Adaptive success control in computerized adaptive testing

JOACHIM HÄUSLER

Abstract

In computerized adaptive testing (CAT) procedures within the framework of probabilistic test theory the difficulty of an item is adjusted to the ability of the respondent, with the aim of maximizing the amount of information generated per item, thereby also increasing test economy and test reasonableness.

However, earlier research indicates that respondents might feel over-challenged by a constant success probability of \( p=0.5 \) and therefore cannot come to a sufficiently high answer certainty within a reasonable timeframe. Consequently response time per item increases, which – depending on the test material – can outweigh the benefit of administering optimally informative items. Instead of a benefit, the result of using CAT procedures could be a loss of test economy.

Based on this problem, an adaptive success control algorithm was designed and tested, adapting the success probability to the working style of the respondent. Persons who need higher answer certainty in order to come to a decision are detected and receive a higher success probability, in order to minimize the test duration (not the number of items as in classical CAT).

The method is validated on the re-analysis of data from the Adaptive Matrices Test (AMT, Hornke, Etzel & Rettig, 1999) and by the comparison between an AMT version using classical CAT and an experimental version using Adaptive Success Control.

The results are discussed in the light of psychometric and psychological aspects of test quality.

Key words: CAT, adaptive testing, test economy, ASC, success control

1 Mag. Joachim Häusler, Dr.G.Schuhfried GmbH, Hyrtlstraße 45, 2340 Mödling, Austria; e-mail: haeusler@schuhfried.at
Adaptive success control in computerized adaptive testing

Theoretical framework

One of the main arguments for adaptive test procedures (CAT, Computerized Adaptive Testing) – which were originally known as Tailored Testing (Lord, 1968) – is the high level of economy promised for tests constructed in this way (Hornke, 1993). Assuming that a sufficiently large item pool is available, each subject receives the specific items appropriate to his or her ability; each item therefore generates the maximum amount of information about the latent trait.

The result is that a pre-defined level of precision can be achieved with a minimum number of items – a situation that at first sight appears very economical. Since, however, CAT can only be used within the framework of a probabilistic test model – usually the 1-PL model of Rasch (Rasch, 1980) or the 2-PL model of Birnbaum (1968) – items must be presented in the form of a pure power test. The subject alone is responsible for selecting the amount of time he or she will invest in working on an item. In consequence the response time per item may increase significantly and this may offset or even outweigh the saving in the number of items that need to be presented (Wild, 1989). Approaches have been devised which attempt to counter this problem with “speed-adapted testing” (Nährer, 1989). This method incorporates an exploratory phase in which a linear speed-accuracy trade-off is estimated; this is used in the test phase to select items with a very high probability of leading to answers within the time limits set for individual items. It therefore provides a means of countering those effects of time limits, which are disadvantageous for probabilistic test models. Adaptive tests based on probabilistic speed and power models (Scheiblechner, 1979; Roskam, 1997) would be able to solve this problem more elegantly, but no such tests are as yet in existence. The Frankfurt Adaptive Concentration Test (FAKT; Moosbrugger & Goldhammer, 2006) could be mentioned as an example of an adaptive speed test. However, the problem described above does not arise with this test, since in FAKT speed and accuracy are aggregated in an unsystematic way into a single measure of performance. In his context “unsystematic” refers to the fact that it could not be shown that the combined scoring of accuracy and speed in fact corresponds to the speed-accuracy trade-off that characterizes this type of task (cf. Häusler, 2004a). The combined scoring therefore could favour a specific working style on the impulsive/reflexive dimension; the test would therefore not meet the criterion of fairness.

Furthermore, CAT is often described by test subjects as over-challenging, too tiring and unduly demotivating (Andrich, 1995). This is in sharp contrast to the original intention, which was to use item selection to create a challenging and optimally motivating situation in which the subject feels neither over- nor under-challenged. The findings of Heckhausen (1989) shed doubt on whether all subjects prefer a success probability of p=0.5. Heckhausen postulates inter-individual differences, with success-motivated subjects preferring a success probability of around 70-80%, while failure-avoidance motivated subjects prefer success probabilities that are minimally informative (very high or very low). In psychological assessment it does of course make no sense to provide subjects with tasks of the latter type. According to Atkinson (1964) the subjective expectancy of success or failure is a determining factor of motivation. This expectancy primarily depends on the difficulty of the item. According to Koestler & McClelland (1990) a moderate difficulty is to be preferred, where success can be achieved by investing enough effort.
The result of these problems is that adaptive tests, despite the innovative principle behind them, still do not get the credit they deserve. The present paper aims at initiating a discussion of how CAT procedures might be modified in order to make them more appropriate.

The method of adaptive success control

The primary goal of Nährer’s “speed-adapted testing” was a test duration that is calculable in advance and as far as possible similar for all subjects. By contrast, the method described in this paper aims at the optimization of test duration in order to improve test economy. The testing process can be viewed as an information-gathering task in which information on the value of a latent trait in a particular individual is acquired. Concepts from the field of information theory can therefore be applied.

CAT maximises the amount of information (I) generated by an item.

\[ I_j = P(+|\xi_j, \sigma_j) \times P(-|\xi_j, \sigma_j) = \frac{e^{\xi_j - \sigma_j}}{1 + e^{\xi_j - \sigma_j}} = \max. \]  

This is achieved when the person ability parameter \( \xi_j \) equals the item difficulty parameter \( \sigma_j \). In this case – in accordance with the model equation of the Rasch model – the success probability is \( P(+|\xi_j, \sigma_j) = \frac{e^{\xi_j - \sigma_j}}{1 + e^{\xi_j - \sigma_j}} = 0.5 \); the corresponding amount of information gathered is I=0.25. If items are presented with a success probability either above or below \( p=0.5 \), the amount of information generated will be correspondingly less. The relation between success probability and information generated is described by the information function. The shape of this function is, however, extremely flat. This means that within reasonable limits there is very little loss of information when the success probability deviates from \( p=0.5 \).

If the success probabilities of a linear test with fixed item sequence are assumed to be distributed uniformly within the interval \( p=[0.2;0.8] \), the optimal adaptive test is only about 13% more informative. If the success probabilities are assumed to have a normal distribution \( (m=0.5; s=0.1) \), the information gain by applying an adaptive test procedure is only around 5%.

Since CAT procedures continue testing until a particular level of reliability has been reached and since the standard error of measurement SEM is a function of the information (I) collected \( (SEM = \frac{1}{\sqrt{I}}) \) equation (1) can be resolved analytically for the required number of items. The number of items is in turn related – via the response time of the individual items – to the total test duration T.

\[ T = \frac{1}{I \times SEM^2} \times T \]  

(2)
The function of test duration is therefore made up of two components, namely the test length and the mean item response time $T$. The test length is the number of items, necessary to reach a given standard error of measurement (SEM) and can be defined by using the information function. The mean item response time on the other hand is an unknown function of the success probability. There follows a demonstration of how the type of this unknown function affects the relation between success probability and the test duration. Figure 1 shows examples of several different forms of relationship and demonstrates that the minimal test duration is not necessarily achieved by means of items with a success probability of $p=0.5$.

A constant relation between success probability and item response time is highly improbable. Decision-making models for stimulus pair comparisons (Vickers, 1970) or multiple choice situations (Häusler, 2004b) postulate that the decision time required increases (e.g. linearly) as success probability decreases. This can be explained theoretically by viewing the solution process as a method of accumulating information that is continued until an inter-individually different required response certainty threshold has been reached. The item is then answered – but not necessarily correctly. It is therefore entirely plausible that relations

---

**Figure 1:** Examples of the effect of different mathematical relations between success probability (represented here by the item difficulty $\sigma$ for a subject of ability $\xi=0$ based on the model equation of the Rasch model $P = \frac{e^{\xi-\sigma}}{1 + e^{\xi-\sigma}}$) and mean item response time on the total test duration $T$. $b_0$ and $b_1$ are the parameters of the different relations.

In the example on the left there is assumed to be a constant relation (the test duration is independent of the success probability) between success probability and item response time. The relation between success probability and total test duration corresponds to the inverse information function. The test is longest when items that are too easy or too difficult are used; it is shortest when items with $p=0.5$ are presented.

In the other two examples there is assumed to be a linear (centre) or exponential (right) relation between success probability and item response time. Here the shortest overall test duration is no longer achieved with items with $p=0.5$, but with more simple items.
between success probability and response time might be modelled in which success probabilities of $p=0.5$ lead to sub-optimal minimization of test duration. In fact the method currently used in CAT is only a marginal solution for the unlikely event that there is a constant relation between success probability and item response time.

With Adaptive Success Control (ASC) not only item difficulty but also success probability is adapted to the individual subject. It is therefore a second-order adaptive process. The success control algorithm operates as an approximation algorithm for estimating the most economical success probability on the basis of the previous item response times. The adaptive algorithm selects an item that offers the subject the previously optimized success probability. Figure 2 shows the closed-loop systems of adaptive testing and adaptive success control.

For many subjects ASC would result in little, if any, shortening of test duration. Adaptive adjustment of success probability is therefore only necessary and useful with test subjects with particular working styles, or in some cases also with particular test materials. The ASC algorithm should therefore only be triggered if a subject’s response times over several items have exceeded a particular threshold.

However, because of the limited data available, actual estimation of the optimal test duration is scarcely feasible. Neither can it be expected that a large amount of the variance of the test duration can be explained by the relation between success probability and response time. Since, however, the aim is only the pragmatic improvement of test economy, it is instead sufficient to formulate an appropriate heuristic adaptation method.

![Figure 2: Adaptive Testing (left) and Adaptive Success Control (right) represented as servoloops.](image)

Adaptive Difficulty Control

<table>
<thead>
<tr>
<th>Ability</th>
<th>Working style</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Item Parameter</td>
<td>50%</td>
</tr>
</tbody>
</table>

Adaptive Testing is a first-order closed-loop system, in which the selection of items is determined by the performance achieved (adaptive difficulty control). By contrast, Adaptive Success Control utilizes an additional loop in order to adapt the desired success probability to the subject’s working style.
**ASC trigger phase**

ASC is only triggered if the response time for several items exceeds a specified threshold. This threshold can be set on the basis of the distribution of the test durations in a test’s norm sample and should represent which test durations are considered acceptable.

When ASC is triggered, 2 starting items are presented.

1) Presentation of an item with a maximum success probability value (for this example \( \text{ASC}_{\text{max}} = 0.8 \)). The maximum success probability is also a constant which must be declared in the ASC framework. ASC will operate in the success probability range between \( p = 0.5 \) and \( p = \text{ASC}_{\text{max}} \).

2) Presentation of an item with a success probability of \( p = (0.5 + \text{ASC}_{\text{max}}) / 2 \).

It is now possible to estimate the test duration for the two marginal conditions (3) and (5), and for the central condition (4).

This example yields three estimates for the test duration: Equation (3) is the estimated test duration if the adaptive algorithm was continued without success control. Equation (5) is the estimated test duration if the test was continued with the maximum success probability defined in the ASC constants (in this case \( p = 0.8 \)). Equation (4) is the estimated test duration if a medium success probability of \( p = 0.65 \) was applied.

\[
T_{50} = \frac{1}{1 * \text{SEM}^2} \times T_{50} = \frac{1}{0.25 * 0.6^2} \times T_{50} = 11.1 \times T_{50}
\]  
(3)

\[
T_{65} = \frac{1}{1 * \text{SEM}^2} \times T_{65} = \frac{1}{0.23 * 0.6^2} \times T_{65} = 12.1 \times T_{65}
\]  
(4)

\[
T_{80} = \frac{1}{1 * \text{SEM}^2} \times T_{80} = \frac{1}{0.16 * 0.6^2} \times T_{80} = 17.4 \times T_{80}
\]  
(5)

**ASC adaptation phase**

As has already been noted above, these estimates are prone to error, so that an exact adaptation would produce only a semblance of precision and could become bogged down in local minima. The method therefore proceeds by always assuming the new success probability to be halfway between the two best preceding success probabilities. Since the adaptation is based on single observations, the value is made fuzzy by adding a normally distributed random error. The standard deviation of the random error would correspond to half of the interval between the two previous best success probabilities. That approach prevents the adaptation process to get stuck in a local minimum.

In the examples let the response times on the trigger phase items be that way, that the estimated test durations are \( T_{50} > T_{80} > T_{65} \). The optimum success probability therefore seems to be between 0.65 and 0.80. The new success probability is therefore calculated as a normally distributed random variable with \( m = 0.725 \) and \( s = 0.075 \).
In this example the new selected success probability is to be 0.78.

\[ T_{78} = \frac{1}{\sqrt{T_{78}}} = \frac{1}{\sqrt{0.247}} \cdot T_{78} = 2.014 \cdot T_{78} \]

After presenting one item with the success probability \( p=0.78 \), the new triplet of observed success probabilities (\( T_{80}, T_{78} \) and \( T_{60} \)) is compared. If \( T_{65} \) and \( T_{78} \) are the two success probabilities with the lowest estimated test duration, they are kept and a new value is generated. The normally distributed random variable has \( m=0.715 \) and \( s=0.065 \).

The ASC procedure is continued in this way by keeping the best two success probabilities out of each triplet and generating a new one until the end of the test. It does not necessarily lead to a convergence; the estimated values are more likely to level out as a gentle oscillation around the optimal success probability, thus approximating to the optimal success probability for the subject.

**Study I**

Using the Adaptive Matrices Test (AMT; Hornke, Etzel & Rettig, 1999) as an example, the currently available adaptive algorithm will be tested by re-analysis of the existing norm sample. Of principal interest is the issue of whether the relations between success probability and item response time hypothesized above can in fact be found, and what form they take.

**Results**

The sample data was gathered between 2003 and 2005 in the research laboratory of Dr. G. Schuhfried GmbH. The sample consists of 392 individuals (40% men, 60% women) aged between 18 and 81. All the EU education levels (EU1: Compulsory schooling not completed; EU2: Completed compulsory schooling; EU3: Completed vocational training; EU4: High-school graduation with university entrance exam; EU5: University or college degree) were represented (EU1: 1%, EU2: 14%, EU3: 34%, EU4: 42%, EU5: 9%).

While the number of items necessary to reach the target reliability of \( \alpha \geq 0.80 \) has a more or less acceptable normal distribution, the test duration is skewed; for some respondents test completion times of more than three hours were recorded. The histogram in Figure 3 shows the total test duration for the AMT, test form S2.

Since (at least during the initial items of an adaptive procedure) a wide range of success probabilities is represented, the attempt can be made to estimate the correlation between success probability and response time without any special adaptation of the AMT. The norm sample of the AMT was used for this purpose, although it was limited to the quartile of subjects with the highest test completion times. Similarly, only the first 10 items were taken into account, since further into the adaptive process it can be assumed that only success probabilities very close to \( p=0.5 \) will be represented.
Distribution of test duration in seconds (without instructions) for the AMT, in order to achieve a target reliability of $\alpha=0.8$ ($m=2375, s=1595, skew=1.8, kurtosis=5.6$). This yields 460 individual item responses which were re-analysed. The re-analysis resulted in a distribution of the success probabilities that are about normal distributed ($m=0.60; s=0.18; skew=-0.290; kurtosis=-0.535$). For the response times the distribution is again sharply positively skewed ($m=167.2; s=134.4; skew=2.617; kurtosis=11.725$). A scatter diagram of the correlation between item success probability and item response time is shown in Figure 4.

The linear approximation of the data is able to explain around 11% of the variance in response times by means of the success probability. This corresponds to a weak linear correlation. Approximations using higher-order polynomials make better prediction possible. This prediction, however, lies within the range of random adaptation because of the additionally required degrees of freedom. A genuine gain in explained variance can be obtained with an exponential modelling (15% explained variance). This too does not offer fewer degrees of freedom, and therefore greater scope for random adaptation, than a linear model. The optimal success probabilities can be derived from the hypothesized correlation functions as minima of the functions. This is shown in Figure 5; the mathematical derivation of these functions will be found in the Appendix.
Figure 4:
Curve fit for the relation between success probability $p$ and item response time $bt$ for tasks with success probabilities in the interval $[0.1; 0.9]$. This results in more than chance correlations for different function types. Higher polynomials do not lead to noticeably better predictions than a linear model. The linear or the exponential relational form seems to provide the most economical description of the data.

<table>
<thead>
<tr>
<th>Relation</th>
<th>$R^2$</th>
<th>df</th>
<th>$F$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>0.114</td>
<td>458</td>
<td>58.65</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Square</td>
<td>0.114</td>
<td>457</td>
<td>29.36</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Cubic</td>
<td>0.115</td>
<td>456</td>
<td>19.80</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Exponential</td>
<td>0.148</td>
<td>458</td>
<td>79.53</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Figure 5:
Correlation between success probability and total test duration projected from the model in Study I for a target SEM <=0.6. The derivations are given in the Appendix.
Study II

The next step is to carry out a feasibility study to test the effectiveness of adaptive success control and investigate the extent to which it is possible to use ASC to shorten the average test duration, even if the average test length is increased.

Two test forms of the AMT were created for this purpose: a classical computerized-adaptive form and a form with adaptive success control (the upper margin of the success control range was set to ASC_{\text{max}}=0.8). For both test forms a standard error of measurement of $SEM \leq 0.6$ was selected as a termination condition. Subjects were assigned at random to one of the two test forms. Item responses and item response times were recorded.

In order to minimize the data collection work while at the same time utilizing the option of equivalence hypotheses, a special form of optimum experimental design was used: the triangular sequential test (Whitehead, 1983; Schneider, 1992; Rasch et al., 2004). In a similar way to older designs for sequential t-tests (Wald, 1947), this method also tests the hypotheses stepwise by blocks of subjects; it is thus more economical in terms of means than traditional optimum sample design. The more relevant advantage, however, is that the optimal sample design enables both the null hypothesis and the alternative hypothesis can be tested with a known statistical risk (Rasch & Kubinger, 2006; Bortz, 1999; Bandemer & Näther (1980)).

The risks of the triangular sequential test were selected at $\alpha = 0.01$ and $\beta = 0.01$. For the purposes of the present investigation a type-I-error has no more disadvantage than a type-II-error; the $\alpha$ and $\beta$ risks were therefore set at the same level. Because of the nature of the investigation and the prior information gained from Study I, the triangular plan is formulated one-sided, so that the specifications for the triangular test for the required sample size yield an expected value of $n = 72$ and a maximum value of $n = 140$. For a test of fixed length an optimum experimental design would suggest a sample size of $n = 96$. The TRIQ module of the CADEMO software package (Rasch, Verdooren & Gowers, 1999) was used for the computation.

Results

The data was collected in 2005 in the research laboratory of Dr. G. Schuhfried GmbH.

A triangular sequential test was carried out on experimental blocks consisting of 5 subjects from each experimental condition. It was possible to terminate data collection after $n=60$ subjects, because the null hypothesis could be rejected.

The resulting sample consists of 39% men, 61% women of EU education levels 2-5 (EU2:4%, EU3:55%, EU4:32%, EU5:9%). The mean age is 56 with a standard deviation of 10 years. Figure 6 shows the graph of the Z value in the triangular sequential test.
Figure 6:
Graph of the one-sided triangular sequential test with the risks $\alpha=0.01$ and $\beta=0.01$. The null hypothesis can be rejected after $n=60$ subjects.

A classical t-test also reveals a significant ($t=3.630; \text{df}=58; p=0.001$) and relevant ($d=0.95; m_{\text{CAT}}=1174 \text{ sec}; m_{\text{ASC}}=720 \text{ sec}$) difference in test duration between the two test forms.

The CAT test form is, as expected, on average some three items shorter than the ASC test form ($t=-5.086; \text{df}=58; p<0.001$). This apparent economy is, however, more than offset by the shorter mean item response time in the ASC test form.

Effects of the item difficulty induced by the type of test presentation were noted; the ability parameters estimated by the two test forms do not differ by more than a chance amount ($t=1.222; \text{df}=8; p=0.227$).

Discussion

The studies described in this paper provide further support for the argument that the saving in items achieved through adaptive testing, in order to arrive at the same degree of reliability, is very slight. If one abides by the commonly used guidelines on scale construction of classical test theory (cf. Bortz & Döring, 2002; Leinert & Raatz, 1998) and selects primarily success probabilities or classical difficulty indices between 0.2 and 0.8, the superiority of adaptive procedures has only a minimal effect on test length. Some papers claim that adaptive testing can shorten tests by 50% or more (cf. Vispoel, Rocklin & Wang, 1994). In this context it should be noted, however, that the linear tests used for comparison were created especially for this purpose. One must therefore suppose that these linear tests were deliberately designed to contain primarily very simple or very difficult items. Such a method is of course not in accordance with the procedure that a careful test developer would employ in constructing a linear test.
Furthermore, the test results under consideration here add weight to the fear that, while adaptive testing does indeed optimize the number of items presented, it is only sub-optimally economical with regard to test duration. The Adaptive Success Control Algorithm proposed here has been shown to be more economical; the difference identified is not only statistically significant but also relevant in terms of content. In addition there were no indications that the change in test presentation had altered its difficulty. It can therefore be concluded that only an individualized approach to test presentation can genuinely fulfill the requirement for improved test economy.

On the basis of the results presented here, the role of adaptive tests needs to be re-evaluated. Adaptive testing should focus less on the opportunity for presenting optimally informative items and instead concentrate on the possibility of varying the way in which the test appears to individual subjects. This feature can be used, firstly, to adapt the success experienced in the test to the motivational needs of the subject (Sommer & Häusler, 2004) and thus to test each subject in an individualized manner under the conditions in which he or she can achieve his best performance, rather than striving for an arbitrary success probability of p=0.5 as a result of misguided considerations of economy.

Secondly, the method can be used to deliberately create standardized over-challenging and under-challenging situations, in which the subject’s behaviour can be observed in the manner of an objective personality test (Kubinger, 2006; Schmidt, 1975). An example of this is a subtest of the Hyperkinetic Syndrome Assessment Method (HKSD; Häusler, 2004c) in which the ability to adapt the effort invested is measured by changes in performance and working style (reflexivity) as the subject moves between under-challenging, optimally challenging and over-challenging tasks.

A third approach could involve using the success probability selected by the subject as an objective measure of the subject’s level of challenge – whether explicitly in a self-adapted test (Rocklin & O’Donnell, 1987) or implicitly through the algorithm of adaptive success control. Such a method of approach would make it possible to convert almost any adaptive ability test into a multi-functional test (Wagner-Menghin, 2006), which measures a personality variable within the framework of an ability test. This approach is not only very economic but as well is accepted by the respondent as a reasonable test – a feature that lacked some other objective personality tests, which applied “nonsense” tasks to cover the personality measurement.

References


Appendix

Derivation of the exponential time function

For the analysis of the minima of the estimated exponential regression function \( T_{exp} \) is differentiated (\( \dot{T}_{exp} \)) and set to zero. This term is solved for the optimal success probability \( P_{extreme} \) for the regression weights \( b_0 \) and \( b_1 \) out of the exponential regression. Two solutions can be found for the quadratic equation, one of them \( (P_1) \) being outside of the definition range of a probability and therefore being discarded. \( P_2 \) is being kept as the minimum of the function derived out of the exponential regression.

\[
T_{exp} = \frac{b_0 * e^{b_1 P}}{SEM^2 *(P - P^2)}
\]

\[
\dot{T}_{exp} = \frac{b_0 * e^{b_1 P} *(1-2P)}{SEM^2 * (P - P^2)^2} + \frac{b_0 * b_1 * e^{b_1 P}}{SEM^2 * (P - P^2)} = 0
\]
Derivation of the linear time function

For the analysis of the minima of the estimated linear regression function $T_{lin}$ is differentiated ($\hat{T}_{lin}$) and set to zero. This term is solved for the optimal success probability $P_{extreme}$ for the regression weighs $b_0$ and $b_1$ out of the linear regression. Two solutions for the quadratic equation can be found, one of them ($P_2$) being outside of the definition range of a probability and therefore being discarded. $P_1$ is being kept as the minimum of the function derived out of the exponential regression.

\[
\begin{align*}
P_{extreme} &= 2+b_1 \pm \sqrt{4+b_1^2} \\
P_{extreme} &= 2 - 1.5146 \pm \sqrt{4 + (-1.5146)^2} \\
&= 0.4854 \pm 2.5088 \\
&= -3.0292
\end{align*}
\]

$P_1 = -0.988$; $P_2 = 0.668$

\[
\begin{align*}
T_{lin} &= \frac{b_0 + b_1 \cdot P}{SEM^2 \cdot (P - P^2)} \\
\hat{T}_{lin} &= \frac{(1 - 2P) \cdot (b_0 + b_1 \cdot P)}{SEM^2 \cdot (P - P^2)^2} + \frac{b_1}{SEM^2 \cdot (P - P^2)} = 0 \\
P_{extreme} &= \frac{-b_0 \pm \sqrt{b_1 \cdot b_0 + b_1}}{b_1}
\end{align*}
\]

\[
\begin{align*}
P_{extreme} &= -4.9022 \pm \sqrt{4.9022 \cdot 4.9022 - 3.6417} \\
&= -4.9022 \pm 2.4858 \\
&= -3.6417
\end{align*}
\]

$P_1 = 0.664$; $P_2 = 2.029$