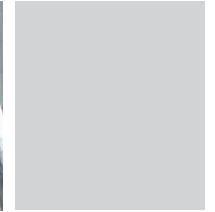




German Federal
Employment Agency



***Searching for the True Confidence Interval of the True Score:
Inference on Person Parameters in
Classical Test Theory and Item Response Theory***

Overview

- Coverage of Wald confidence intervals
- A refined method for asymptotic confidence intervals
- Exact interval estimation
- Bayesian credible intervals
- Norm-referenced (interval) estimation:
Referring to the observed or to the latent distribution?

Notes and Notation:

- Considerations apply to IRT and CTT
- $X \hat{=} \hat{\theta}$
- $\tau \hat{=} \theta$
- Regression confidence intervals: Bayesian credible intervals

Interval estimation: Why and how?

In contrast to points estimates, interval estimates

- ▶ ...are not prone to be misunderstood as a characteristic of the tested person
- ▶ ... allow for inference on the plausibility of each possible value
- ▶ ... give a good idea of the precision of measurement

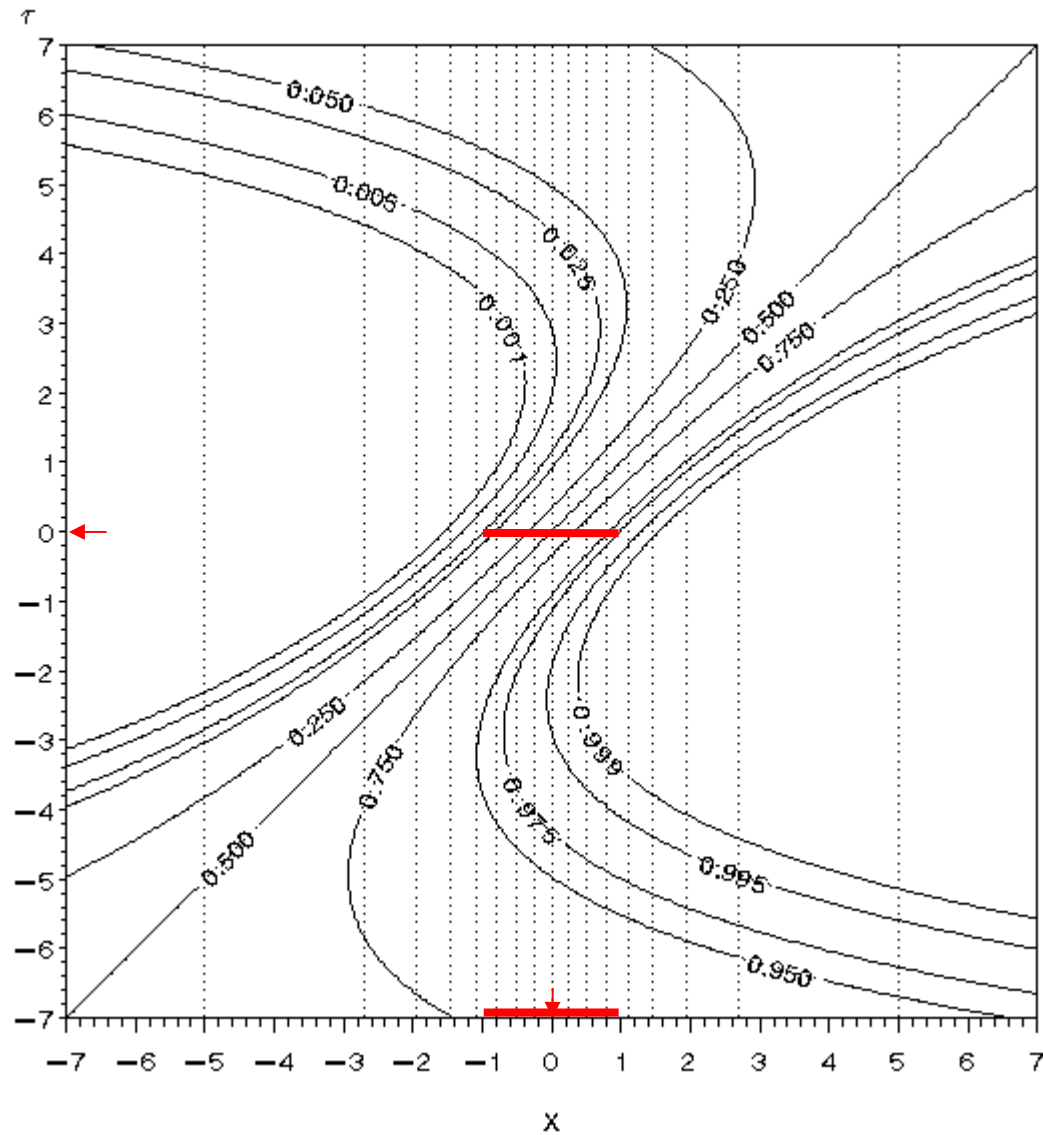
Requirements for interval estimates in an individual diagnostic setting:

- ▶ Should allow for a probabilistic statement that is true for the given individual
- ▶ Probability of Overestimation ($\tau < LCL$) and underestimation ($\tau > UCL$) should be equal
- ▶ The estimate should only depend on the response pattern.



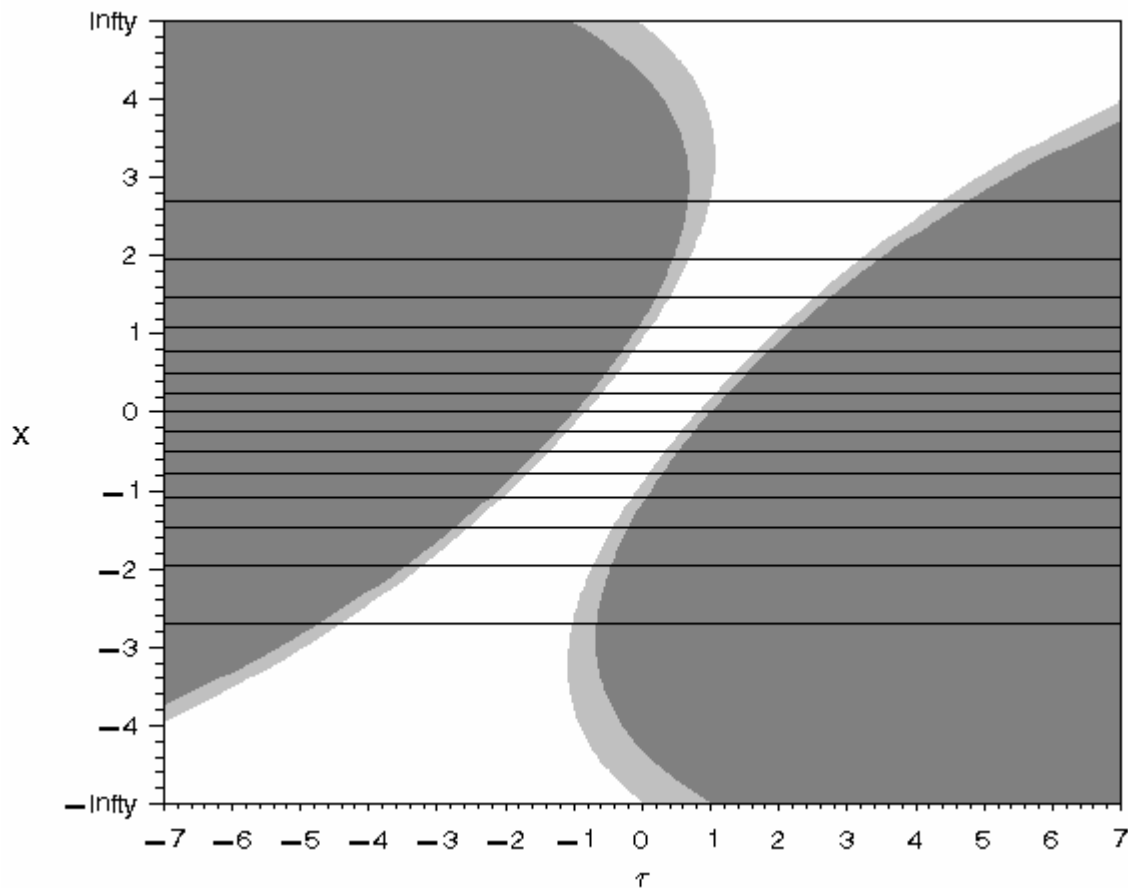
How to calculate confidence intervals for individual assessment?

- Quick and dirty (Wald): point estimate and its standard error
- Correct:
Include all values in the confidence interval that would not lead to a significant deviation from the data, if they were taken as null hypothesis
 - Asymptotic confidence sets based on Fisher information and asymptotic normality



Asymptotic quantiles of the conditional distribution of (the person parameter estimate) X for a Rasch-homogenous item pool of 16 items with identical item parameters ($= 0$). The dotted lines indicate possible values of X .

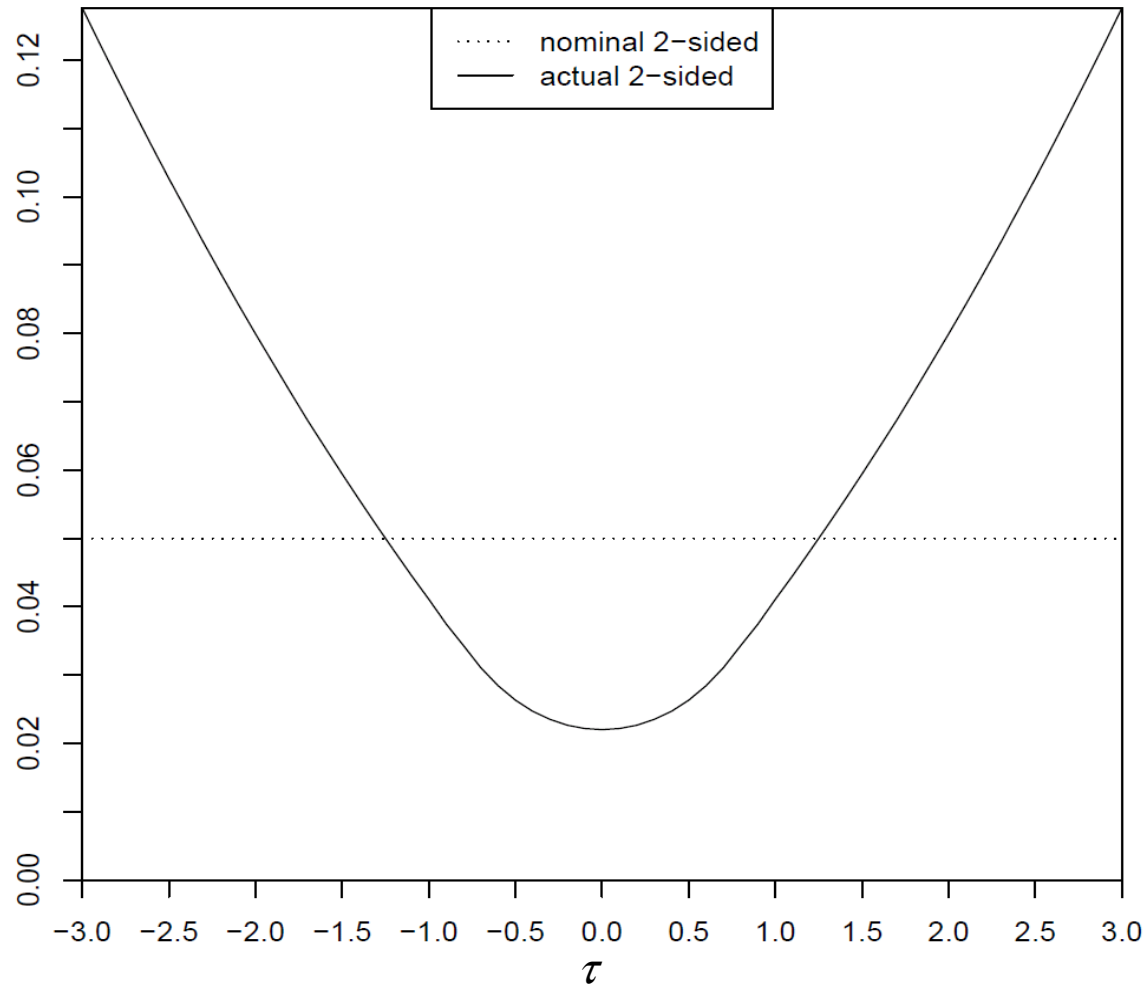
Two-sided 90% (white area) and 95% (white and light gray area) quick and dirty (Wald) confidence intervals that result by adding $\text{probit}(\alpha/2)$ and $\text{probit}(1-\alpha/2)$ times the standard error of measurement to the point estimate



Straight lines indicate possible values of X .

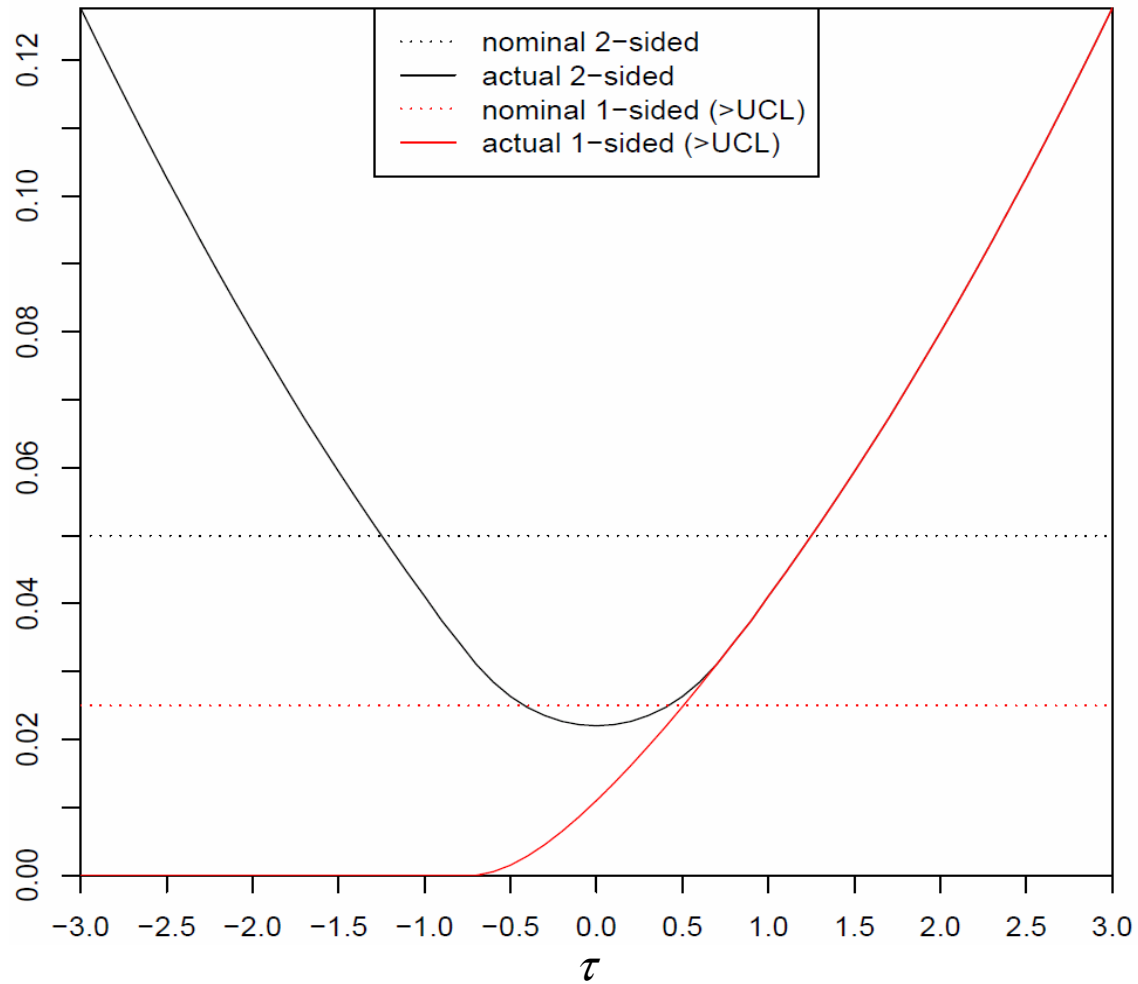
Actual error rates of Wald confidence intervals (distributional assumptions are met)

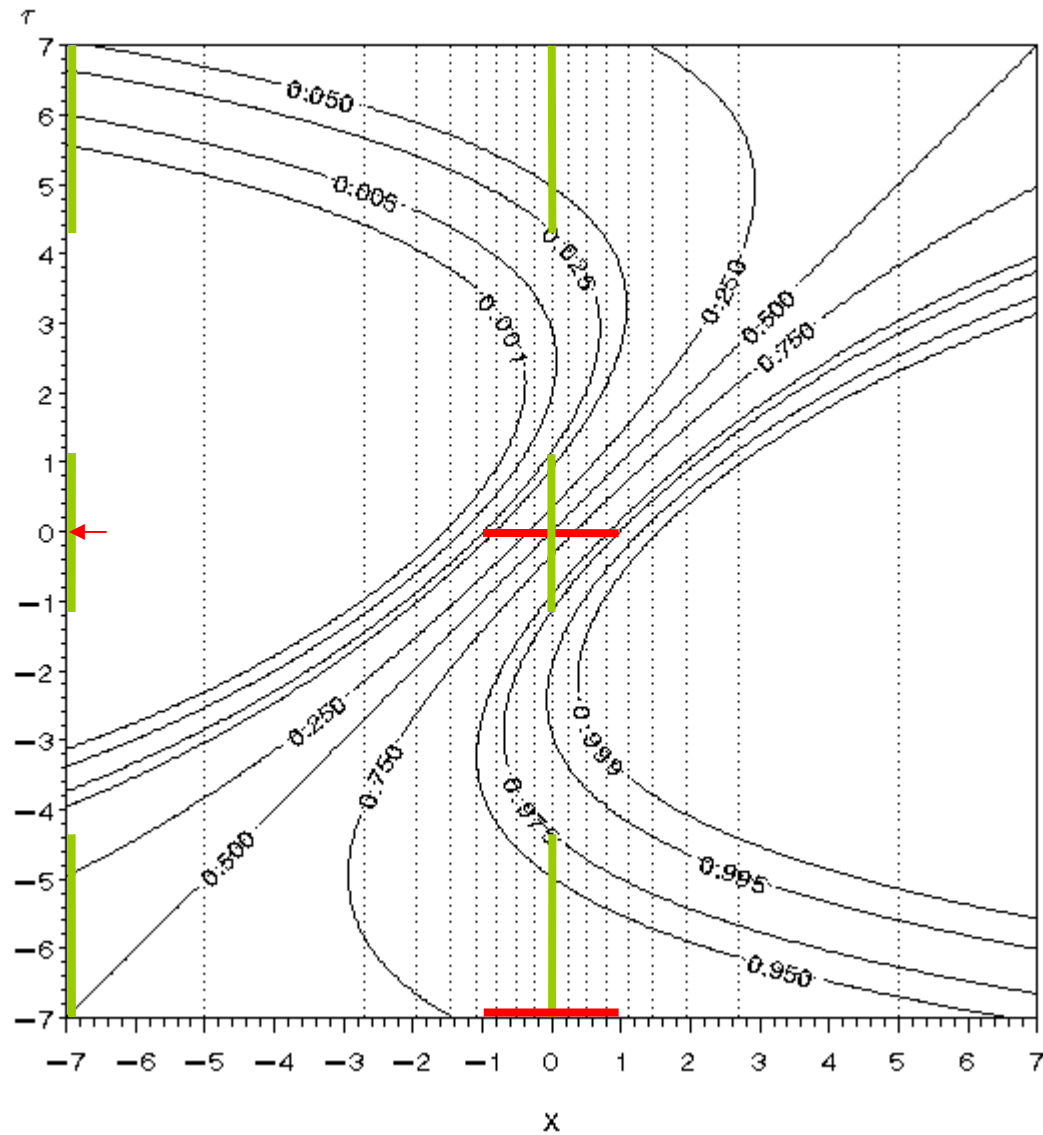
$P(\alpha)=1$ -coverage



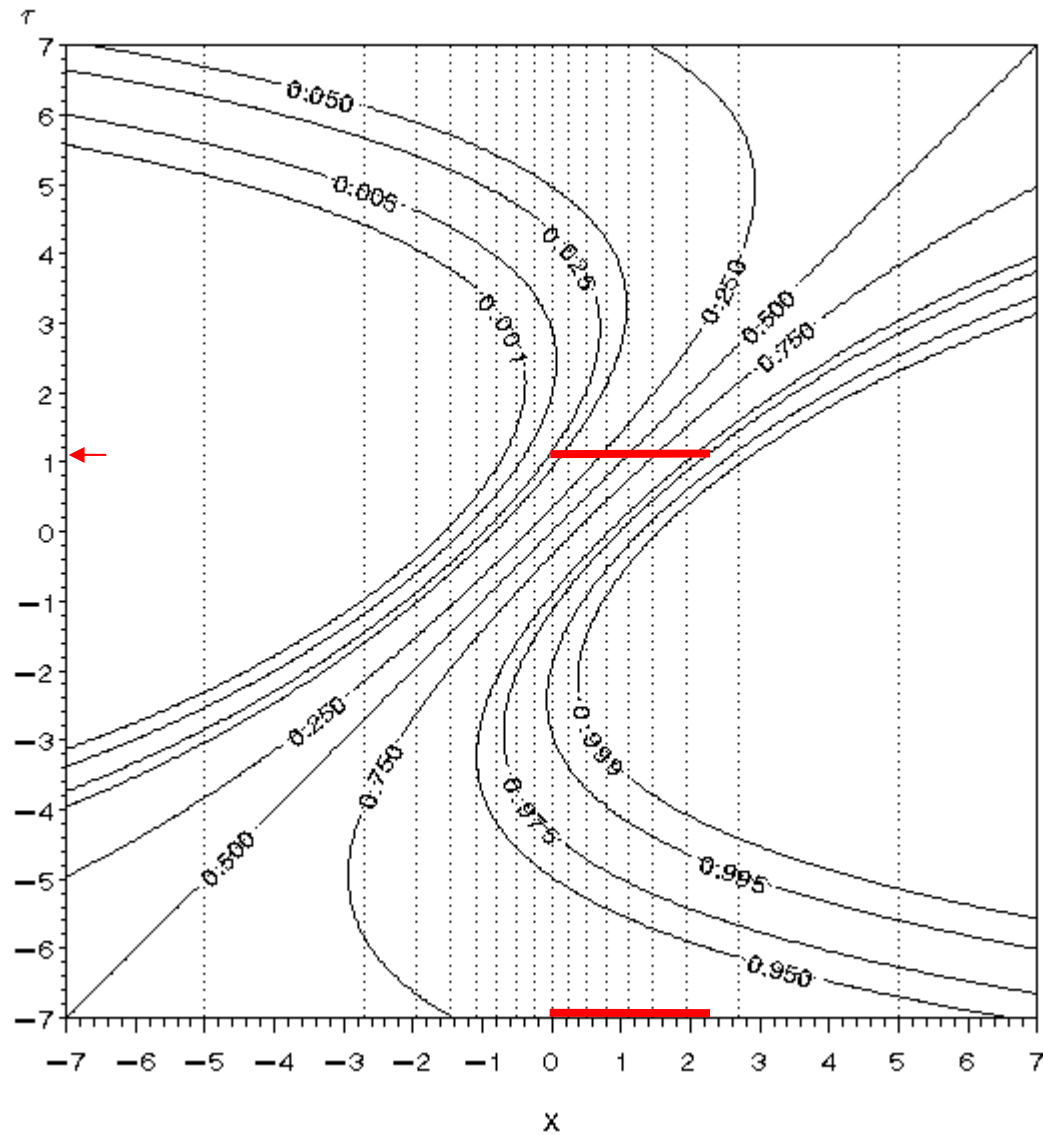
Actual error rates of Wald confidence intervals (distributional assumptions are met)

$P(\alpha)=1$ -coverage

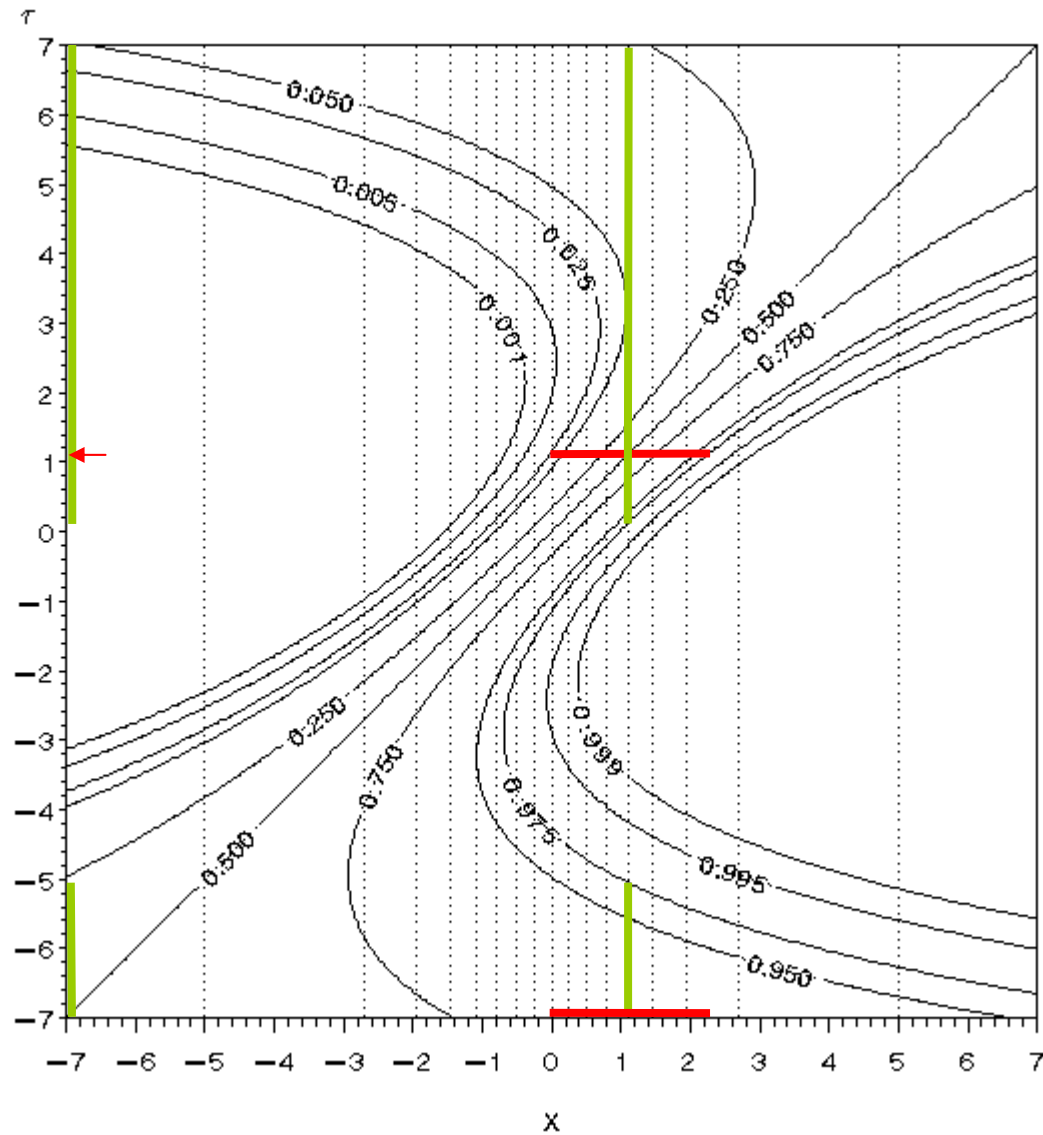




Asymptotic quantiles of the conditional distribution of (the person parameter estimate) X for a Rasch-homogenous item pool of 16 items with identical item parameters ($= 0$). The dotted lines indicate possible values of X .

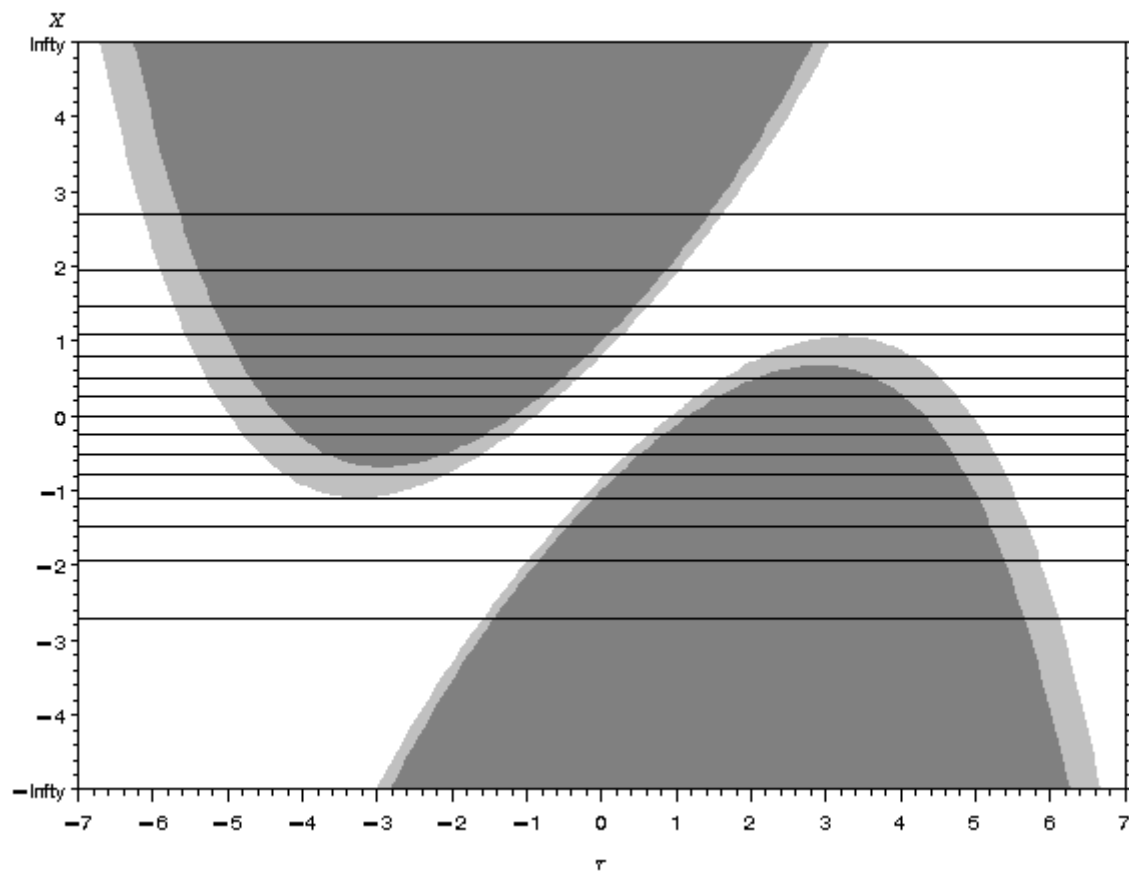


Asymptotic quantiles of the conditional distribution of (the person parameter estimate) X for a Rasch-homogenous item pool of 16 items with identical item parameters ($= 0$). The dotted lines indicate possible values of X .



Asymptotic quantiles of the conditional distribution of (the person parameter estimate) X for a Rasch-homogenous item pool of 16 items with identical item parameters ($= 0$). The dotted lines indicate possible values of X .

Asymptotic 90% and 95% confidence intervals based on the Fisher information and asymptotic normality





How to calculate confidence intervals for individual assessment?

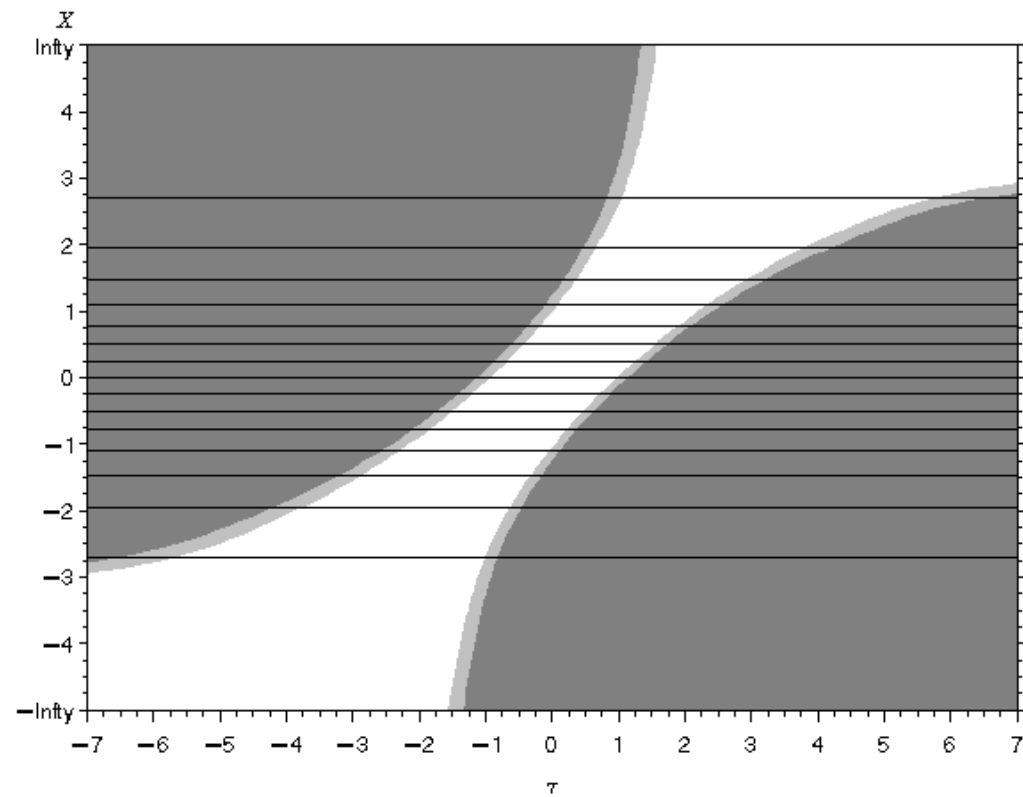
- Quick and dirty: point estimate and its standard error
- Correct:
Include all values in the confidence interval that would not lead to a significant deviation from the data, if they were taken as null hypothesis
 - Asymptotic confidence sets based on Fisher information and asymptotic normality
 - usually consist of two or three nonoverlapping intervals (that include $-\infty$ and ∞)
 - artefact of using finite tests
 - just take the interval that contains the point estimate



How to calculate confidence intervals for individual assessment?

- Quick and dirty: point estimate and its standard error
- Correct:
Include all values in the confidence interval that would not lead to a significant deviation from the data, if they were taken as null hypothesis
 - Asymptotic confidence sets based on Fisher information and asymptotic normality
 - usually consist of two or three nonoverlapping intervals (that include $-\infty$ and ∞)
 - artefact of using finite tests
 - just take the interval that contains the point estimate
 - Conservative confidence intervals based on the exact (usually discrete) distribution of a statistic

Conservative confidence intervals based on the exact discrete conditional distribution of X



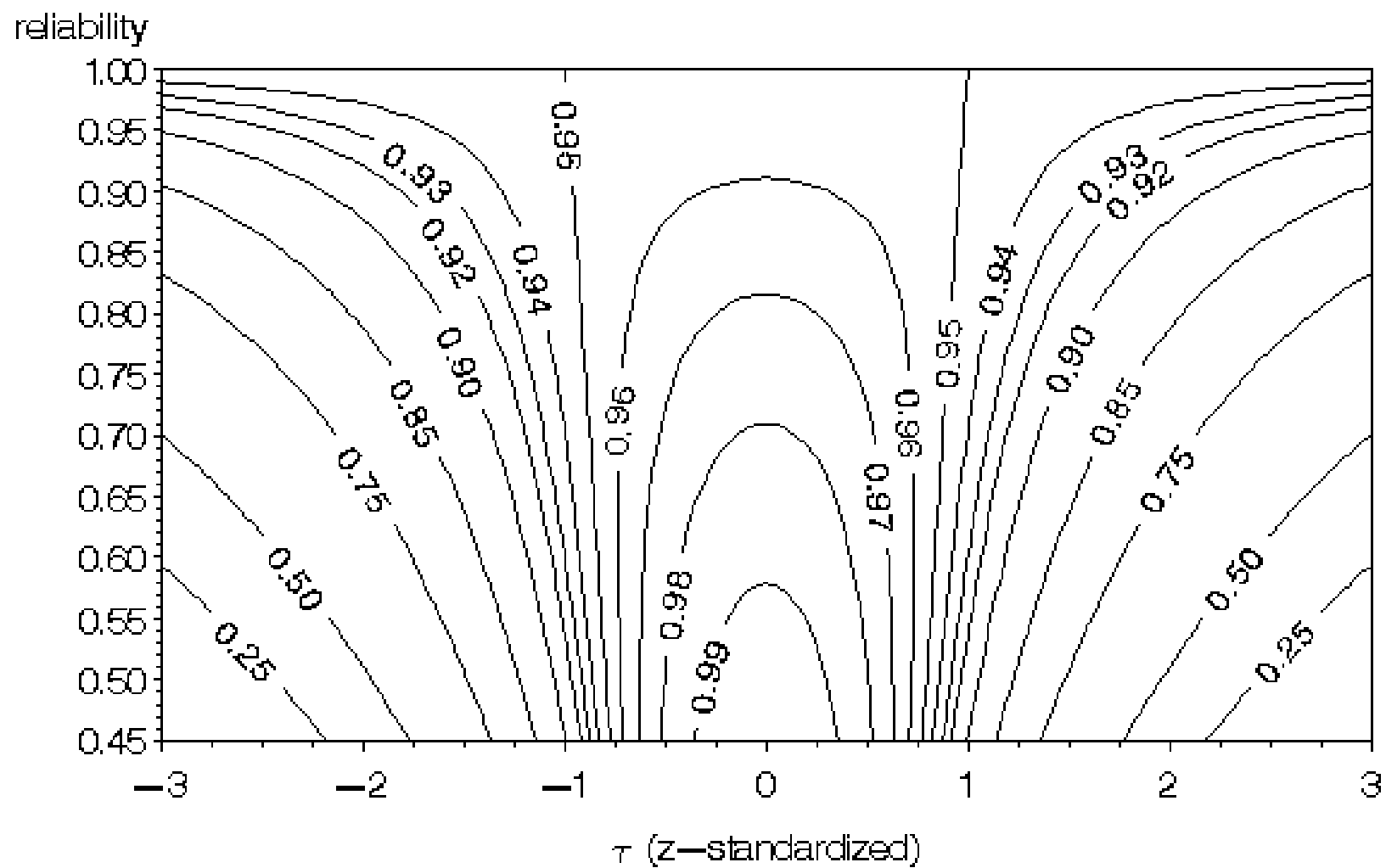
How to calculate confidence intervals for individual assessment?

- Quick and dirty: point estimate and its standard error
- Correct:
Include all values in the confidence interval that would not lead to a significant deviation from the data, if they were taken as null hypothesis
 - Asymptotic confidence sets based on Fisher information and asymptotic normality
 - usually consist of two or three nonoverlapping intervals
 - artefact of using finite tests
 - just take the interval that contains the point estimate
 - Conservative confidence intervals based on the exact (usually discrete) distribution of a statistic
 - Exact confidence intervals based on the randomized continuous distribution of a statistic
 - uniformly most accurate c.i. for Rasch Models (Klauer, 1991)
 - c.i. that balance the probability of over- and underestimation could be developed for any IRT-Model with monotonically increasing ICCs, thresholds or TCC

What about Bayesian credible intervals (regression confidence intervals)?

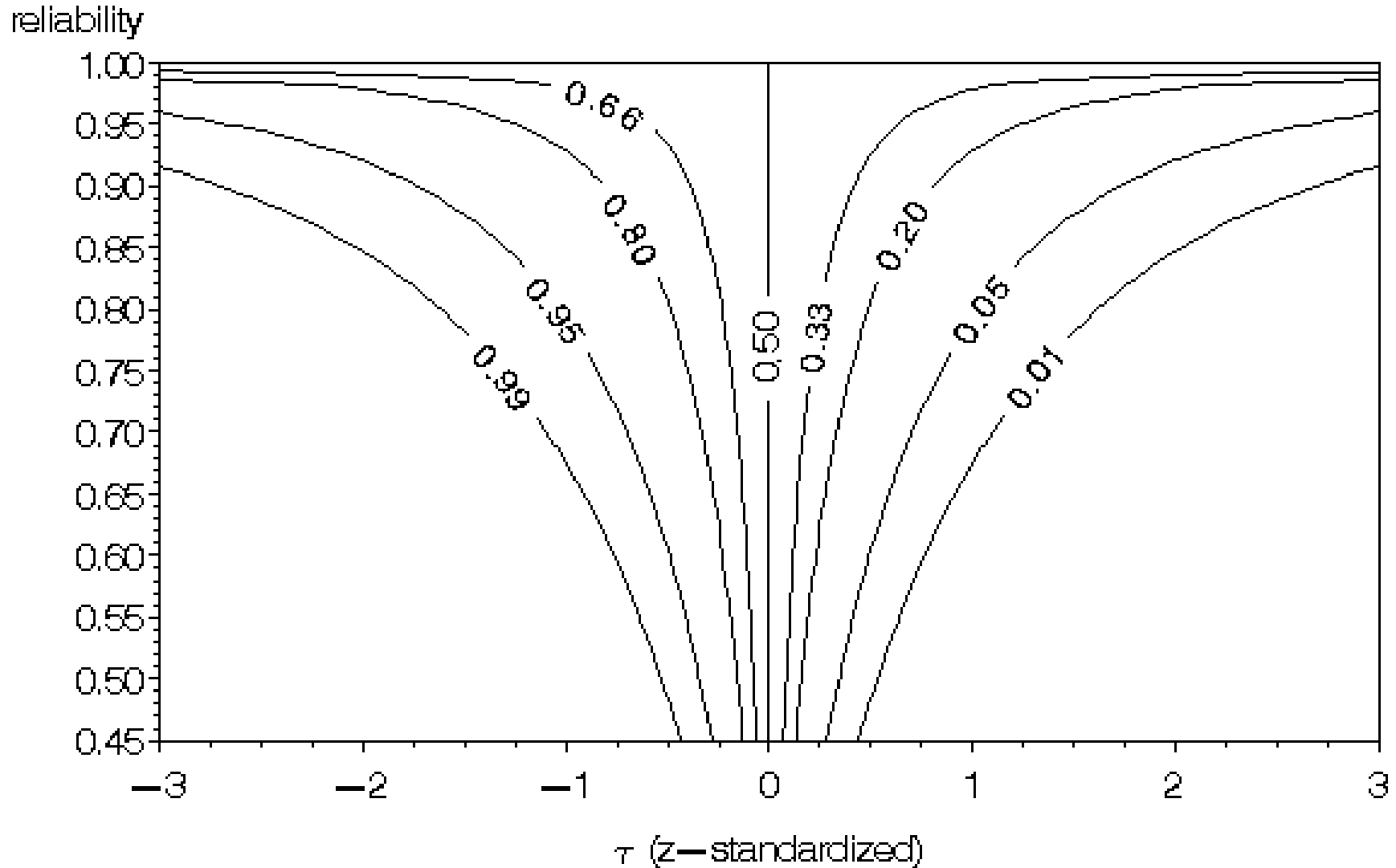
- Not suited for individual assessment (for the purpose of counselling or expertise)
 - Do not Allow for a probabilistic statement that is true for the given individual!
 - How should background variables be dealt with?
- Best choice if utility is to be maximized over many descisions about many individuals (e.g. personel selection of a company or the military).
- Bayesian credible interval are smaller, but they have a greater tendency to overlap (i.e. the regression towards the mean outweighs the lower spread)

Coverage of 95% credible intervals in case of bivariate normality of X and τ



Conditional probability of overestimation for 95% credible intervals
that do not include τ (bivariate normality of X and τ assumed)

$$P(\tau < \text{LCL} | \tau \notin [\text{LCL}, \text{UCL}])$$



Should norm-referenced scores refer to the latent proficiency distribution?

- Yes, because we want to estimate the position of the individual in the very same distribution.
- Relating the estimate of the latent proficiency (i.e. the observed scores) to the observed score distribution does not allow for a statistically sound statement about how proficient somebody is in comparison to the targeted population!
- ▶ However, using norm-referenced scale scores that relate to the latent proficiency distribution leads to a peculiarity:
 - We tell more than 5% of the testees they are (estimated to be) better than 95% of the respective population.
 - We tell more than 5% of the testees they are (estimated to be) worse than 95% of the respective population.

Conclusion: How should test scores be reported?

- Refer to the latent distribution
- Use interval estimation
- Bayesian credible intervals only if overall utility is to be maximised (across persons)
- Wald confidence intervals only in case of a flat information function
- Asymptotic confidence sets
 - Not for short scales
 - Only the central interval of the confidence set
 - C.i. should contain Infinity if there is no variance in the response pattern
- Exact confidence intervals (randomisation) if it is defensible that examinees with same response get different results
- Otherwise: Conservative confidence intervals

Thank you very much for your attention!

For questions, comments or a manuscript:

safir.yousfi@arbeitsagentur.de