large (standard deviation equal to 0.18), compared to the Kingsbury strategy (SD = 0.02).

The results from Kingsbury (1996) and Wise et al. (1999) cannot be considered conclusive. Neither study looked at real people implementing the strategies in a real testing situation. People do not answer an item either by knowing an answer without doubt or by guessing totally at random. Evidence from multiple choice tests suggests that people tend to give confidence ratings for each response when they are unsure of the correct answer (Ben-Simon & Ben-Shachar, 2000). Without results from real people attempting to implement the Kingsbury strategy, it is impossible to know if and how the strategy works.

The results of the Kingsbury strategy depend crucially on the item selection algorithm. A correct answer always leads to a higher estimated ability, and under MIIS, a higher estimated ability almost always leads to an item with higher difficulty being selected. Under SIIS, the relationship between item responses and changes in item difficulty is less predictable. Therefore, MIIS is more susceptible to the Kingsbury strategy than SIIS. This study will examine the effects of the Kingsbury and generalized Kingsbury strategies on a CAT implementing SIIS.

Method

Four simulation studies were completed, all using the same item pool and adaptive algorithms. The first study looked at the bias and standard error on a CAT using SIIS without review. This was used as a baseline for the remaining three studies. Study 2 looked at changes in bias and standard error from allowing review on a CAT using SIIS. Study 3 looked at the results of the Wainer strategy under SIIS. Study 4 looked at the results of the Kingsbury and generalized Kingsbury strategies under SIIS.

Item pool and test characteristics

The item pool consisted of 600 items drawn from 10 forms of a college entrance mathematics test. The items were calibrated with the 3PL item response theory model, using large samples of real data. The formula for the 3PL model is given in equation 1.

$$\Pr(X_i = 1 | q, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{\exp[1.7a_i(q - b_i)]}{1 + \exp[1.7a_i(q - b_i)]} \qquad (1)$$

where X_i=1 indicates a correct response to item *i*, θ is the ability of the examinee, and a_i, β_i, and c_i are the discrimination, difficulty, and pseudoguessing parameters, respectively, for item *i*. Discrimination parameters ranged from 0.25 to 2.2, with a mean of 0.99. Difficulty parameters averaged 0.42, with a range from –3.8 to 4.0. Pseudo-guessing parameters ranged from 0.03 to 0.50, with a mean of 0.19. All simulated tests had a fixed length of 35 items. Provisional ability estimates were generated with EAP, while the final ability estimator varied (see below). Item exposure rates were controlled using a Hybrid method of exposure control (Fan, Thompson, & Davey, 1999). Items were selected to be no more than 1 unit in difficulty different from the current ability estimation (ε=1), so that simulated examinees were not given items that were too easy or difficult.

Specific information target

An inverse U-shaped target was used, with the target chosen to match the average precision under MLE of the 10 forms of the mathematics test from which the items were selected. This target was selected to provide first- and second-order equity between the paper-and-pencil and CAT versions of the mathematics test (Lord, 1980; Davey & Thomas, 1996; Thompson, Davey, & Nering, 1998). Intermediate information targets were proportional to the test information target, so that after n items, the information target was n/35 of the final information target. Figure 1 shows the intermediate and final information targets for the entire range of ability.

Estimators

Four estimators of final ability were examined, MLE, WLE, EAP, and MAP. For both MLE and WLE, ability estimates were bounded at –5 and 5. Precision for MLE and WLE was measured by estimating the standard error with the asymptotic standard error. For the EAP and MAP, the prior distribution was a standard normal distribution with 31 quadrature points spread evenly between –3 and 3. The standard error was estimated by the standard deviation of the posterior distribution (De Ayala, Schafer, & Sava-Bolesta, 1995).

Study 1: Simulation of standard CAT without item review

Method

1000 simulees were generated at each of 11 equally-spaced levels of true ability θ_T between –5 and 5 inclusive. A CAT using SIIS was simulated with item responses from the 3PL model with $\theta = \theta_T$. Final ability estimates, $\hat{\theta}_T$, were calculated using each of the four estimators. The amount of bias, $\text{Bias}_T = \hat{\theta}_T - \theta_T$, and the standard error, SE_T were also calculated.

Results

Figure 2 graphs the bias, Bias_T, for each of the four estimators, averaged over the 1000 simulees at each ability level. There is a substantial amount of bias, especially in the tails, and particularly for the Bayesian estimators. For EAP, MAP, and WLE, the results are consistent with standard bias results for the estimators on paper-and-pencil tests, which also tend to have inverted U-shaped information functions. MLE is less biased than is typically found (Lord, 1983), but this is because of ceiling and floor effects that result in the inward bias near the extremes. The ceiling and floor effects also increase the magnitude of bias with WLE, although in the same direction in which the estimator is naturally biased.

Figure 3 graphs the standard error, SE_T for each of the four estimators. The results are consistent with typical results for paper-and-pencil tests. For MLE, the standard error is very close to the inverse of the SIIS target.

Study 2- Simulation of changes in bias and precision with review

Method

Item review was simulated by assigning to each simulee two ability levels, a before-review ability, θ_b, and an after-review true ability, θ_T. Since most examinees benefit a small amount from review (Benjamin et al., 1984; Waddell & Blankenship, 1994), the change was simulated with a baseline condition of a small increase in ability of

0.1 units, where $\theta_T = \theta_b + 0.1$. Variability in the amount of change was simulated with three additional levels of change from before-review ability to after-review ability, -0.1, 0.4, and 1.0.

1000 simulees were generated at each of 11 equally spaced θ_T levels between –5 and 5 inclusive, and the corresponding θ_b levels were calculated according to the level of change in ability. A CAT was simulated under SIIS with item responses generated from a 3PL model with $\theta = \theta_b$. For each estimator, a before-review estimate of ability was calculated, $\hat{\theta}_{b,b}$, where the first subscript gives the ability which is being estimated, and the second subscript gives the ability on which the initial item responses and item selection were based. The resulting before-review bias, $\text{Bias}_{b,b} = \hat{\theta}_{b,b} - \theta_b$, and before-review standard error, $\text{SE}_{b,b}$, were also calculated. After-review responses to the previously selected items were then generated under a 3PL model with $\theta = \theta_T$. The after-review ability estimate, $\hat{\theta}_{T,b}$, was calculated for each estimator, along with the resulting after-review bias, $\text{Bias}_{T,b} = \hat{\theta}_{T,b} - \theta_T$, and after-review standard error, $\text{SE}_{T,b}$. A simulation was also run with $\theta = \theta_T$ (equivalent to Study 1), resulting in the ability estimate $\hat{\theta}_{T,T}$, bias $\text{Bias}_{T,T}$, and standard error $\text{SE}_{T,T}$.

Three methods of judging the effects of item review on bias were examined, with $|\cdot|$ indicating the absolute value:

1. $\text{Bias}_{T,b}$, the amount of bias after review
2. $|\text{Bias}_{T,b}| - |\text{Bias}_{b,b}|$, the change in magnitude of bias from before review to after review
3. $|\text{Bias}_{T,b}| - |\text{Bias}_{T,T}|$, the change in magnitude of bias from before review to after review, controlling for changes in the ability being estimated

For all three, the results were averaged over the 1000 simulees at each ability level.

Three methods of judging the effects of item review on precision were examined:

1. $\text{SE}_{T,b} - \text{SE}_{b,b}$, the magnitude of change in standard error from before review to after review
2. $(\text{SE}_{T,b} - \text{SE}_{b,b}) / \text{SE}_{b,b} * 100\%$, the percentage change in standard error from before review to after review
3. $(\text{SE}_{T,b} - \text{SE}_{T,T}) / \text{SE}_{T,T} * 100\%$, the percentage change in standard error, controlling for changes in the ability being estimated

For all three, the results were averaged over the 1000 simulees at each ability level.

Although there are many ways to simulate item responses with increased ability from item review, the method of adjusting the before-review ability and regenerating all item responses was chosen for its simplicity in programming. The method overestimates the number of item changes, but the results can be considered a worst-case scenario of answer changes. With 1000 simulees at each ability level, the effects of the answer changes on the ability estimate should average out, leaving the average level of bias mostly unaffected by this choice of simulation method. The standard error after review, however, will be higher than would be expected with a more realistic simulation of answer changes.

Results

Figure 4a, 4b, 4c, and 4d are graphs of the bias, $\text{Bias}_{T,b}$, for a true ability that is, respectively, -0.1, 0.1, 0.4, and 1.0 theta units higher with review. The horizontal axis for these graphs and all other graphs in this paper is the true ability, Θ_T. There is a substantial amount of bias, especially in the tails, and particularly for the Bayesian estimators. For EAP, MAP, and WLE, the results are consistent with standard bias results for the estimators on paper-and-pencil tests, which also tend to have inverted U-shaped information functions. MLE is less biased than is typically found (Lord, 1983), but this is because of ceiling and floor effects that result in the inward bias near the extremes. The ceiling and floor effects also increase the magnitude of bias with WLE, although in the same direction in which the estimator is naturally biased.

Of more concern is the change in the magnitude of bias from before to after review, $|\text{Bias}_{T,b}| - |\text{Bias}_{b,b}|$, graphed in Figures 5a, 5b, 5c, and 5d for -0.1, 0.1, 0.4, and 1.0 increases in ability, respectively. For the most part, with positive changes in ability from review, the change in the magnitude of bias is negative for low abilities, positive for high abilities, and negligible for abilities around 0. For the negative change in ability after review, the pattern is reversed.

Most of this pattern is because of expected changes in bias resulting from changes in ability. There is an expected difference in bias between the before and after ability levels independent of item review. Because of this expected change in bias, a more informative comparison is $|\text{Bias}_{T,b}|$ - $|\text{Bias}_{T,T}|$, which controls for changes in ability. Figures 6a, 6b, 6c, and 6d graph this difference in the magnitude of bias. For most ability levels and estimators, there is a small change in the magnitude of bias, generally between -0.1 and 0.1. Only for two values for moderately high ability levels with large after-review ability increases under MLE is the difference in magnitude substantially larger than 0.2.

The change in magnitude of bias is generally highest for MLE. WLE, EAP, and MAP for the most part display the same pattern of differences in bias, with none of the three being consistently higher or lower than the others.

Figures 7a to 7d graph the change in standard error from before to after review, $SE_{T,b}$ - $SE_{T,T}$. For EAP and MAP, there is little change in the standard error, with all changes between -0.12 and 0.11. For WLE, changes are generally small, but with a maximum of about 0.33 and a minimum of -0.45, both for an increase in ability of 1.0 from review. For MLE, there are large increases, particularly for high ability examinees with large ability increases from review.

The level of the standard error complicates interpretation of the changes in magnitude of standard error. More informative is the percentage change in standard error from before to after review, ($SE_{T,b}$ - $SE_{b,b}$) / $SE_{b,b}$ * 100%, graphed in Figures 8a to 8d. For the 0.1 decrease in ability from review, there is a 5 to 20 percent increase in standard error for moderately low abilities (-2 and –1), and a large percent increase in standard error for very high abilities with MLE. For the increases in ability from review, there is usually a reduction, and at worst a very small percentage increase in standard error for abilities below 0, and a loss in precision for positive abilities. Exceptions to this are for very low abilities with large increases in ability from review for MLE, and slight decreases in standard error for high abilities with EAP. The pattern of changes in standard error is expected, because of the inverted U-shaped information target. Therefore, as with the analysis of changes in bias, $SE_{T,b}$ is compared to $SE_{T,T}$ to remove effect of changes in ability.

Figures 9a to 9d graph the percentage changes in standard error, controlling for changes in ability, for the four levels of ability change from review, ($SE_{T,b}$ - $SE_{T,T}$) / $SE_{T,T}$ * 100%. For the decrease in ability, the largest increase in standard error for EAP, MAP, and WLE is under 9%. For MLE, there are increases above 10% for abilities between –1 and –3, and very large increases for high ability examinees. For the small increase in ability resulting from review, there are no increases in standard error greater than 10% except for high ability examinees with MLE. For an ability increase of 0.4 from review, there are increases in standard error greater than 10% for MLE for ability level –3 and above about 1.2. For WLE, the maximum is 17% at an ability of 3, but otherwise the increase is below 10%, while for EAP and MAP, the maximum is under 8%. For ability gains of 1.0 resulting from review, there are increases in standard error greater than 10% for all positive abilities for all estimators. The pattern of change, greatest for MLE, followed by WLE, MAP, and EAP, is fairly consistent over all four changes in ability.

Study 3: Simulation of Wainer strategy

Method

1000 simulees were generated at each of 11 equally spaced true ability levels, θ_T, between –5 and 5 inclusive. The CAT was simulated based on SIIS, but with each item answered incorrectly before review. Review consisted of generating standard item responses from the 3PL model to the already administered items with $\theta = \theta_T$. The ability estimate, $\hat{\theta}_{T,W}$, was calculated, where the first subscript indicates the ability being estimated, and the second subscript indicates the ability on which item selection was based, with W indicating item selection based on the Wainer strategy (that is, all items answered incorrectly before review). The bias, $\text{Bias}_{T,W}$, and standard error, $SE_{T,W}$ were also calculated. A simulation was also run for a standard CAT without using the Wainer strategy (Study 1). Ability estimates for this condition, $\hat{\theta}_{T,T}$, were also generated, along with $SE_{T,T}$.

The effects of the Wainer strategy were measured with:

1. $\text{Bias}_{T,W}$, the amount of bias from implementing the Wainer strategy
2. $\hat{\theta}_{T,W}$ - $\hat{\theta}_{T,T}$, the amount of benefit from using the Wainer strategy
3. $(SE_{T,W} - SE_{T,T}) / SE_{T,T} * 100\%$, the percentage increase in standard error from using the Wainer strategy

For all three, the results were averaged over the 1000 simulees at each ability level.

This simulation overestimates the effects of the Wainer strategy, because it is assumed that examinees are able to choose an incorrect answer perfectly. This is certainly not the case in reality (Vispoel et al., 1999), so this simulation can be considered a worst-case scenario.

Results

Figure 10 graphs the bias from using the Wainer strategy, $\text{Bias}_{T,W}$, for the four estimators. This graph is similar to the expected bias curve. More interesting is the increase in ability estimate from using the Wainer strategy compared to answering the items as modeled, $\hat{\theta}_{T,W}$ - $\hat{\theta}_{T,T}$, graphed in Figure 11. MLE results in the greatest increase in estimated ability from the Wainer strategy, for a true ability of 3. For the other estimators there is a negative impact over most ability levels from using the Wainer strategy. The exceptions are a small increase of under 0.1 with EAP and MAP for abilities of –3 and below, and a very small increase of under 0.015 with WLE for abilities of -1 and 0.

Figure 12 graphs the percentage increase in standard error from using the Wainer strategy rather than responding to the items as modeled, $[SE_{T,W}-SE_{T,T}] / SE_{T,T} * 100\%$. There are large increases in standard error for all estimators, with especially large increases for positive abilities.

Study 4: Simulation of Kingsbury strategy

Method

The Kingsbury strategy was simulated with a 2x3x2 design. The first facet is the answers marked for possible change, with two conditions, generalized Kingsbury vs. Kingsbury strategy. The latter will be called the pure Kingsbury strategy to differentiate it from the generalized Kingsbury strategy. For the pure Kingsbury strategy, following Kingsbury (1996), all items for which the difficulty was at least one unit higher than the true ability were assumed to be guessed. Only guessed items were identified for possible change. This restriction was not used for the generalized Kingsbury strategy, that is, all item responses were identified for possible change.

The second facet is difficulty change recognition, with three conditions, optimal vs. real vs. not relevant. For the optimal condition, examinees were assumed to be able to recognize perfectly changes in item difficulty of at least 0.5 units, so that an answer identified for possible change was changed if the subsequent item was 0.5 units easier, indicating an incorrect answer. For the real condition, following Wise et al. (1999), an examinee was able to recognize successfully changes in item difficulty of at least 0.5 units with probability 0.75. Therefore, an answer identified for possible change was changed with probability 0.75 if the subsequent item was 0.5 units easier, and changed with probability 0.25 if the subsequent item was 0.5 units harder. The optimal condition can be considered a worst-case scenario, while the real condition is likely closer to real-world conditions. For the not relevant condition, the answer to any item identified for possible change was changed regardless of changes in item difficulty. This gives the effects, according to the 3PL model, of changing answers without implementing a strategy involving item review.

The third facet is answer changing behavior, with two conditions, perfect vs. imperfect. In the perfect change condition, if an incorrect item response was changed, it was changed to a correct response. In the imperfect change condition, the item response was changed to a correct response with probability 0.34, equivalent to guessing from among the remaining three options. In both cases, if a correct item response was changed, it was changed to incorrect. The perfect change condition can be considered a worst-case scenario and the imperfect change condition a best-case scenario from the test-administrators perspective, with reality somewhere in between, and likely closer to the imperfect change condition.

The estimated ability before implementing the strategy, $\hat{\theta}_T$, was calculated. Also calculated was the estimated ability after implementing the strategy, $\hat{\theta}_{xxx}$, where the three subscripts indicate the condition for each of three facets. Table 1 summarizes the conditions and labels for each of the subscripts.

The effects of the two types of Kingsbury strategies were measured in three ways:

1. $\hat{\theta}_{xxx}$ - $\hat{\theta}_T$, the amount of benefit from implementing the strategy

2. $\hat{\theta}_{xnx}$ - $\hat{\theta}_T$, the modeled effects of answer changing, which gives the effect of changing answers without implementing an item review strategy
3. IE_{xxx}, the benefit of the strategy independent of the modeled effects for condition xxx, where

$$IE_{xxx} \begin{array}{ll} = \hat{\theta}_{xxx}-\hat{\theta}_T - \max\{0,\hat{\theta}_{xnx}-\hat{\theta}_T\} & \text{if } \hat{\theta}_{xxx}-\hat{\theta}_T > 0 \\ = 0 & \text{otherwise} \end{array} \qquad (2)$$

Since the irrelevant condition yields results from changing answers without implementing an item review strategy, the formula for IE gives the benefits from implementing the strategy independent of the benefits from changing answers without implementing a strategy. This can be considered the true gains from the pure or generalized Kingsbury strategy.

Results

Figures 13a to 13d show the results for the four pure Kingsbury strategy conditions, $\hat{\theta}_{Kxx}$ - $\hat{\theta}_T$, and Figures 13e to 13h graph the results for the generalized Kingsbury strategy, $\hat{\theta}_{gxx}$ - $\hat{\theta}_T$. For the generalized Kingsbury strategy, there is essentially a monotonic decrease in the benefits of the strategy as ability increases. The largest benefits are approximately 0.3 for EAP and MAP and 0.7 for MLE and WLE under the perfect change condition, and approximately 0.1 for all four estimators under the imperfect change condition. For the pure Kingsbury strategy, as compared to the generalized Kingsbury strategy, the effects of the strategy drown out as ability increases, as fewer items are administered that are assumed to be guessed, and therefore candidates for answer changing. There is little difference between the real and optimal conditions, indicating little relationship between changes in item difficulty and item responses. As expected, the benefits of the strategy are higher under the perfect change condition than under the imperfect change condition.

The set of conditions that can be considered arguably closest to reality is the imperfect change with real recognition of difficulty changes. In this case, under the pure Kingsbury strategy, there is an increase in ability estimate resulting from use of the strategy for examinees with very low ability, between –5 and about –2.8. This increase is never greater than 0.1. Examinees with ability between about –2.8 and 0 are penalized for use of the strategy, with a maximum decrease in ability estimate of slightly more than 0.1 for MLE, at an ability level of approximately –2. Examinees with ability above 0 have no effect from the strategy because they are not administered items more than 1 unit more difficult than their true ability.

The results are similar for the generalized Kingsbury strategy, with very low ability examinees benefiting from the strategy at a maximum of 0.1. Examinees with ability above –3 do not benefit from the strategy.

Under the not relevant item difficulty recognition condition, there is no effect from an item review strategy, only model-based effects of answer changing. That is, all changes in ability estimates are a result of modeled effects of answer changing rather than strategic effects. Figures 14a to 14d graph the changes in ability estimate for the irrelevant condition, $\hat{\theta}_{xnx}$ - $\hat{\theta}_{T}$. The graphs tend to be exaggerated versions of the real and optimal condition graphs.

Judgment of the benefits of the Kingsbury strategy and generalized Kingsbury strategy are best made with the modeled effects removed. This was calculated with equation 2, and graphed in Figures 15a to 15d for the pure Kingsbury strategy, and Figures 15e to 15h for the generalized Kingsbury strategy. For the perfect answer changes condition, there is never a benefit from implementing the strategy above and beyond the benefits from changing all answers marked for possible change. For the imperfect answer changing condition, under both the pure and generalized Kingsbury strategies, and under both the real and the optimal difficulty change recognition, the independent strategic effect is negative except for the ability level of –3. For a -3 ability, there is only a slight advantage for employing the strategy rather than changing all answers marked for possible change under the strategy, under 0.04 for all conditions and all estimators. This effect is even smaller for the real condition than for the optimal condition, with all effects under 0.02. Across estimators, the benefit from the pure and generalized Kingsbury strategies is slightly larger for MLE and WLE than EAP and MAP.

Discussion

Bias in estimation

Increases in bias do not seem to be a problem for item review on a CAT using SIIS. With only a few exceptions the increase in the magnitude of bias with review is smaller than the change in ability from review. This indicates that there is a net gain in accuracy from allowing review. Most of the exceptions to this result are with MLE, so choosing an estimator other than MLE weakens this argument against allowing item review.

Decreased precision

Because of the effects of SIIS on targeting, there should be smaller changes in precision with SIIS than with MIIS. The results of this simulation included greater changes in precision than have been found in other studies using CATs with MIIS (e.g. Lunz, & Bergstrom, 1995; Vispoel et al., 2000). The explanation for this surprising result probably lies in the simulation method. During review, new item responses were generated for all items. There was thus a much larger number of answers changed than is consistently found in the literature (Benjamin et al., 1984; Waddell & Blankenship, 1994). This increases the amount of change in standard error expected as a result of item review.

Even so, there is generally not a very large increase in standard error. As Lunz, et al. (1992) point out, it would be possible to administer more items to make up for this loss of precision. With SIIS, however, a better option is available. The final information target can be increased to make up for the loss in precision from item review, so that after item review, examinees at a given ability level will have approximately the same amount of information as the original no-review information target.

Wainer strategy

The results on increased ability estimates for the Wainer strategy under SIIS are similar to the results from CATs using MIIS (Vispoel, et al., 1999). For the estimators that are biased inward, i.e., EAP, MAP, and WLE, there is on average no gain or a loss from using the strategy. Gains occur only when, by chance, an examinee happens to get more (easy) items correct than expected.

The increase in standard error means that the chance of getting more items correct than expected is higher under the Wainer strategy than with no strategy. The examinees who are most likely to get more right than expected are those for whom the standard error is highest. These examinees, whose ability is 2 or higher, are the same examinees for whom the Wainer strategy leads on average to a loss in ability estimate. Nonetheless, the increased standard error across all ability levels means that some examinees will benefit from the Wainer strategy.

The increase in standard error for the Wainer strategy under SIIS is similar in shape to the increase under MIIS (Vispoel et al., 1999). However, the increase should be less under SIIS than under MIIS, although this expectation was not directly tested in this research. Under MIIS, an examinee who gets all items incorrect will get the very easiest items available in the item pool. This leads to a test that is as far from perfect targeting as possible (in the easy direction), so that there is as large an increase in standard error as possible. Under SIIS, this ‘perfect’ mistargeting does not occur. Even though the intermediate ability estimates are very low, the examinee will probably not be administered all of the easiest items. The test will still have too little information for high ability examinees, but the precision loss will be smaller.

The fact that the results of the Wainer strategy under SIIS are so similar to the results from MIIS says that the choice between these two item selection algorithms makes little difference on the effects of the Wainer strategy. However, this simulation overestimated the benefits of the Wainer strategy, because it was assumed that examinees answered all the items incorrectly at first. Only about three-fourths of examinees in Vispoel et al. (1999) were able to do this, and that was under MIIS, for which the very easiest items are administered. It is likely that fewer examinees would be able to answer all the items incorrect at first on a CAT using SIIS instead of MIIS, because the items administered will not be as easy.

Preventing the use of the Wainer strategy is not possible when review is allowed. Some examinees will inevitably feel that they have found a way to beat the system by using the Wainer strategy, even though increases in ability estimates are unlikely. The test administrator can discourage the use of the strategy by choosing an appropriate estimator, which, based on these results and those from Vispoel et al. (1999), is any estimator that is biased inward. In addition, the Wainer strategy is very easy to detect, so a test administrator can flag any response patterns that indicate the use of the strategy.

Kingsbury strategy

There is some benefit to using the Kingsbury and generalized Kingsbury strategies. However, almost all the benefits are the result of changing answers according to the 3PL model, not from implementation of an item review strategy. Only for one ability level, -3, with imperfect answer changing, is there any benefit to selecting items to change based on item difficulty changes. This benefit is small, under 0.04 in all cases, and, under the assumption of standard normally distributed ability, fewer than 2% of examinees are expected to be in the range of abilities that benefit from the strategy. Based on these results, the Kingsbury strategy should not be an important concern for a CAT using SIIS.

Under MIIS, there is a close relationship between item responses and changes in item difficulty. It is on this relationship that the Kingsbury strategy is based. As Kingsbury (1996) showed, with perfect difficulty change recognition, there is a substantial benefit to implementing the strategy under MIIS. However, under SIIS, the relationship between item responses and changes in item difficulty is much reduced. These results show that the relationship is reduced to such a small level that the Kingsbury strategy does not work.

The results from the not relevant condition show that a good strategy for very low ability examinees is to answer all the items as well as they can, then change all of the answers to guesses among the remaining options. By using this strategy, which does not require item review, very low ability examinees are more likely to answer correctly items which have pseudo-guessing parameters below the level of chance. For example, suppose an extremely low ability examinee answers an item with four options that has a pseudo-guessing parameter of 0.2. If the examinee answers to the best of his or her ability, he or she will get the item correct with probability 0.2. If the examinee follows the strategy, that probability rises to approximately (1 - 0.2) * 0.33 = ~0.28 (the probability of getting the item wrong initially times the probability of guessing correctly among the remaining options). The idea of answering all items as well as the examinee can, then changing them, seems ridiculous, yet the increase in estimated ability is a prediction of the 3PL model. Despite the greatest desires of psychometricians, people respond to items not according to a model, but according to what they think the right answer is. Models are only approximations, and results that depend so much on the model may not be true in real life. No study has looked at real people attempting to implement the Kingsbury strategy. While the results of this study are certainly promising for allowing of item review on a CAT using SIIS, they cannot be considered conclusive.

Other review options

Many of the problems with allowing item review on a CAT stem from the fact that the initial selection of items on the test becomes less than optimal when answers are changed. A simple solution to this problem is to administer additional items after item review is completed, but there are many obvious problems with this technique, not the least of which is giving more items after the examinee thought the test was over.

Three methods for reducing the optimal item selection problems associated with item review are limited review, block review, and stimulus-based review. Stocking (1997) simulated results for each of these under the Wainer strategy, which yields the worst-case scenario for selection of items.

Limited review is when an examinee may look back at all of the items, but may change only a limited number of answers. Stocking (1997) found that, unless the number of answer changes allowed was extremely small (2 on a 28 item test), there was substantial bias using MLE, and there was a large loss of precision. Although it is true that examinees tend not to change many answers during review, creating artificial limits to the number of answer changes will probably not satisfy examinee’s desire for review and may increase complaints about review.

Block review is when an examinee gets full review within a block of items, but cannot review across blocks. For example, an examinee may be given full review of the first ten items on a test after those items have been administered. After completing review, the next ten items are administered, at which point review is allowed on the most recently administered block of ten items, but not on previous blocks of items. Stocking (1997) found that the bias was very small, even when a 28-item CAT was divided into only two blocks of 14 items each. Precision was virtually the same as with no review when there were 4 blocks, with only a small decrease in precision with 2 blocks.

Stimulus-based review is very similar to block review, but with varying length blocks. Review is allowed within blocks of items that share a common stimulus, but not across blocks. For tests on which many items share a common stimulus, such as a reading test, stimulus-based review leads to results indistinguishable from block review. For tests on which few items share a common stimulus, stimulus-based review leads to results very similar to a no review condition.

Vispoel et al. (2000) randomly assigned subjects to one of several conditions, no review, full review, and three different-sized block review conditions. In all review conditions, examinees were satisfied with their review options, even when the block size was small (5 items). As block size increased, examinees marked more items for review, changed more answers, spent more time in reviewing answers, and tended to spend more total time on the test. However, there was no discernible pattern to ability estimates; that is, ability estimates did not seem to be related to the review condition, although across all review conditions, average ability estimate was higher than in the no review condition. While there was substantial variability across review conditions in the results of Vispoel et al., (2000), it suggests that block review may function just as well as full review, and may be a good way to reduce item selection problems associated with review.

Conclusion

Computerized adaptive tests are a relatively new concept, and have not been implemented widely. It may be argued that the advantages of allowing item review will not exist as more examinees become used to CATs. The desire for item review may stem mostly from the fact that examinees are accustomed to having review on paper-and-pencil tests. As more people get used to a no review condition on a CAT, fewer examinees will desire review. Also, the increased accuracy from allowing review will be eliminated as the test-taking population gets used to having only one chance to answer an item correctly. This argument has merits, but it is certainly not reasonable to expect these drastic changes. CATs are not widely used, and probably will not be for many years. The use of paper-and-pencil tests will continue into the foreseeable future. It is very unlikely that examinees will adjust to conditions on a CAT to such a degree that item review is unnecessary.

If the problems with review are strong enough then it should not be allowed. This study suggests that, by using SIIS rather than MIIS, several of the problems are reduced or eliminated. More research remains to be done to conclude with certainty that review should be allowed, particularly into the use of the Kingsbury strategy. These results suggest that, by choosing an appropriate item selection algorithm and ability estimator, and perhaps by implementing block review, item review could be allowed without fear of increased bias, decreased precision, or susceptibility to the Wainer and Kingsbury strategies.

References

Benjamin, L. T., Cavell, T. A., & Shallenberger, W. R. (1984). Staying with initial answers on objective tests: Is it a myth? *Teaching of Psychology, 11*, 133-141.

Ben-Simon, A., & Ben-Shachar, G. (2000, April) Cognitive aspects of partial knowledge as expressed in multiple-choice tests. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Bowles, R. (2001). *An examination of item review on computer adaptive tests*. Manuscript in preparation, University of Virginia. Available: http://kiptron.psyc.virginia.edu/ProjectPapers/ryan/ryanpaper.html.

Davey, T., & Fan, M. (2000, April). Specific information item selection for adaptive testing. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Davey, T., Pommerich, M., & Thompson, T. D. (1999, April). Pretesting alongside an operational CAT. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Davey, T., & Thomas, L. (1996, April). Constructing adaptive tests to parallel conventional tests. Paper presented at the annual meeting of the American Educational Research Association, New York.

De Ayala, R. J., Schafer, W. D., & Sava-Bolesta, M. (1995). An investigation of the standard errors of expected *a posteriori* ability estimates. *British Journal of Mathematical and Statistical Psychology, 47*, 385-405.

Fan, M., Thompson, T., & Davey, T. (1999, April). Constructing adaptive tests to parallel conventional programs. Paper presented at the annual meeting of the National council on Measurement in Education, Montreal.

Gershon, R., & Bergstrom, B. (1995, April). Does cheating on CAT pay: Not! Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L, & Reckase, M. K. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 4*, 347-360.

Kingsbury, G. G. (1996, April). Item review and adaptive testing. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Linacre, J. M., & Wright, B. D. (2000). Winsteps- Bigsteps- Ministep manual on-line [On-line]. Available: http://www.winsteps.com/winman/index.htm.

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

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48*, 233-245.

Lord, F. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Education Measurement, 23*, 157-162.

Lunz, M. E., & Bergstrom, B. (1995). Computerized adaptive testing: Tracking candidate response patterns. *Journal of Educational Computing Research, 13*, 151-162.

Lunz, M., Bergstrom, B, & Wright, B. D. (1992). The effect of review on student ability and test efficiency for computerized adaptive tests. *Applied Psychological Measurement, 16*, 33-40.

Stocking, M. L. (1997). Revising item responses in computerized adaptive tests: A comparison of three models. *Applied Psychological Measurement, 21*, 129-142.

Stone, G. E., & Lunz, M. E. (1994). The effect of review on the psychometric characteristics of computerized adaptive tests. *Applied Measurement in Education, 7*, 211-222.

Thompson, T., Davey, T., & Nering, M. L. (1998, April). Constructing adaptive tests to parallel conventional programs. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Vispoel, W. P. (1998). Reviewing and changing answers on computer-adaptive and self-adaptive vocabulary tests. *Journal of educational Measurement, 35*, 328-347.

Vispoel, W. P., Hendrickson, A. B., & Bleiler, T. (2000). Limiting answer review and change on computerized adaptive vocabulary tests: Psychometric and attitudinal results. *Journal of Educational Measurement, 37*, 21-38.

Vispoel, W., Rocklin, T., Wang, T., & Bleiler, T. (1999). Can examinees use a review option to obtain positively biased ability estimates on a computerized adaptive test? *Journal of Educational Measurement, 36*, 141-157.

Waddell, D. L., & Blankenship, J. C. (1994). Answer changing: A meta-analysis of the prevalence and patterns. *Journal of Continuing Education in Nursing, 25*, 155-158.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice, 12*, 15-20.

Wang, T., Hanson, B. A., & Lau, C. A. (1999). Reducing bias in CAT trait estimation: A comparison of approaches. Applied Psychological Measurement, 23, 263-278.

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 109-135.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450.

Wise, S. L. (1996, April). A critical analysis of the arguments for and against item review in computerized adaptive testing. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Wise, S. L., Finney, S. J., Enders, C. K., Freeman, S. A., & Severance, D. D. (1999). Examinee judgments of changes in item difficulty: Implications for item review in computerized adaptive testing. *Applied Measurement in Education*, *12,* 185-198.

Table 1: Summary of the conditions and subscripts for the Kingsbury strategy

Facet	Condition	Description	Subscript
Answer marked for possible change	pure Kingsbury	Only guessed items are marked for possible change	θ_{Kxx}
	generalized Kingsbury	All items are marked for possible change	θ_{gxx}
Difficulty change recognition	optimal	Difficulty changes are recognized perfectly	θ_{xox}
	real	Difficulty changes are recognized imperfectly	θ_{xrx}
	not relevant	Answers are changed regardless of difficulty changes	θ_{xnx}
Answer change behavior	perfect	Incorrect answers are changed to correct with probability 1	θ_{xxp}
	imperfect	Incorrect answers are changed to correct with probability 0.34	θ_{xxi}

Figure 1: SIIS information targets

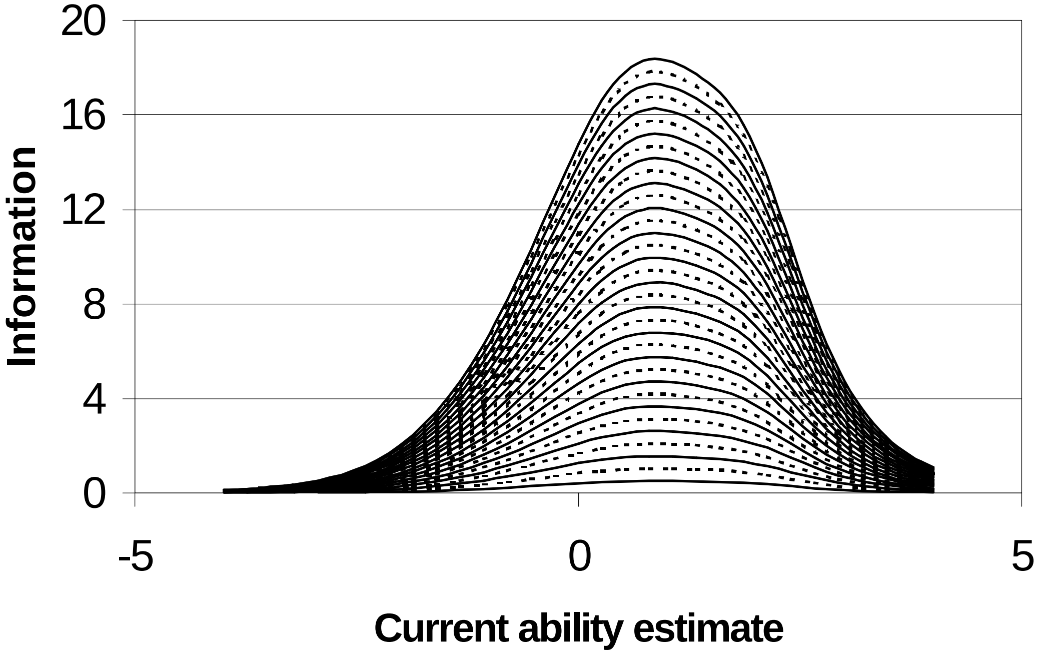


Figure 2: Bias_T

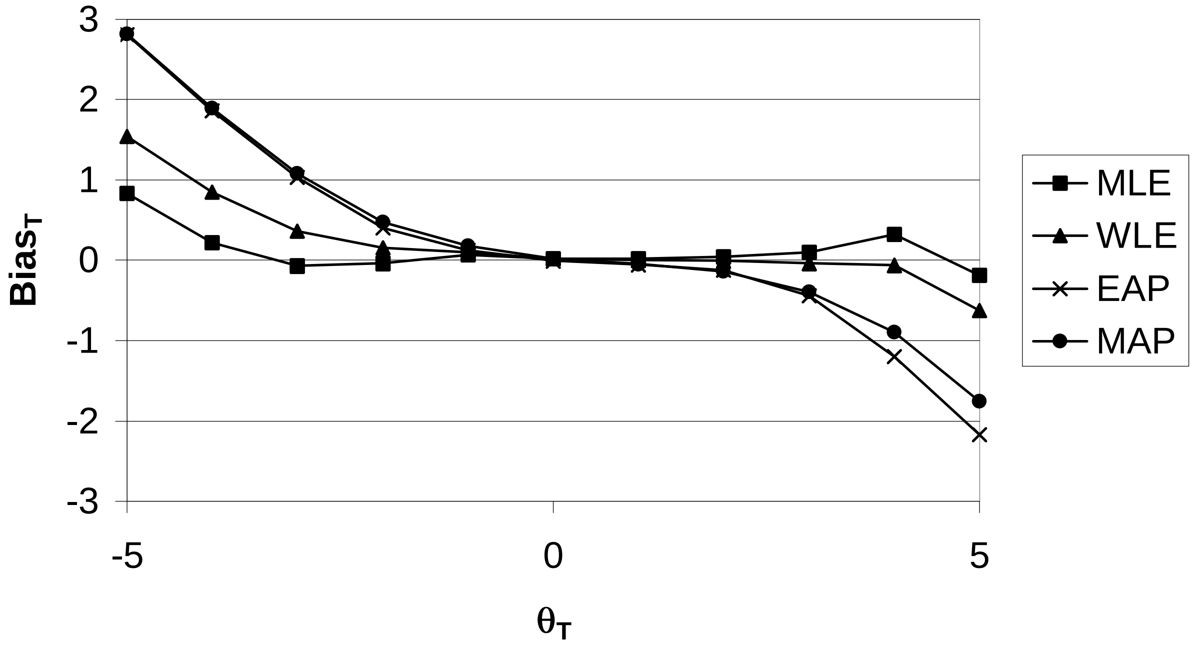


Figure 3: SE_T

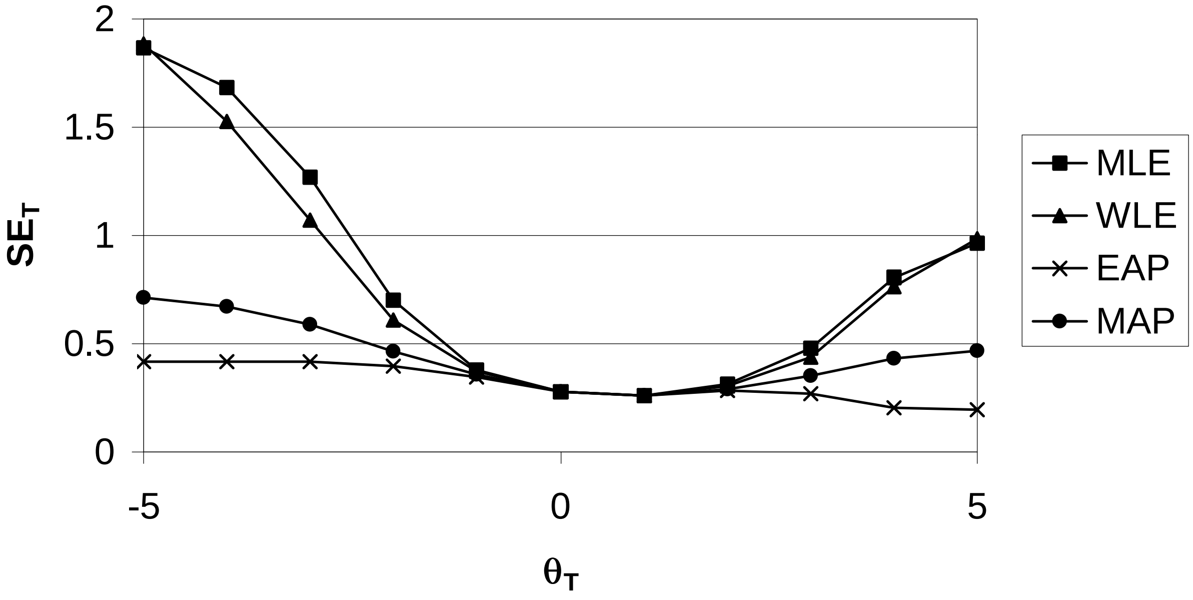


Figure 4a: Bias$_{T,b}$; $\theta_T=\theta_b-.1$

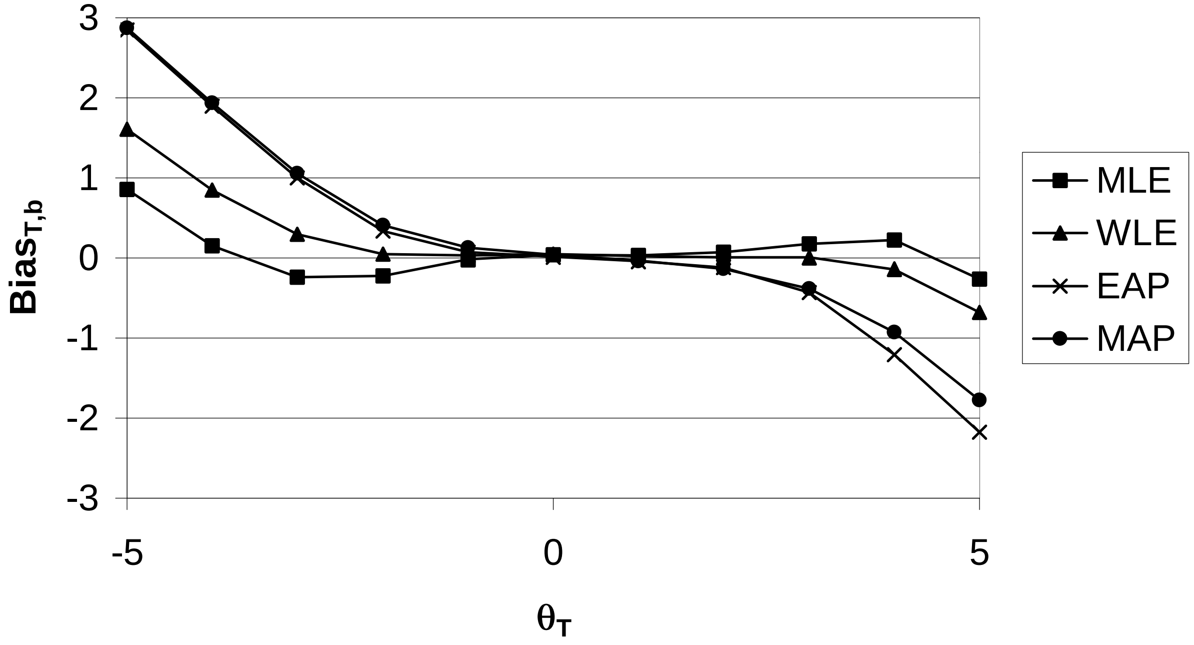


Figure 4b: Bias$_{T,b}$; $\theta_T=\theta_b+.1$

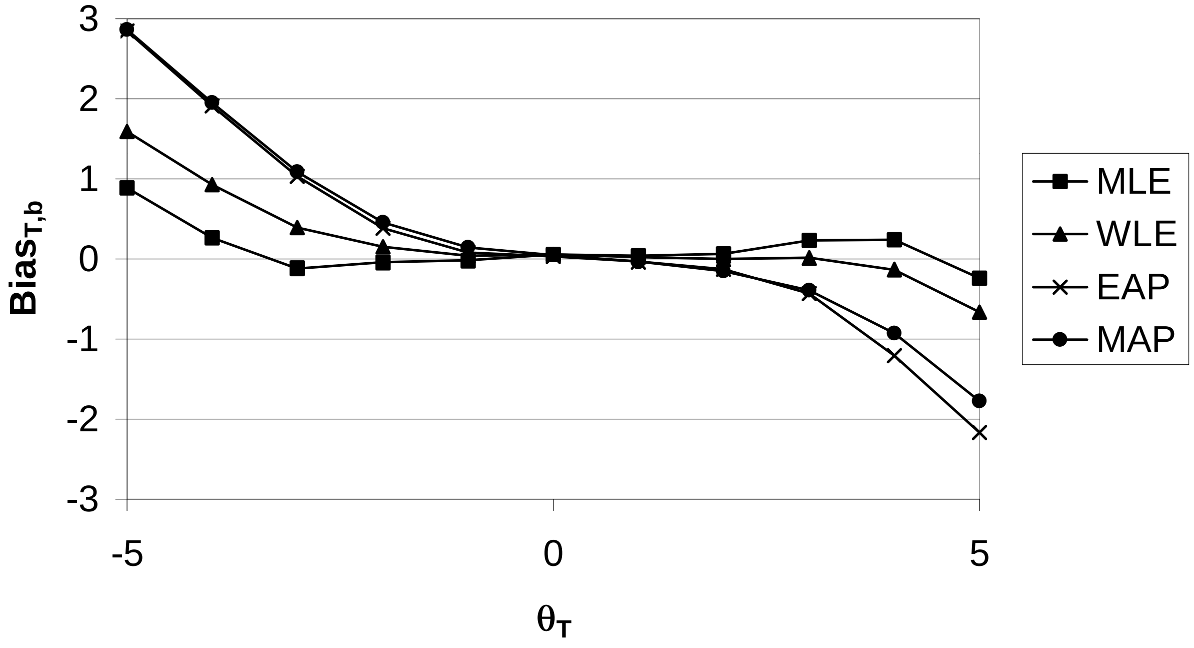


Figure 4c: Bias$_{T,b}$; $\theta_T=\theta_b+.4$

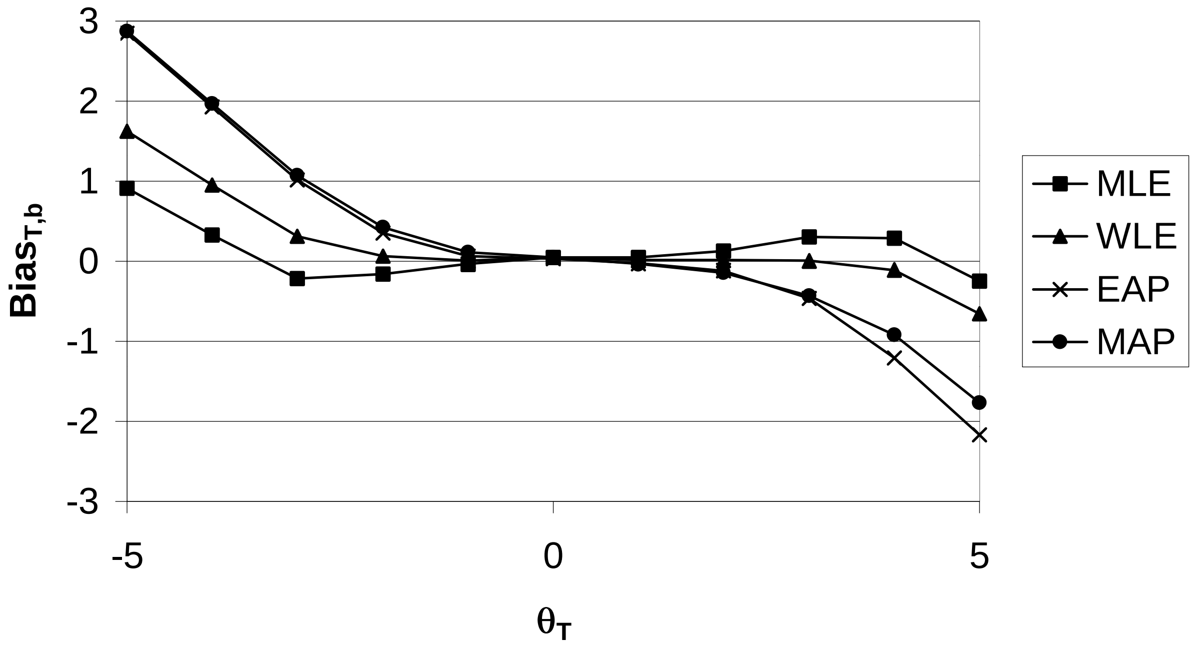


Figure 4d: Bias$_{T,b}$; $\theta_T=\theta_b+1.0$

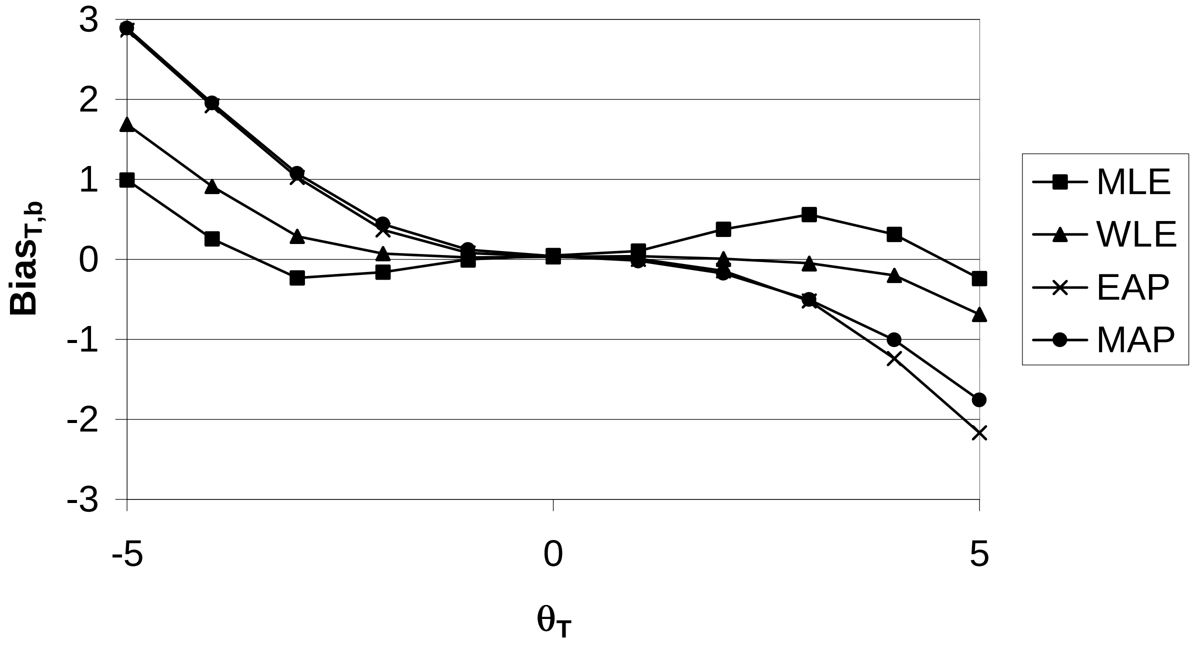


Figure 5a: $|Bias_{T,b}|$-$|Bias_{b,b}|$; θ_T=θ_b-.1

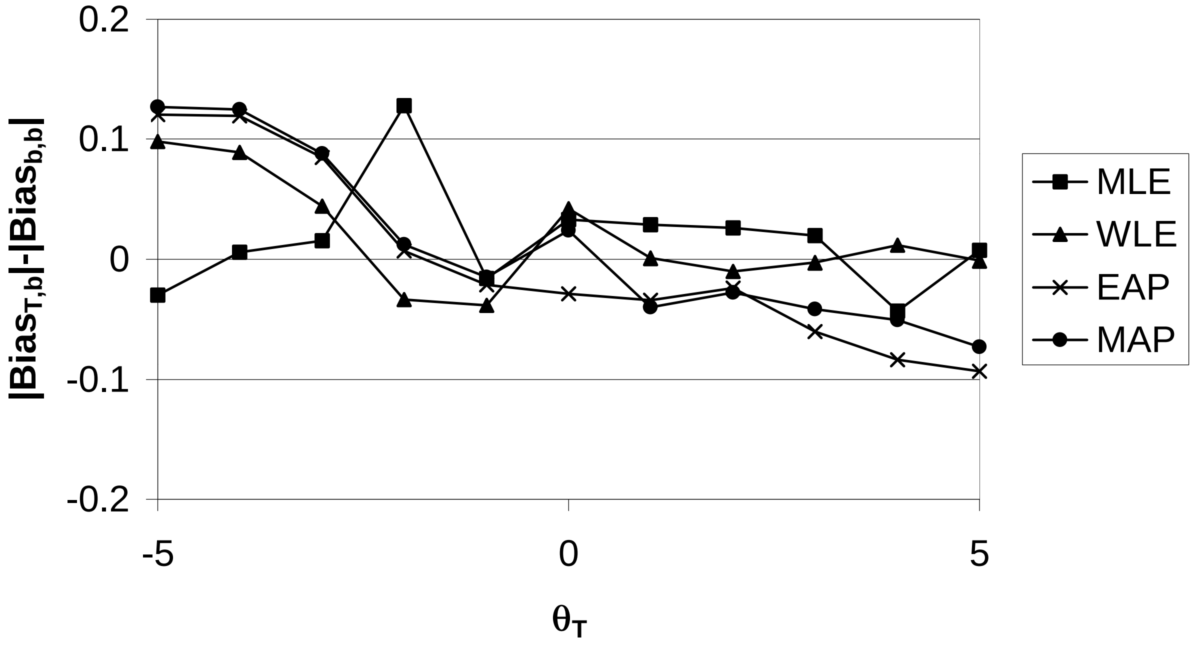


Figure 5b: $|Bias_{T,b}|$-$|Bias_{b,b}|$; θ_T=θ_b+.1

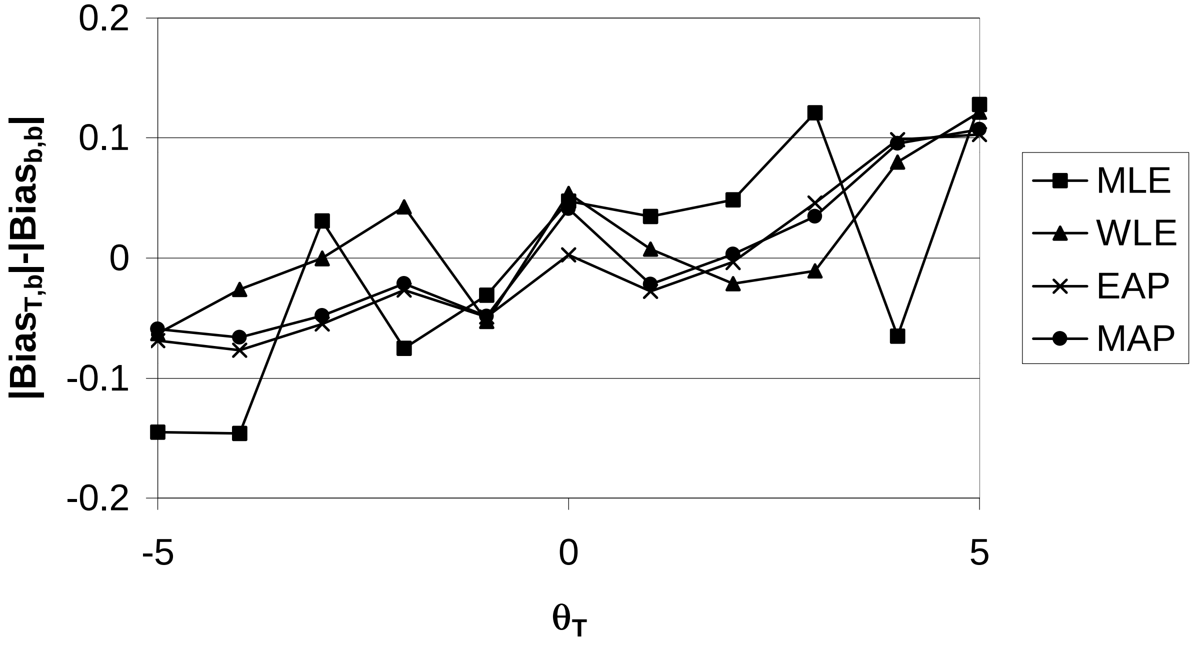


Figure 5c: $|Bias_{T,b}|$-$|Bias_{b,b}|$; $\theta_T=\theta_b+.4$

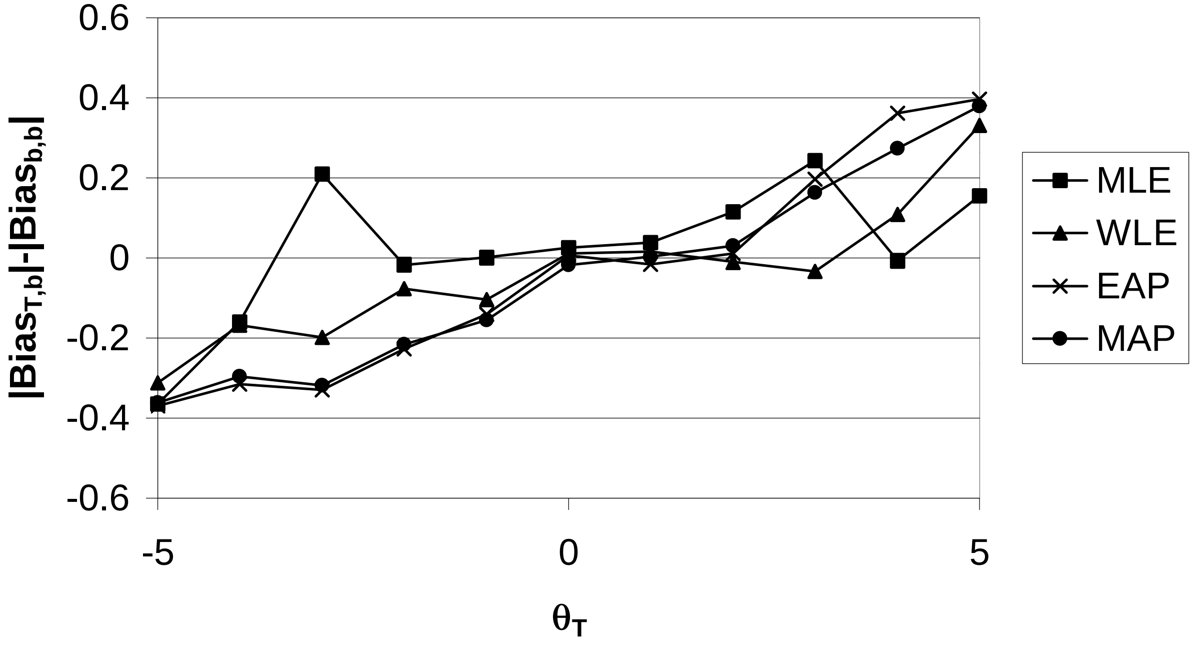


Figure 5d: $|Bias_{T,b}|$-$|Bias_{b,b}|$; $\theta_T=\theta_b+1.0$

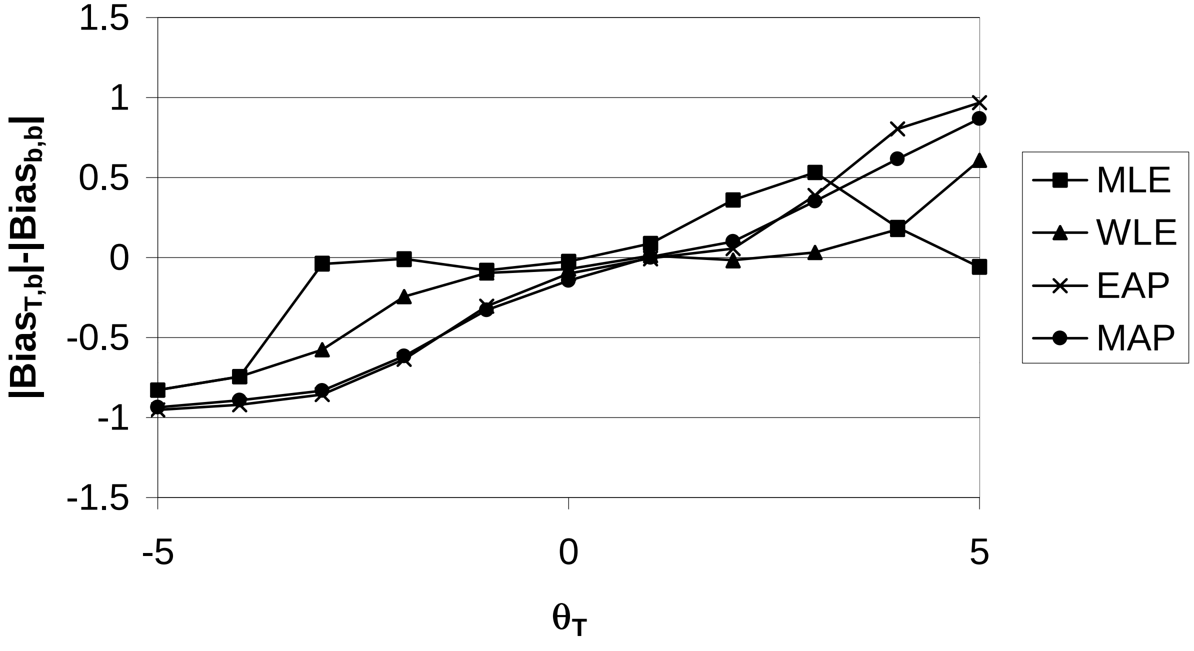


Figure 6a: $|Bias_{T,b}|-|Bias_{T,T}|$; $\theta_T=\theta_b-.1$

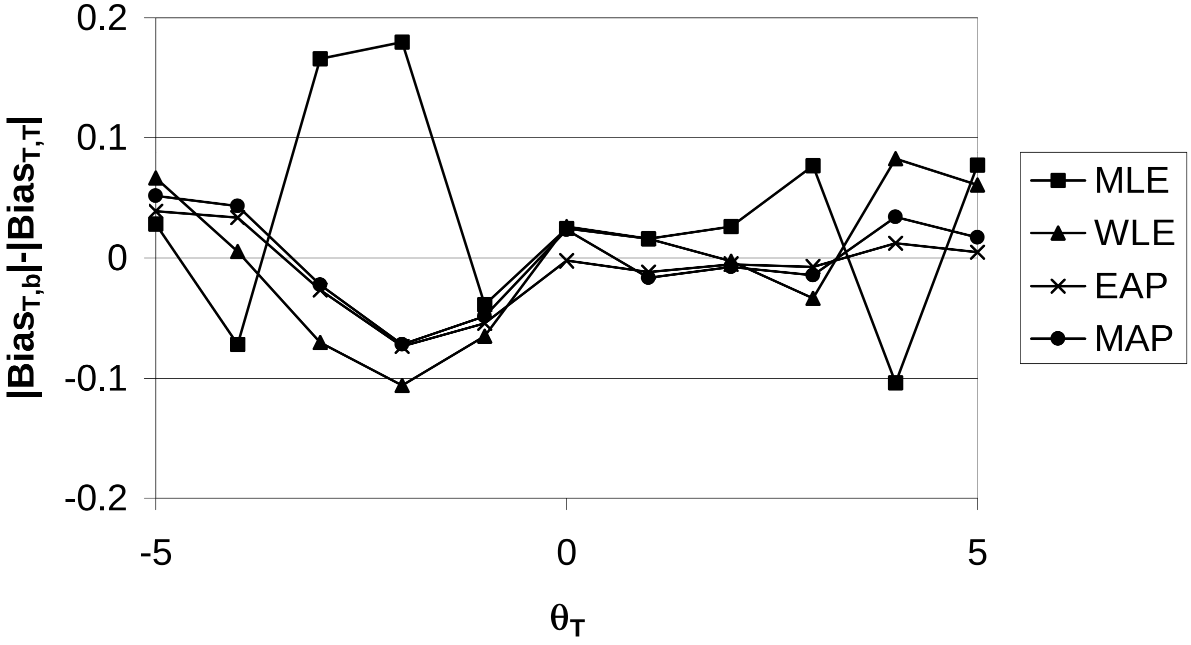


Figure 6b: $|Bias_{T,b}|-|Bias_{T,T}|$; $\theta_T=\theta_b+.1$

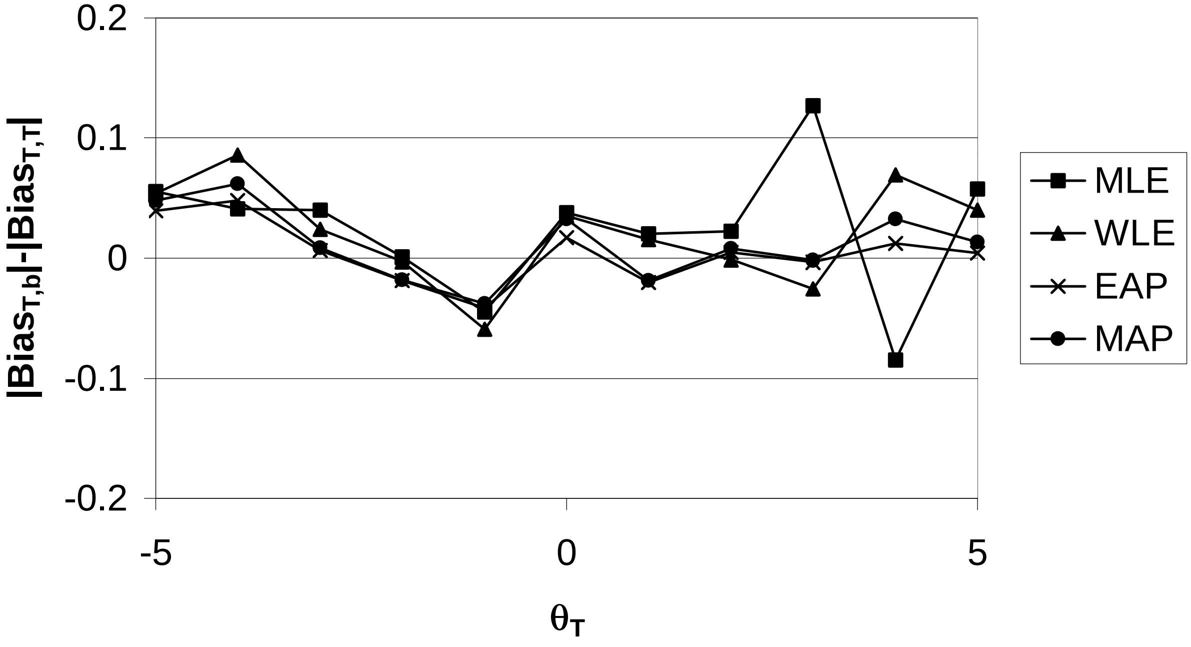


Figure 6c: $|\mathrm{Bias}_{T,b}|$-$|\mathrm{Bias}_{T,T}|$; θ_T=θ_b+.4

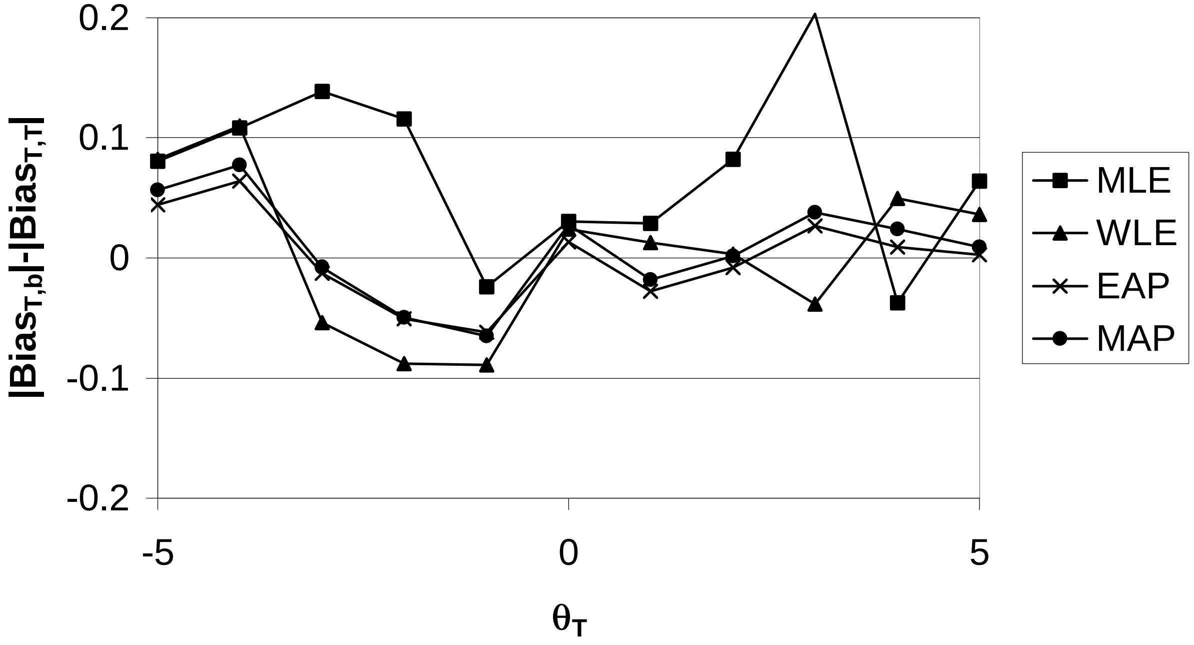


Figure 6d: $|\mathrm{Bias}_{T,b}|$-$|\mathrm{Bias}_{T,T}|$; θ_T=θ_b+1.0

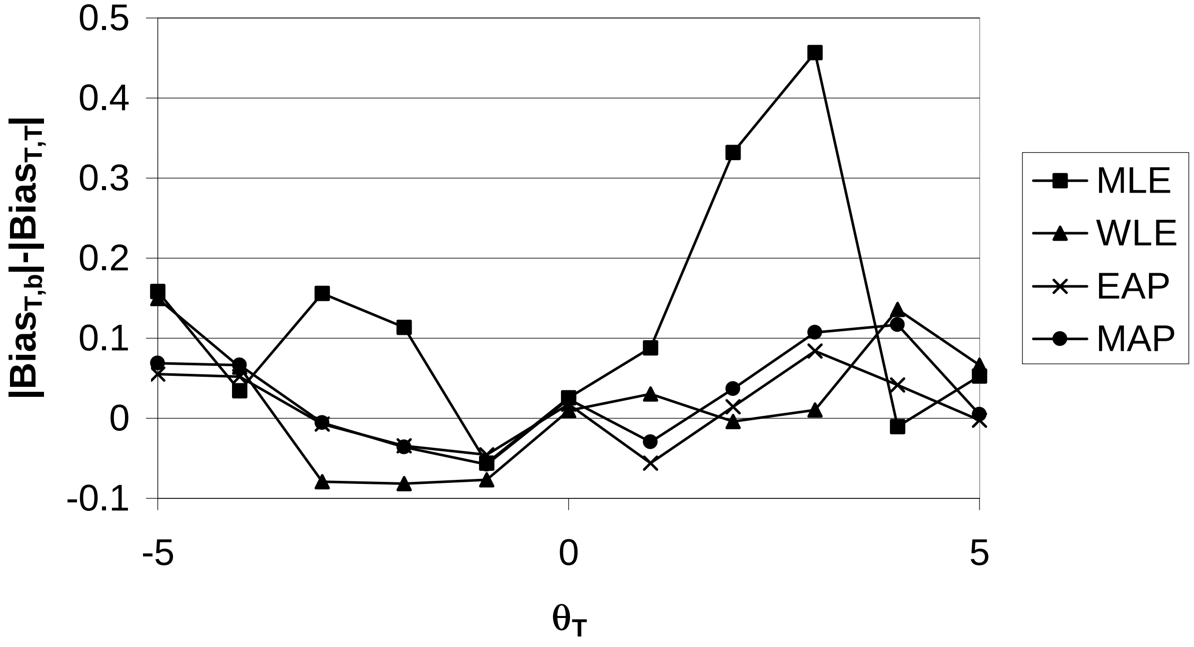


Figure 7a: $SE_{T,b}-SE_{b,b}$; $\theta_T=\theta_b-.1$

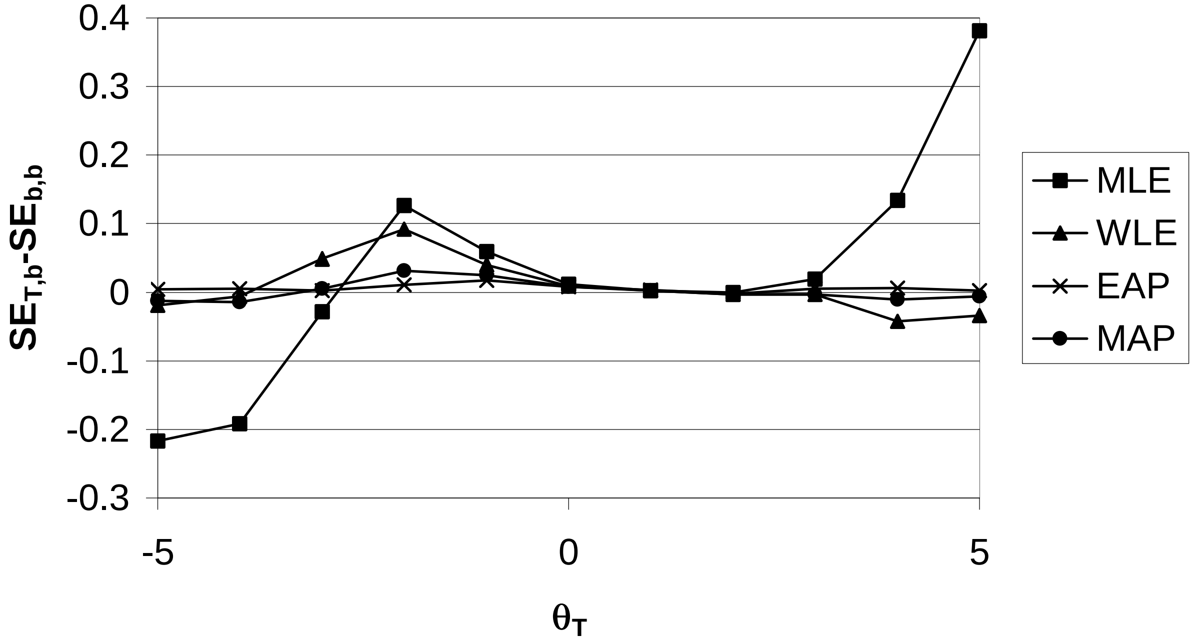


Figure 7b: $SE_{T,b}-SE_{b,b}$; $\theta_T=\theta_b+.1$

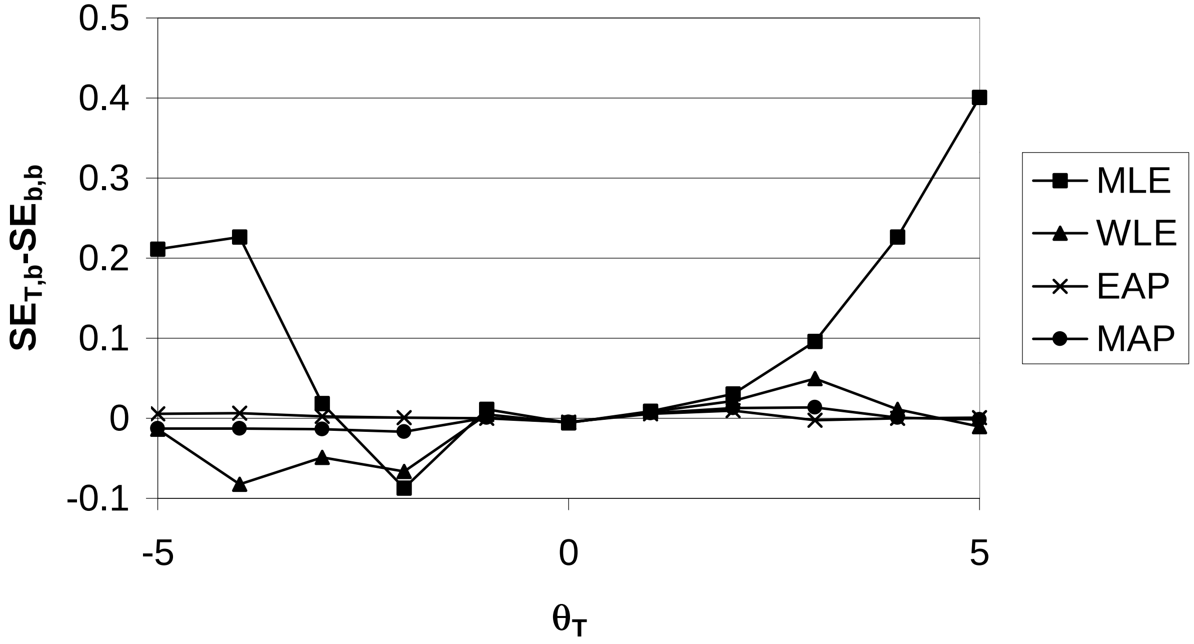


Figure 7c: $SE_{T,b}$-$SE_{b,b}$; $\theta_T=\theta_b+.4$

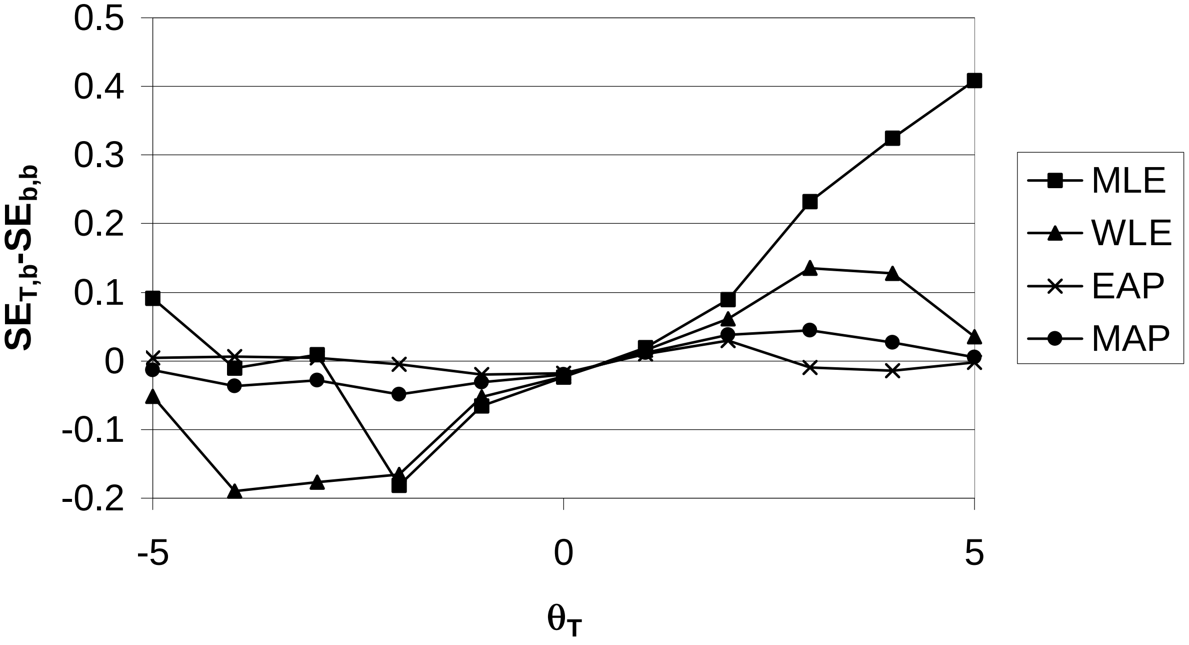


Figure 7d: $SE_{T,b}$-$SE_{b,b}$; $\theta_T=\theta_b+1.0$

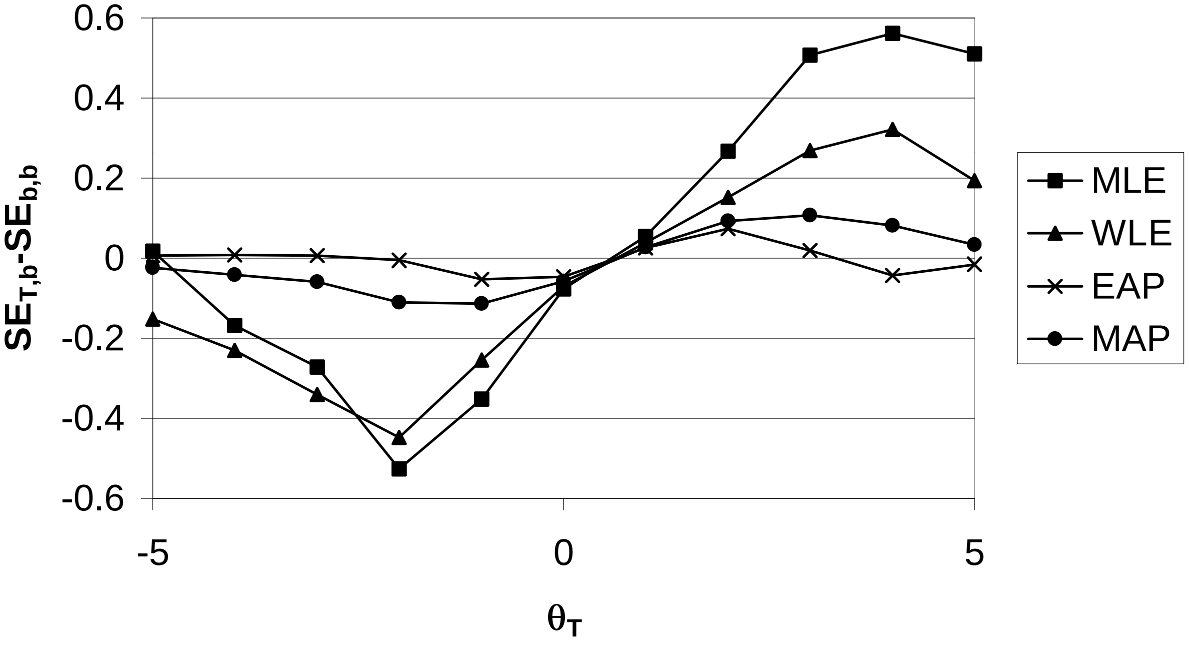


Figure 8a: $(SE_{T,b}-SE_{b,b})/SE_{b,b}*100\%$; $\theta_T=\theta_b-.1$

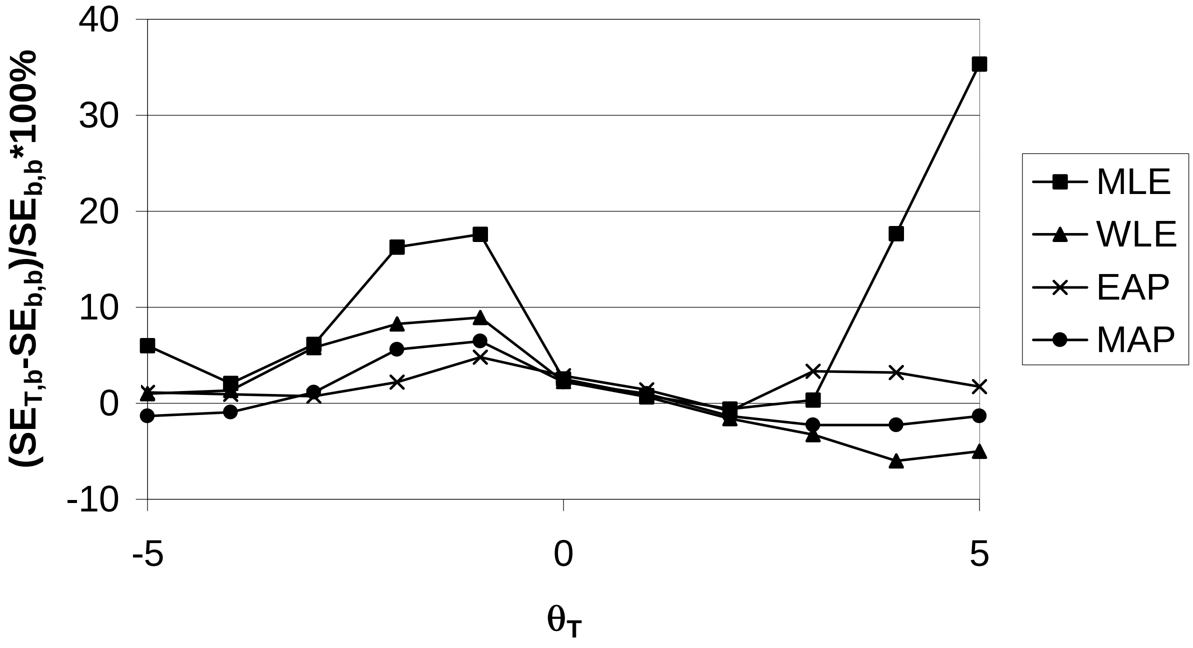


Figure 8b: $(SE_{T,b}-SE_{b,b})/SE_{b,b}*100\%$; $\theta_T=\theta_b+.1$

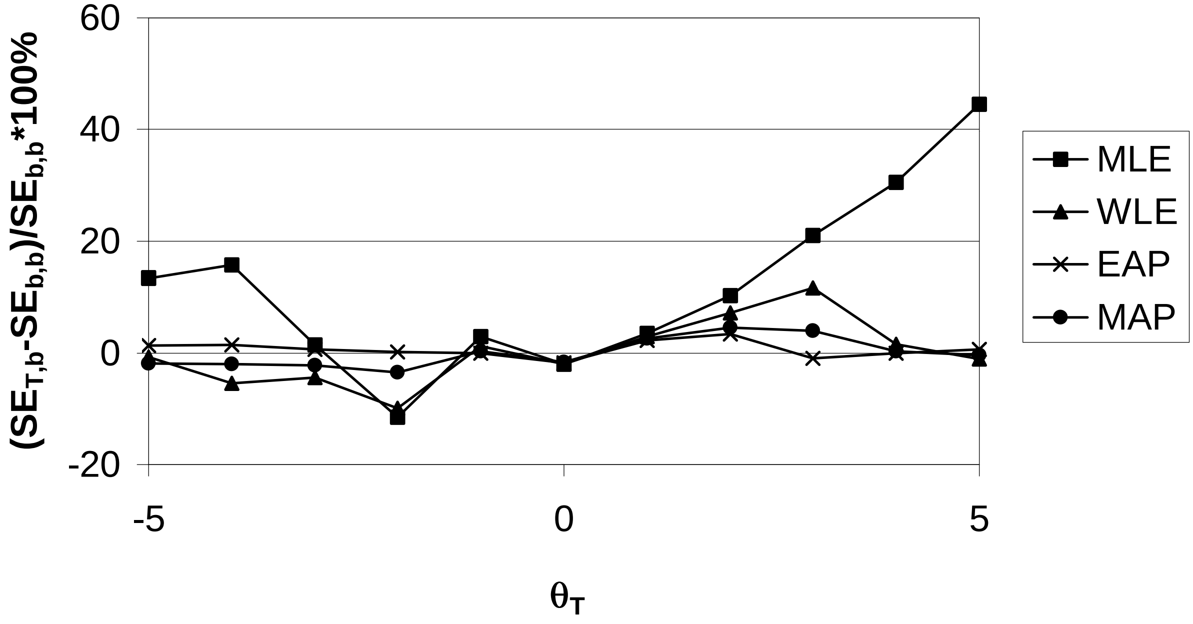


Figure 8c: $(SE_{T,b}-SE_{b,b})/SE_{b,b}*100\%$; $\theta_T=\theta_b+.4$

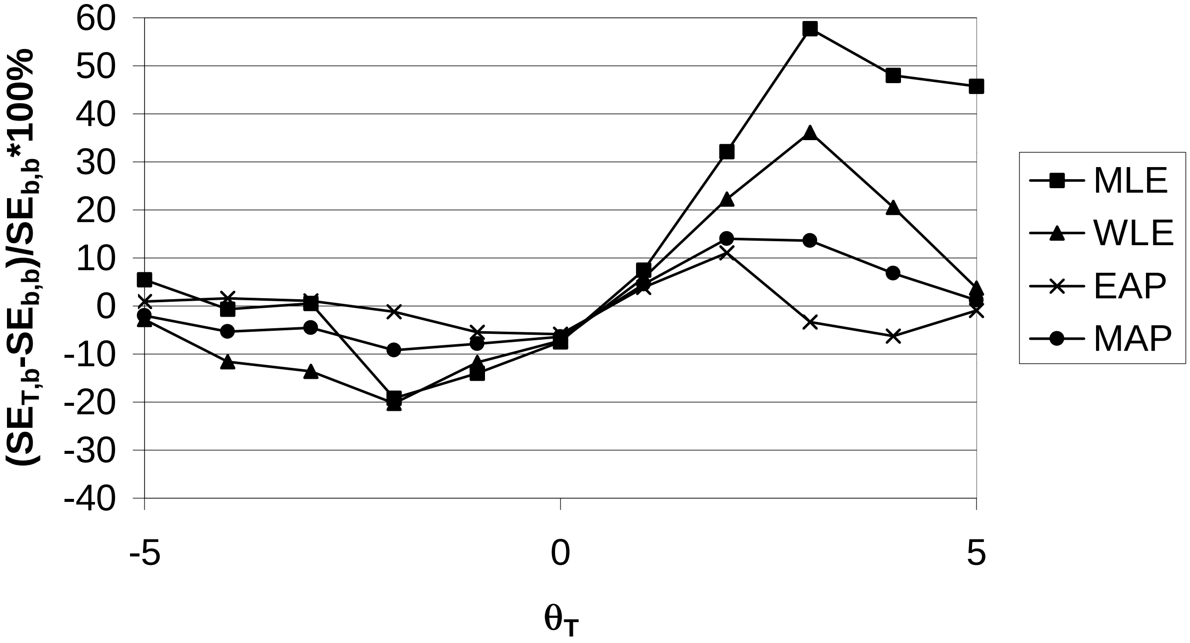


Figure 8d: $(SE_{T,b}-SE_{b,b})/SE_{b,b}*100\%$; $\theta_T=\theta_b+1.0$

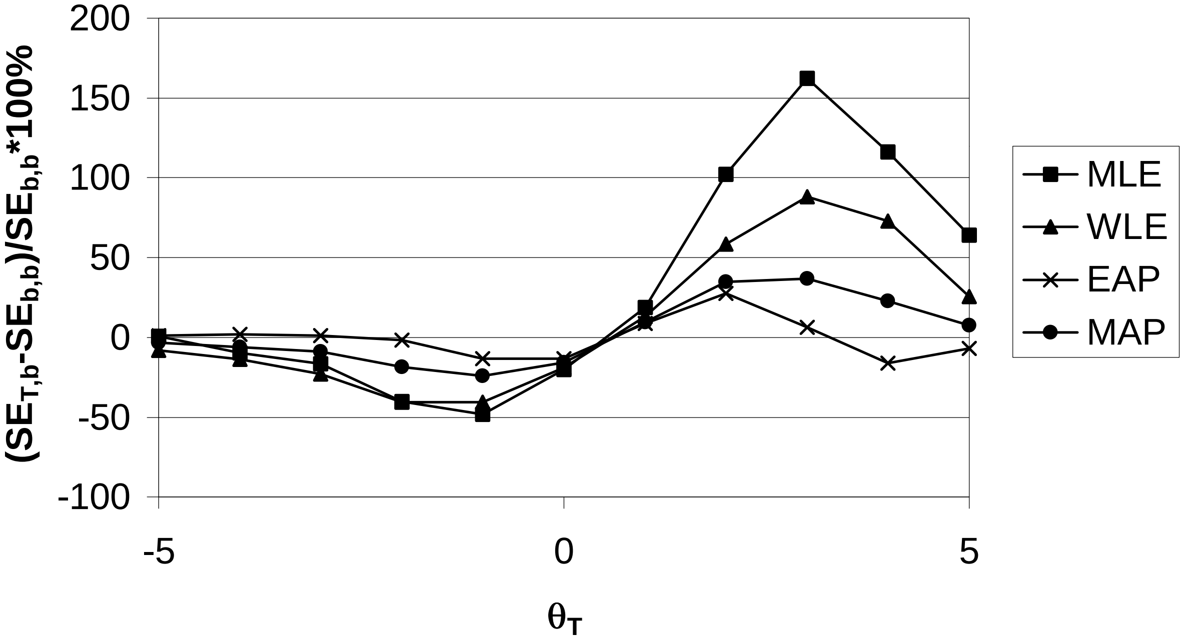


Figure 9a: $(SE_{T,b}-SE_{T,T})/SE_{T,T}*100\%$; $\theta_T=\theta_b-.1$

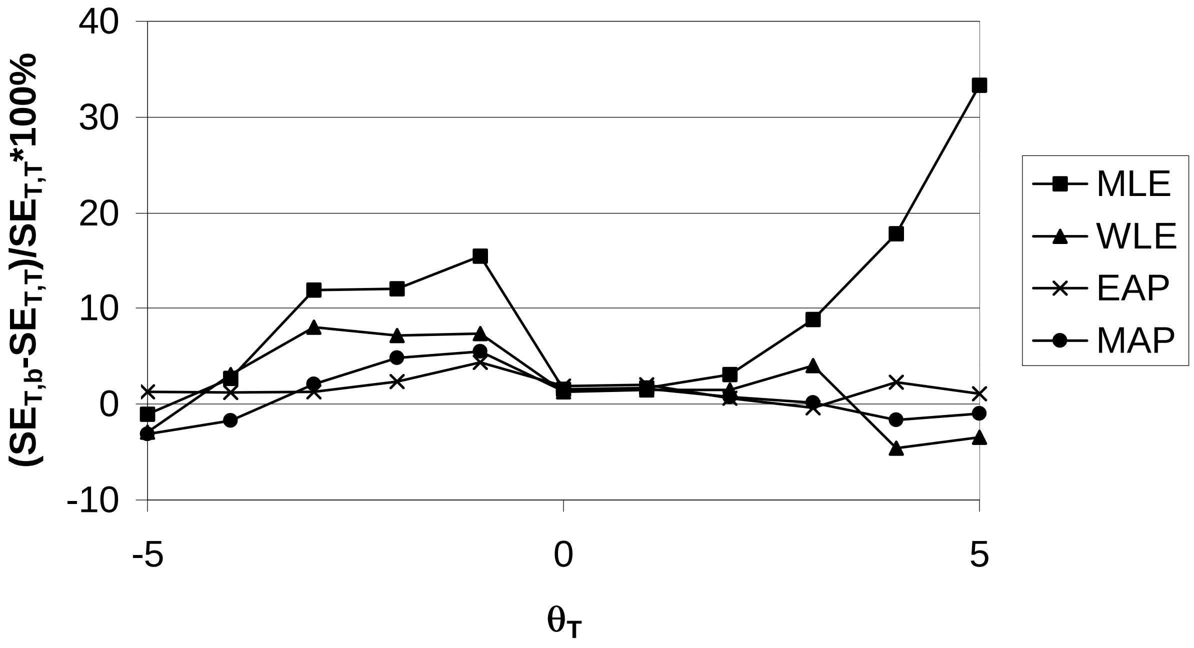


Figure 9b: $(SE_{T,b}-SE_{T,T})/SE_{T,T}*100\%$; $\theta_T=\theta_b+.1$

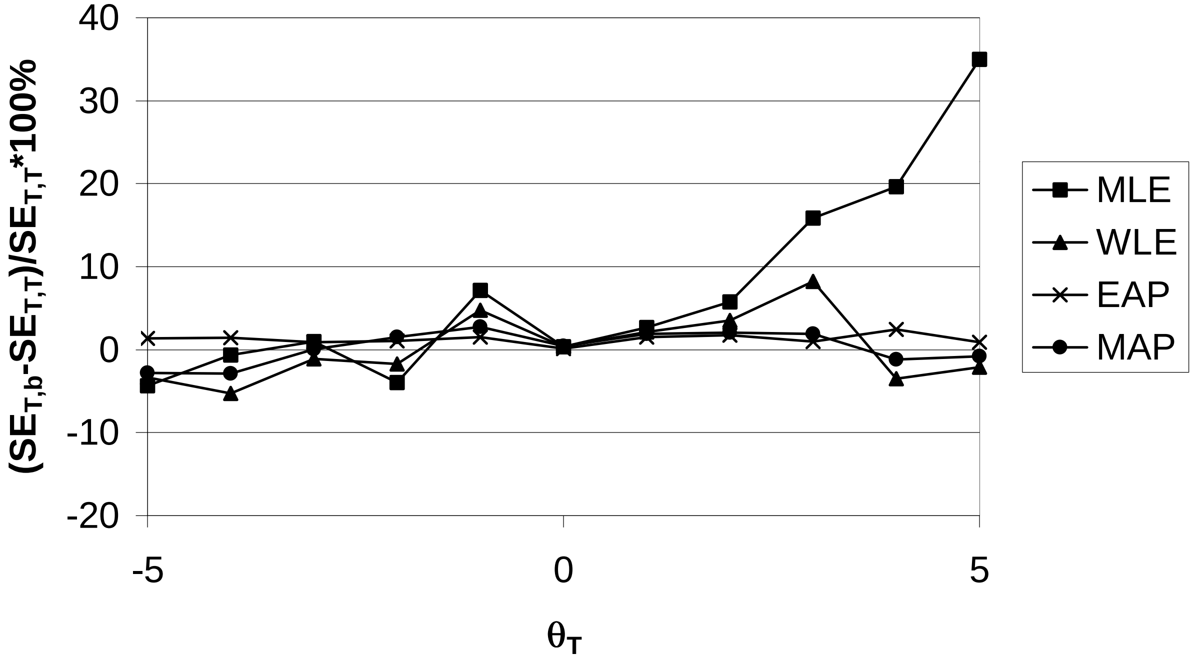


Figure 9c: $(SE_{T,b}-SE_{T,T})/SE_{T,T}*100\%$; $\theta_T=\theta_b+.4$

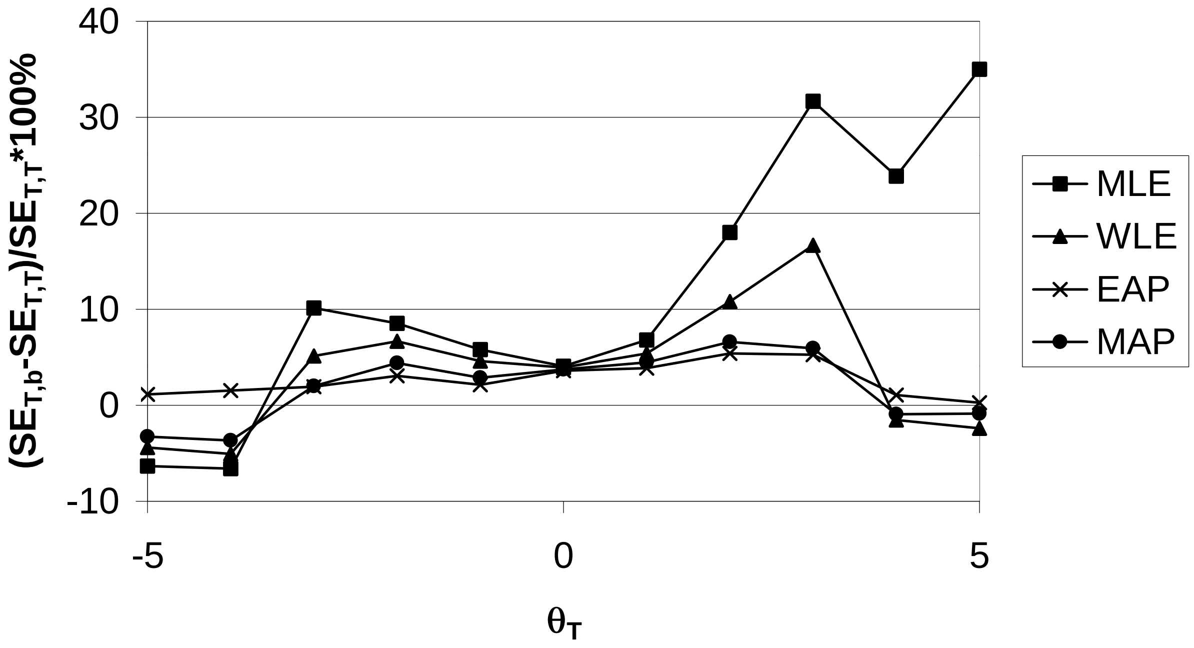


Figure 9d: $(SE_{T,b}-SE_{T,T})/SE_{T,T}*100\%$; $\theta_T=\theta_b+1.0$

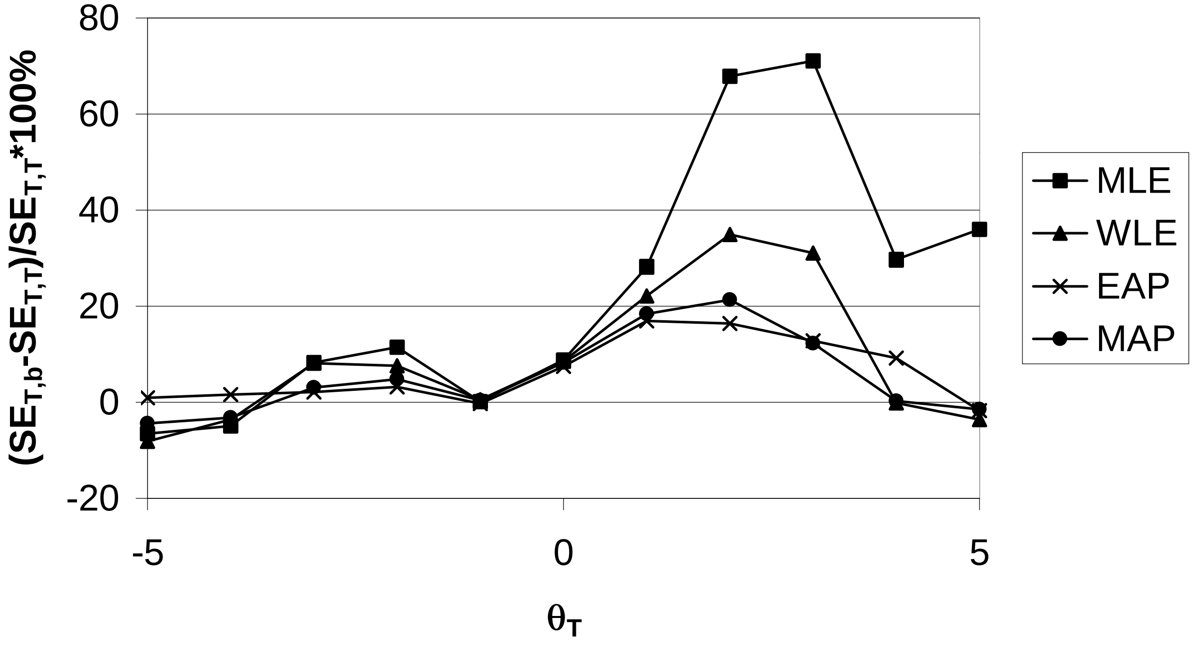


Figure 10: $\text{Bias}_{T,W}$

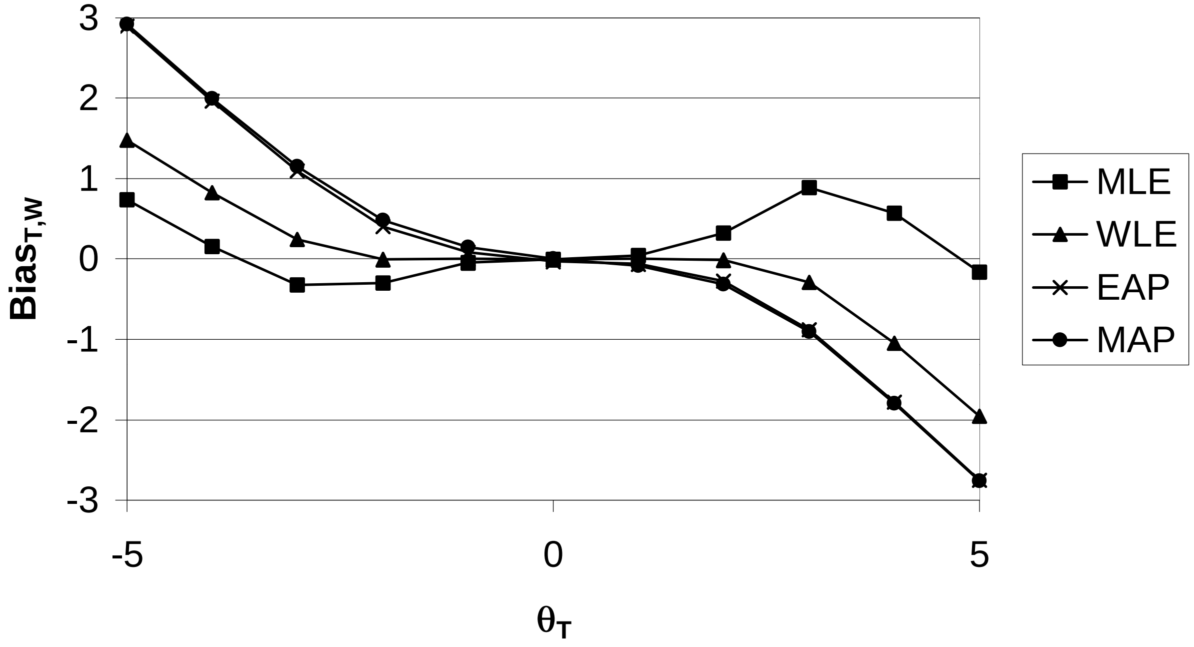


Figure 11: $\theta\text{hat}_{T,W}-\theta\text{hat}_{T,T}$

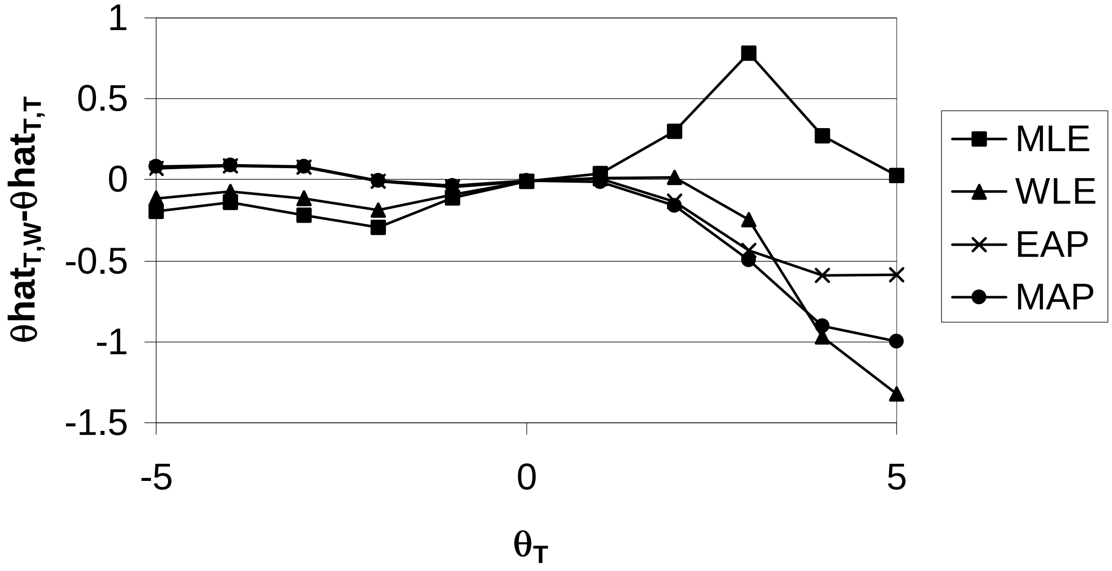


Figure 12: $(SE_{T,W}-SE_{T,T})/\ SE_{T,T}*100\%$

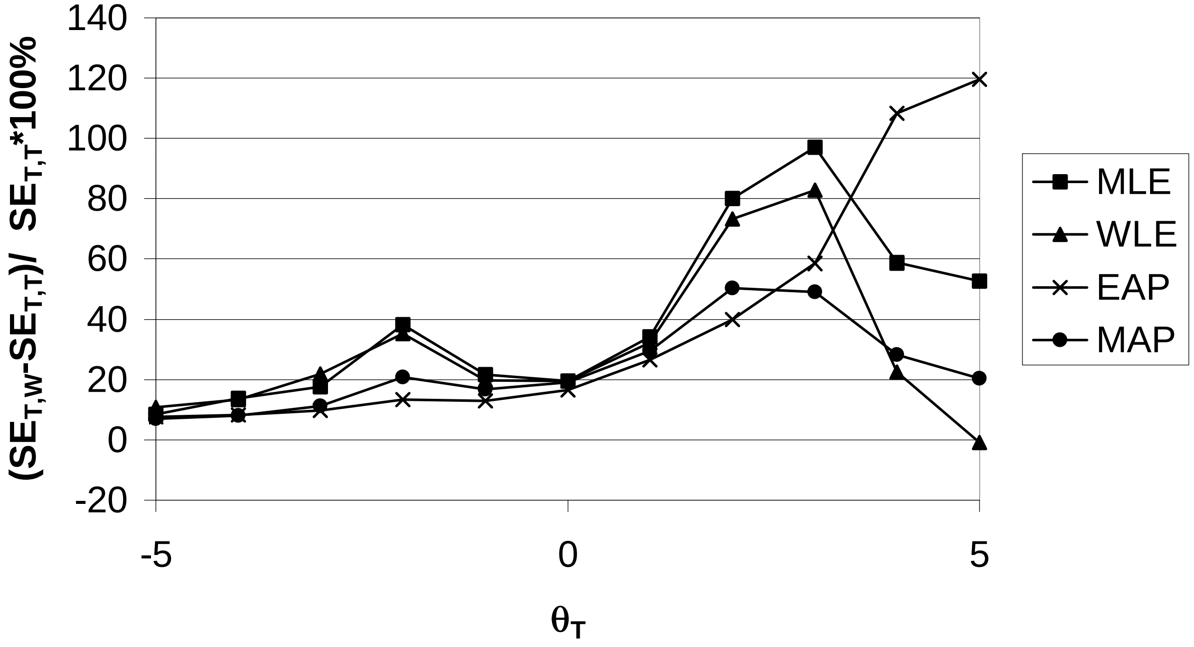


Figure 13a: ^{q} hat$_{Kop}$-^{q} hat$_{T}$

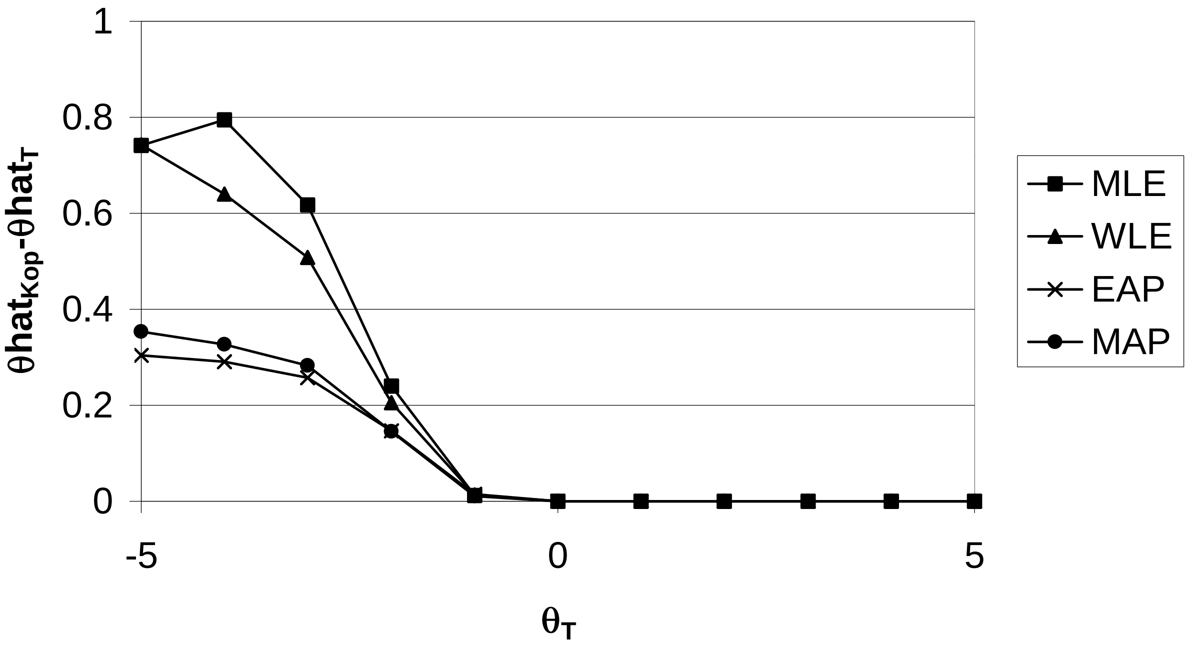


Figure 13b: θhat$_{Koi}$-θhat$_{T}$

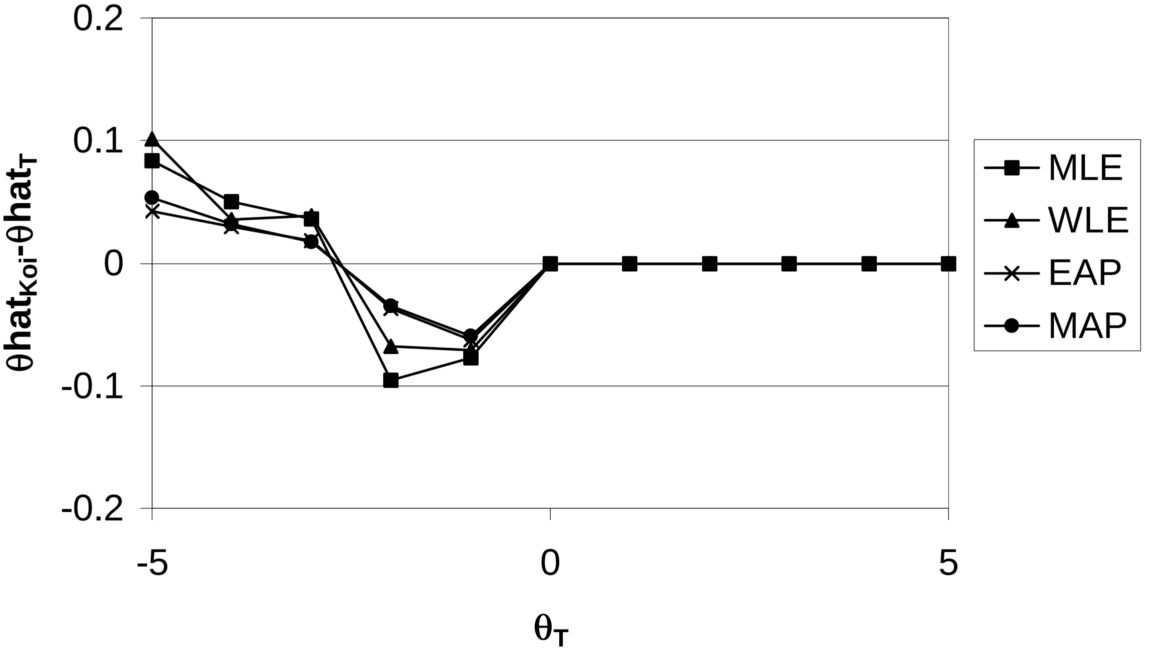


Figure 13c: $\theta hat_{Krp}-\theta hat_{T}$

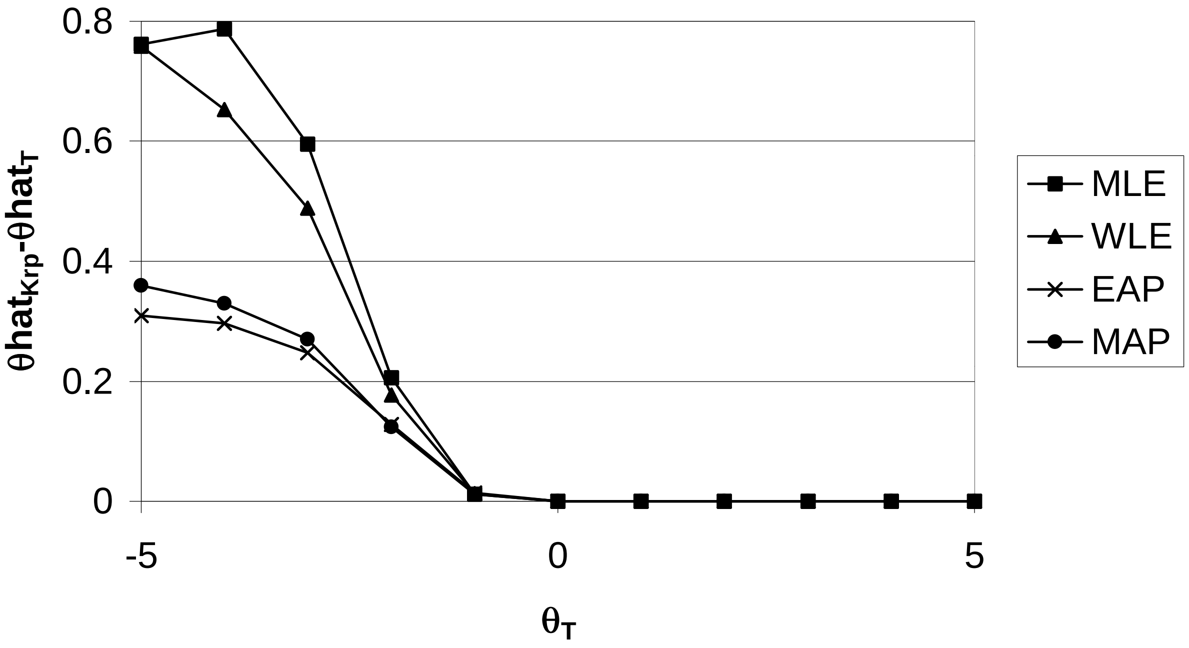


Figure 13d: $\theta hat_{Kri}-\theta hat_{T}$

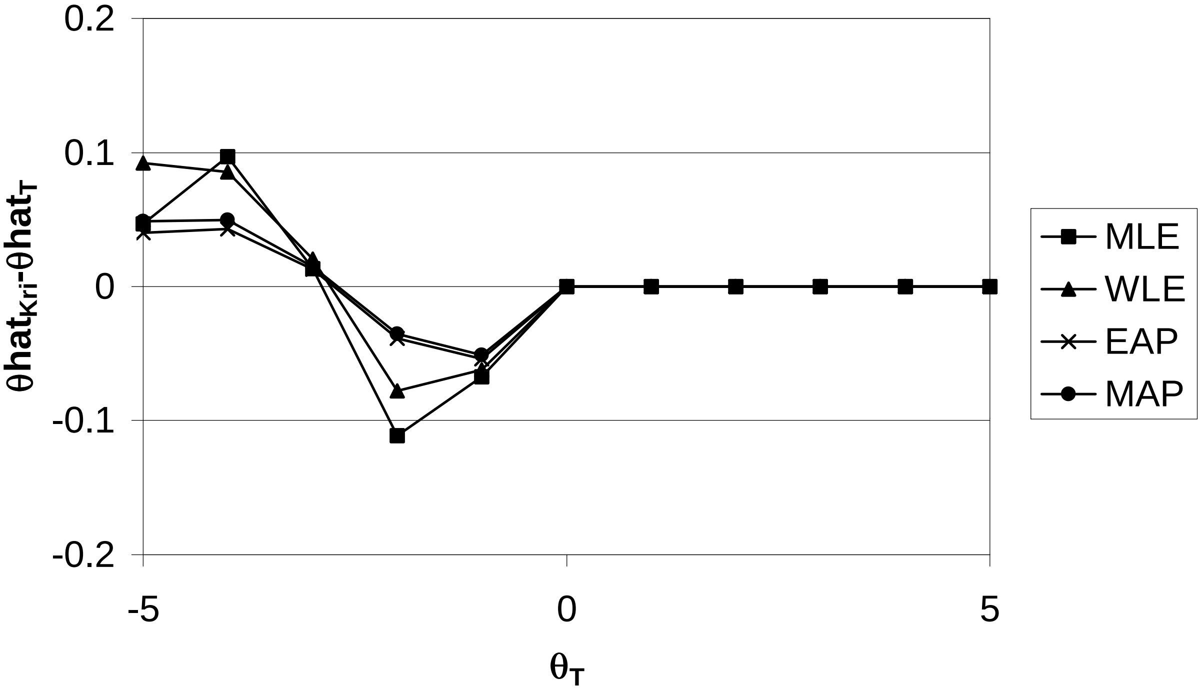


Figure 13e: θhat_{gop}-θhat_T

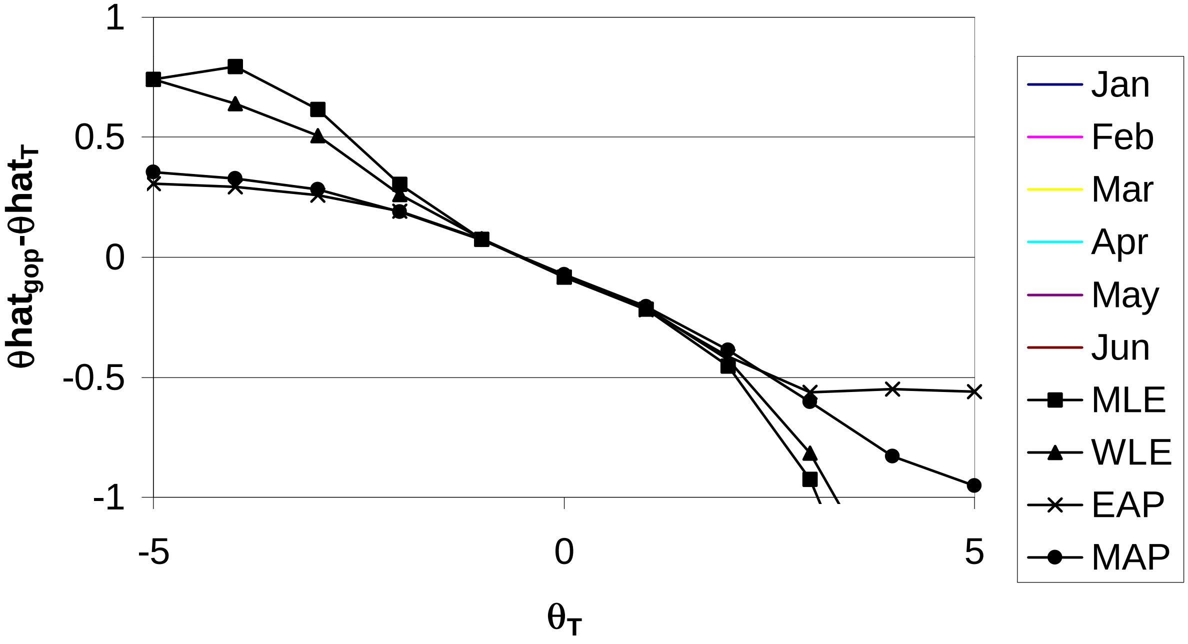


Figure 13f: θhat_{goi}-θhat_T

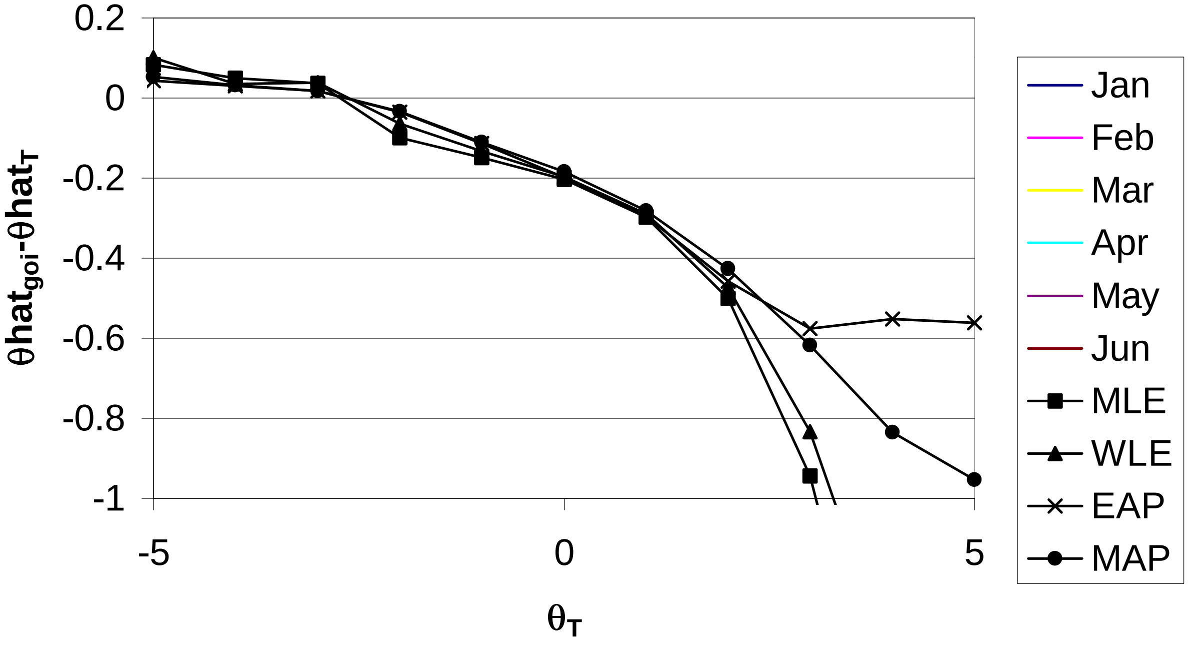


Figure 13g: θhat$_{grp}$-θhat$_{T}$

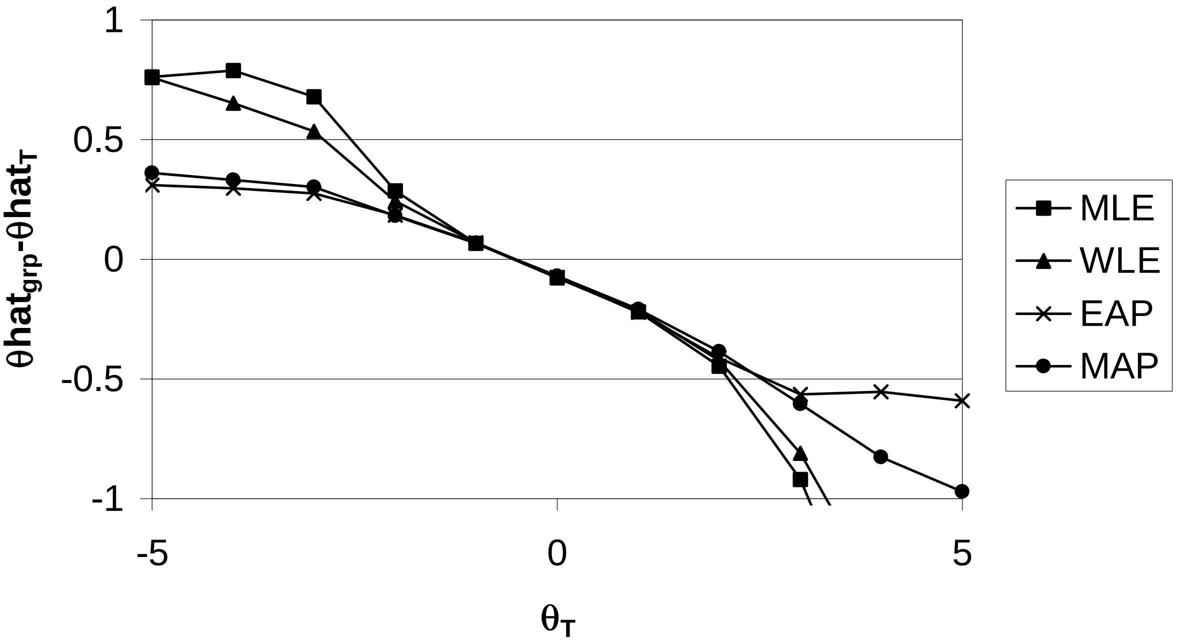


Figure 13h: θhat$_{gri}$-θhat$_{T}$

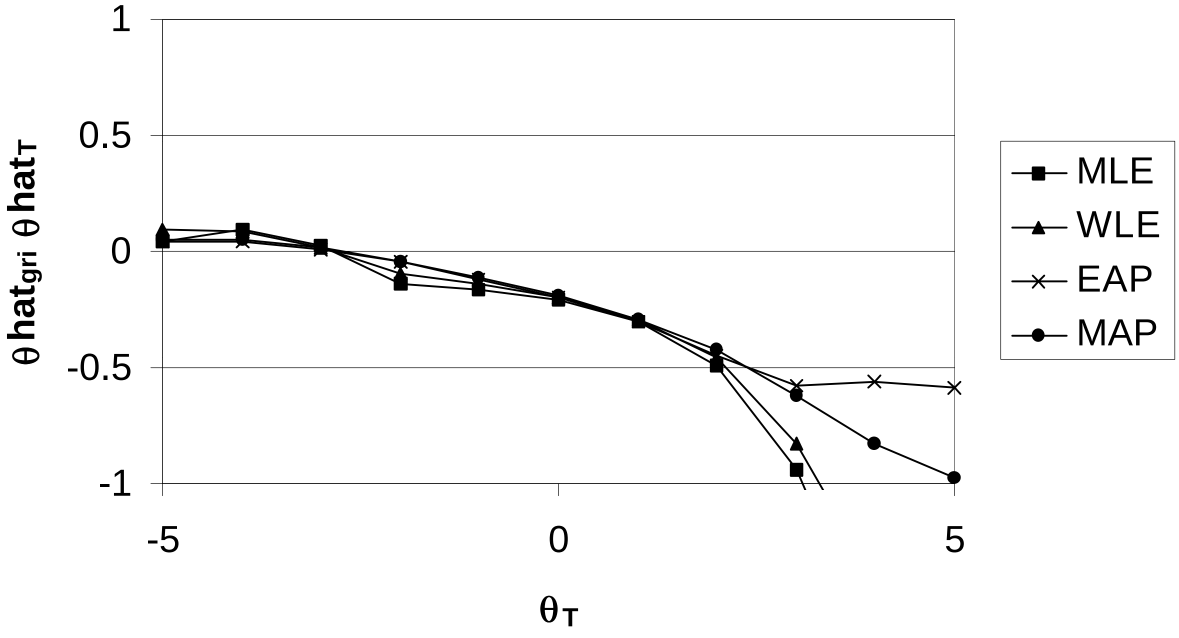


Figure 14a: θhat_{Knp}-θhat_{T}

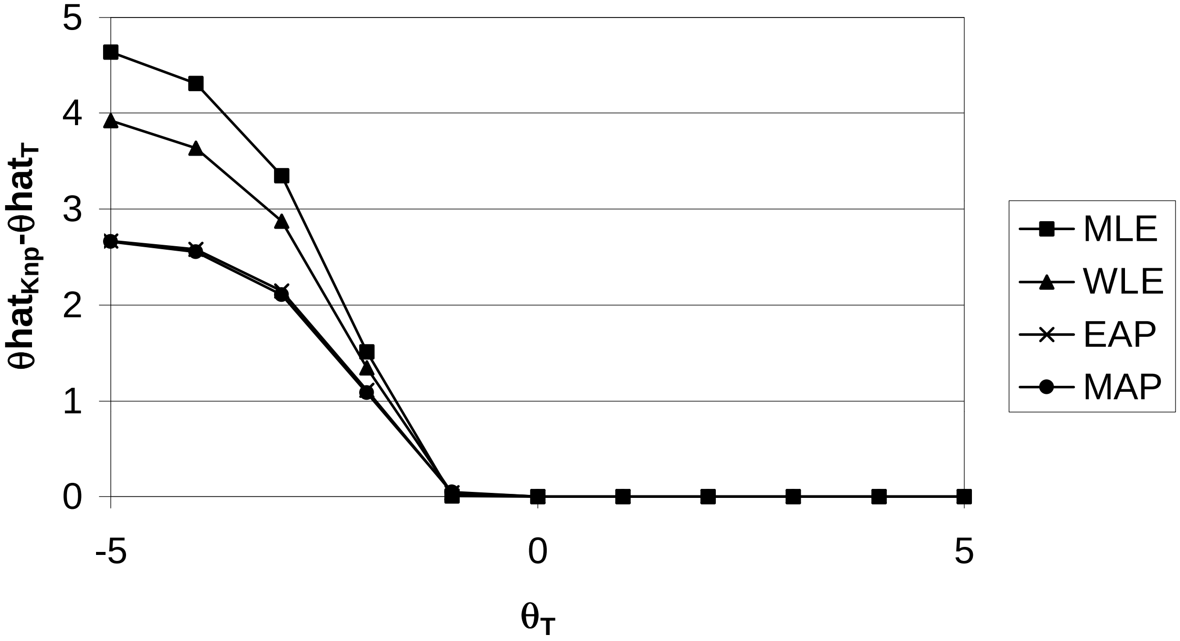


Figure 14b: θhat_{Kni}-θhat_{T}

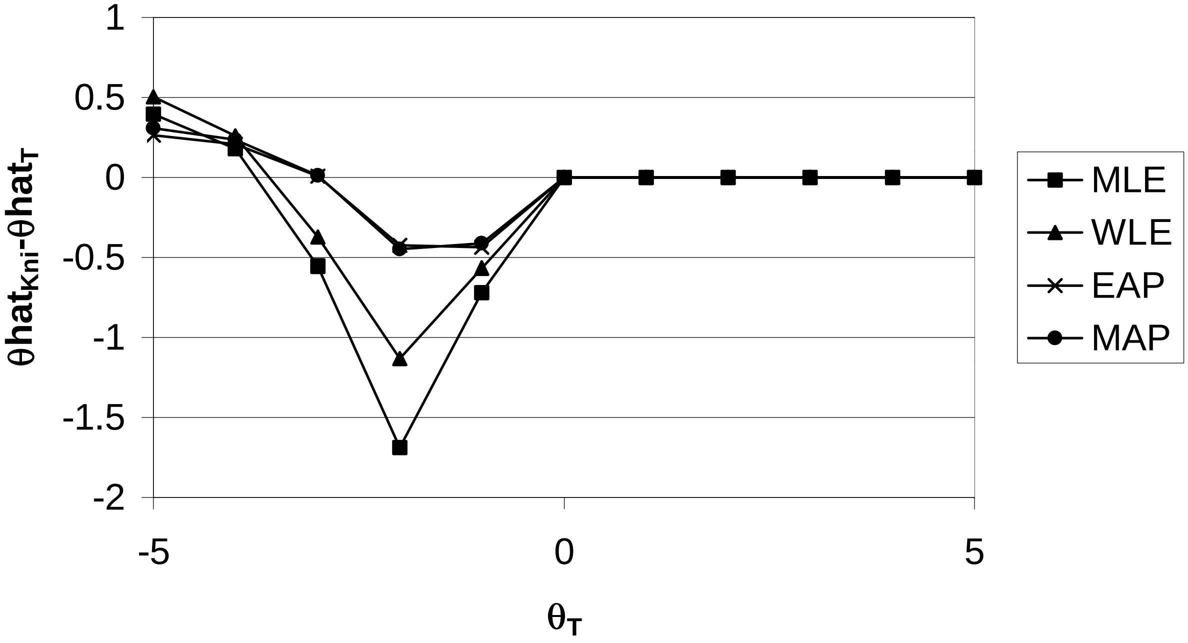


Figure 14c: θhat_{gnp}-θhat_T

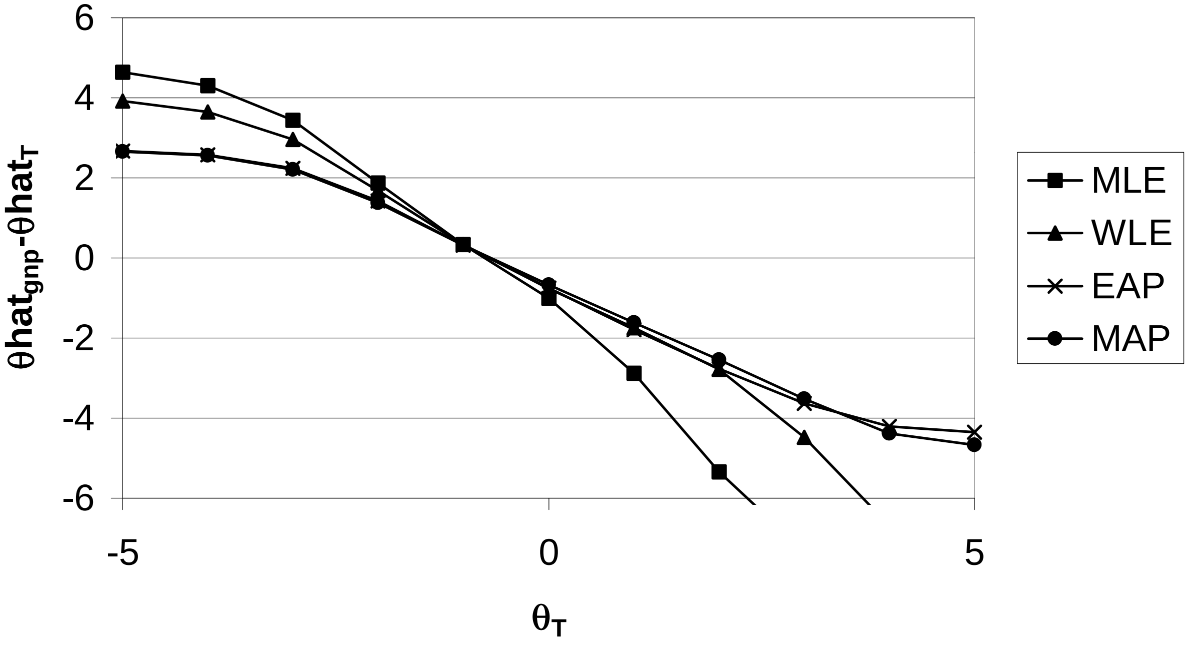


Figure 14d: θhat_{gni}-θhat_T

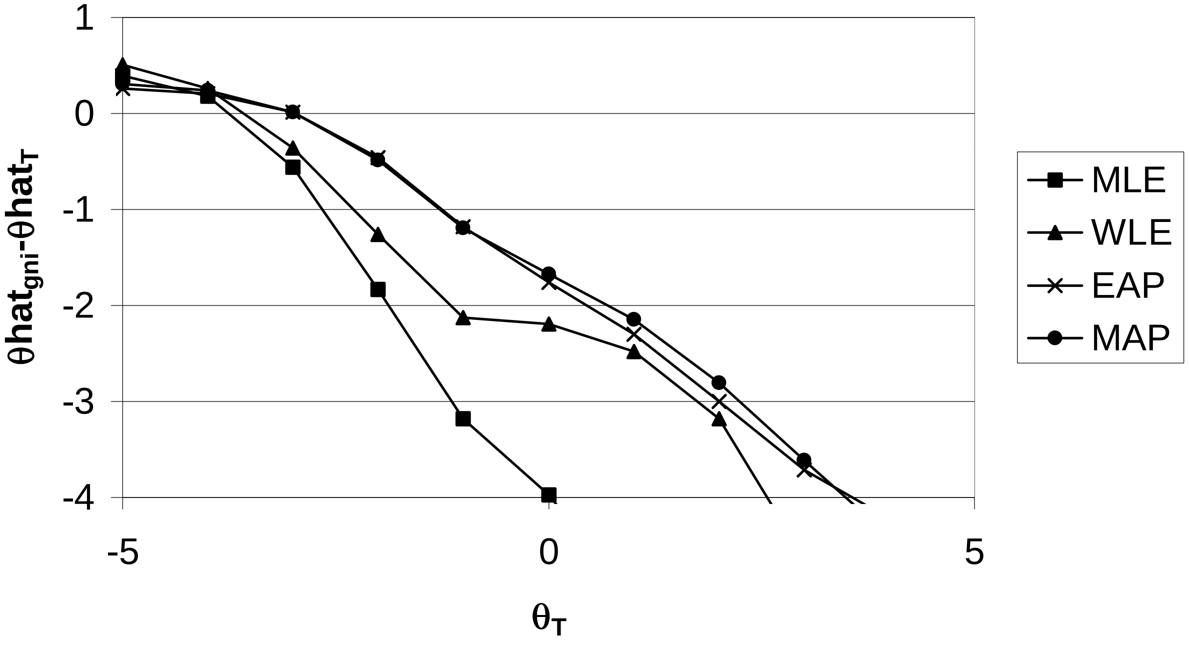


Figure 15a: IE_{Kop}

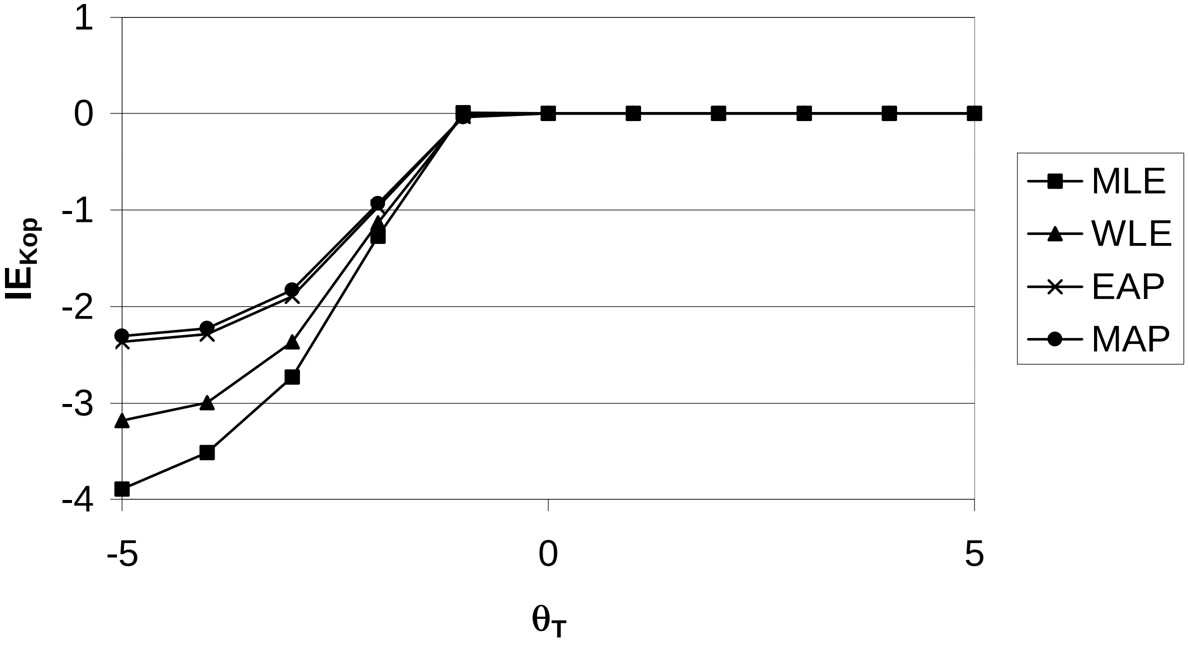


Figure 15b: IE_{Koi}

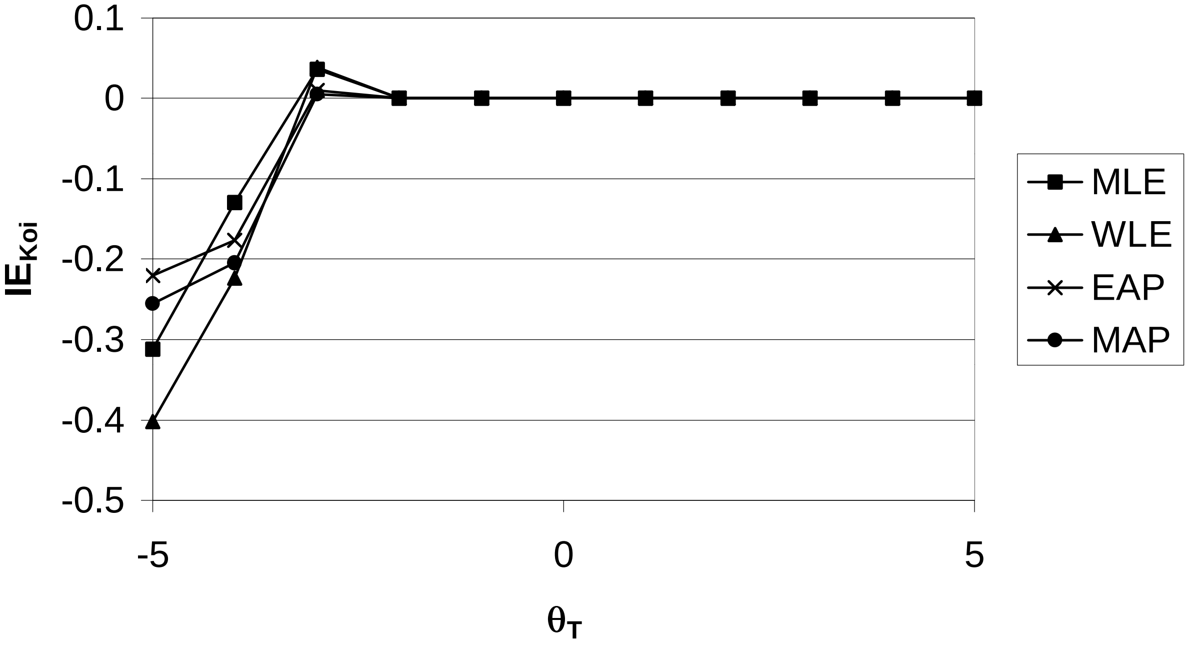


Figure 15c: IE_{Krp}

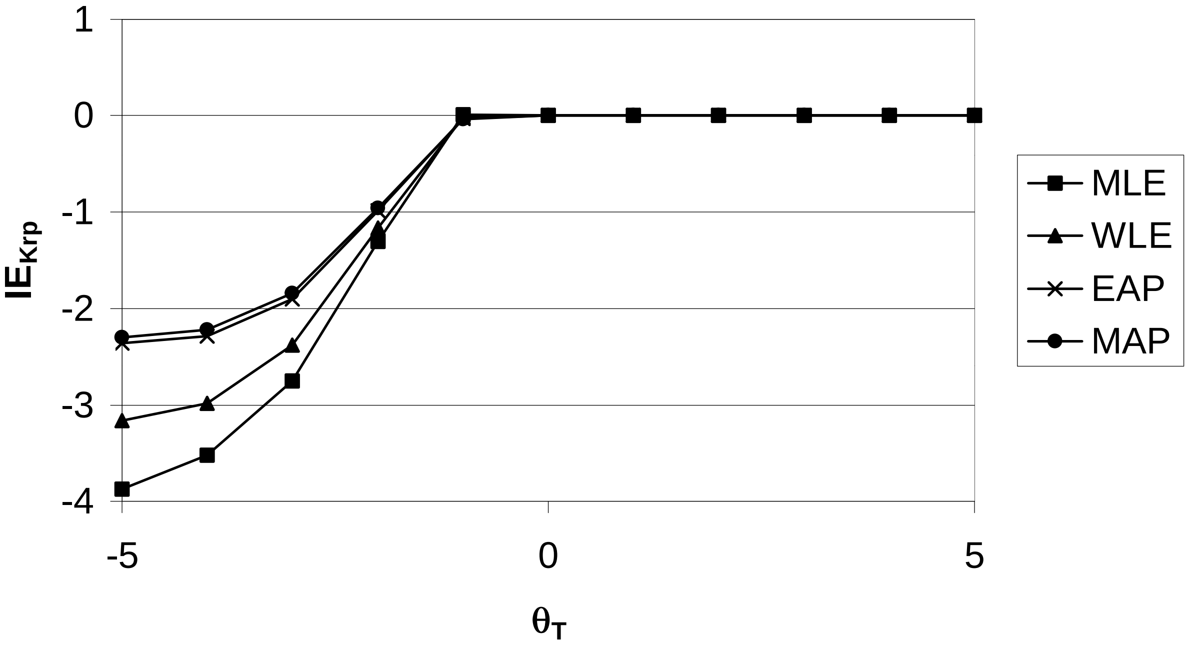


Figure 15d: IE_{Kri}

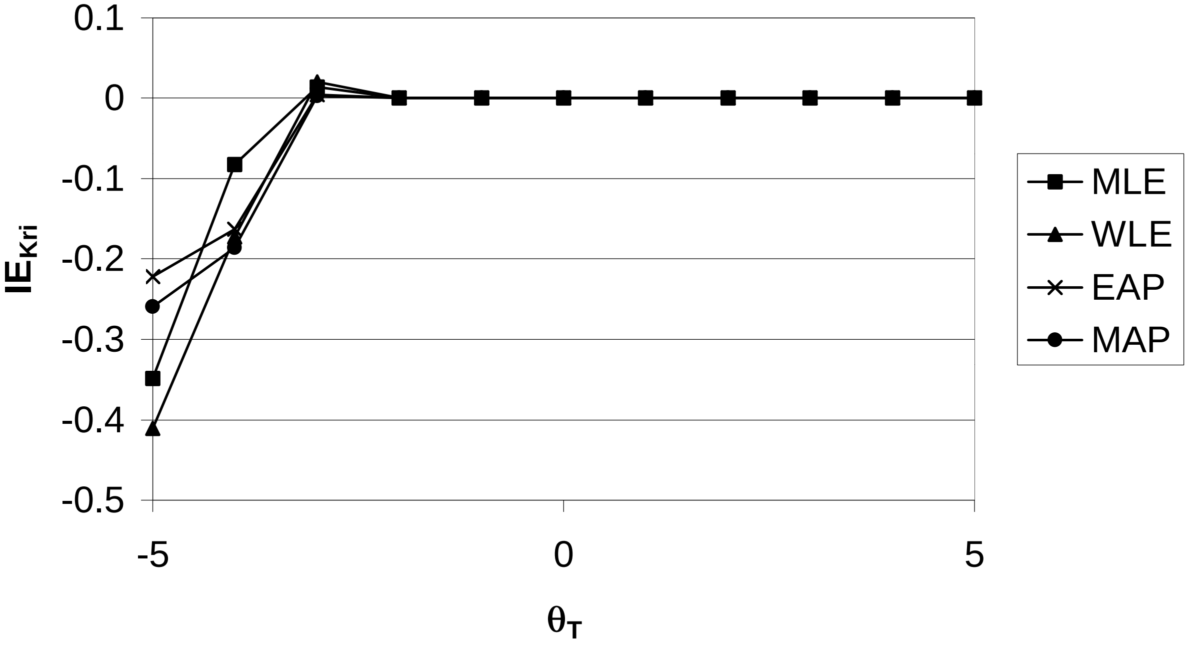


Figure 15e: IE_{gop}

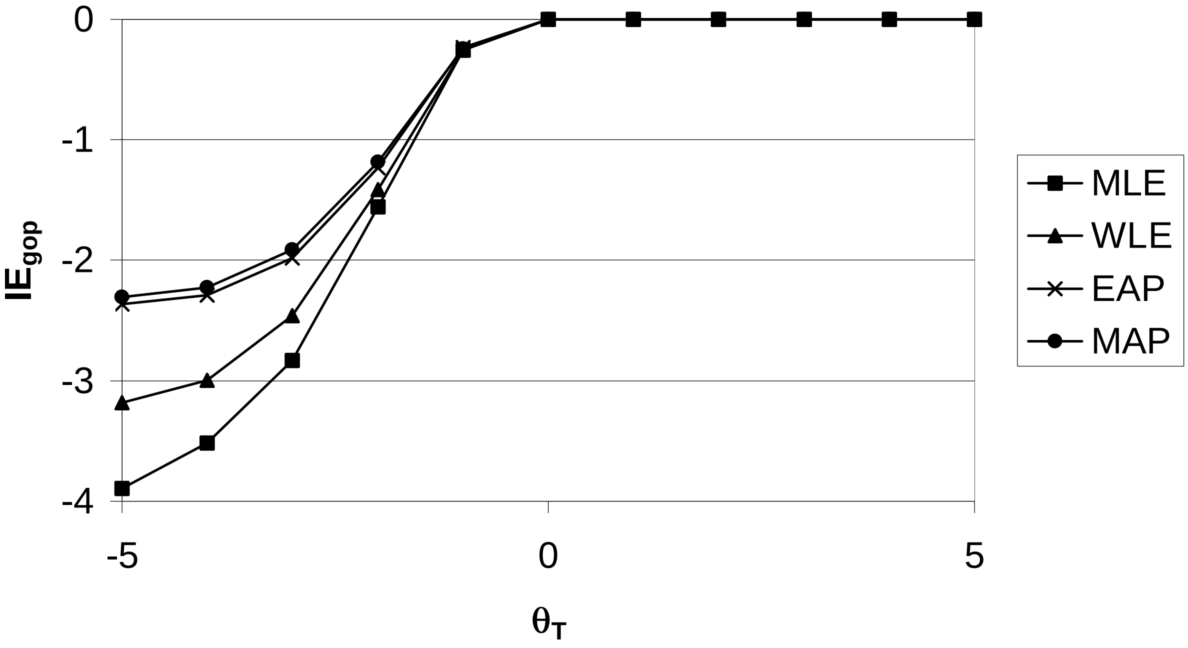


Figure 15f: IE_{goi}

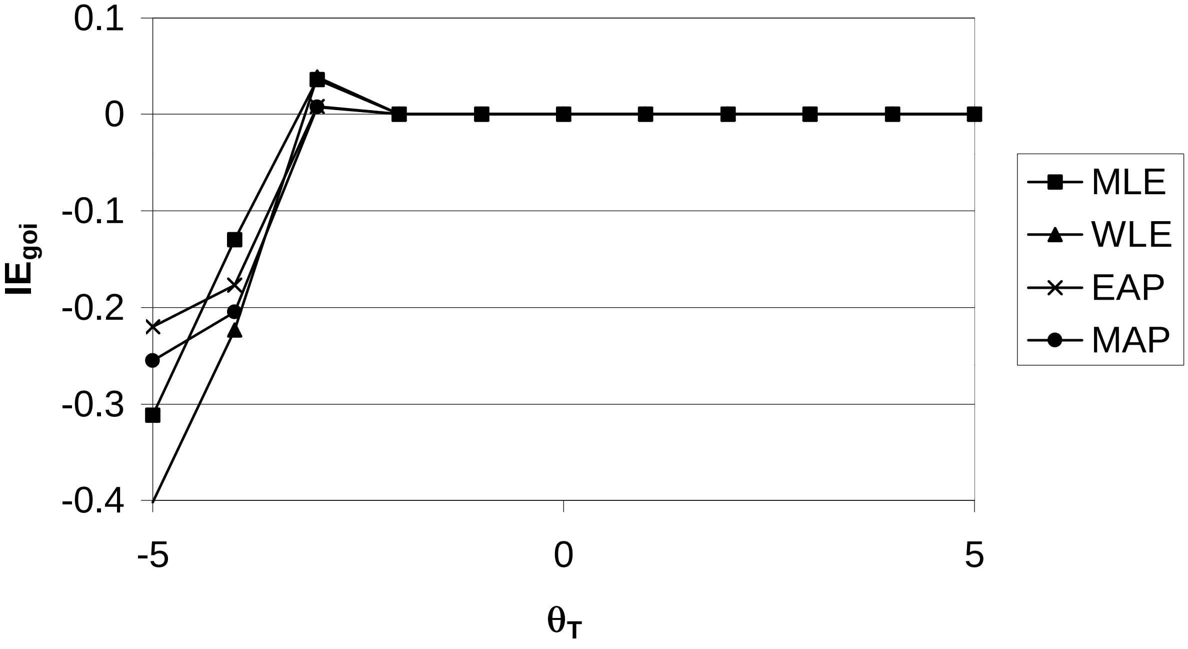


Figure 15g: IE_{grp}

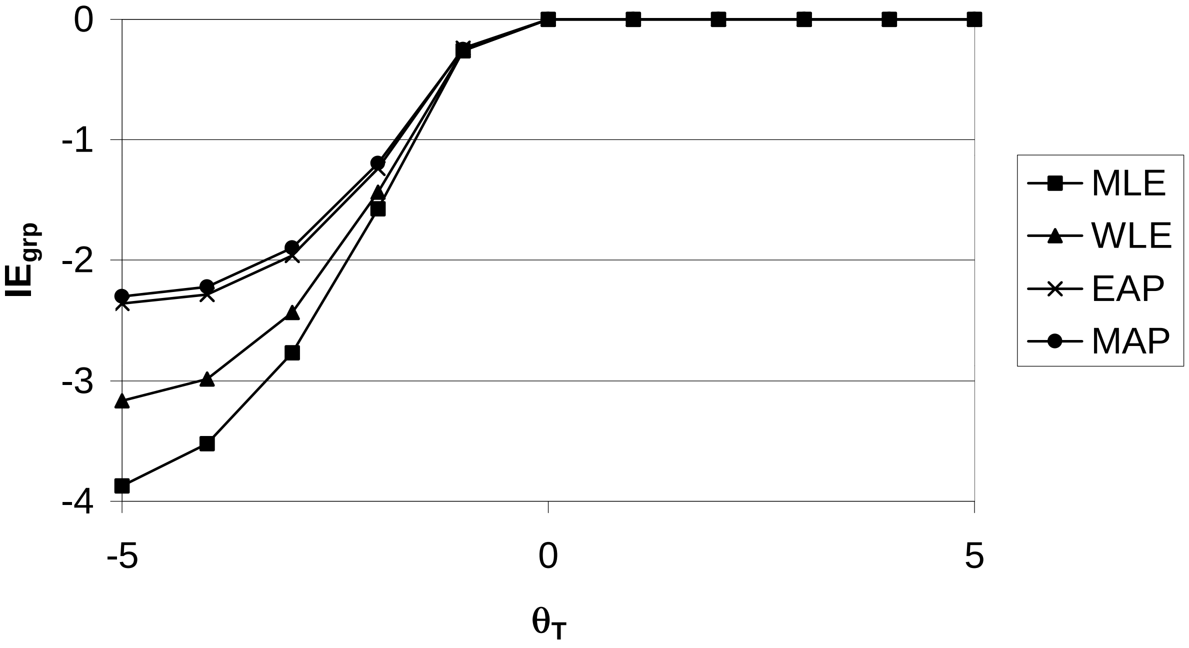
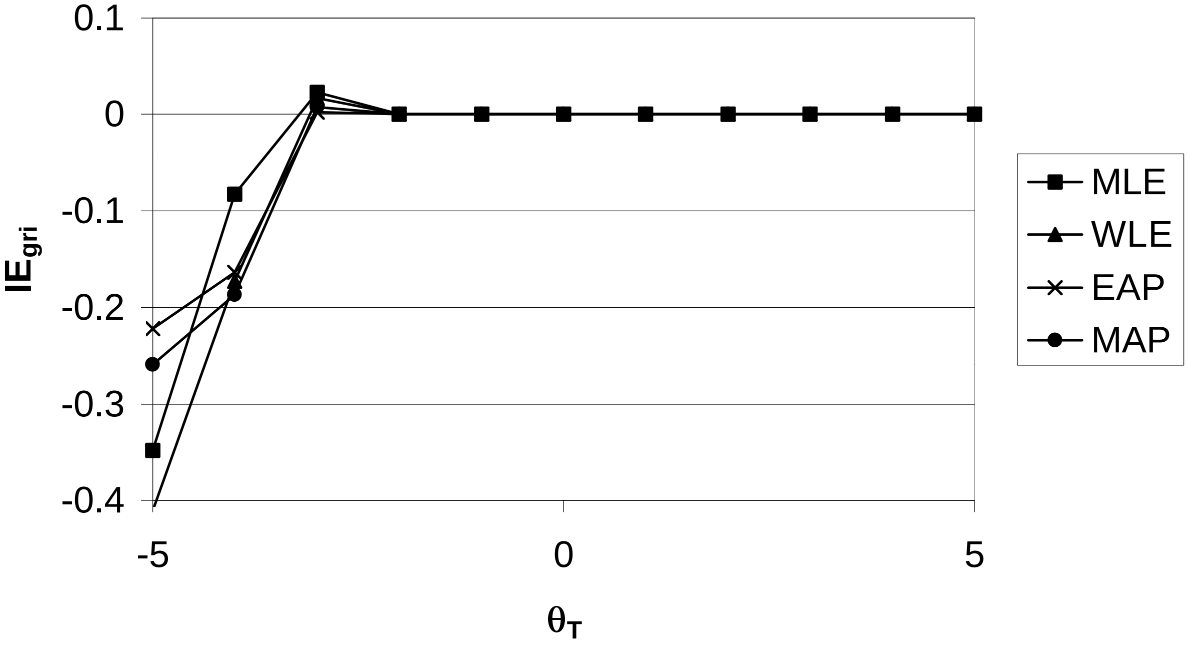


Figure 15h: IE_{gri}


IE_{gri}
θ_T
MLE
WLE
EAP
MAP