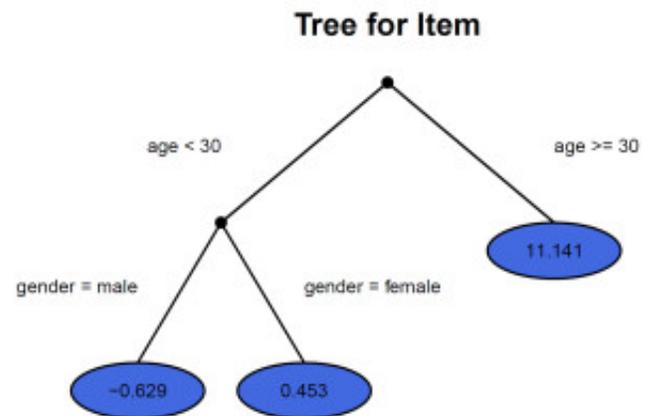# Test items may be unfair

Intelligence tests and other achievement tests always aim at measuring an ability that is not directly observable. What one can observe is how a person performs when trying to solve a collection of items (master a collection of tasks). From the pattern of solved items one can infer on the underlying ability, for example, the intelligence of a person. In modern item response theory the ability is considered a latent trait that determines the performance in tests but is not to be confused with the solution of items. The latter are just observations and therefore indicators but not the latent trait itself.

One problem with the measurement of latent abilities is that tests may be unfair, that is, they may favor specific groups. If the difficulty of an item depends on the membership of the person who tries to solve it to a racial, ethnic or gender subgroup one speaks of *Differential Item Functioning*. More concrete, *Differential Item Functioning* is present if the probability of a correct response among equally able persons depends not only on the intelligence but on the membership to a subgroup. So far, *Differential Item Functioning* has been considered with reference to subgroups only. More generally, the difficulty can also depend on a continuous indicator as age, and it is not known which thresholds are relevant. If the difficulty of an item depends on age a change in difficulty can occur at 30 years of age but also at 40 or a different threshold. A way to determine simultaneously which items from a collection of items show *Differential Item Functioning* and which subgroups are relevant are tree-based methods. Trees have the advantage that the obtained results are easily accessible through the visualization as a tree structure. Figure 1 shows a specific item from an intelligence test that shows *Differential Item Functioning* with respect to two indicators. The strongest difference is found for age. The item is much easier (left branch) for persons younger than 30 years of age and harder for persons older than 30 years (right branch, high difficulty 11.137). For younger individuals there is an additional gender specific difference. The task is easier for males than for females. For males the difficulty is very low (-0.626), for males much higher (0.456).

When developing and using tests that measure latent abilities one should be aware of the

phenomenon of *Differential Item Functioning*. Ideally tests should not contain suspicious items. If this cannot be obtained one should at least know which problem items are in the set and possibly correct for their effects.

Gerhard Tutz
Ludwig-Maximilians-Universität, Munich, Germany

Publication

[Item-focussed Trees for the Identification of Items in Differential Item Functioning.](http://atlasofscience.org)
Tutz G, Berger M
 Psychometrika. 2015 Nov 23