Journal of Open Humanities Data

A Comparison of Statistical Tests for Likert-Type Data: The Case of Swearwords

RESEARCH PAPER

ROALD EISELEN
GERHARD B. VAN HUYSSTEEN
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

There has been a long-standing debate about the applicability of different statistical tests to Likert-type data. This work contributes to this debate by comparing the results of three statistics, Analysis of Variance, Kruskal-Wallis *H* test, and ordinal logistic regression, in the context of self-reported attitude and usage of swearwords. The aims of this comparison are to determine how similar the results of the different statistics are, which of the statistical test are most appropriate for sociopragmatic linguistic investigations, and how the results influence the interpretation possibilities of the same data. An analysis of the results for the different statistics shows some overlap between the three statistics, but that the parametric ANOVA is substantially more conservative in identifying significant relationships between sociodemographic factors and the usage and attitude toward swearwords, while the *H* test and ordinal logistic regression models are mostly identical, except where the assumptions of the regression model are violated. Based on our results, we argue that the Kruskal-Wallis *H* test, with the associated post-hoc test, is the most appropriate test for Likert-type data, with sufficient descriptive power to allow for detailed and informed analysis of this data.

CORRESPONDING AUTHOR:
Roald Eiselen
Centre for Text Technology (CTexT), North-West University, Potchefstroom, South Africa
Roald.Eiselen@nwu.ac.za

KEYWORDS:
Likert-type data; Swearwords; Afrikaans; Statistical comparisons

TO CITE THIS ARTICLE:
Eiselen, R., & van Huyssteen, G. B. (2023). A Comparison of Statistical Tests for Likert-Type Data: The Case of Swearwords. *Journal of Open Humanities Data*, 9: 18, pp. 1–13. DOI: https://doi.org/10.5334/johd.132

DOI: 10.5334/johd.132

(1) INTRODUCTION

Despite the substantial increase in the use of advanced and sophisticated statistical methods in linguistic research over the last decade (Gries, 2015; Larsson et al., 2022), one of the more controversial areas of statistical analysis (in general and in linguistics) is the analysis of Likert-type data. There has been a longstanding disagreement on which statistical methods are valid for the analysis of this type of data. The main point of disagreement is whether Likert-type data can be assumed to be interval data or not, i.e., whether the different points on the scale can be considered equidistant, or whether they should be considered ordinal data. This distinction is important since parametric statistics, like analysis of variance and mean calculation, are only applicable to continuous or interval data, while ordinal data requires the use of non-parametric statistics, such as the Kruskal -Wallis *H* test and calculation of the median. The selection of the type of statistic is not only integral to how the results are analysed, but also to the validity of the conclusions.

We contribute to this debate by comparing the results of three statistical tests to assess the degree of agreement between the different tests. Of special interest to us is the appropriateness and interpretability of the respective tests in a typical sociopragmatic investigation (Culpeper, 2021; Reiter & Placencia, 2005). We therefore use data that was collected as part of a project that investigates various sociopragmatic aspects of swearwords (see section (3)).

Our analysis (section (4)) comprises a comparison of Analysis of Variance (ANOVA), Kruskal-Wallis *H* test (KW-H), and ordinal logistic regression (OLR) to ascertain how the results compare across various response and predictor variables. The aim is to answer, through a discussion of the results (section (5)), the following questions:

1. Do the different statistical tests produce comparable results?
2. Do different tests allow for different interpretations of the same data?
3. Which statistical test is most appropriate and robust for interpreting Likert-type data in sociopragmatic linguistic investigations?

(2) CONTEXTUALISATION AND RELATED WORK

The use of Likert and Likert-type items to measure and quantify typically qualitative data is widely used not only in sociopragmatic research (Reiter & Placencia, 2005), but also in other areas in the Humanities and Social Sciences (Dubois, 2013; Larsson et al., 2022). Likert (1932) originally developed these Likert items such that a single qualitative attribute is measured multiple times and then averaged across the related questions for use with parametric statistical tests to remain valid. The use of individual Likert-type items has become increasingly widespread, and in turn caused longstanding disagreement between researchers on what the appropriate set of statistical tests for these Likert-type data should be (Boone & Boone, 2012; Endresen & Janda, 2017; Harpe, 2015; Jamieson, 2004; Sullivan & Artino Jr, 2013). The main point of contention is about the nature of the data: Is Likert-type data **interval data**, which allows for parametric and basic descriptive statistics (such as means and ANOVA) to be used in the analysis of the data; or is it **ordinal data** that should be analysed using those statistical tests specifically designed for ordinal data, which typically have substantially more complex assumptions that should be adhered to when performing statistical analysis.

The main argument for the use of parametric statistics is that parametric statistics are robust enough to handle data that is not necessarily normally distributed, and that is not continuous or interval data. Various power analysis investigations (Hsu & Feldt, 1969; Norman, 2010; Sangthong, 2020) have been published where the veracity of this claim has been investigated, and in general confirmed the robustness of such tests on Likert-type data, especially when there is a large enough set of respondents. Van Hecke (2012) however shows that ANOVA is less powerful than KW-H in the case of asymmetrical distributions and is likely to be more conservative in attributing significant effects. Other researchers (e.g., Harpe, 2015) also argue that extending the scale beyond five options (to seven or nine), or that labelling only the extreme edges of the scale with nominal values while intermediary values are labelled numerically, further reduces the discrepancy between interval and ordinal data.

Specifically related to the domain of swearing, several studies have used Likert-type items to determine offensiveness, tabooness, self-perceived frequency of usage, and other related aspects of taboo language (e.g., Beers Fägersten, 2007, 2012; Dewaele, 2004, 2018; Janschewitz, 2008; Jay, 1992). Most of these studies only made use of ANOVA and descriptive statistics, with only one study using KW-H. A summary of some of the most prominent studies, including the kinds of statistical tests they used, is presented in Table 1.

PUBLICATION	TASK	ASPECTS MEASURED	SCALE	# USABLE RESPONDENTS	STATISTICAL TESTS
Beers Fägersten (2007)	Word ratings	Offensiveness	1–10 (Not offensive – Very offensive)	60	Descriptives (mean, mode, SD); One-way ANOVA test for significance
Beers Fägersten (2012)	Multiple choice questions	Frequency of swearing	Never, Rarely, Sometimes, Often	60	Descriptives (percentages)
	Likelihood judgements	Frequency of swearing	0–100 (Not likely at all – Most likely possible); 1–9 (Never heard at all – Heard very frequently)	53; 59	Descriptives (mean)
Jay (1992)	Situation ratings	Offensiveness and tabooness	1–9 (Not offensive/ obscene at all – Most offensive/ obscene word imaginable)	52; 59; 90†	Descriptives (mean, rank-order); ANOVA
Janschewitz (2008)	Word ratings	Personal use; familiarity; offensiveness; tabooness; valence; arousal; imageability	1–9 (Positive/low – Negative/high)	78	One-way ANOVA; 2×5 mixed ANOVA; post hoc *t* test with Bonferroni corrected alphas
Dewaele (2004)	Multiple choice questions	Perception of emotional force	1 = Does not feel strong; 2 = Little; 3 = Fairly; 4 = Strong; 5 = Very strong	1039 + 50††	Multivariate analyses of variance (MANOVA); Scheffe´ post-hoc test; linear regression analysis
Dewaele (2018)	Multiple choice questions; word ratings	Familiarity with meaning; offensiveness; frequency	0–5 (Very low – Very high)	2347	One-sample Kolmogorov-Smirnov test; Mann-Whitney test; Kruskal-Wallis *H* test

Table 1 Summary of statistical tests used in previous studies on swearwords.
Note: † = Three different experiments. †† = Two modes of collection (online and paper-based).

Endresen and Janda (2017) previously compared the results of five statistical models on acceptability judgments with a similar aim of determining which of the models are most appropriate to apply to Likert-type items. In their comparison of the models, they base their conclusions on the application of three predictor variables on the Likert-type acceptability ratings. They conclude that all five models are appropriate to a greater or lesser degree, and that the outcomes of both parametric and non-parametric tests provide comparable results, even though only two of their models significantly correlate the same predictor variables. They also crucially do not include the non-parametric equivalent of the ANOVA, KW-H test, in their experiments.

Although our study has similar aims to their investigation, there are several important differences. Firstly, the scope of their study was relatively limited in terms of the number of test cases based on predictor and outcome variables. This makes it difficult to draw definitive conclusions about the similarity and differences between the results of the different models, and consequently the applicability of the different models for Likert-type data. We address this by conducting a substantially larger set of experiments, 80 hypotheses tested in total, to get a clearer indication of how the results for the different statistics compare. Furthermore, the data in our experiments shows greater variance in the distributions of the outcomes, covering both highly skewed and relatively normal distributions. To simplify the comparisons between different statistics, we only consider a single predictor for each outcome, rather than the multiple predictors used in their study. Lastly, our application is in a different domain, i.e., tabooness ratings, rather than acceptability judgments, although we not expect the domain to have a substantial impact on the applicability of our results.

(2.1) STATISTICAL TESTS

For purposes of this investigation, we will focus on three statistical tests often used in the analysis of Likert-type data, namely ANOVA, KW-H, and OLR. Each of the tests have different underlying assumptions: from the relatively simple assumptions for ANOVA, to the substantially more complex set of assumptions for OLR (for explanations of these assumptions, see most statistics textbooks, e.g., Verma and Abdel-Salam (2019)). Additionally, all these tests allow for potentially differing interpretations of the data and statistic results.

ANOVA is a parametric test used to determine if there are significant differences between two or more categorical groups (levels) by determining whether variance is only due to chance (within-group variation), or variance is due to chance and an effect, the predictor variable, (between-group variation). If the between-group variation is more dominant than the within-group variation, the effect is considered significant. If there are more than two levels, ANOVA does not specify which levels differ significantly; however, those can be calculated with post-hoc tests where the variance between each level is calculated, typically using Tukey's HSD (honestly significant difference) (Tukey, 1977). ANOVA has two primary assumptions related to how the data fits the ANOVA model: (1) the normal distribution assumption, and (2) the homogeneity of variances assumption. As mentioned above, various power analysis studies have shown that ANOVA is robust against violation of the normality assumption if the levels have a sufficiently large set of respondents, typically when there are between two and nine levels with more than 15 respondents in each level.

KW-H (a.k.a. Kruskal-Wallis test by ranks, or one-way ANOVA on ranks (Laerd Statistics, 2018)) is a non-parametric counterpart of ANOVA, which does not make any assumptions about the distribution of the responses. It was originally designed to analyse ordinal, rather than interval or continuous data – especially where more than three levels are being compared.[1] Although the test does not make direct assumptions about the distributions of data, the interpretation of the results does require visual inspection of the distribution of scores for each level of the predictor variable. If their distributions have the same shape, the test can interpret median values; alternatively, only statements about the distributions of each level can be made. As with ANOVA, KW-H does not directly identify which levels differ, but the Dunn post-hoc test (Dunn, 1961), which indicates which levels differ significantly from each other, is typically used to this end. Both Tukey's HSD and the Dunn post-hoc test assess the significance of difference between pairs of group means. Unlike the omnibus tests (ANOVA and KW-H), these post hoc tests can indicate the specific groups where the means differ. It is important to note that although either ANOVA or KW-H may indicate a significant difference in means across all the levels, post-hoc tests may not indicate that there is a significant difference between any two groups.

The final statistical test that will be considered here is OLR, a parametric test that is specifically used when one has an ordinal response variable and one or more predictor variables that are continuous or categorical (Laerd Statistics, 2018). Like KW-H, it also does not require normally distributed data but does, however, require proportional odds between the levels (in addition to the assumption that two or more of the continuous independent variables are not highly correlated with each other – which is not applicable to our data). As a regression model, OLR is designed to make predictions of future behaviour, given a particular predictor variable on a particular response variable. The resultant model can be interpreted as an odds ratio: What are the odds that a member from level A will choose a higher or lower ordinal value than a member from level B? This is done by determining whether a more complex model (i.e., a model that includes the effect) is a better fit for the data than an intercept-only model. OLR can also consider multiple effects to improve the model, though that is not the focus of this study.

(3) DATA CONTEXTUALISATION

To compare and assess these three statistical techniques, specifically to determine which of them is most efficient for interpreting typical sociopragmatic data, we use a small dataset from the project *Swearing in South Africa: Multidisciplinary research on language taboos*

1 It is more typical to use the Mann-Whitney *U* test when two levels are being compared. However, in the case of our data, only one of the sociodemographic variables consist of two levels, and for the sake of simplicity and consistency, we therefor prefer to use KW-H across all variables.

(Van Huyssteen, 2021). Broadly speaking, this project aims to get insight into a range of aspects related to the usage and perception of, and attitudes toward taboo language.

For example, in one of the subprojects, the aim is to empirically obtain offensiveness ratings for as many Afrikaans swearwords as practically possible through Short Word Surveys (SWS). This aim is comparable to the aims of studies done with L1 speakers of English (e.g., Beers Fägersten, 2012; Dewaele, 2015; Ipsos MORI, 2016; Janschewitz, 2008; Jay, 1992), and Dutch (e.g., Van Sterkenburg, 2019), as well as LX speakers of English (e.g., Dewaele, 2016a, 2016b), and other multilinguals (e.g., Dewaele, 2012). Although these kinds of offensiveness ratings are of academic interest to linguists, psychologists, and other researchers, our end goal is also more applied, namely, to develop a publicly accessible, evidence-based, online dataset of swearwords and their offensiveness ratings. This dataset can then be applied in various settings where the use of swearwords is either restricted by law or may have financial impact in a business setting.

This dataset is available (Van Huyssteen & Eiselen, 2023)[2] with a full description of the data and data collection procedure in Van Huyssteen et al. (2023). In short, participants register on the vloek.co.za website and provide sociodemographic information on 21 factors. The participants are then prompted to complete Short Word Surveys (SWS) where each participant rates their perceived attitudes for a particular swearword on a nine-point Likert scale for each of the following attitudinal dimensions: Production frequency, Perception frequency, Offensiveness (self), Tabooness (others), Emotionality, Conspicuousness, Familiarity, and Sex of referent.

In South Africa, it is by law for the task of the Film and Publication Board (FPB) to provide content and age advisories for films, computer games, and certain publications that are released/published in South Africa. One of the criteria relates to "strong language", which is defined as "crude words, threats, abuse, profanity or language that amounts to prejudice" (Republic of South Africa, 2022, p. 8). Such strong language should be categorised as "of a mild, moderate, strong or very strong impact". However, this offensiveness scale is not operationalised (not even by means of examples), and authors, publishers, film makers, and other content creators are therefore left without any practical guidelines.

For example, a publisher may want to include a potentially taboo word in the title of a novel but would like to determine how the FPB might respond to such a title. Instead of doing their own (expensive and time-consuming) market research, they could rather consult said dataset as a dashboard, adjust the variables to fit their target audience (e.g., in terms of age, gender, level of qualification, income group, religious views, etc.), and get results on the perception of the word, and the impact it might have on that specific audience. This could then be presented as arguments in an appeal against a decision of the FPB.

However, for more sophisticated and reliable interpretations using sociodemographic variables, the project needs an appropriate and reliable statistical technique to analyse and report on these ratings in ways that are not only clear and informative, but also robust enough to implement as part of an automated workflow (i.e., an implementation that does not require a human in the loop).

The current study focuses only on four attitudinal dimensions (Production frequency; Perception frequency; Offensiveness; and Tabooness) of a selection of four swearwords (*feeks*; *piele*; *moffie*; *jissis* – see below)[3] using five sociodemographic factors (age group; sex group; religious view; political view; world view). The four words were chosen to represent a variety of morphosyntactic, semantic, and pragmatic aspects (marked in bold below) of swearwords, so that the findings of this article should be applicable to the analysis of a variety of other swearwords as well. The four sources of Tabooness and the four chosen words are:

- **Misogyny:** *feeks* ('harridan') is a rather **old-fashioned** word, used as an **epithet** for a strong-willed woman;
- **Sexuality:** *piele* ('penises'; 'fucking A') is a **noun**, denoting more than one penis; however, in the context asked in the SWS, it is used as an **adjective** or **interjection** to indicate that something is good, fantastic, perfect;

2 https://doi.org/10.25388/nwu.23708229.

3 Since the focus of this article is on the statistics related to swearword ratings, we don't provide detailed descriptions of the etymology, semantics, and pragmatics of these words. Below we provide cursory notes on their meaning and usage.

- **Bigotry:** *moffie* ('faggot') is a **slur** for a gay man, but as was the case for the English word *gay*, *moffie* is in the process of being **reappropriated** by the gay community; and
- **Blasphemy:** *jissis* ('jeezuz') is a **deformation** of the **religious** name *Jesus*, and it is mostly used as an **interjection**.

The four attitudinal dimensions considered in this study primarily relate to the usage of and attitude to the swearwords. **Production** refers to how often the respondent says or writes a swearword, while **Perception** indicates how often the respondent hears or reads a swearword. **Offensiveness** relates to how offensive the respondent finds a word themselves, and **Tabooness** equates to how offensive the respondent thinks other people perceive the swearword to be. In the subsequent tests, the attitudinal dimensions are considered as the response variables, while sociodemographic factors act as predictor variables. Based on the responses from between 147 and 179 participants (see below), the median values of the four response variables for each of the four words are summarised in Table 2.

Table 2 Median Likert scale (1–9) values of the four response variables for the four swearwords.

WORD	PRODUCTION	PERCEPTION	OFFENSIVENESS	TABOONESS
feeks	3	5	2	3
piele	1	3	5	7
moffie	2	4	7	7
jissis	1	4	9	8

The selection of attitudinal dimensions and sociodemographic factors are also done in order to cover a range of predictor and response variables that are aligned closely with the source of tabooness of the word under consideration, viz.:

1. **Misogyny:** We expect words like *feeks*, which might be use pejoratively in reference to women, to be more offensive to female respondents. This expectation is based on the assumption that terms of abuse will be more abusive (and hence offensive and perceived as taboo) to the abusee, than to the abuser. Furthermore, since the word is old-fashioned, one can expect that younger respondents will use the word less often and find it less offensive.
2. **Sexuality:** Numerous studies have indicated that men use strong swearwords much more often than women (see Güvendir, 2015 for a brief summary of findings). We therefore hypothesise that male respondents will report that they use and encounter words like *piele* more often than female respondents. It is also expected that older men produce and perceive these types of words more often than younger men.
3. **Bigotry:** Since homophobia is a sociological phenomenon where hegemonic masculinity is purported as superior to alternate masculinities (see, among others, Kimmel, 2005, ch. 2), we expect that a word like *moffie* will be perceived as less offensive by older, conservative male respondents (i.e., people representing hegemonic masculinity).
4. **Blasphemy:** It is expected a priori that blasphemous words will be more offensive to more religious and conservative people.

Although not all registered users completed the SWSs for all four words (see Section (4)), the total number of respondents for each of the words constitute a relatively large sample, ranging between 147 and 179 participants. Table 3 summarises the distribution of respondents for each of the four words across sociodemographic factor levels.

Two further details regarding the respondents are relevant to our analyses:

1. Respondents who did not choose any of the levels listed above (e.g., who chose to not answer a question), were excluded from this analysis, since respondents who do not specify one of the levels could potentially belong to any of the levels. The inclusion of these respondents would violate the independence assumption underpinning all of the statistical methods under consideration, and potentially invalidate the reported results.
2. For both the political and world view factors, only two respondents identified as Very Conservative, which are too few respondents to make valid statistical inferences. In all of the subsequent hypothesis testing, the Very Conservative level is excluded from consideration.

SOCIODEMOGRAPHIC FACTOR	LEVEL	*FEEKS* (*n* = 147)	*PIELE* (*n* = 167)	*MOFFIE* (*n* = 179)	*JISSIS* (*n* = 152)
		n (%)			
Age group	18–39	55 (37.41)	65 (38.92)	67 (37.43)	63 (41.45)
	40–59	62 (42.18)	66 (39.52)	76 (42.46)	59 (38.82)
	60+	30 (20.41)	36 (21.56)	36 (20.11)	30 (19.74)
Sex	Male	55 (37.41)	63 (37.72)	73 (40.78)	59 (38.82)
	Female	92 (62.59)	104 (62.28)	106 (59.22)	93 (61.18)
Religious views	Very religious	30 (20.41)	35 (20.96)	34 (18.99)	33 (21.71)
	Religious	56 (38.10)	65 (38.92)	65 (36.31)	58 (38.16)
	Moderate	19 (12.93)	23 (13.77)	26 (14.53)	24 (15.79)
	Not really	18 (12.24)	18 (10.78)	21 (11.73)	17 (11.18)
	Not at all	24 (16.33)	26 (15.57)	33 (18.44)	20 (13.16)
Political views	Very Conservative	2 (1.36)	2 (1.20)	2 (1.12)	4 (2.63)
	Conservative	11 (7.48)	14 (8.38)	13 (7.26)	12 (7.89)
	Moderate	61 (41.50)	75 (44.91)	75 (41.90)	69 (45.39)
	Liberal	34 (23.13)	39 (23.35)	43 (24.02)	36 (23.68)
	Very liberal	39 (26.53)	37 (22.16)	46 (25.70)	31 (20.39)
World view	Very Conservative	2 (1.36)	2 (1.20)	2 (1.12)	2 (1.32)
	Conservative	17 (11.56)	22 (13.17)	19 (10.61)	19 (12.50)
	Moderate	37 (25.17)	44 (26.35)	44 (24.58)	46 (30.26)
	Liberal	34 (23.13)	44 (26.35)	43 (24.02)	36 (23.68)
	Very liberal	57 (38.78)	55 (32.93)	71 (39.66)	49 (32.24)

Table 3 Sociodemographic summary of respondents for the four swearwords.

(4) METHODOLOGY

For the purpose of comparing the different tests, each of the three tests is applied to all of the combinations of words, response and predictor variables. In total, 80 different hypotheses are tested (i.e., five sociodemographic factors x four attitudinal dimensions x four words) using the three tests. The aim of these hypothesis tests is not to identify significant effects, but rather to compare the hypotheses that are considered significant by each of the statistics to ascertain whether the identified predictors and response variables are different. Significance is considered at the 95% confidence level.

For ANOVA, eight of the hypotheses showed a significant effect for the predictor variable on the response variable (see Table 4). Both KW-H and OLR also showed significant effects between levels for these eight hypotheses. An analysis of the post-hoc tests for both ANOVA and KW-H identified the same levels where statistically significant effects are observed.

WORD	RESPONSE	PREDICTOR	*n*	MEDIAN	ANOVA *p*	KW-H *p*	OLR *p*
feeks	Production	Age	147	3	.010	.004	.005
feeks	Offensiveness	Sex	147	2	.038	.009	N/A
piele	Perception	World view	165†	3	.022	.009	.001
piele	Offensiveness	Age	167	5	.001	.001	.016
piele	Offensiveness	Sex	167	5	<.0005	<.0005	<.0005
piele	Offensiveness	Political view	165†	5	.038	.028	.007
piele	Offensiveness	World view	165†	5	.013	.012	.001
piele	Tabooness	Sex	167	7	.007	.005	.004

Table 4 Significant ANOVA results.
Note: N/A = Violation of the assumption of proportional odds for OLR; † = Respondents from "Very conservative" level excluded since there were fewer than five respondents.

In addition to the eight hypotheses with significant effects identified by ANOVA, KW-H identified a further 17 significant effects for a total of 25 effects (see Table 5). There is also substantial overlap between these hypotheses and the OLR results for the same hypotheses, with only a single instance where OLR does not identify a significant effect, namely *piele* | Production | Age (marked in bold in the last column in Table 5). However, there are six hypotheses where the assumption of proportional odds for OLR is violated (marked "N/A" in the last column in Table 5).

Table 5 Significant KW-H results.
Note: N/A = Violation of assumption of equal variance for ANOVA, or assumption of proportional odds for OLR; † = Respondents from "Very conservative" level excluded since there were fewer than five respondents.

WORD	RESPONSE	PREDICTOR	*n*	MEDIAN	ANOVA *p*	KW-H *p*	OLR *p*
piele	Production	Age	167	1	**.088**	.043	**.251**
piele	Production	Sex	167	1	N/A	<.0005	<.0005
piele	Production	Political view	165†	1	**.069**	.008	N/A
piele	Production	World view	165†	1	N/A	.0134	.001
piele	Perception	Political view	165†	3	**.074**	.034	N/A
piele	Tabooness	Age	167	7	**.058**	.004	.015
moffie	Production	Sex	179	2	N/A	.008	.002
moffie	Offensiveness	Sex	179	7	N/A	.004	.002
jissis	Production	Sex	152	1	N/A	.001	.002
jissis	Production	Political view	148†	1	N/A	.001	N/A
jissis	Production	Religious view	152	1	N/A	<.0005	<.0005
jissis	Production	World view	150†	1	N/A	<.0005	N/A
jissis	Offensiveness	Sex	152	9	**.0526**	.044	.031
jissis	Offensiveness	Political view	148†	9	N/A	.001	N/A
jissis	Offensiveness	Religious view	152	9	N/A	<.0005	N/A
jissis	Offensiveness	World view	150†	9	N/A	<.0005	<.0005
jissis	Tabooness	Religious view	152	8	**.054**	.016	.010

Note that in most of the 17 hypotheses identified by KW-H, the assumption of homoscedasticity for ANOVA is violated (marked "N/A" in the ANOVA column in Table 5). Although the *p* values for ANOVA that do not violate this assumption would be significant at the 90% level, they are not significant at the more accepTable 95% level.

It is also noticeable that KW-H seems to identify hypotheses with significant effects even with highly skewed distributions, i.e., where median values are at the very extreme ends of the Likert scale. In contrast, ANOVA mainly identifies hypotheses where the median is closer to the centre of the scale, although this is not always the case (e.g., see *moffie* | Offensiveness | Sex), as discussed in more detail in the following section.

Lastly, in addition to the hypotheses identified as significant by ANOVA and KW-H, three additional effects are identified by OLR only (see Table 6). OLR therefore identified a total of 20 effects (i.e., seven in Table 4, ten in Table 5, and three in Table 6). In all these cases, the proportional odds assumption for OLR has not been violated.

Table 6 Significant OLR results.

WORD	RESPONSE	PREDICTOR	*n*	MEDIAN	ANOVA *p*	KW-H *p*	OLR *p*
piele	Perception	Religious view	167	3	.207	.125	.040
piele	Offensiveness	Religious view	167	5	.132	.086	.016
moffie	Perception	Political view	177	4	.338	.304	.030

(5) RESULTS AND DISCUSSION

Our first research question (see section (1)) was whether the different statistical tests produce comparable results. From the results presented in the previous section, the two tests that were designed specifically to work with ordinal data (KW-H and OLR) not only identified the same

eight significant hypotheses as ANOVA, but also additional significant effects: KW-H identified a total of 25 hypotheses and OLR 20. Given the fact that ANOVA is not necessarily simpler to run or interpret than KW-H – both require post-hoc tests to determine the specific levels between which significant effects are observed – there does not seem to be a good reason for using a parametric test (ANOVA) for Likert-type data, especially when the value distributions are highly skewed.

When choosing between KW-H and OLR, there seems to be a high degree of agreement between them when considering a single predictor variable. On the downside for OLR though, is that a substantial number of hypotheses are not permissible when considering the OLR assumption of proportional odds, along with the fact that the test is substantially more complex to apply and interpret. It would therefore seem advisable to use KW-H for Likert-type data of this kind. One does however also need to consider how the results for each of these tests can be interpreted and how the interpretation possibilities address the fundamental aims of a particular study.

The second research question concerned the descriptive power of the respective tests and the interpretation possibilities for each. Both ANOVA and KW-H only indicate that there are statistically significant differences between one or more of the levels under consideration, and both require a post-hoc test to indicate which levels differ significantly. It is possible for an overall test to indicate that there are statistically significant differences between levels, but that the post-hoc test does not show a significant difference between any two levels. In the data used in this study, however, there are at least two levels that show significant differences for all of the tests identified by either the ANOVA or KW-H.

As an example of the results of these post-hoc tests, Tables 7 and 8 show the post-hoc test results when comparing the different individual Age levels for the Production of the word *feeks*. Both post-hoc tests indicate that there is statistically significant difference between Production for the 18–39 and 40–59 age levels, while only the Dunn test indicates that there is a difference between 18–39 and 60+ levels. These tests still don't provide further insight into the nature of the difference, i.e., whether people between 18 and 39 produce *feeks* more or less often. Further analysis is therefore required.

AGE GROUP	18–39	40–59	60+
18–39	1.000	**.006**	**.037**
40–59	**.006**	1.000	1.000
60+	**.037**	1.000	1.000

Table 7 Dunn's post-hoc test results for Production of the word *feeks*.

AGE GROUP	18–39	40–59	60+
18–39	1.000	**.014**	.053
40–59	**.014**	1.000	.999
60+	.053	.999	1.000

Table 8 Tukey's HSD test results for Production of the word *feeks*.

As part of the review of KW-H, it is necessary to do a visual inspection of the box plots of the different levels in order to determine whether the distributions of the various levels are the same (or at least very similar), since this determines how the results can be interpreted. If the distributions are similar, one can make statements about the median of the respective level, for example the median of 18–39 level is one point lower than that of the 40–59 level. If the distributions are not similar, one can only make statements about the most likely response of members of a particular level, for example a randomly selected member of the 18–39 group is more likely to assign a lower Production value than a member from the 40–59 group. Visual inspection of the box plot in Figure 1 shows that the distribution of the three Age levels is not similar enough to adhere to KW-H assumption and therefore we can only interpret the results in terms of the distribution, not the median. In this case, a randomly selected member of the 18–39 group will assign a lower Production value than either the 40–59 or 60+ groups statistically significantly.

DOI: 10.5334/johd.132

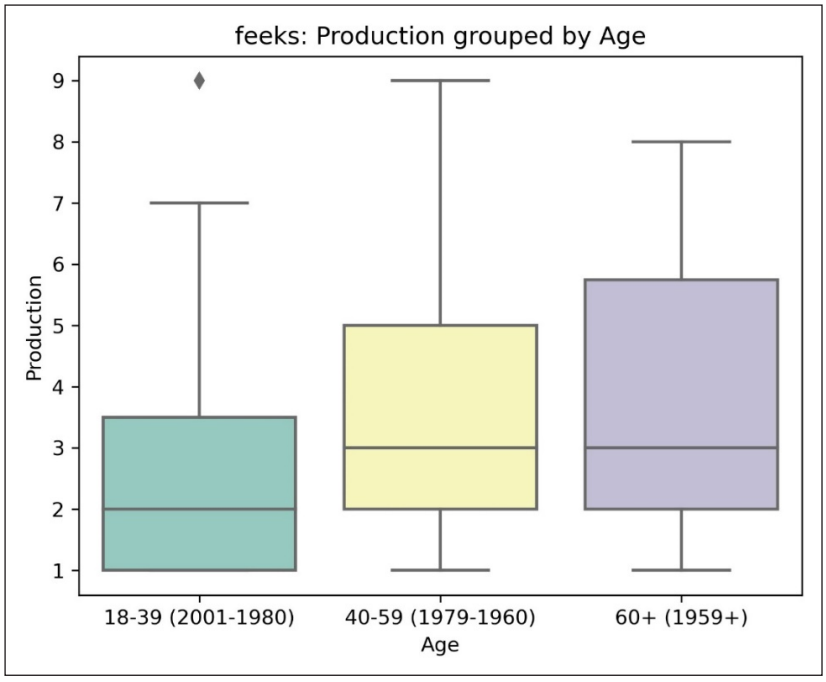

Figure 1 Box plot of different Age levels for the Production of *feeks*.

OLR models allow for a slightly different interpretation of the results, as the test is specifically designed to make predictions of expected future behaviour in the parameter estimates. OLR gives a representation of the significance ($p < 0.05$) of the predictor variable (e.g., Age) in relation to a response variable (e.g., Production) for a reference level (e.g., 18–39), as well as an odds-ratio which indicates how much more likely one level is to assign a higher or lower value than the reference level. Tables 9 and 10 provide an overview of the results for two levels (40–59 and 60+) compared to the reference level (18–39) when considering the Production of *feeks* with Age as the predictor variable.

AGE GROUP	18–39	40–59	60+
18–39	1.000	**.001**	**.008**
40–59	**.001**	1.000	.983
60+	**.008**	.983	1.000

Table 9 OLR significance values for the Production of *feeks*.

AGE GROUP	18–39	40–59	60+
18–39	1.000	.333	.335
40–59	.333	1.000	1.008
60+	.335	1.008	1.000

Table 10 OLR odds ratio values for the Production of *feeks*.

These results indicate that a person from the 18–39 group is significantly more likely to assign a lower value (odds ratio < 1) than people from either the 40–59 ($p = .001$) and 60+ ($p = .008$) levels at ratios of 3.0 (1/.333) and 2.99 (1/.335) to one. These ratios should be interpreted as follows: three people in the 18–39 age group will assign a lower value for each person in the 18–39 level that assigns a higher value. Between the 40–59 and 60+ there is not a significant difference ($p = .983$) as substantiated by the odds ratio of 1.008, which is almost identical.

The answer to the last research question (which statistical test is most appropriate and robust for interpreting the data?) serves as conclusion of the article.

(6) IMPLICATIONS AND APPLICATIONS

There has been longstanding disagreement within the statistical research community on the appropriate statistical test to apply to Likert-type data. Some proponents of the more straightforward ANOVA argue that the test is more easily interpretable and is sufficiently robust to

deal with data that violate the main assumptions of the test, *viz.* following a Gaussian distribution and constituting interval data (Endresen & Janda, 2017; Norman, 2010). Opponents argue that the violation of the assumptions is serious enough to cast doubt on the results and conclusions drawn from the parametric tests, and that one should rather use non-parametric tests specifically designed for ordinal data, such as KW-H or the substantially more complex parametric OLR.

In this work, we show that although there is some overlap in the results between the different statistical tests when analysing Likert-type data, KW-H in general performs best when analysing this type of data, especially when the responses are heavily skewed, as is the case of controversial subjects such as offensive language. Beyond the fact that KW-H validates a substantially larger group of effects, which in and of itself does not indicate a superior test statistic, the test shows almost identical behaviour to ANOVA where ANOVA does indicate statistical significance. Further analysis of KW-H results indicate that those effects identified, but not identified by ANOVA, do in fact show effects that indicate a significant difference between levels, and is corroborated by OLR. Since ANOVA and KW-H are similarly interpretable, the results indicate that it should be preferable to use KW-H, rather than ANOVA, when working with Likert-type data.

OLR is a substantially more complex statistical model to generate given the required assumptions associated with the test. The results of the OLR models largely agree with KW-H, and apart from the fact that the test allows for a different interpretation model, the odds-ratio, which gives a more powerful description of the differences between levels, does not provide additional qualities which would necessitate its use over that of KW-H, unless one is specifically interested in reviewing multiple predictor variables, for example combining Age and Gender, in which case this would be the most appropriate test.

ACKNOWLEDGEMENTS

We would like to acknowledge Cornelius van der Walt (BlueTek Computers) and Jaco du Toit (NWU) for their technical and creative work in the implementation of the Vloekmeter. A comprehensive list of all the co-workers, collaborators, and students on the project is published on vloek.co.za/oor-ons.

None of the results and/or opinions in this paper can be ascribed to any of the people or organizations mentioned above.

Ethical clearance for the research project was obtained through the Language Matters Ethics Committee of the NWU (ethics number: NWU-00632-19-A7).

FUNDING INFORMATION

This research is partially funded by the Suid-Afrikaans Akademie vir Wetenskap en Kuns, and partially made possible through barter agreements with BlueTek Computers, Afrikaans.com, and WatKykJy.co.za. The Woordeboek van die Afrikaanse Taal (WAT), Handwoordeboek van die Afrikaanse Taal (HAT), and Centre for Text Technology (CTexT) of the North-West University (NWU) are hereby also acknowledged for generously supplying the project with material from their respective databases.

COMPETING INTERESTS

The second author is a director of the not-for-profit company, Viridevert NPC (CIPC registration number: 2016/411799/08), who owns and manages the website vloek.co.za. This website was developed specifically for this project, and this conflict of interest has been approved by the NWU.

AUTHOR CONTRIBUTIONS

Roald Eiselen: Conceptualisation; Formal Analysis; Investigation; Methodology; Software; Writing – original draft; Writing – review and editing.

Gerhard B. van Huyssteen: Conceptualisation; Data curation; Funding acquisition; Investigation; Methodology; Writing – original draft; Writing – review and editing.

AUTHOR AFFILIATIONS

Roald Eiselen orcid.org/0000-0002-8612-5175
Centre for Text Technology (CTexT), North-West University, Potchefstroom, South Africa
Gerhard B. van Huyssteen orcid.org/0000-0003-1705-1747
Centre for Text Technology (CTexT), North-West University, Potchefstroom, South Africa

REFERENCES

Beers Fägersten, K. (2007). *A sociolinguistic analysis of swearword offensiveness*. Retrieved from https://www.researchgate.net/publication/265009714_A_sociolinguistic_analysis_of_swearword_offensiveness (last accessed: 12 October 2023).

Beers Fägersten, K. (2012). *Who's swearing now? The social aspects of conversational swearing*. Cambridge Scholars Publishing.

Boone, H. N., & **Boone, D. A.** (2012). Analyzing Likert data. *Journal of Extension*, *50*(2), 1–5. DOI: https://doi.org/10.34068/joe.50.02.48

Culpeper, J. (2021). Sociopragmatics: Roots and definition. In M. Haugh, D. Z. Kádár, & M. Terkourafi (Eds.), *The Cambridge Handbook of Sociopragmatics* (pp. 15–29). Cambridge University Press. DOI: https://doi.org/10.1017/9781108954105

Dewaele, J.-M. (2004). Blistering barnacles! What language do multilinguals swear in?! *Estudios de Sociolinguistica*, *5*(1), 83–106. Retrieved from http://hdl.handle.net/10315/2489 (last accessed: 12 October 2023). DOI: https://doi.org/10.1558/sols.v5i1.83

Dewaele, J.-M. (2012). "Christ fucking shit merde!" Language preferences for swearing among maximally proficient multilinguals. *Sociolinguistic Studies*, 4(3), 595–614. DOI: https://doi.org/10.1558/sols.v4i3.595

Dewaele, J.-M. (2015). British 'Bollocks' versus American 'Jerk': Do native British English speakers swear more – or differently – compared to American English speakers? *Applied Linguistics Review*, *6*(3), 309–339. DOI: https://doi.org/10.1515/applirev-2015-0015

Dewaele, J.-M. (2016a). Self-reported frequency of swearing in English: do situational, psychological and sociobiographical variables have similar effects on first and foreign language users? *Journal of Multilingual and Multicultural Development*, *38*(4), 330–345. DOI: https://doi.org/10.1080/01434632.2016.1201092

Dewaele, J.-M. (2016b). Thirty shades of offensiveness: L1 and LX English users' understanding, perception and self-reported use of negative emotion-laden words. *Journal of Pragmatics*, *94*, 112–127. DOI: https://doi.org/10.1016/j.pragma.2016.01.009

Dewaele, J.-M. (2018). "Cunt": On the perception and handling of verbal dynamite by L1 and LX users of English. *Multilingua*, *37*, 53–81. DOI: https://doi.org/10.1515/multi-2017-0013

Dubois, D. (2013). Statistical reasoning with set-valued information: Ontic vs. epistemic views. In *Towards Advanced Data Analysis by Combining Soft Computing and Statistics* (pp. 119–136). Springer. DOI: https://doi.org/10.1007/978-3-642-30278-7_11

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*(293), 52–64. DOI: https://doi.org/10.1080/01621459.1961.10482090

Endresen, A., & **Janda, L. A.** (2017). Five statistical models for Likert-type experimental data on acceptability judgments. *Journal of Research Design and Statistics in Linguistics and Communication Science*, *3*(2), 217–250. DOI: https://doi.org/10.1558/jrds.30822

Gries, S. (2015). Quantitative linguistics. In J. Wright (Ed.), *International Encyclopedia of the Social and Behavioral Sciences* (2nd ed., Vol. 19, pp. 725–732). Elsevier. DOI: https://doi.org/10.1016/B978-0-08-097086-8.53037-2

Güvendir, E. (2015). Why are males inclined to use strong swear words more than females? An evolutionary explanation based on male intergroup aggressiveness. *Language Sciences*, *50*, 133–139. DOI: https://doi.org/10.1016/j.langsci.2015.02.003

Harpe, S. E. (2015). How to analyze Likert and other rating scale data. *Currents in Pharmacy Teaching and Learning*, *7*(6), 836–850. DOI: https://doi.org/10.1016/j.cptl.2015.08.001

Hsu, T.-C., & **Feldt, L. S.** (1969). The effect of limitations on the number of criterion score values on the significance level of the F-test. *American Educational Research Journal*, *6*(4), 515–527. DOI: https://doi.org/10.3102/00028312006004515

Ipsos MORI. (2016). *Attitudes to potentially offensive language and gestures on TV and radio*. Retrieved from https://www.ofcom.org.uk/__data/assets/pdf_file/0022/91624/OfcomOffensiveLanguage.pdf (last accessed: 12 October 2023).

Jamieson, S. (2004). Likert scales: How to (ab) use them? *Medical Education*, *38*(12), 1217–1218. DOI: https://doi.org/10.1111/j.1365-2929.2004.02012.x

Janschewitz, K. (2008). Taboo, emotionally valenced, and emotionally neutral word norms. *Behavior Research Methods*, *40*(4), 1065–1074. DOI: https://doi.org/10.3758/BRM.40.4.1065

Jay, T. B. (1992). *Cursing in America: A psycholinguistic study of dirty language in the courts, in the movies, in the schoolyards and on the streets*. John Benjamins. DOI: https://doi.org/10.1075/z.57

Kimmel, M. S. (2005). *The gender of desire: essays on male sexuality*. State University of New York Press. Retrieved from https://ebookcentral.proquest.com/lib/northwu-ebooks/detail.action?docID=3407756 (last accessed: 12 October 2023). DOI: https://doi.org/10.1353/book4893

Laerd Statistics. (2018). *Statistical tutorials and software guides*. Retreived from https://statistics.laerd.com/ (last accessed: 12 October 2023).

Larsson, T., Egbert, J., & **Biber, D.** (2022). On the status of statistical reporting versus linguistic description in corpus linguistics: A ten-year perspective. *Corpora*, *17*(1), 137–157. DOI: https://doi.org/10.3366/cor.2022.0238

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*, 1–55.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*(5), 625–632. DOI: https://doi.org/10.1007/s10459-010-9222-y

Reiter, R. M., & **Placencia, M. E.** (2005). Research methods in sociopragmatics. In R. M. Reiter & M. E. Placencia (Eds.), *Spanish Pragmatics* (pp. 213–230). Palgrave Macmillan UK. DOI: https://doi.org/10.1057/9780230505018_6

Republic of South Africa. (2022). *Films and Publications Act (65/1996): Classification guidelines for the classification of films, interactive computer games and certain publications*. Government gazette Retrieved from https://www.fpb.org.za/wp-content/uploads/2022/07/Classification-Guidelines-Effective-1-August-2022.pdf (last accessed: 12 October 2023).

Sangthong, M. (2020). The effect of the Likert point scale and sample size on the efficiency of parametric and nonparametric tests. *Thailand Statistician*, *18*(1), 55–64. Retrieved from https://ph02.tci-thaijo.org/index.php/thaistat/article/view/228886 (last accessed: 12 October 2023)

Sullivan, G. M., & **Artino Jr, A. R.** (2013). Analyzing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education*, *5*(4), 541–542. DOI: https://doi.org/10.4300/JGME-5-4-18

Tukey, J. W. (1977). Exploratory data analysis. In *Addison-Wesley Series in Behavioral Science: Quantitative Methods* (Vol. 2, pp. 131–160). Addison-Wesley.

Van Hecke, T. (2012). Power study of ANOVA versus Kruskal-Wallis test. *Journal of Statistics and Management Systems*, *15*(2–3), 241–247. DOI: https://doi.org/10.1080/09720510.2012.10701623

Van Huyssteen, G. B. (2021). Swearing in South Africa: Multidisciplinary research on language taboos. *Proceedings of the International Conference of the Digital Humanities Association of Southern Africa 2021, South Africa* (Virtual). DOI: https://doi.org/10.55492/dhasa.v3i01.3854

Van Huyssteen, G. B., & **Eiselen, R.** (2023). *Afrikaans swearword scores: Dataset* (Version 2.0), North-West University. DOI: https://doi.org/10.25388/nwu.23708229

Van Huyssteen, G. B., Eiselen, R., & **Du Toit, J.** (2023). A dataset of self-reported attitudes to Afrikaans swearwords. *Journal of Open Humanities Data, 9*(14), 1–8. DOI: https://doi.org/10.5334/johd.127

Van Sterkenburg, P. G. J. (2019). *Rot lekker zelf op: Over politiek incorrect en ander ongepast taalgebruik*. Scriptum. DOI: https://doi.org/10.5117/NEDTAA2021.1.006.OOMS

Verma, J. P., & **Abdel-Salam, A.-S. G.** (2019). *Testing statistical assumptions in research*. John Wiley & Sons Inc. DOI: https://doi.org/10.1002/9781119528388

TO CITE THIS ARTICLE:
Eiselen, R., & van Huyssteen, G. B. (2023). A Comparison of Statistical Tests for Likert-Type Data: The Case of Swearwords. *Journal of Open Humanities Data*, 9: 18, pp. 1–13. DOI: https://doi.org/10.5334/johd.132

Submitted: 17 August 2023
Accepted: 02 October 2023
Published: 30 October 2023

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.