Practical Advances in Invariance Testing: Effect Sizes in Invariance Testing

Johannes Karl & Ronald Fischer

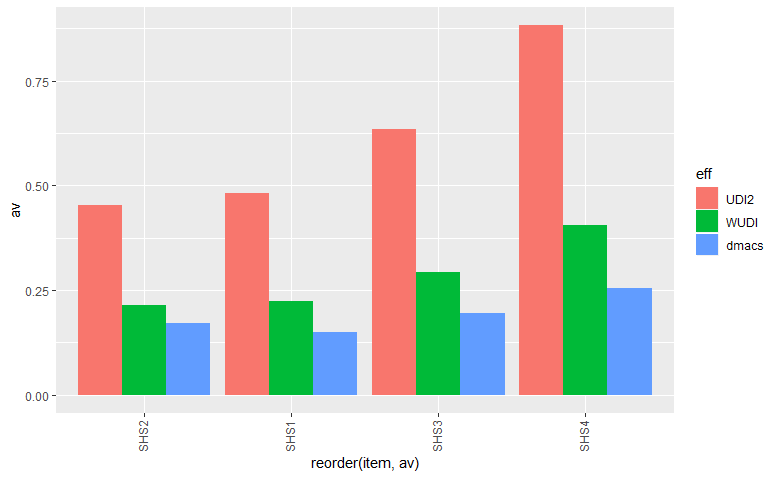
Traditionally invariance testing was concerned with the question whether measures can be compared across groups more as a necessity to enable further analysis rather than a research topic in itself (Fischer & Karl, 2019). Recent advancements in quantifying the degrees of invariance of items between groups open up interesting new research avenues about the potential sources of invariance on an item level. One such advancement is the introduction of effect sizes that quantify the degree of invariance of items between groups (Gunn, Grimm, & Edwards, 2019; Nye & Drasgow, 2011). These effect sizes do not only allow researchers to get a finer grained perspective on metric invariance in their data, but can themselves be targets of research. Similar to other studies that used invariance on a structural level to determine sources of non-invariance across cultures.

One barrier to a wider spread adoption of these effect sizes is the limited implementation in commonly used statistical programs (Gunn et al., 2019). In our current article we aim to address this by providing an applied example of an implementation of these effect sizes in the R language (R Core Team, 2018). We further developed a number of functions that address current shortcomings in the use of effect sizes for invariance testing, such as the absence of confidence intervals and the lack of multi-group applications. In the current article we want to provide the reader with a detailed description how to obtain effect sizes for their data and interpret the output. We also aim to provide suggestions for potential lines of research using these methods.

**Effect Sizes in Invariance Testing**

**Practical Example**

In the following section we want to show a practical example how the investigation of effect sizes can yield new insights into the cross-cultural relationship of measures. First we examined the effect of item wording in a multi-national data set examining happiness and a multi-national data set examining the cross-cultural appropriateness of a mindfulness scale. In both cases, we found that the negatively worded items consistently produced larger average non-invariance effect sizes that positively worded items. This indicates that even after translation, item wording effects remain that lead to substantial differences in responding.



**This was not only true for the effect size itself, but also for the standard deviation.**

