

A Teacher-friendly Quality Checklist for L2 Quantitative Reports

Joseph P. Vitta and Christopher Nicklin

Sample Design				
<i>Area</i>	<i>Positive Check</i>	<i>Positive Check</i>	<i>Positive Check</i>	<i>Negative Check</i>
1. Sample Size Planning	<i>Search the document for 'power' where an effect has been referenced and a suitable sample size has been determined and then fulfilled (a priori power)</i>	<i>Read the participants section of the methods and see if an earlier study has had its sample size referenced as a benchmark to meet (heuristics).</i>	The research admits <i>in methods</i> that they could only have access to x number of participants and they use this number to determine the minimal effect size that they can detect (sensitivity analysis)	Sample size is presented without any overtly described planning process.
2. Selecting Participants by Chance (Randomized Sampling)	Participants are randomly chosen from the population or in a probabilistic manner (very rare in our field, e.g., Hiver & Al-Hoorie, 2020). <i>Search for 'random' / 'probability' + 'sampling'</i>	Researchers <i>in methods</i> have selected multiple sites from the intended population.	<i>Intentionally blank</i>	Single-Site Sample
3. If experimental design, conditions and groups are constructed by chance (Random assignment)	<i>Search document for 'random' / 'randomly' / 'random assignment' / randomly assigned</i> It should be clearly stated that participants were assigned to different groups by chance	Such assignment is also acceptable at the class level (Vitta & Al-Hoorie, in press)	Researcher built groups to maintain equality in relation to proficiency or the language outcome <i>(purposeful assignment but less than ideal; see Fisher, 1935)</i>	No such overt description of random assignment and/or purposeful assignment

Measurements & Testing			
<i>Area</i>	<i>Positive Check</i>	<i>Positive Check</i>	<i>Negative Check</i>
Validity of Measurements	The researcher argues for the appropriateness of measurements or items/tests employed. And/Or Research can present a 'scale validation' process	The researcher references past studies' use of target words/items and/or tests (or similar ones).	No such validity account
Reliability of Measurements	Search methods for 'reliability' look for Greek letters such as α (Cronbach's alpha), ω (Omega), κ (Cohen's kappa). If you see these, the researcher has reported reliability	The researcher presents an argument for the consistency of their measurements (less ideal).	No such reliability account
Evidence of Inferential Testing for Generalizations (Research Question / Hypothesis focus)	a. exact p values and/or BF10 values (even Bayesian probability) b. Look for the reporting of test statistics such as F , t , etc. c. 'significant'	p values approximated ($p < .05$) Test statistics are abridged (less ideal)	Generalizations from descriptive statistics only (major cause for concern)
Overt Reporting of Effect Sizes	a. overt labelling of effect size metrics: ex: d , g , r , β , η_p^2 , η^2 b. qualification of the effect size (e.g., a large association [Plonsky & Oswald, 2014] was observed ($r = .7$. $p < .001$) c. nonsignificant effect sizes are reported d. an overall feeling of complete reporting (e.g., M , SD , etc.	<i>Intentionally blank</i>	No effect sizes are presented; only p values