**The TOST As A Method Of Similarity Testing In Linguistics**

*computational and experimental approaches, methodology, equivalence testing, data simulation*

**INTRODUCTION** Classical analyses typically test for differences and their null hypotheses state that the compared samples come from the same population. If negative, the outcome is *insufficient evidence* to assume a difference between the samples; which is not, though, sufficient to assume *equivalence* (Altman and Bland, 1995), or similarity for that matter. Linguistics heavily relies on classical tests (e.g. all 16 experimental talks at the LSA 2013 used classical tests). However, they are insufficient for many linguistic questions. Consider RQ<sub>1,3</sub> (p.2). Negative results for RQ<sub>1,3</sub> would probably go unreported. This disincentivises such research (Bakker, van Dijk, and Wikkerts, 2012) and the field might miss out. An similarity test would be more suitable.

**THE TOST** The TOST, attributed to Westlake (1976), is one of the most common similarity tests (Richter and Richter, 2002). It performs two one-sided t-tests and the null hypotheses are (H₀₁): the difference in means of the two samples is bigger than a pre-set boundary δ and (H₀₂): the difference is smaller than −δ.

H₀₁: \( \mu₁ - \mu₂ > \delta \) \hspace{1cm} H₀₂: \( \mu₁ - \mu₂ < -\delta \)

A positive outcome (rejecting both nulls) denotes *similarity within the range* δ. The researcher sets δ based on her knowledge of previous research. However, this leaves room for subjectiveness (Clark, 2009). Hence, our goal is to find an objective way to set δ.

**DATA SIMULATION** The “right” δ-value is the value that gives a positive test outcome (indicating similarity) with statistical power at 1−α = 95% and 1−β = 80%. To observe how the desired δ-values behave for different data, we simulate a “two-samples-one-population” setting for various datasets (24 in total; p.2) over various Ns (3 to 50000). In the simulations, we “TOSTed” random pairs of subsets from a dataset, over and over again. In total, we simulated ~2.1×10¹² data points.

**PREDICTING AND VALIDATING δ** We found a relationship between observed δ (δ<sub>obs</sub>; from our simulations) and the subsets’ pooled standard deviation (s<sub>p</sub>). This relationship is near-constant for N<sub>p</sub> (pooled from each pair of subsamples) and we call its quotient τ (the *Tübingen Quotient*; τ comes from δ<sub>obs</sub> thus τ<sub>obs</sub>; see f₁).

\[
\begin{align*}
  f₁: & \quad \tau_{\text{obs}} = s_p ÷ \delta_{\text{obs}} \\
  f₂: & \quad \tau_{\text{pred}} = (\sqrt{N_p}) ÷ 4.581 \\
  f₃: & \quad \delta_{\text{pred}} = s_p ÷ \tau_{\text{pred}}
\end{align*}
\]

Fig. 1 shows τ<sub>obs</sub> over increasing Nₛ<sub>p</sub>. Curve-fitting τ<sub>obs</sub> led to f₂, which predicts τ (τ<sub>pred</sub>). f₁ and the 4.581 are our critical findings, because: *by reversing f₁ to f₃ f₂ can be used to objectively set δ (δ<sub>pred</sub>).* In a validation phase, we then compared τ<sub>obs</sub> to τ<sub>pred</sub>. For large parts, they match within ±0.1% (Fig. 2). Further simulations indicate that our results also apply to non-linguistic data.

**CONCLUSION** In our view, the TOST similarity test is a useful tool in a linguist’s repertoire, allowing to investigate research questions that ask for similarity. So far, the lack of instructions to objectively set δ might have been a barrier to use this test. The present work outlined such guidelines and we hope that they will help boost similarity testing in linguistics.

*word count (body): 489*
Additional Materials

RQ\textsubscript{1,3}
RQ\textsubscript{1}: Can highly experienced L2 learners attain a native-like level of language production?
RQ\textsubscript{2}: At which age do teenagers typically reach adult-like reading times?
RQ\textsubscript{3}: Are resumptive pronouns perceived as equally bad across modalities?

THE DATASETS
Source: authors or colleagues (all 24 datasets). Areas: syntax (13), phonetics (8), psycho-linguistics (3). Units: Likert-Scale data (13), normalised Likert-Scale data (4), Hz (4), ms (3). Aggregation: aggregated (18), non-aggregated (6). Size of Datasets: 42 to 152, mean = 85.79.

REFERENCES