Journal of Communication Disorders 69 (2017) 44–57

Contents lists available at ScienceDirect

Journal of Communication Disorders

journal homepage: www.elsevier.com/locate/jcomdis

The paired *t* test and beyond: Recommendations for testing the central tendencies of two paired samples in research on speech, language and hearing pathology

Toni Rietveld[*], Roeland van Hout

Centre for Language Studies, Radboud University, Erasmusplein 1, 6525 HT Nijmegen, The Netherlands

A R T I C L E I N F O

Keywords:
Tests for paired samples
Assumptions for nonparametric statistics
Recent developments in the analysis of data obtained in paired samples designs
Statistical interaction and correlation in paired samples designs

A B S T R A C T

Purpose: In this tutorial we review current practice in the analysis of data obtained in designs involving two dependent samples and evaluate two conventional statistics: the *t test for paired samples* and its non-parametric alternative, the *Wilcoxon Signed Ranks test* (WSR). It is a sequel to our tutorial on the analysis of designs with two independent samples on the basis of non-count data (Rietveld & van Hout, 2015). The frequency with which these statistics are used is assessed on the basis of publications on disordered communication in *Clinical Linguistics & Phonetics, Journal of Communication Disorders and Journal of Speech, Language and Hearing Research* for the time interval 2006–2015. We conclude with a number of recommendations for the analysis and presentation of data.

Conclusions: Researchers should more consistently present the relevant characteristics of their data (means, medians, SD, skewness, tailedness, outliers etc.) and explicitly consider the assumptions that apply to their statistical methods, such as correlations between data obtained on two occasions, interactions between participants and treatment, and the symmetry of difference scores, many of which are hardly ever reported or even tested. Two recommendations are particularly relevant. First, the WSR is not a proper test for central tendencies as a replacement of the conventional *t* test for paired samples whenever assumptions about the dependent variable are in doubt. Second, researchers should choose statistical procedures on the basis of the null hypothesis (H0) to be tested and not primarily on the basis of the type of data (ordinal or interval). Two relevant H0's in the field of speech-language pathology are: (1) $\mu_1 = \mu_2$ (the mean obtained in condition 1 is equal to the mean in condition 2) and (2) $p = 0.5$, which says: the probability to obtain (for instance) higher scores in condition 2 than in condition 1 is 0.5. We recommend the permuted *t* test for paired samples to test the first H0 and the permuted Brunner-Munzel rank test to test the second.

1. Introduction

When two sets of non-count data are obtained in a design with two related, matched or dependent samples (the three terms are used interchangeably) many researchers use a *t test for paired samples* (*Tp*). However, quite often a *non-parametric* alternative is chosen, such as the well-known *Wilcoxon Signed Ranks test* (*WSR*), which is also known as *Wilcoxon Matched Pairs, Signed Rank(s) test.* A procedure is considered non-parametric if it is used for sets of data without regard to the shape of the distribution (expressed in

[*] Corresponding author.
E-mail address: A.Rietveld@let.ru.nl (T. Rietveld).

http://dx.doi.org/10.1016/j.jcomdis.2017.07.002
Received 8 December 2016; Received in revised form 19 July 2017; Accepted 23 July 2017
Available online 25 July 2017

terms of parameters like normality, variance, skewness, etc.). The term 'non-parametric' suggests that no assumptions are made on the characteristics (parameters) of the distributions from which the samples are taken (Siegel & Castellan, 1988). We have to emphasize that the term non-parametric does not imply 'free of assumptions', as we will see later. In our tutorial, we will discuss the balance between the parametric *t* test, its variants and the non-parametric alternatives in research on speech-language pathology.

Paired samples can appear in different designs:

a) The same participants are measured at two points in time, for instance before and after treatment; this design is also called a repeated measures design in the context of Analysis of Variance (ANOVA) or a model with one fixed (the two points of measurement) and one random effect (the participants) in linear mixed models.
b) The participants measured at different times are matched; they are pairwise similar in aspects considered relevant for the investigation, for instance IQ or age.
c) The participants are 'naturally matched', for instance twins, or couples where responses of the male partner are compared with those of the female counterpart.

We will concentrate on the design mentioned under a). Participants in research on speech and language disorders are very often measured at two successive moments in time (with equal intervals) in order to assess the success of a specific treatment; to simplify we will call these occasions *T1* and *T2*. When time is involved in a treatment design, it is tempting to interpret the changes observed in a causal way (for example, most people recover from a cold after they take cold medication). Understanding the design of paired samples more precisely helps to avoid a post hoc fallacy (e.g., most people recover from a cold after a couple of days). We discuss two phenomena that require more attention in paired samples designs: the correlation between moments in time and the interaction between participants and the variable time.

To assess current statistical practice in the domain of speech-language pathology, we counted the frequency of parametric and non-parametric statistics in designs with paired samples in three representative journals from 2006 to 2015 (see also Rietveld & van Hout, 2015). We restricted our counts to straightforward examples of comparing two means or medians, excluding post-hoc comparisons. Two criteria were used to choose these journals: a) They should belong to the first 15 of a ranked list of journals in Audiology and Speech-Language Pathology, based on a 5-year Impact factor, with ranks evenly distributed over this list, and b) the topics covered by these journals should be varied and comprehensive, covering large parts of the field rather than specific topics such as fluency or aphasia. Three journals, listed in Table 1 along with the outcomes of our review, met these selection criteria. We searched for all tests available for paired samples designs and only found instances of the conventional *t test for paired samples,* the *Wilcoxon Signed Rank Test* and the *Sign Test.* In three cases bootstrapping (see Section 6) was used.

Overall, the percentage of use of parametric tests outnumbers those of non-parametric tests. Non-parametric tests occur more frequently in *Clinical Linguistics & Phonetics* than in the other two journals. We found a similar percentage of parametric tests in the first five years (2006–2010) compared to the last five years of our time window: 65.9% (123/183) versus 72.6% (159/219). There is no significant trend towards more parametric or non-parametric testing ($\chi^2(1, N = 402) = 1.383, p > 0.05$). We counted the arguments given for the 60 non-parametric tests used in the three journals over the last five years (2011–2015). In 19 cases the arguments that were given were: scale type (4), non-normality of the scores (5), small sample size (6), unequal variances (1), the presence of outliers (1), presence of 'null scores' (1) and 'not all criteria met for a parametric test' (1).

On the basis of our review, current practice and many 'conventional' manuals, we prudently presume that most researchers use, consciously or not, a straightforward two-step decision procedure: (1) When there is reason to assume that the data are not interval, apply *WSR*, otherwise apply *Tp*; (2) When the difference scores of interval data are not normally distributed, apply *WSR*, otherwise apply *Tp.* In current practice any further explication or more formal testing is hardly ever seen.

In this tutorial, a summary of the characteristics of a number of frequently used statistics is given and a number of questions are discussed which are relevant for the characterization of data used in paired samples designs. We will give reasons for why two specific statistical procedures should be used for testing hypotheses which are often very relevant in speech and language research. Researchers should choose statistical procedures as a function of the null hypothesis (H0) to be tested and not primarily by the type of data (ordinal or interval), as we will argue below. In Section 2, we will present possible hypotheses on the basis of designs with paired samples, in Section 3, details of the three conventional tests for paired samples (*Tp*, *WSR* and *ST*) are given and in Section 4, we address the relevance and use of tests to assess assumptions underlying tests for two paired samples. In Section 5, special problems occurring in designs with paired samples are addressed; we also give the results of a number of preliminary tests carried out on a small artificial dataset. In Section 6, we discuss new developments, including randomization and bootstrapping, that have become available in the last decade. Finally, in Section 7 we give our conclusions, including our recommendation on how to approach the

Table 1
Numbers of articles in which the parametric *t*-test for paired samples and/or its conventional non-parametric alternatives were applied on data obtained in paired samples designs, in three journals over the period 2006–2015.

Journal	*t test for paired samples (Tp)*	*Wilcoxon Signed Rank Test (WSR)*	*Sign Test (ST)*	*% parametric*
Clinical Linguistics & Phonetics	53, including 1 with bootstrapping	43	4	53.0%
Journal of Communication Disorders	51	15	1	76.1%
Journal of Speech, Language, and Hearing Research	168, including 2 with bootstrapping	54	3	74.7%