GHENT UNIVERSITY

FACULTY OF SCIENCES

Measuring gender identity: A comparison between the Likert scale and the Fuzzy scale.

Aisa Burgwal

Master dissertation submitted to

obtain the degree of

Master of Statistical Data Analysis

Promotor: Prof. Dr. Christophe Ley

Co-Promotor: Prof. Dr. Joz Motmans

Department of Applied Mathematics, Computer Science and Statistics

Academic year 2019-2020

GHENT UNIVERSITY

FACULTY OF SCIENCES

Measuring gender identity: A comparison between the Likert scale and the Fuzzy scale.

Aisa Burgwal

Master dissertation submitted to

obtain the degree of

Master of Statistical Data Analysis

Promotor: Prof. Dr. Christophe Ley

Co-Promotor: Prof. Dr. Joz Motmans

Department of Applied Mathematics, Computer Science and Statistics

Academic year 2019-2020

Foreword

The dataset is the result of an online and anonymous survey that was distributed by myself, with the help of a number of people who shared the research (Sam van Hamme, Eva Decaesstecker, Michael Bracke, Lut Joris, Tim D'Hondt, Felien Laureys, Ruth Karuranga, Johanna Swinkels, Felix de Bie, etc.). KoBoToolbox was used to develop the survey, R was used to analyze the data. The data is completely anonymous and confidential. Secondary use of the dataset is possible with permission of the author and promoter.

Opening this master thesis, I want to display my gratitude to some great people that helped me with the realization of this project. Initially, I would like to thank my promotor Prof. Dr. Christophe Ley, who not only came up with the idea of the master thesis and guided me throughout the project with great advice, he also allowed me to combine my interest in gender minority groups with the great educational basis I obtained from the master program in Statistical Data Analysis. The latter allowed me to indulge myself in the interesting available research conducted for different types of scales to measure gender identity, especially the Fuzzy set scale, which I knew nothing about at the beginning of the project. This motivated me to explore the yet undiscovered techniques to analyze Fuzzy set scales and to search for a way to complete this project. I perceive the insights obtained in this master thesis as an addition to research already conducted, and as an opening of various doorways for further exploration and investigation.

Secondly, I want to thank my co-promotor Prof. Dr. Joz Motmans, who has taught me so much the last four years. I cannot even begin to explain how he helped me with my scientific growth, in terms of research, writing skills, and the ability to successfully finish the whole process of data collection to publication. He has been the best mentor, colleague, and friend. At last, I want to thank Ruth Karuranga, the best support I could wish for. Two years ago, we started the master together, and hopefully will finish together as well.

The biggest thank you for all your support and help!

Aisa

Table of contents

Table of Figures

Table of Tables

Abstract

In general there is often unclarity about how to investigate gender identity. Sex assigned at birth and gender identity are still regularly used as interchangeable concepts in research. However, gender identity is more than a binary variation of male or female. Over the past decades, various studies have tried to measure gender identity. Some studies used continuous scales, others a categorical question. This paper presents two types of scales to measure gender identity. Using a quantitative approach, a 5-point Likert scale and a Fuzzy scale were used to measure gender identity. Both scales were compared and groups were formed based on sex assigned at birth and self-reported transgender identity in order to study differences in gender identity between groups. The quantitative data indicated that the Fuzzy scale captured a higher subjectivity and variability in responding, compared to the Likert scale. However, when tests performed with a 5-point Likert scale were compared with tests performed with a Fuzzy scale, the Fuzzy scale did not prove superiority over the 5-point Likert scale. Both scales did not detect any differences in gender identity between respondents assigned male at birth and assigned female at birth, and both scales detected similar differences in gender identity between cisgender and transgender respondents. These results, in combination with qualitative data regarding the use of a Fuzzy scale, led to the preference of a 5-point Likert scale when measuring gender identity. Based on these results, many other studies are to be developed, with a focus on finding the best scale to measure gender identity. Also, future research should focus on developing statistical methods to compare the Likert scale and the Fuzzy scale, rather than comparing them descriptively.

Introduction

In an increasingly complex and diverse society, many population groups are characterized by diversity. Minority groups based on gender identity, often referred to by the term transgender people, have become an important target group within equal rights policies, both nationally and internationally (Equality Act 2010, 2010; Supreme Court of the United States, 2020; Vlaamse Overheid, 2009). Transgender is an umbrella term that includes those people whose gender expression and/or gender identity differs from conventional expectations based on the physical sex assigned at birth (SAAB) (Goffman, 1963). A gender identity is a person's intrinsic feeling of being male (a boy or a man), female (a girl or a woman) or an alternative gender (Bockting, 1999). Research states that it is very difficult to say how many people in our society are transgender, and refers to a great diversity within the transgender community (Collin, Reisner, Tangpricha, & Goodman, 2016; Gates, 2011; Homans, 2014; Monro, 2020). Also, the various definitions of transgender affects number and proportion estimates. Most studies focus on individuals seeking or receiving transgender and gender nonconforming (TGNC) related care, other studies focus on individuals meeting the criteria for transgender and gender nonconforming diagnoses, and other studies focus on self-reported TGNC identity. The first type of research finds the lowest count of respondents who meet the criteria for this definition of transgender. Estimates generally range between 1 and 30 per 100,000 individuals, with self-reported TGNC identity to be orders of magnitude more frequent. A recent literature review reported a range of people who self-identify as TGNC from 100 to 2000 per 100,000 or 0.1% to 2% among adults (Collin et al., 2016; Goodman et al., 2019).

In addition to the increasing public attention to potential diversity in gender identity, scientific insights into researching and measuring gender identity have also evolved considerably. However, there are no general guidelines for questioning and examining gender diversity within quantitative research. There are guidelines and protocols for the psychological screening of gender diversity, but there is currently no accepted research protocol on how to study gender diversity (for a detailed overview of existing methods and guidelines, see Motmans, Burgwal, & Dierckx, 2020).

In this paper, we ought to compare two different techniques of analyzing gender identity. The first technique, based on a Likert scale, has been regularly used within survey research, as opposed to other techniques. These other techniques, such as a categorical methodology using a question with three or more answer options ('male', 'female', 'transgender', …), proved to be problematic and will not be analyzed within this paper (Motmans et al., 2020). Literature on analyzing Likert data is also open to a lot of discussion, and questioning gender identity using a Likert scale does not appear to be the best approach (some people want to choose a point between the different answer options or may not agree with the options provided). Therefore, a second technique is discussed within this paper, based on a Fuzzy scale. This scale will be evaluated to possibly provide a better solution when measuring gender identity in general and/or in population-specific surveys.

In a first section, a literature review is provided in which the social context and statistical background are discussed. The methodology used to compare both techniques is discussed in a second section, together with the characteristics of the data sources. The third section will focus on data cleaning, exploration and analysis. The paper is concluded in a final section and suggestions for future research are offered.

1. Literature Review

In this section existing literature concerning measurements of gender identity in general and within population-specific studies is covered. Basic concepts and technical terms will be discussed briefly in order to avoid confusion.

In general there is often unclarity about how to investigate gender diversity. For example, sex assigned at birth and gender identity are still regularly used as interchangeable concepts in research (Angus, 2012). In most quantitative research gender identity is then simply conceptualized in the same way as sex assigned at birth and questioned by means of a closed question with a binary answer option 'man' and 'woman'. Respondents are thus forced to choose one of these two categories and then gender identity and sex assigned at birth are often compared to one another (Lorber, 2006; Ritz et al., 2017; Westbrook & Saperstein, 2015). This results in no distinction between sex assigned at birth and the more complex concept of gender identity. Those who identify themselves outside the conventional gender binary of men and women (e.g. those who identify as gender non-binary[1]) are incorrectly distinguished within research. Those respondents are wrongly placed within the category of a female/male gender identity, or treated as missing because they leave the question unanswered. Also for cisgender people (those whose sex assigned at birth does not differ from their gender identity), a simplistic binary question for gender identity leads to unrealistic data. Gender identity is more than the M/F on a birth certificate or identity document. Over the past decades, psychology and sociology have developed an increasingly broader view of gender diversity (Westbrook & Saperstein, 2015). Different arguments provide evidence that gender is more than a binary variation of male or female.

First, the common dichotomy, based on the biological binary sex assigned at birth, has many exceptions (Blackless et al., 2000; Callens, Longman, & Motmans, 2017; Fausto-Sterling, 2000). This gender variation is often labeled under the umbrella term intersex (Davis, 2014; Motmans & Longman, 2017; Reis, 2007). The binary birth categories 'man' and 'woman' are therefore much less natural as people often think.

Second, we see that the term gender identity itself is anything but fixed and unchangeable. The term was first used by Stoller (1968) and has since been widely used to make a distinction between, on the one hand, the physical body and the sex assigned at birth and, on the other

[1] A person who does not identify with the gender they were assigned at birth and also identifies outside the gender binary of male or female (e.g. identifies neither/and male nor/and female) (Richards et al., 2016).

hand, the psychological identity of whether or not one feels more like a man or a woman. This identity is not a fixed fact, but is formed through a socialization process in which certain gender expressions and gender roles are taught. These gender expressions and roles are not static or unchangeable, but can vary greatly over time and the place where people grow up.

Third, interpreting gender identity in a binary way has ideological consequences. Such boundaries create insiders and outsiders, those who are 'normal' and those who are 'not normal' (Lorber, 2006). By interpreting gender as something layered and fluid, this rigid and binary thinking ends and individuals are recognized in their true gender identity. The heterogeneity within both the transgender and cisgender population also becomes visible.

At last, people with a non-binary gender identity often do not fit in a binary gender identity model. A simplistic binary approach to gender identity has no regard for those who feel neither male nor female (Herdt, 1996). The study of Bockting (2008) shows that a significant proportion of American respondents could not agree with a female or male identity. Belgian research also shows that an underestimated part of the population does not identify itself as a woman or as a man (Dierckx, Meier, & Motmans, 2017). Also, transgender individuals who identify with a binary gender identity score very differently on health and mental well-being, compared to non-binary individuals (Burgwal et al., 2019; Fundamental Rights Agency, 2020; Harrison, Grant, & Herman, 2012; Warren, Smalley, & Barefoot, 2016).

Current methods in quantitative research are no longer in line with existing knowledge about gender identity. Improving methods to investigate gender identity is therefore important.

1.1 Existing methods to measure gender identity

Various studies have tried to measure gender identity. Some studies use continuous scales (see e.g. Åhs et al., 2018; Bakker & Vanwesenbeeck, 2007; Bockting, Benner, & Coleman, 2009; Dierckx et al., 2017; Kuyper & Wijsen, 2014; Schoonacker, Dumon, & Louckx, 2009; Van Caenegem et al., 2015), others a categorical question (see e.g. Australian Bureau of Statistics, 2016; Balarajan, Gray, & Mitchell, 2011; Conron, Scott, Stowell, & Landers, 2012; Crissman, Berger, Graham, & Dalton, 2017; Flores, Brown, & Herman, 2016; Grant et al., 2011; Keuzenkamp, 2012; Statistics New Zealand, 2015; Westat, 2017).

For measuring gender identity with a continuous scale, previous research has used a 7-point Likert scale (see e.g. Bockting et al., 2009), or a 5-point Likert scale (see e.g. Bakker & Vanwesenbeeck, 2007). In a recent study regarding romantic relationships, two 7-point Likert scales were used (Bockting et al., 2009), where respondents had to indicate to what extent they felt male and to what extent they felt female (Leroy, 2019). On the basis of those two questions and a question about sex assigned at birth, two groups were created: cisgender persons, and gender incongruent persons. Cisgender persons were respondents assigned male at birth (AMAB) (A1 = 1) who indicated that they felt mainly male (score on A2a ≥ 4) and did not feel female (score on A2b ≤ 3), gender-congruent men. Analogously, respondents assigned female at birth (AFAB) (A1 = 2) who indicated that they felt mainly female (score on A2b ≥ 4) and did not feel male (score on A2a ≤ 3), gender-congruent women. The respondents who met these conditions were the members of the cisgender group in this study. The other respondents were the gender incongruent members. The additional question 'Have you already told at least one person that your gender identity is different from your sex assigned at birth?' (T1; Motmans, Wyverkens, and Defreyne (2017)) was used as a subsidiary question to determine which persons within the gender incongruent group were considered transgender and not gender non-binary. Two of the final three groups (cisgender, transgender, and gender non-binary persons) were used as a predictor variable for gender identity in subsequent analyses, to study sexual health of cisgender and transgender persons. However, this categorical approach is less nuanced than a continuous approach that better reflects the conceptualization of gender identity as a spectrum in which people can report gender incongruent and gender non-binary feelings (Castleberry, 2019; Van Caenegem et al., 2015). When using a continuous approach to gender identity, there is more room to also consider non-binary persons (non-binary persons were recoded as missing within the study of Leroy (2019)), a group that often remains underexposed in transgender research (Burgwal et al., 2019). The Likert scale was thus not considered as a continuous outcome variable, as in this paper. The study of Leroy (2019) did show that gender identity is much more complex than a gender binary suggests, but the boundaries to distinguish groups were still arbitrary. Due to the lack of a well-founded scientific rationale, the distinction between cisgender and transgender persons became more of a functionally motivated but ultimately subjective decision. Also, people who strongly feel female/male opposite to their sex assigned at birth do not necessarily have to identify themselves with the term transgender. Whether the different groups created within the study of Leroy (2019) are therefore a realistic representation of reality cannot be determined.

For this paper, a new survey was developed, asking the same question 'how feminine/masculine do you feel' twice, using a different scale each time (Bakker & Vanwesenbeeck, 2007). First, a 5-point Likert scale was presented. Second, a Fuzzy scale was used. Before describing methodology and research questions, these two different scales and their underlying analysis will be discussed.

1.2 Statistics with a Likert scale

When rating traits or attributes that cannot be measured directly (such as satisfaction, attitude, gender identity, ...), different scales have been considered. The best-known scales in this setting are the discrete ones, which consist of choosing the most appropriate 'values' within a class according to the rater judgement (such as Likert-type scales) .

Likert scales range from a group of categories - least to most - that ask people to indicate how much they agree or disagree, approve or disapprove, or believe statements are true or false. The main consideration is to include at least five response categories (Gil & González-Rodríguez, 2012; Likert, 1932). There have been debates among the users of Likert scales about its best possible usability in term of reliability and validity of number of points on the scale (Colman, Norris, & Preston, 1997; Cox, 1980; Preston & Colman, 2000). Several advantages and disadvantages for a 5- and 7-point scale have been identified in various studies. When considering reliability of the responses from participants in a survey, a 7-point scale may perform better compared to a 5-point scale. The 7-point scale provides more varieties in options which in turn increase the probability of capturing the objective reality of people (Cox, 1980). A respondents' agreement with a specific topic may lie in between two descriptive options provided on a 5-point scale. A 7-point scale may eliminate this problem up to an extent, by eliciting retrieval beyond the utmost level of agreement provided by a 5-point scale (Finstad, 2010; Komorita & Graham, 1965). However, the validity of the Likert scale is driven by the applicability of the topic concerned, in the context of respondents' understanding and judged by the creator of the response item. When the topic concerned is not relevant to the respondents' everyday context, the provision of more options may reduce content and construct validity of the scale. Providing options more close to the original view of the respondent reduce the role of ambiguity in the responses (Finstad, 2010; Lubiano, de la Rosa de Sáa, Montenegro, Sinova, & Gil, 2016).

When Likert-type scale data are analyzed for statistical purposes, they can be treated as categorical variables, with the consequence that techniques for analyzing them are quite limited. They can also be coded continuously by consecutive integer numbers, which to some extent increases the number of possible procedures. Then the choice is to be made within a continuum, so the variability, diversity and subjectivity are ensured. Statistical conclusions are also reliable and generally no relevant information is lost.

In general, mean and standard deviation are invalid parameters for descriptive statistics when data is on an ordinal scale, as are all parametric analyses based on the normal distribution. However, the question is how robust Likert scales are for deviations from linear, normal distributions (Norman, 2010). Gaito (1980) solves this problem by turning it into someone else's problem, but not the statistician's problem. Gaito indicates that there is no relationship between type of scale and statistical techniques used, in contrast to what many textbooks discuss (Blalock, 1997; Schmidt, 1979). He stated that although the interpretation given to results does take into account the origin of the numbers, this aspect is irrelevant for statistical purposes. This means that even if conceptually a Likert scale is ordinal, we cannot theoretically guarantee that the actual distance between 1 = 'Absolutely disagree' and 2 = 'Disagree' is the same as 4 = 'Agree' and 5 = 'Absolutely agree', but this is not relevant to the analysis because the computer cannot confirm or deny it in any way. There are no independent observations to verify or refute the problem. And all the computer can do is draw conclusions about the numbers themselves. So if the numbers are fairly distributed, we can draw conclusions about their averages, differences, etc. Strictly speaking, we cannot draw any further conclusions about differences in the underlying, latent characteristic reflected in the Likert scale, but this does not invalidate the conclusions about the numbers. The person interpreting the analyses must decide whether the analysis of the numbers reflects the underlying construct.

For measurement and analysis of gender identity with a Likert-type scale, different considerations can be made. On the one hand, while the average could be calculated for any set of numbers, some papers on the Likert scale indicate that calculating an average is difficult to interpret (Jamieson, 2004). The responses are not on a simple linear scale. The nature of the scale prevents the calculation of a valid standard deviation. On the other hand, other authors write in favor of using parametric tests when analyzing Likert-scale data (Norman, 2010), showing power superiority over non-parametric tests such as the Kruskal-Wallis test. Norman (2010) provides convincing evidence that parametric tests can be used with data from Likert scales. That is, parametric tests tend to give 'the correct answer' even when statistical assumptions, such as normal distribution of data, are even extremely violated. Other suggestions for null hypothesis testing include the use of non-parametric procedures such as the Kruskal-Wallis test (Allen & Seaman, 2007). However, this does not test a meaningful hypothesis. The conventional analysis approach commonly used involves splitting the data and looking at the proportion of responses that fall above or below a given cut-off point. Tabulating confidence intervals, using a bootstrapping procedure, can also be interesting when measuring gender identity. How each of these approaches would function when analyzing a concept as gender identity will be discussed first.

1.2.1 R analysis for Likert scales

One of the difficult aspects of the Likert scale in calculating the mean when measuring gender identity is the potential to confuse a series of neutral responses with a series of extreme responses. To illustrate this problem, a simulation study is performed on a fictional dataset, where gender identity is measured with a 5-point Likert scale. The choice of a 5-point Likert scale is made because gender identity is, for the majority of the population, not an everyday concept. Since a 7-point scale may make the concept of gender identity to complex, a 5-point scale is used (Finstad, 2010; Lubiano, de la Rosa de Sáa, et al., 2016).

The analyzes that will be applied to the data gathered for this paper will be discussed here. Three samples were simulated. The first simulated sample gives an equal chance that a respondent feels strongly masculine or feminine, with nothing in between. The second sample has no extreme reactions. In the third, there is an equal chance of choosing one of the five answers. The distributions of each sample are visualized in Figure 1.

```
set.seed(7)
library(ggplot2)
n <- 100
x1 <- sample(c(1,5), n, replace = TRUE)
x2 <- sample(c(2,3,4), n, replace = TRUE)
x3 <- sample(c(1,2,3,4,5), n, replace = TRUE)
x <- c(x1, x2, x3)
q <- rep(c("Sample_1", "Sample_2", "Sample_3"), each = n)
d <- data.frame(q,x)
g0 <- ggplot(d, aes(x=x))
g0 + geom_bar() + facet_wrap("q") + xlab("Likert score")
```

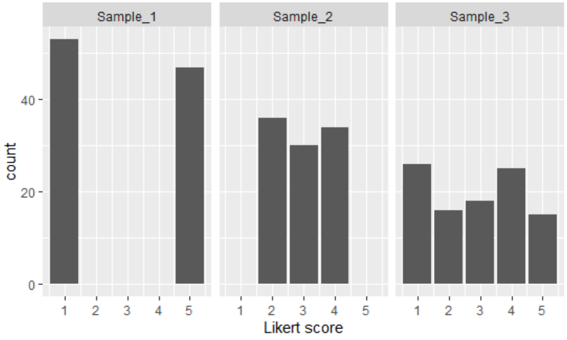


Figure 1. Distributions of simulated data

```
library(magrittr)

library(rlang)

library(dplyr)

d %>% group_by(q) %>% summarise(mean = mean(x), median = median(x)) -> dd

library(knitr)

kable(dd)
```

Table 1. Mean and median of simulated samples

	M	*Md*
Sample 1	2.88	1
Sample 2	2.98	3
Sample 3	3.87	3

Note: M = Mean, Md = Median.

While such an extreme pattern is unlikely to occur in practice, the simulation illustrates the problem. The pattern of responses is very different, but they all have similar average scores (see Table 1).

A test to use would be a one way analysis of variance.

```
mod <- lm(x ~ q)

anova(mod)

Analysis of Variance Table

Response: x

           Df Sum Sq Mean Sq F value Pr(>F)

q           2   0.74   0.370  0.1636 0.8492

Residuals 297 671.83   2.262
```

The ANOVA shows no significant differences between the mean scores for the three samples.

The Kruskal-Wallis test is based on ranks. The null hypothesis being tested is that the location parameters of the distribution of the scores are the same in each sample.

```
kruskal.test(d$x ~ d$q)

Kruskal-Wallis rank sum test

data:  d$x by d$q

Kruskal-Wallis chi-squared = 0.56168, df = 2, P-value = 0.7551
```

This also shows no significant difference between the three samples. However, it is clear that there are differences in the pattern of responses that are not picked up by both procedures.

Another way is to simplify the data in classes and look at the number of responses that fall in each class. The measure used can be, for example, the proportion of respondents who feel masculine (1-2), feminine (4-5) or and/neither masculine and/nor feminine (3).

```
d$x1[d$x == 1 | d$x == 2] = 1

d$x1[d$x == 3] = 2

d$x1[d$x == 4 | d$x == 5] = 3

tb <- table(d$x1, d$q)

round(prop.table(tb, margin = 2)*100, 1)

chisq.test(tb)

Pearson's Chi-squared test

data:  tb

X-squared = 48.75, df = 4, P-value = 6.583e-10
```

Now there are very clear differences between the samples. Although the mean itself is sometimes difficult to interpret, it is possible to produce intervals for the mean using bootstrapping. Because it is impossible to measure a concept in an entire population, this procedure is used to determine the value of a parameter and its interval by statistical sampling. This includes resampling from an existing sample with data replacement. In other words, if a sample has only five respondents who give the scores 1, 5, 4, 5, 2, a random sample can be taken and will very occasionally produce 5, 5, 5, 5, 5 or 1, 1, 1, 1, 1. Typically, it will produce a mixture of the values. If we repeat the resampling thousands of times and exclude the extreme values that are very rare, we can get a bootstrapped confidence interval for the mean by calculating it for all random samples. This approach will occasionally break down for small samples (for example, when all values are identical), but in general it is quite robust and will never yield values beyond the limits of the data. To demonstrate how this can be used with data for thirty samples that show underlying differences in the pattern of responses, data will be simulated by varying both the number of respondents and the response pattern.

```
sim_sample <- function(i){
  sample <- paste("sample", i, sep = "_")
  n <- sample(30:100,1)
  GI <- sample(d$x, n, replace = TRUE)
  c <- data.frame(sample = sample, GI = GI)
  c
}
c <- do.call("rbind", lapply(1:30, sim_sample))
boot_mean <- function(x){
  n <- length(x)
  x <- replicate(1000, mean(sample(x, n, replace = TRUE)))
  round(quantile(x, c(0.025, 0.5, 0.975)), 2)
}
c %>% group_by(sample) %>%
  summarise(n = n(),
            mean = boot_mean(GI)[2],
            lwr = boot_mean(GI)[1],
            upr = boot_mean(GI)[3],
            ) -> cc
cc <- cc[order(-cc$mean), ]
cc$sample = factor(cc$sample, levels = cc$sample[order(cc$mean)], ordered =
TRUE)
g0 <- ggplot(cc, aes(x = sample))
g0 <- g0 + geom_point(aes(y = mean), colour = "red")
g0 <- g0 + geom_hline(yintercept = mean(d$nss), col = "green") +
  xlab("Mean Likert score with bootstrapped 95% confidence intervals")
g1 <- g0 + geom_errorbar(aes(ymin = lwr,ymax = upr)) + coord_flip()
g1
```

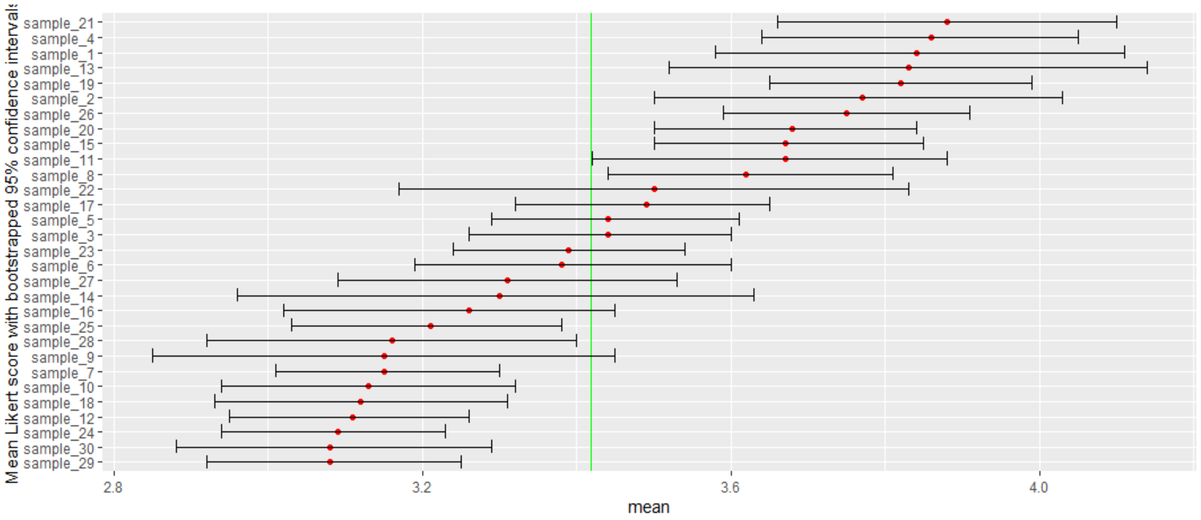


Figure 2. Bootstrapped mean confidence intervals (95%)

Note. Green line = overall mean in original sample, Red dot = mean in simulated sample

In Figure 2, the simulated sample means are given by the red dots. This is an estimate of the true mean μ of the underlying distribution. To make the confidence interval we need to know how much the distribution of $\bar{x}$ varies around μ. That is, we would like to know the distribution of $\delta = \bar{x} - \mu$. If we knew the distribution we could find $\delta_{.025}$ and $\delta_{.975}$ the 0.025 and 0.975 quantiles of δ. Then we would have $P(\delta_{.025} \leq \bar{x} - \mu \leq \delta_{.975} \mid \mu) = 0.95 \leftrightarrow P(\bar{x} - \delta_{.025} \geq \mu \geq \bar{x} - \delta_{.975} \mid \mu) = 0.95$ which gives a 95% confidence interval of $[\bar{x} - \delta_{0.025}, \bar{x} - \delta_{.975}]$. With confidence intervals, the probabilities computed are probabilities concerning the statistic $\bar{x}$ given that the true mean is μ. The bootstrap principle offers a practical approach to estimating the distribution of $\delta = \bar{x} - \mu$. We approximate it by the distribution of $\delta^* = \bar{x}^* - \bar{x}$ where $\bar{x}^*$ is the mean of a bootstrap sample. Since δ^* is computed by resampling the original data, we simulate δ^* many times (in this example 1000 times). Hence, by the law of large numbers, the distribution of δ^* can be estimated with high precision. The quantiles $\delta_{.025}$ and $\delta_{.975}$ can be approximated by $\delta^*_{.025}$ and $\delta^*_{.975}$. Every bootstrap 95% confidence interval for μ is then $[\bar{x} - \delta^*_{0.025}, \bar{x} - \delta^*_{.975}]$. In this way, we provide a confidence interval that contains the true population mean with 95% chance.

Within this paper, different samples will be formed based on sex assigned at birth and on whether or not respondents identify with a transgender identity. The previously discussed techniques will be performed and discussed in terms of their value.

1.3 Statistics with a Fuzzy scale

Statistical reasoning is a specific and relevant instance of approximate reasoning, under uncertainty. It refers to the analysis of collective phenomena, namely phenomena which are defined with reference to a collection of empirical observations. The observed data may be affected by various types of uncertainty: 1) the measuring system; 2) the way of expressing the assessment of their measure (e.g. numerically, linguistically, by means of numerical intervals, etc.); 3) the way they are eventually selected from a larger population; 4) the possible vagueness in defining the underlying concepts through the use of observable variables (e.g. measuring an opinion by means of a visual analogue scale) (Coppi, D'Urso, & Giordani, 2006). Due to the heterogeneity within the concept gender identity, the last option seems to be the case when measuring gender identity. Some theoretical informational ingredients can be affected by uncertainty as well. For example, we may doubt the Gaussian assumption when studying the distribution of a given quantitative variable. Due to uncertainty, also the conclusions of the process are uncertain. Fuzziness may be adopted as a tool for coping with some types of uncertainty affecting statistical data (e.g. in case 4 of the above mentioned list). The notion of fuzzy random variable has been introduced to model random mechanisms generating imprecisely-valued data which can be properly described by means of fuzzy sets (Colubi, Coppi, D'Urso, & Gil, 2007). Observed data can be jointly affected by two sources of uncertainty: fuzziness (due to imprecision, vagueness), and randomness (due to sampling or measurement errors of stochastic nature). Suitable probability models constitute the usual tool for dealing with randomness due to sampling from finite or infinite populations. Randomness and fuzziness may act separately or jointly on the various informational ingredients of a statistical reasoning process (Zadeh, 1995). From a mathematical viewpoint they are respectively managed by means of probability and fuzzy sets theories. For this paper, we will use the notion of fuzzy random variables in the sense of Puri and Ralescu (1986).

It should be mentioned that there are other approaches to model fuzzy data in a random context. The concept of fuzzy random variable was first introduced by Kwakernaak (1978, 1979), and later formalized in a slightly different way by Kruse and Meyer (1987). Although mathematical conditions in the model stated by Puri and Ralescu and in that by Kwakernaak/Kruse and Meyer coincide for some relevant cases, the situations to be modelled essentially differ. Fuzzy random variables in Kwakernaak/Kruse and Meyer’s sense formalize either fuzzy perceptions or fuzzy descriptions of existing real-valued data generated by a random mechanism, and most of the statistical analysis refers to parameters and characteristics of the distribution of the original real-valued data. Instead, fuzzy random variables in Puri and Ralescu's sense were conceived to formalize random mechanisms which directly assign fuzzy values/labels, with no underlying original real-valued process behind. Of course, results and statistical methods using this last formal notion can be applied for the first one as well, but aim and scope frequently differ: the interest in Puri and Ralescu's approach will be usually focused on the parameters/characteristics of the (sometimes fuzzy) distribution of the fuzzy random mechanism.

As already mentioned above, one can frequently come across an underlying imprecision due to the vagueness of a concept such as gender identity. This imprecision can be modelled by means of fuzzy sets. Judgements like 'strongly disagree', 'disagree', 'neutral', 'agree', and 'strongly agree' are data that can be frequently encountered in many real-life situations in which randomness is involved in obtaining data. Most of these labels, which are essentially imprecise, can be suitably modelled by means of fuzzy sets of the space of real numbers. The methodology to be illustrated in this paper will allow us to manage these values/categories by exploiting all the information contained in their meaning, instead of only considering whether these values are or are not different or whether they occupy different positions in a ranking (as it is usually done in traditional statistics with categorical and ordinal data).The full exploitation of the information in imprecise values is achieved through the use of convenient distances between fuzzy sets.

1.3.1 Formalizing fuzzy data

fuzzy-valued data within this paper will be those belonging to the class

$$F_c(\mathbb{R}) = (U \mid \mathbb{R} \rightarrow [0, 1] \mid U_\alpha \text{ is a compact interval for all } \alpha \in [0,1]),$$

where U_α denotes the α-level of fuzzy set U, that is

$$U_\alpha = (x \in \mathbb{R} \mid U(x) \geq \alpha)$$

if $\alpha \in [0,1]$. This definition results in a trapezoidal or triangular fuzzy-valued number, consisting of four or three data points. The lowest and highest data point form the support (supp, U_0), the central data point(s) are considered as the core (U_1),

$$U_0 = cl(supp) = cl(x \in \mathbb{R} \mid U(x) > 0),$$

with cl(supp) representing the limits of the support for all the fuzzy-valued data. U(x) represents the degree of compatibility of x with the property defining U, or the degree of possibility of x being U (Colubi et al., 2007). Figure 3 provides more clarity on the different parameters of importance in estimating these fuzzy-valued data.

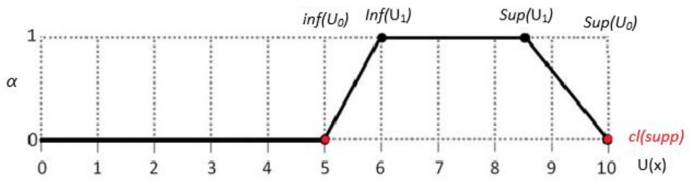


Figure 3. A fuzzy-valued trapezoidal datum with relevant parameters
Note. Inf(U_0) = infimum of level 0-set (support), Inf(U_1) = infimum of level 1-set (core), Sup(U_1) = supremum of level 1-set (core), Sup(U_0) = supremum of level 0-set (support), cl(supp) = limits of the support.

The space of fuzzy numbers which will model data is denoted by $F_c(\mathbb{R})$. In this way each fuzzy datum U ϵ $F_c(\mathbb{R})$ to be considered can be characterized by means of the family of compact intervals $([\inf U_\alpha, \sup U_\alpha])_{\alpha \epsilon [0,1]}$, with α being fixed at 0 and 1. Many distances between fuzzy data can be found in the literature (Diamond & Kloeden, 1994; Klement, L., & Ralescu, 1986; Puri & Ralescu, 1986). In order to analyze probabilistic and statistical aspects of fuzzy random variables, a metric which has been shown to be very convenient and easy to interpret is the D^{φ}_{W} distance (Bertoluzza, Corral, & Salas, 1995; Lubiano, Gil, Lopez-Dıaz, & Lopez-Garcıa, 2000). For U, V ϵ $F_c(\mathbb{R})$ the D^{φ}_{W} distance between these two fuzzy-valued numbers U and V is given by

$$D^{\varphi}_{W}(U,V) = \sqrt{\int_0^1 \int_0^1 [f_U(\alpha,\lambda) - f_V(\alpha,\lambda)]^2 dW(\lambda) d\varphi(\alpha)}$$

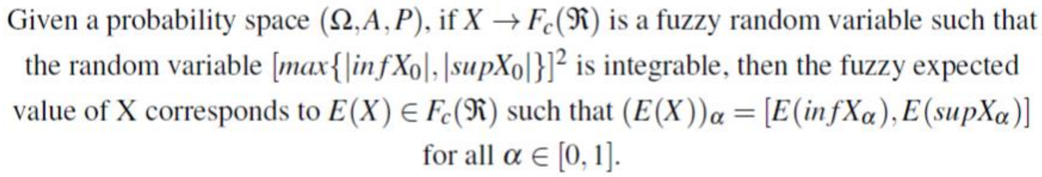
Where $f_U(\alpha,\lambda) = \lambda supU_\alpha + (1-\lambda) infU_\alpha$ for the fuzzy-valued number U, for the $f_V(\alpha,\lambda) = \lambda supU_\alpha + (1-\lambda) infU_\alpha$ fuzzy-valued number V, and W and φ are weighting measures which can be identified with probability measures on the measurable space ([0, 1], B[0, 1]): W being associated with a non—degenerate distribution, whereas φ has a distribution function which is strictly increasing on [0, 1]. ([0, 1], B[0, 1]) refers to the set of possible outcome values (ranging from 0 to 1) and a Borel σ-algebra, which refers to a collection Σ of subsets of the outcome set. Conditions for W and φ are imposed to guarantee that D^{Ψ}_{W} is in fact a metric, but it should be noted that the associated weights have not a stochastic meaning. Given a probability space (Ω, A, P) that models the considered random experiment, an associated random fuzzy number is a mapping X: Ω → $F_c(\mathbb{R})$ such that for all α ϵ [0, 1] the α-level mapping X_α is a compact random interval (that is, for all α ϵ [0,1] the real-valued mappings inf X_α and sup X_α are random variables) (Puri & Ralescu, 1986). Based on Colubi, Domınguez-Menchero, Lopez-Dıaz, and Ralescu (2001), if X: Ω → $F_c(\mathbb{R})$ is a fuzzy random number, then it is a D^{φ}_{W}-Borel-measurable mapping with respect to the Borel σ-field generated on $F_c(\mathbb{R})$. The

Borel measurability will enable to consider trivially the induced distribution of a random fuzzy number as well as the independence of fuzzy random numbers.

Since fuzzy data correspond to [0, 1]-valued mappings on $\mathbb{R}$, they are in fact functional data with a very intuitive interpretation. However, whereas functional data are often assumed to be Hilbert space-valued, the space of fuzzy data with the usual arithmetic is not linear. As a consequence, in the development of many probabilistic and statistical results one should take special care to guarantee that involved operations do not lead to elements out of $F_c(\mathbb{R})$ (Colubi et al., 2007; Klement et al., 1986). For this reason, the case of fuzzy data often requires a specialized analysis.

1.3.2 Relevant parameters of the distribution of a fuzzy random variable

The best known fuzzy parameter is the Aumann-type mean of a fuzzy random variable which is defined as a fuzzy-valued measure of the central tendency of a fuzzy random variable and defined as follows (Puri & Ralescu, 1986):



Given a probability space (Ω, A, P), if $X \rightarrow F_c(\Re)$ is a fuzzy random variable such that the random variable $[max\{|infX_0|, |supX_0|\}]^2$ is integrable, then the fuzzy expected value of X corresponds to $E(X) \in F_c(\Re)$ such that $(E(X))_\alpha = [E(infX_\alpha), E(supX_\alpha)]$ for all $\alpha \in [0, 1]$.

The use of fuzzy random variables can be best illustrated with an example. General practitioners are classified by patients in accordance with their ‘degree of agreeableness’. The labels assigned to general practitioners, namely, $\tilde{x}_1$ = ‘very low degree of agreeableness’, $\tilde{x}_2$ = ‘low degree of agreeableness’, $\tilde{x}_3$ = ‘medium degree of agreeableness’, $\tilde{x}_4$ = ‘high degree of agreeableness’, $\tilde{x}_5$ = ‘very high degree of agreeableness’, have not been assigned on the basis of an underlying real-valued magnitude, but rather on the basis of subjective judgement/perceptions of the patients. As a consequence, the classification process can be viewed as a fuzzy random variable *X* which takes on five values, $\tilde{x}_1$, $\tilde{x}_2$, $\tilde{x}_3$, $\tilde{x}_4$, and $\tilde{x}_5$ which can be described, for instance, in terms of *S*-curves as those in Figure 4.

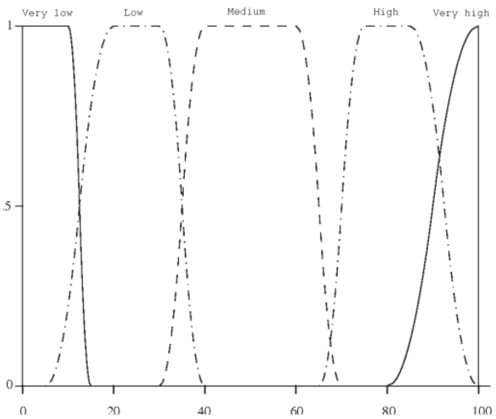


Figure 4. Values of the 'degree of agreeableness' of general practitioners

Healthcare insurance companies could be interested in the ‘mean degrees of agreeableness’ of general practitioners in a given area, that will be denoted by ζ_1, ζ_2 and ζ_3. For this purpose, they consider an overall sample of, for example, $n = 133$ patients, and observe fuzzy random variable X on three sub-samples of sizes $n_1 = 37$, $n_2 = 47$ and $n_3 = 49$. In this example, for illustrative purposes, the sample data will be considered as finite populations with the following distributions:

Table 2. Distribution of the degree of agreeableness of patients in three sub-samples in a given area

	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$	$\tilde{x}_4$	$\tilde{x}_5$
ζ_1	.19	.30	.29	.13	.09
ζ_2	.19	.22	.36	.19	.04
ζ_3	.14	.24	.25	.27	.10

Note: ζ_i = proportion of agreeableness in that area, $\tilde{x}_i$ = specified area

As an illustration of the idea of fuzzy mean, Figure 5 represents the distributions of Table 2 (which can be understood as multinomial distributions with fuzzy values).

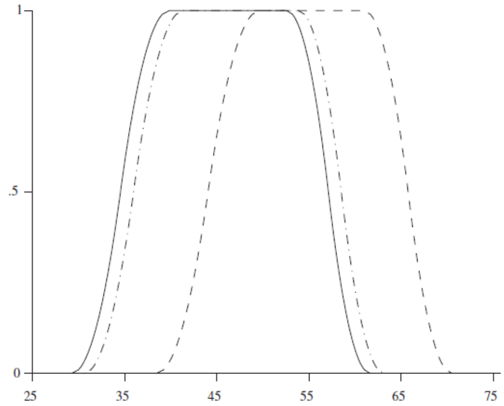


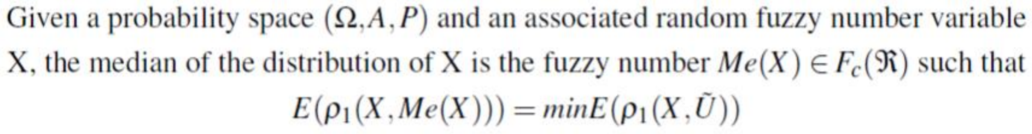
Figure 5. Mean (fuzzy) values of the 'degree of agreeableness' of general practitioners for the three considered distributions in the example

As can be seen in Figure 5, the mean agreeableness corresponding to the sample from ζ_1 (represented by means of a continuous curve ——) could be interpreted as to be 'rather low to slightly moderate'. The mean agreeableness corresponding to the sample from ζ_2 (represented by means of a dash-dot curve - . - .) could be interpreted as to be 'slightly low to rather moderate'. The mean agreeableness corresponding to the sample from ζ_3 (represented by means of a dashed curve ---) could be interpreted as to be 'moderate to rather high'. The distributions have a trapezoidal shape, which is typical for fuzzy data.

The Aumann-type mean value is the most common used value to get some idea about the central tendency of a sample or population of fuzzy data. Nevertheless, one should know that the Aumann-type fuzzy mean also inherits from the real-valued case the sensitivity of the mean to the existence of extreme values (outliers). Questioning gender identity often involves high values (very masculine, very feminine), so discussing and calculating the median can also be interesting. The median of a real-valued random variable is usually defined in two equivalent ways, namely: either as a middle position value with respect to a specified ranking, or as a value minimizing the mean distance to the distribution of the variable through an L^1-type metric. Since fuzzy numbers cannot be ranked through a universally acceptable total ordering, the second definition will be considered based on the metric ρ_1. This metric is called the 1-norm distance between fuzzy numbers and can be defined as follows:

$$\rho_1(\tilde{U},\tilde{V}) = \frac{1}{2}\int_0^1 (|inf\tilde{U}_\alpha - inf\tilde{V}_\alpha| + |sup\tilde{U}_\alpha - sup\tilde{V}_\alpha|)d\alpha$$

This distance ρ₁ is shown to be easy-to-handle for purposes of extending the notion of the median. Following Sinova, Gil, Colubi, and Van Aelst (2012), when specifying trapezoidal fuzzy data, the sample 1-norm median can be defined as follows:



Given a probability space (Ω, A, P) and an associated random fuzzy number variable X, the median of the distribution of X is the fuzzy number $Me(X) \in F_c(\Re)$ such that

$$E(\rho_1(X, Me(X))) = minE(\rho_1(X, \tilde{U}))$$

The lack of a universally acceptable total ordering between fuzzy numbers is overcome by using this L^1 type distance between fuzzy numbers based on the 1-norm and on the infimum/supremum (or on the support function) of the fuzzy numbers. The support function refers to all the points with nonzero membership. Unlike the median for random variables, the median for any random fuzzy number does not necessarily match any of the observed random fuzzy numbers. Calculations involved can be performed using specific R functions. As an example that confirms this claim and illustrates the calculation of the median, let's consider the random fuzzy number associated with the 'overall assessment' of agreeableness of one general practitioner, but now of 31 randomly selected patients (fictitious) for whom values are shown in Table 3. Trapezoidal fuzzy numbers for the median (Tra()) are calculated for each patient (Pt). The outermost two numbers refer to the support of the fuzzy number, the inner two numbers represent the core.

Table 3. Trapezoidal responses to the overall rating of agreeableness of 31 patients

Pt	**Tra()**	**Pt**	**Tra()**	**Pt**	**Tra()**
1	50-60-70-78	12	31-51-51-81	22	65-70-75-80
2	44-47-51-68	13	50-60-70-80	23	72-79-88-92
3	44-50-70-77	14	46-56-56-66	24	80-90-90-100
4	86-90-96-99	15	57-66-74-100	25	69-76-85-89
5	50-60-70-80	16	0-1-7-15	26	60-70-80-90
6	39-49-59-69	17	60-70-90-100	27	66-95-95-100
7	35-45-55-66	18	77-82-87-92	28	50-60-70-78
8	54-60-64-77	19	57-60-64-67	29	60-70-70-80
9	60-65-65-70	20	51-61-61-71	30	50-60-60-70
10	91-96-96-100	21	18-28-28-38	31	40-50-50-60
11	60-70-70-80				

Note: Pt = Patient, Tra() = Trapezoidal fuzzy number.

The corresponding median can be estimated using a large number of levels, according to the ideas of Trutschnig, Lubiano, and Lastra (2013), and is shown in Figure 6.

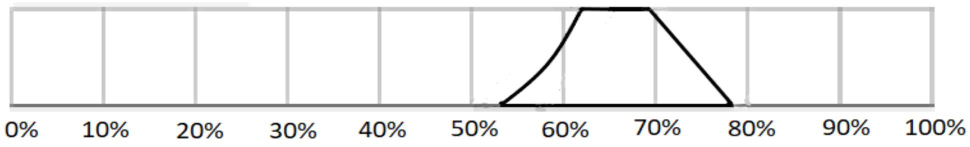


Figure 6. Median of the random fuzzy number overall rating of agreeableness

The fuzzy median associated with the data in Table 3 shown in Figure 6 does not match any of the data. Formally proven in Sinova et al. (2012), the 1-norm median is more robust than the average since it is somewhat less influenced by possible 'outliers' (in this case, some high values). Also this example shows that the mean is less robust than the median. For example, if the response of the 16th patient (who clearly represents an outlier in the sample) is removed from the dataset, the median hardly varies, while the mean increases by about 2 units (more specifically, the mean response for the 31 patients is Tra(54.07, 62.33, 69.69, 78.58), while once the 16th answer is removed the mean is equal to (55.06, 64.30, 71.50, 80.61).

Another fuzzy parameter is the absolute variation of a fuzzy random variable, which can be obtained, for instance, by expressing how much 'in error' a number is expected to be as a description of variable values. This error can be quantified in a natural way as follows (Körner, 1997):

Let $X : \Omega \to F_c(\Re)$ be a fuzzy random variable so that $[max\{|infX_0|,|supX_0|\}]^2$ is integrable, then, the variance of X is given by

$$Var(X) = E([D^{\varphi}_{W}(X, E(X))]^2 = \int_{\Omega} [D^{\varphi}_{W}(X(\omega), E(X))]^2 dP(\omega)$$

For the distributions in the example the absolute variation of X in the different finite populations can be quantified. For example, the variations are Var($X|(\omega_{1,1},\ldots,\omega_{1,37}))$ = 731.09, Var($X|(\omega_{2,1},\ldots,\omega_{2,47}))$ = 746.51, Var($X|(\omega_{3,1},\ldots,\omega_{3,49}))$ = 813.14. Since the scales for the fuzzy mean values are similar, absolute variations are comparable, hence it can be concluded that patients from Ω_1 and Ω_2 show close variations, whereas (in the absolute sense) X is slightly more variable over the patients from Ω_3 than in the previous two ones.

1.3.3 Estimation/testing on relevant parameters associated with fuzzy random variables

The sample fuzzy mean value $\bar{X}_n = 1/n\ (X_1 + \ldots + X_n)$ discussed in the previous section is an unbiased and consistent estimator of the fuzzy parameter E(X). The fuzzy mean of the fuzzy-values estimator $\bar{X}_n$ over the space of all random samples equals E(X) and $\bar{X}_n$ converges in probability almost-surely to E(X). Since the distributions in the example above correspond to random samples of patients, the samples can be utilized as samples from unknown populations Ω_1, Ω_2 and Ω_3. Thus, the (sample) mean values in Figure 5 are fuzzy estimates of the population mean values.

In testing the null hypothesis H_0: E(X) = U $\epsilon_c(\mathbb{R})$ at the nominal significance level $\alpha \,\epsilon\, [0,1]$, H_0 should be rejected whenever

$$\frac{[D^{\varphi}_{W}(\bar{X}_n, U)]^2}{\hat{S}^2_{(W,\varphi)_n}} > z_\alpha$$

where z_α is the 100(1 – α) fractile of the bootstrap distribution of

$$T_n = \frac{[D^{\varphi}_{W}(\bar{X}^*_n, \bar{X}_n)]^2}{\hat{S}^{*2}_{(W,\varphi)_n}}$$

with

$$\bar{X}_n = \frac{1}{n}(X_1 + ... + X_n)\ , \quad \hat{S}^2_{(W,\varphi)_n} = \sum_{i=1}^{n} \frac{[D^{\varphi}_{W}(X_i, \bar{X}_n)]^2}{(n-1)}\ , \quad \bar{X}^*_n = \frac{1}{n}(X^*_1 + ... + X^*_n)\ , \quad \hat{S}^{2*}_{(W,\varphi)_n} = \sum_{i=1}^{n} \frac{[D^{\varphi}_{W}(X^*_i, \bar{X}^*_n)]^2}{(n-1)}$$

(Colubi et al., 2007). For example, health insurance companies are interested in checking whether or not the ‘mean degree of agreeableness’ of general practitioners in each area is ‘medium/high’, where this value is assumed to be described by means of the fuzzy set described in Figure 7.

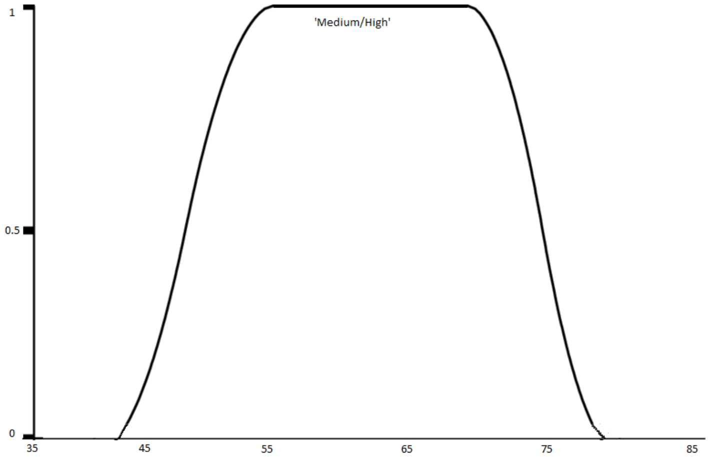


Figure 7. fuzzy value medium/high

By applying bootstrap techniques to test such a hypothesis on the basis of the available sample data, we get the *P*-values for Ω_1, Ω_2, and Ω_3, which are 0.004, 0.006 and .11. This means that, at significance level 0.05, the 'mean degree of agreeableness' in area Ω_3 can be accepted to be 'medium/high', whereas this is not sustainable in Ω_1 and Ω_2.

Besides inferences on the mean, inferences on the variance can also be interesting when testing hypotheses. Consider a fuzzy random variable $X : \Omega \rightarrow F_c(\mathbb{R})$ associated with the probability space (Ω, A, P) $[max\{|infX_0|, |supX_0|\}]^2$ and such that is integrable. Let $X_1,\dots,X_n$ be fuzzy random variables which are independent and identically distributed as X. Then the corrected sample variance

$$\hat{S}^2_{(W,\varphi)} = \frac{1}{n}\sum_{i=1}^{n}[D^{\varphi}_{W}(X_i, \bar{X}_n)]^2$$

is an unbiased and consistent estimator of Var(X): that is, the mean of the real-valued estimator $\hat{S}^2_{(W,\psi)}$ over the space of all random samples equals Var(X) and $\hat{S}^2_{(W,\psi)}$ converges in probability almost-surely to Var(X) (Colubi et al., 2007).

In testing the null hypothesis H_0: $Var(X) = \delta_0 \epsilon \mathbb{R}$ at the nominal significance level $\alpha \epsilon [0, 1]$, H_0 should be rejected whenever

$$\frac{\sqrt{n}|\frac{1}{n}\sum_{i=1}^{n}[D^{\varphi}_{W}(X_i, \bar{X}_n)]^2 - \delta_0|}{\widehat{S^2_{(X,\varphi)_n}}} > z_{\alpha}$$

,

where z_α is the $100(1 - \alpha)$ fractile of a N(0,1) distribution, where

$$\widehat{S^2_{(W,\varphi)_n}} = \frac{1}{n}\sum_{i=1}^{n}([D_W^{\varphi}(X_i,\bar{X}_n)]^2 - \frac{n-1}{n}\hat{S}^2_{(W,\varphi)_n})^2$$

.

Assume that health insurance companies believe that an absolute variation in the 'degree of agreeableness' of less than or equal to 730 can be permitted. The last test can be performed to test whether the variation in Ω_1, Ω_2, and Ω_3 fulfills such a condition. The corresponding *P*-values are given by .91, .61 and .13, hence the hypothesis that the variability is being admissible for each of the three areas can be accepted at the usual significance levels.

1.3.4 R analysis for Fuzzy scales

The implementation of Fuzzy linear regression methods in R can be facilitated by a number of designed packages, such as FuzzyNumbers, fuzzyreg, and SAFD (Gagolewski & Caha, 2019; Skrabanek & Martinkova, 2018; Trutschnig et al., 2013). The package FuzzyNumbers provides an excellent introduction into fuzzy numbers and offers great flexibility in designing trapezoidal fuzzy numbers. However, in order to implement Fuzzy linear regression with the package fuzzyreg, triangular fuzzy numbers (TFN) are needed. So the trapezoidal fuzzy numbers have to be simplified to vectors of length 3. The first element in these vectors specify the central value x_c, where the degree of membership is equal to 1. The second element is the left spread, which is the distance from the central value to a value x_l where the degree of membership is 0 ($x_l < x_c$). The left spread is thus equal to $x_c – x_l$. The third element is the right spread, i.e. the distance from the central value to a value x_r where $x_r > x_c$. The central value x_c is the core, and the interval (x_l, x_r) is the support of the TFN. Since we are working with non-symmetric fuzzy numbers (different spreads on the left and the right of the central value) Fuzzy least squares (FLS) will be used as method of analysis because it supports a simple Fuzzy linear regression for a non-symmetric triangular fuzzy variable. This probabilistic-based method calculates the fuzzy regression coefficients using least squares. The fuzzy regression models can be used to predict new data within the range of data used to infer the model. A model with three regression functions can be calculated: one for the central tendency of the fuzzy regression model, one for the lower boundary of the model support interval, and one for the upper boundary of the model support interval.

In testing the hypotheses for equal trapezoidal means and medians, functions will be written. In these functions, two new test statistics were calculated, not introduced in the previous section. These are the ‘degree of acceptance of the H_0’

$$(D(P > S))$$

and ‘the degree of rejection of the H_0’

$$(D(S > P) = 1 - D(P > S)$$

in accordance with the definition from Parchami, Taheri, and Mashinchi (2010). When the degree of rejection is higher than 0.05, we reject the null hypothesis. When the degree of acceptance is higher than 0.05, we cannot reject the null hypothesis. An accompanying fuzzy *P*-value was calculated as well, which can also have the shape of a trapezoidal fuzzy number (Parchami, Taheri, & Mashinchi, 2012). The shape refers to the interval within which the P-value lies.

At last, a one-way analysis of variance (ANOVA) model for fuzzy data is introduced with an F-test for fuzzy data (see Lin, Arbaiy, and Hamid (2017) for a detailed introduction). Two F-test statistics are calculated here: an F-statistic for the central point of the observations, and an F-statistic for the range of observations. For each of these F-statistics, a *P*-value can be calculated (under the F-distribution) and an overall *P*-value can be calculated as

$$P = (P_o + P_l) / 2,$$

With P_o the *P*-value for central point o and P_l the *P*-value for range l.

2. Research question and methodology

2.1 Research question and hypotheses

This study will examine a Likert scale and a Fuzzy scale when measuring gender identity, focusing on two predictors of gender identity, sex assigned at birth and identification with a cisgender/transgender identity. Previous research showed that gender identity is not a concept that can be easily categorized. For example, when respondents are asked how they identify themselves and different options are offered, a large number of respondents indicate that they cannot choose a specific option, they want to choose multiple options, or they want to choose a point between two options (Lubiano, de la Rosa de Sáa, et al., 2016; Meier & Motmans, 2020). Previous research has already shown that the use of a scale for measuring gender identity may therefore offer a solution ((Bakker & Vanwesenbeeck, 2007; Bockting et al., 2009; Dierckx et al., 2017; Schoonacker et al., 2009). The Likert scale, on the other hand, has a number of disadvantages that can complicate the use of this scale in the context of measuring gender identity. The Fuzzy scale, another type of scale that takes into account the uncertainty of a concept, might be a better option in research on gender identity, both at a population level and at a more population-specific level. This paper therefore examines the following research question: **Is the Fuzzy scale more suitable for mapping gender identity than the Likert scale?**

The use of a Fuzzy rating scale could yield somewhat different statistical conclusions. To investigate this, we will compare multiple groups with each other, and differences based on the type of scale will be examined. One the one hand, groups will be compared based on their sex assigned at birth (respondents AMAB versus AFAB). On the other hand, groups will be compared based on their identification with a cisgender or transgender identity (cisgender versus transgender respondents). Since a Fuzzy set scale provides much more information than a Likert scale (four points instead of one option are provided based on this type of scale), we therefore predicted that differences between groups will be more visible when a Fuzzy scale is used. In line with past work, we hypothesize that the spread (variance, standard deviation) of gender identity will be higher when using a Fuzzy set scale, in comparison to a Likert scale. Also, if differences between these groups exist, we hypothesize that differences will be more easily detected with a Fuzzy scale than with a Likert scale.

We aim to compare the values of the fuzzy descriptive measures in the preceding section with their counterparts for Likert-type data, and to conclude about the differences between the use of the two response scales. Different tests will be performed to evaluate the means, medians, and spreads of the different groups under investigation. The paper will study if these measures are similar over the two types of response scales, and between groups within the same type of response scale (with groups formed on the basis of SAAB and cisgender/transgender identity). Managing data on the same respondents makes it possible to perform valuable investigations and researchers can apply a variety of modeling techniques to explore important issues.

2.2 Methodology

The study was approved by the Committee for Medical Ethics of Ghent University Hospital. Social media was used (Facebook) to motivate respondents to participate in the study. Since the topic is a sensitive topic for some people, the survey was not promoted in closed Facebook groups. However, a separate, public Facebook page was created to promote the study. Several people shared the survey, in order to reach a sample as representative as possible.

The online and anonymous survey was constructed using KoBoToolbox. This is a free and open source suite of tools for data field collection. It allows to construct a questionnaire, with different types of questions (categorical, numeric, string). The survey was online for 1 month (between May 18 and June 17). The survey consisted of seven questions for each respondent (see Appendix A). The first question asked for informed consent. A short description of the study, followed by the requirement that a respondent had to meet (being at least 18 years of age), was given. The second question assessed sex assigned at birth, meaning the assigned sex on their original birth certificate. The third and fourth question questioned gender identity, on the one hand with a Likert scale, on the other hand with a Fuzzy scale. Since a previous, small-scale, pilot study already showed that a number of people indicated that the question with the Fuzzy scale was difficult to understand, the fifth question asked to what extent respondents found the use of a Fuzzy scale difficult. This was a categorical question with three answer options ('not difficult at all', 'somewhat difficult', 'very difficult'). An open question was also included so that respondents could add anything if they wanted to. The last question asked if the respondent identified (or had ever identified) as transgender. In that way, hypotheses regarding differences between cisgender and transgender respondents could be studied. To compare the values of descriptive measurements (such as the mean, median and variance) and to draw conclusions about the differences between the use of the scales, the survey made use of both scales.

The second question used a 5-point Likert scale, with the options ‘Very masculine’, ‘Somewhat masculine’, ‘Not masculine nor feminine’, ‘Somewhat feminine’, and ‘Very feminine’.

For the third question, a guideline for the mechanism to draw the value that better expresses a response according to a Fuzzy rating scale was followed (Hesketh, Pryor, & Hesketh, 1988):

1) A reference bounded interval was first considered. This is often chosen to be [0,10] or [0,100]. Within this paper, an interval of [0,10] with two decimal places was used. The endpoints are often labeled in accordance with their meaning, referring to the degree of agreement, satisfaction, quality, and so on. For this article, the endpoints were feminine and masculine depending on the sex assigned at birth. For example, if a respondent indicated that they were assigned female at birth, the endpoints were 0 = masculine to 10 = feminine (see Figure 8).

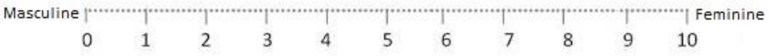


Figure 8. Step 1: A reference bounded interval with endpoints

2) Subsequently, the *support*, or 0-level set, was determined. It corresponds to the interval consisting of the actual values within the reference that are considered 'somewhat compatible' with the response (see Figure 9).

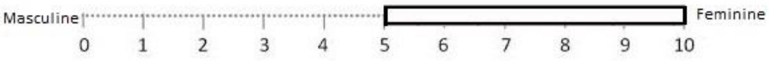


Figure 9. Step 2: The support, associated with the response

3) Then the *core*, or 1-level set, that belongs to the response was determined. It corresponds to the interval consisting of the actual values within the reference that are considered 'fully compatible' with the response (see Figure 10).

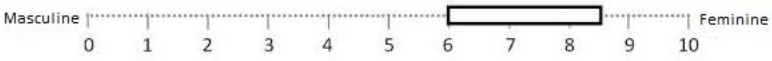


Figure 10. Step 3: The core, associated with the response

4) Ultimately, the two intervals can be ‘linearly interpolated’ to get a trapezoidal fuzzy number. This last step was not visible to respondents, because it is an optional step only to make the scale clearer.

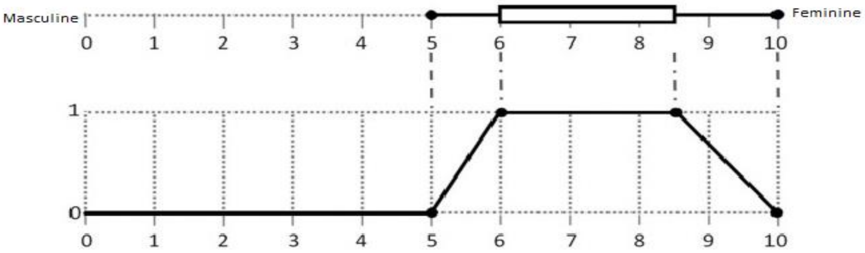


Figure 11. (Optional) step 4: trapezoidal fuzzy number

After data collection, data cleaning and data analysis was performed using R, version 3.6.3 (RStudio Team, 2020). Different statistical methodologies are needed to analyze responses from a Likert scale-based and a Fuzzy rating scale-based question. For some analyses, the items had to be reverse-scored. For example, when using the Likert scale, cisgender respondents AMAB and respondents AFAB were compared on a five-point Likert scale. To be able to investigate if these two groups identified differently (do respondents AMAB identify as strongly as masculine as respondents AFAB identify as feminine), the scores for respondents AMAB were reverse-scored (someone identifying as very masculine (score 1) will get a score of 5). In that way, a high score corresponds to strongly identifying with the sex assigned at birth, a low score corresponds to identifying with the gender opposite to the sex assigned at birth. R-packages called Likert (Bryer, 2016), FuzzyNumbers (Gagolewski & Caha, 2019), fuzzyreg (Skrabanek & Martinkova, 2018, 2019) and SAFD (Trutschnig et al., 2013) have been recently designed to help perform computations with Likert and random fuzzy sets. Statistical conclusions will differ depending on the considered scale, and practical implications from this fact will be discussed.

2.4 Data preparation and initial cleaning

The dataset used within this study can be found in Appendix B. The data-cleaning process excluded respondents who did not give their consent (*n* = 1), who terminated the survey after the question about sex assigned at birth (*n* = 27), and who did not provide four numbers for the Fuzzy scale (*n* = 8).

3. Results

The final sample consisted of *n* = 145 respondents, of which 54 assigned male at birth (AMAB; Q2_a. = Male) and 91 assigned female at birth (AFAB, Q2_a. = Female), 127 having a cisgender identity (Q7. = No) and 18 having a transgender identity (Q7. = Yes) (see Appendix A). On the basis of these two questions (Q2 and Q7), the four groups for analysis were formed. Respondents within the transgender group are those respondents who identify with a transgender identity.

First, participants' gender identity measured with a Likert scale was analyzed. Using a function to provide various statistics about Likert type items, we could see that more cisgender respondents AMAB chose for options in between or opposite to their sex assigned at birth than cisgender respondents AFAB (see Table 4).

Table 4. Respondents AMAB and AFAB choosing a category for their gender identity (Likert scale) (%)

	AMAB	**AFAB**
Masculine	86.96	5.00
Both/Neither	8.70	5.00
Feminine	4.35	90.00

Note: AMAB = assigned male at birth, AFAB = assigned female at birth.

When comparing cisgender respondents with transgender respondents, more transgender respondents chose the option in between masculine and feminine (see Table 5).

Table 5. Cisgender and transgender respondents choosing a category for their gender identity (Likert scale) (%)

	Cis	**Trans**
Masculine	34.92	44.44
Both/Neither	6.35	16.67
Feminine	58.73	38.89

Note: Cis = cisgender, Trans = transgender.

Means and standard deviations for each of the four groups are presented in Table 6. For respondents AMAB and AFAB, scores range from 1 (very masculine) to 5 (very feminine). For cisgender and transgender respondents, scores range from 1 (strong identification with gender opposite to the SAAB) to 5 (strong identification with SAAB).

Table 6. Means and standard deviations, measured with a 5-point Likert scale

	AMAB	AFAB	Cis	Trans
M	1.65	4.29	3.32	1.89
SD	0.82	0.78	0.79	0.90

Note: AMAB = assigned male at birth, AFAB = assigned female at birth, Cis = cisgender, Trans = transgender.

To permit a visualization of the Likert items, density plots were created that treat the Likert variable as a continuous variable (see Figure 12). As expected, the figures belonging to the respondents AMAB and AFAB show an opposite pattern. Respondents AMAB mostly identify with a masculine identity and respondents AFAB mostly identify with a feminine identity. The figures belonging to the cisgender and transgender respondents have been recoded first, in such a way that a low score represents not identifying with the sex assigned at birth and high scores represent identifying with the sex assigned at birth. This will make comparisons between groups more meaningful. Cisgender respondents scored their gender identity mostly in alliance with their sex assigned at birth, unlike transgender respondents, who more often identified their gender identity as opposite to their sex assigned at birth.

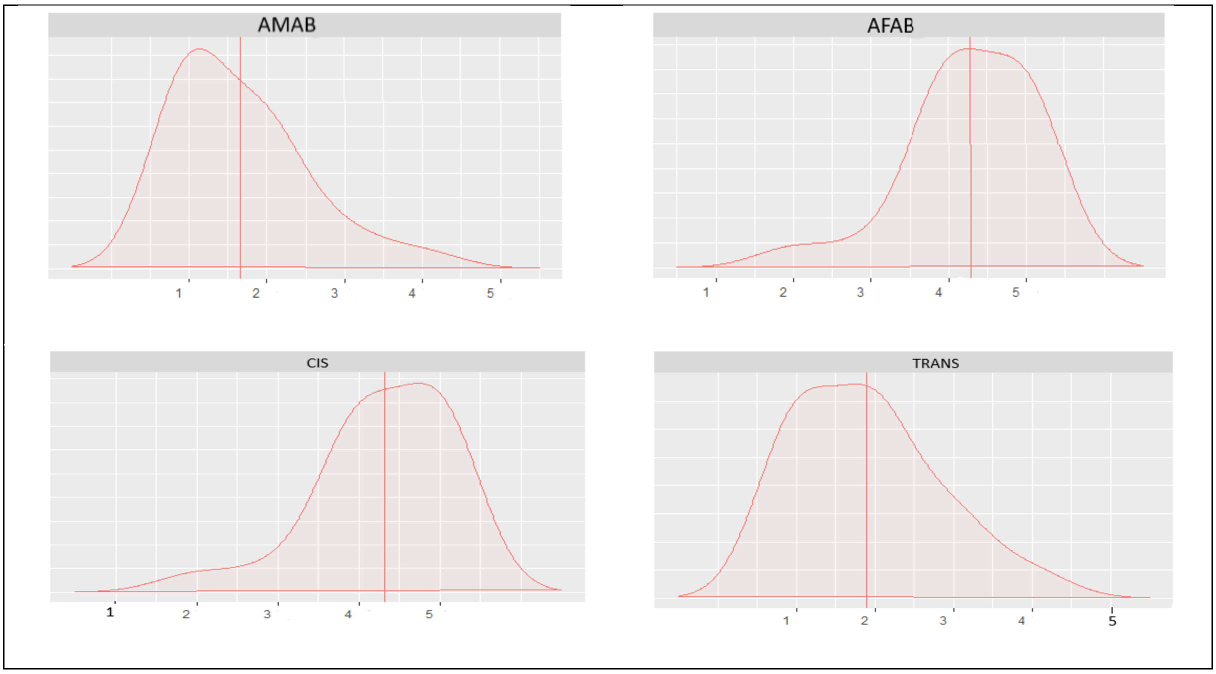


Figure 12. Likert density plots for each of the groups

Note: red line = mean, AMAB = assigned male at birth (top left), AFAB = assigned female at birth (top right), CIS = cisgender (bottom left), TRANS = transgender (bottom right).

Second, participants' gender identity was measured with a Fuzzy scale. Four numbers were obtained to form a trapezoidal fuzzy number, which means that the function's graph forms a trapezoid with the [0-1]-axis (see Figure 11). With the mathematics discussed in the literature study, means, standard deviations and medians could be calculated for each of the four points. fuzzy means, standard deviations, and medians can be found in Table 7. A score of 10 on the Fuzzy scale corresponds to strong identification with sex assigned at birth, a score of 0 corresponds to a gender identity opposite to sex assigned at birth.

Table 7. Trapezoidal means (M), standard deviations (SD) and medians (Md) for every group

	AMAB	**AFAB**	**CIS**	**TRANS**
M	Tra(6.67, 7.48, 8.69, 9.24)	Tra(6.43, 7.51, 8.45, 9.17)	Tra(6.52, 7.50, 8.54, 9.20)	Tra(2.30, 3.57, 5.24, 6.09)
SD	Tra(2.24, 1.89, 1.48, 1.43)	Tra(1.83, 1.51, 1.24, 1.09)	Tra(1.22, 1.66, 1.33, 1.22)	Tra(2.55, 3.02, 3.15, 2.94)
Md	Tra(7.05, 7.97, 8.99, 9.64)	Tra(6.95, 7.88, 8.61, 9.56)	Tra(7.00, 7.92, 8.72, 9.60)	Tra(1.74, 3.00, 4.74, 5.94)

Note: AMAB = assigned male at birth, AFAB = assigned female at birth, CIS = cisgender, TRANS = transgender.

These results were graphically interpolated in order to visually provide a clear view of the variation of gender identity within each group. An overview of the mean trapezoidal fuzzy number, as well as the fuzzy median within each group, is shown in Figure 13.

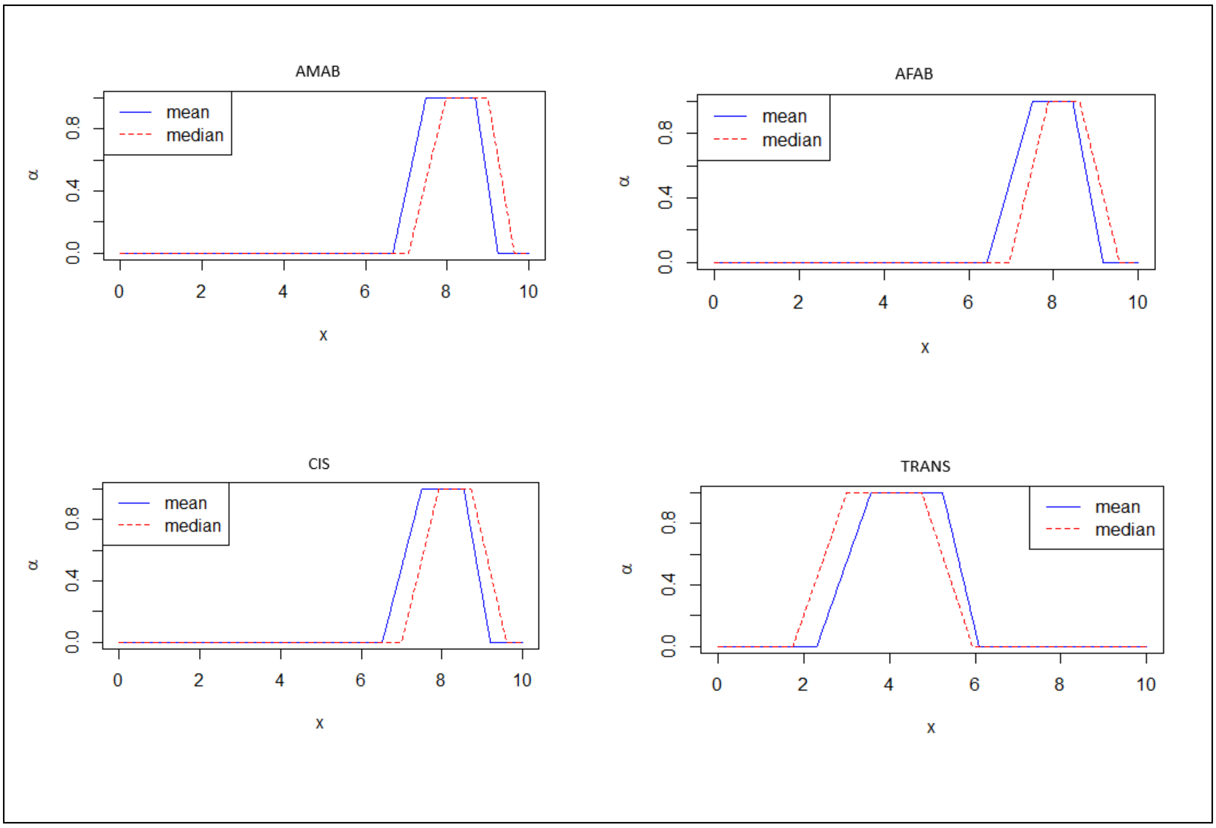


Figure 13. Trapezoidal mean and median for each of the groups

Note: AMAB = assigned male at birth (top left), AFAB = assigned female at birth (top right), CIS = cisgender (bottom left), TRANS = transgender (bottom right).

Next, in order to evaluate differences between groups, various tests were applied depending on the used scale. For the Likert type scale, the following tests were applied depending on the hypotheses of interest: a Levene's test, an Analysis Of Variance (ANOVA), a Kruskal-Wallis test, a Pearson Chi-Squared test, and a bootstrapping approach. For the Fuzzy scale, Fuzzy linear regression was applied first with a Fuzzy least squares method in order to provide a regression model able to predict gender identity in the AMAB and AFAB population, as well as in the transgender and cisgender population. Also, a function was designed in order to test for differences between means and medians. This function was based on the Aumann-type mean and the 1-norm median discussed in the literature study, and tested the hypothesis H_0 of equal means/medians between groups against the alternative hypothesis H_a of an absolute difference of means/medians > 1 (Parchami et al., 2012). For every respondent, trapezoidal fuzzy means and medians were calculated, with corresponding standard deviations. With significance $\alpha =$ .05 and sample sizes taken into account, a trapezoidal p-value could be plotted, together with a degree of acceptance/rejection of the H_0 in order to decide to accept the H_0 or the H_a. At last, an F-test was used to compare the between group variation with the within group variation.

3.1 AMAB - AFAB comparisons

To be able to compare cisgender respondents assigned male at birth and assigned female at birth on the Likert type scale, scores for respondents AMAB were reverse scored. In that way, a high score on the Likert scale indicates identifying as very masculine when AMAB and as very feminine when AFAB. First, a Levene's test to test homogeneity of variances was performed. The null hypothesis assuming equal variances could not be rejected ($F(1,124) = .26$, $p = .610$). The analysis of variance did not show a significant difference between the mean scores of both samples ($F(1,124) = .17$, $p = .683$). However, the residuals from the ANOVA model did not seem to follow a normal distribution. A Kruskal-Wallis test was applied to obtain more reliable results. No significant difference between the location parameters of the distributions of the scores between the samples could be found ($X^2(1) = .39$, $p = .532$). To perform a Pearson chi-squared test with the Likert scale treated as a categorical variable, three groups were formed for each of the five Likert scores. The first group was formed as those respondents identifying very or somewhat opposite to their sex assigned at birth, the second group were the respondents identifying with both a feminine and masculine or neither a feminine nor masculine identity. The third group was based on those respondents identifying somewhat or very with their sex assigned at birth. Again, no significant difference between the two groups could be found here ($X^2(2) = .68$, $p =$ .710). However, three cells had an expected frequency of less than five, which is one of the assumptions of a Chi-square test in order to obtain reliable results. Therefore a Fisher exact test was performed because it does not make the assumption of at least 5 expected frequencies in each cell. Also here, no significant difference between respondents AMAB and AFAB could be found (*p* = .748). Last, a bootstrapping approach was applied, because gender identity has not yet been clearly mapped in the population to this day. Since no differences were found between cisgender respondents AMAB and AFAB, a general bootstrap procedure could be used to estimate the mean value and distribution of gender identity in the population. The results of the bootstrapping procedure are presented in Figure 14. When 30 samples were randomly selected from the cisgender original sample, an overall mean of 4.31 was found. Between the 30 samples, the mean ranged between 4.15 and 4.45. The intervals ranged from 3.98 to 4.59. As expected with cisgender respondents, the average ranged between identifying somewhat or very with the sex assigned at birth and with 95% probability the true population average lies between these intervals.

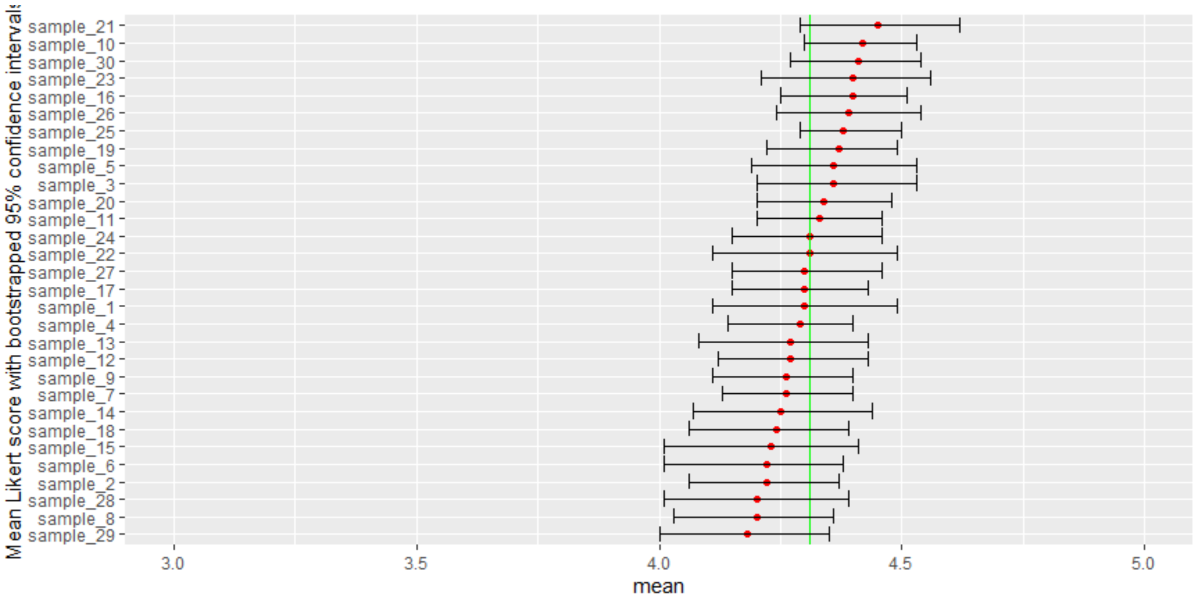


Figure 14. Bootstrapped mean confidence intervals (95%) for respondents AMAB and AFAB together

Note: Green line = overall mean in the original sample, Red dot = mean in simulated sample.

The package FuzzyNumbers (Gagolewski & Caha, 2019) provided an excellent way to construct a trapezoidal Fuzzy scale dataset. However, in order to use the package fuzzyreg (Skrabanek & Martinkova, 2018), a special case of fuzzy numbers was required, the triangular fuzzy numbers (TFN). The conversion from an object of the class FuzzyNumber to a TFN used in fuzzyreg required adjusting the core and the support values of the FuzzyNumber object to the central value and the spreads. After approximating TFNs for respondents AMAB and

AFAB, a Fuzzy least squares method was applied in order to estimate a model for gender identity. The following models were provided:

$$(1)\ Y_c = 7.48 + 0.03 * x$$

$$(2)\ Y_l = 6.67 - 0.24 * x$$

$$(3)\ Y_r = 9.24 - 0.07 * x$$

with (1) being the fuzzy regression model for the central tendency, (2) being the model support interval for the lower boundary and (3) the model support interval for the upper boundary. X is a dummy variable for sex assigned at birth, with AMAB as the reference category. With these models, it is predicted that persons AFAB score on average 0.03 units higher on the gender identity Fuzzy set scale, with a lower boundary on the left (- 0.24) and a lower boundary on the right (- 0.07), in comparison to persons AMAB.

In testing the hypotheses that the mean and median values are the same for respondents AMAB and AFAB, a function was written. With this function, significance was provided on the basis of 'the degree of acceptance/rejection), with α = 0.05 (Parchami et al., 2010). The H_0 stated that the absolute difference in trapezoidal means (and trapezoidal medians) between groups was 0. The alternative hypotheses (H_a) stated that the absolute difference was bigger than 0. For the absolute mean difference of the trapezoidal fuzzy numbers, a degree of acceptance *D(P>S)* = 0.99 was found. Therefore, the H_0 could not be rejected. There is no significant difference in the trapezoidal mean fuzzy values for respondents AMAB and AFAB. For the absolute median difference, a degree of acceptance *D(P>S)* = 0.95 was found. Again, the H_0 assuming equal trapezoidal medians, could not be rejected. Accompanying trapezoidal *P*-values are provided in Figure 15.

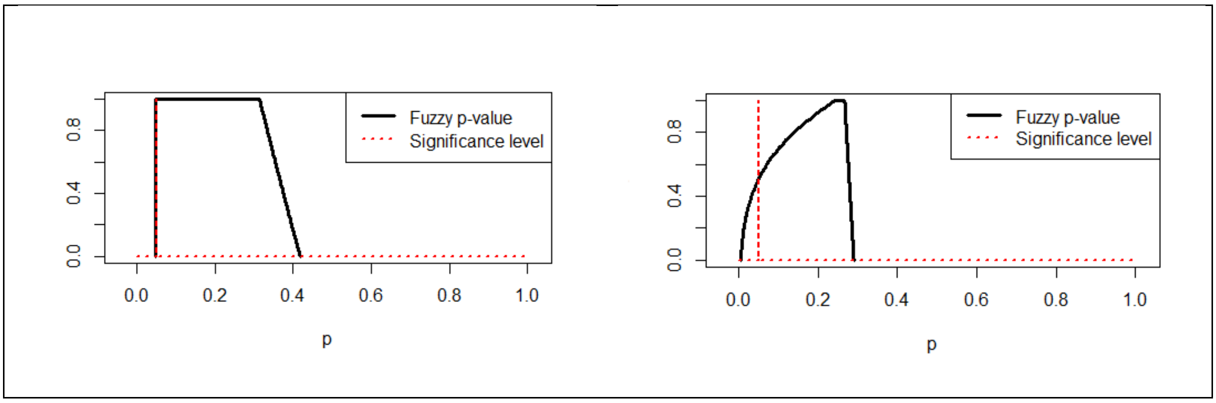


Figure 15. Fuzzy trapezoidal P-values for the difference in means (left) and medians (right) for respondents AMAB versus AFAB

At last, a One-Way ANOVA test was performed, comparing the between group variation with the within group variation of both AMAB and AFAB groups. The hypothesis H_0 assumes that there is no significant difference between both groups. This means that the between group variation is much smaller than the within group variation. For the central point of observations, we got the F-statistic $F_o(1,125) = .24$, and the P-value is $P_o = .628$. For the range of observations, we got the F-statistic $F_l(1,125) = 0.36$, and the P-value is $P_l = .552$. We get $P = (P_o + P_l) / 2 = .590 > 0.05$ under a 95% significance level. Hence, we do not reject the H_0. Between respondents AMAB and AFAB, a significant difference in between group variation, compared to within group variation, could not be observed.

3.2 Cisgender – Transgender comparisons

The same tests as in the previous section were used to compare the cisgender and transgender group. Again, AMAB respondents were recoded so that the respondents with a high score identify with their sex assigned at birth, and the respondents with a low score do not identify with their sex assigned at birth, regardless of a transgender or cisgender identity. A Levene's test was first applied to see if the variances between the two samples were equal. The Levene's test showed no significant result ($F(1,142) = .11$, $p = .745$). Then an ANOVA test was applied to look for a significant difference in the mean values for gender identity. Here the null hypothesis, which assumes equal mean values, could be rejected ($F(1,142) = 144.83$, $p <.001$). Because the distribution of the residuals was not normally distributed, a Kruskal-Wallis test was performed to check for a possible false positive. However, again a significant difference in gender identity was found between both groups ($X^2(1) = 45.53$, $p <.001$). The Likert scale was then categorized to see whether differences were still present when respondents were divided into three groups (a group that does not identify with their sex assigned at birth, a group that identifies with a male and/nor female sex assigned at birth, and the group that strongly identifies with their sex assigned at birth). The Chi-square test also showed that there was a significant association between Likert scale scores and identifying with a transgender or cisgender identity ($X^2(2) = 77.34$, $p <.001$). However, two cells contained less than five expected frequencies. Therefore a Fisher exact test was performed. The conclusion of the Fisher exact test was the same, there is a difference between the three groups for cisgender and transgender respondents ($p <.001$). Because a significant difference in gender identity was found with each test, a bootstrap procedure was used to estimate the mean for transgender respondents, with corresponding intervals with a 95% probability (see Figure 16). For the bootstrap results for cisgender respondents, see the previous section.

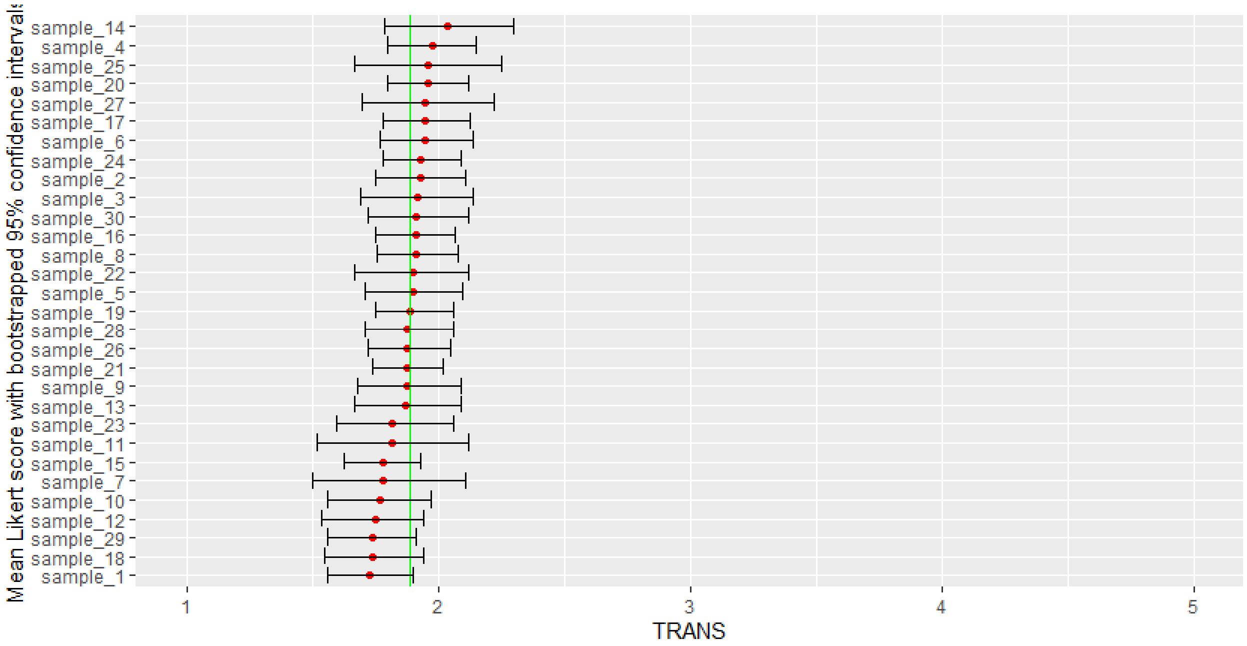


Figure 16. Bootstrapped mean confidence intervals (95%) for transgender respondents

Note: Green line = overall mean in the original sample, Red dot = mean in the simulated sample.

As predicted, the estimated mean of cisgender individuals is more on the side of identifying with the assigned sex at birth, while the estimated mean of transgender people is more on the side of identifying with the gender opposite to their sex assigned at birth. The estimated mean across all samples for transgender people ranged between 1.61 and 2.08, with 95% confidence intervals ranging from at least 1.42 to at most 2.41.

The package FuzzyNumbers (Gagolewski & Caha, 2019) and the package fuzzyreg (Skrabanek & Martinkova, 2018) was used again to apply a Fuzzy least squares method. After approximating TFNs for cisgender and transgender respondents, the FLS method was applied in order to estimate a model for gender identity. The following models were provided:

$$\text{(1) } Y_c = 3.57 + 3.93 * x$$

$$\text{(2) } Y_l = 2.3 + 4.22 * x$$

$$\text{(3) } Y_r = 6.09 + 3.11 * x$$

with (1) being the fuzzy regression model for the central tendency, (2) being the model support interval for the lower boundary and (3) the model support interval for the upper boundary. X is a dummy variable for transgender identification (yes/no), with transgender (yes) as the reference category. With these models, it is predicted that cisgender persons score on average

3.93 units higher on the gender identity Fuzzy set scale (with a range from 0-10), with a higher boundary on the left (+ 4.22) and a higher boundary on the right (+ 3.11), in comparison to transgender persons.

In testing the hypotheses that the mean and median values are the same for cisgender and transgender respondents the same function as discussed in the previous section was applied, with α = 0.05 (Parchami et al., 2010). The H_0 stated that the absolute difference in trapezoidal means (and trapezoidal medians) was 0. The alternative hypotheses (H_a) stated that the absolute difference was bigger than 0. For the absolute mean difference of the trapezoidal fuzzy numbers, a degree of rejection $D(S > P) = 1$ was found. Therefore, the H_0 could be rejected. There is a significant difference in the trapezoidal mean fuzzy values for cisgender versus transgender respondents. For the absolute median difference, a degree of rejection $D(S > P) = 1$ was found. Again, the H_0 stating equal trapezoidal medians, could be rejected. Accompanying trapezoidal *P*-values are provided in Figure 17.

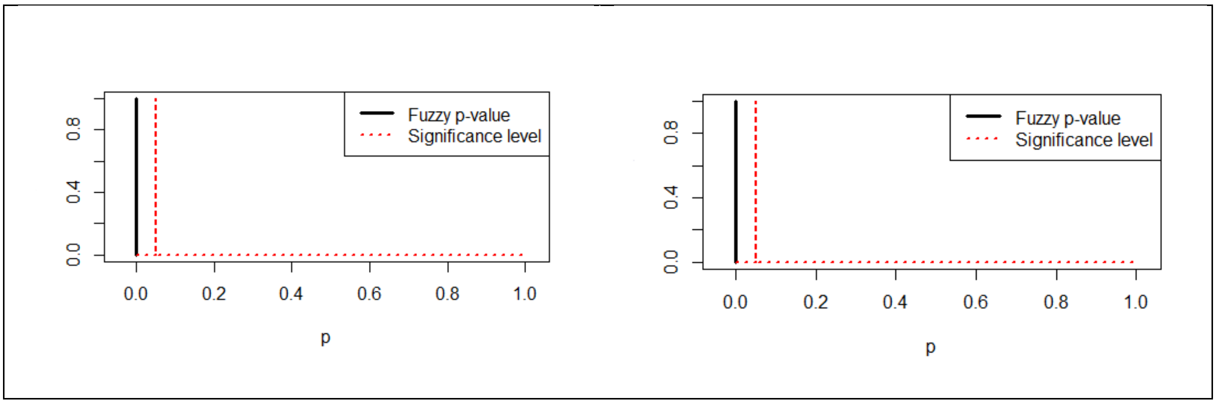


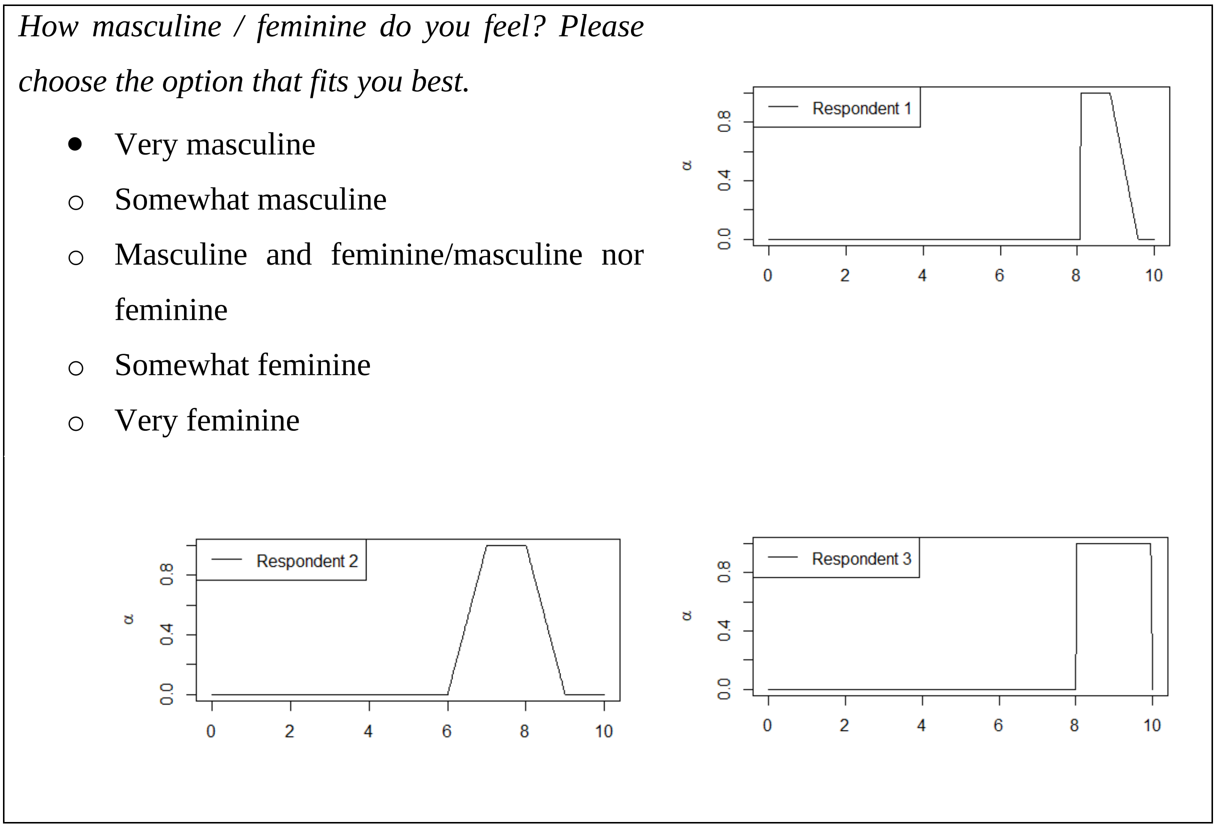
Figure 17. Fuzzy trapezoidal P-values for the difference in means (left) and medians (right) for cisgender versus transgender respondents

To finish, a One-Way ANOVA test was performed, comparing the between group variation with the within group variation of both cisgender and transgender respondents. The hypothesis H_0 assumes that there is no significant difference between both groups. This means that the between group variation is much smaller than the within group variation. For the central point of observations, an F-statistic of $F_o(1,143) = 80.41$ was found, with a *P*-value of $P_o < .001$. For the range of observations, an F-statistic of $F_l(1,143)= 6.63$ was found, with a *P*-value of $P_l = .011$. We get $P = (P_o + P_l) / 2 = .006 < 0.05$, under a 95% significance level. Hence, we do reject the H_0. Between transgender and cisgender respondents, a significant difference in between group variation, compared to within group variation, was observed.

4. Discussion

Quantitative data analysis is promising in terms of investigating gender identity. However, current methods in quantitative research are no longer in line with existing knowledge about gender identity. Most quantitative studies simply conceptualize gender identity in the same way as sex assigned at birth and question the concept through the same closed question with a binary answer option "male" and "female". Uncertainty about how gender diversity should be researched remains apparent. This paper aimed to compare the values of descriptive measures of fuzzy data with their counterparts for Likert-type data, and to descriptively conclude about the differences between the use of the scales, by using two types of responses. These two responses, measured with a 5-point Likert scale and a Fuzzy scale, were used to map gender identity. Four groups, on the one hand, respondents AMAB and AFAB, on the other hand, cisgender and transgender respondents, were compared on both scales. A number of factors limit our ability to make a conclusive determination of the results. Some of these factors are internal to the study (small transgender sample size, issues of representativeness), whereas others have to do with the lack of research on gender identity in general and the dearth of research among transgender and gender non-binary communities in particular.

Regardless of these limitations, some results confirmed our initial hypotheses. Firstly, standard deviations measured with a Fuzzy scale were indeed higher than when measured with a Likert scale, which indicated that the Fuzzy rating scale is much richer and more expressive, and it captures a higher subjectivity and variability in responding, than Likert ones. To support and illustrate this assertion, one can consider a combined graphical display of a double response to the questions about gender identity. The Likert scale-based response chosen by three cisgender respondents AMAB corresponded to Q3 = ‘Very masculine’, and the fuzzy rating scale-based responses for the same respondents are definitely different (see Figure 18). This evidence supports the use of the Fuzzy rating scale since the richness and diversity/variability/subjectivity of the available information clearly increase w.r.t the Likert scale.



How masculine / feminine do you feel? Please choose the option that fits you best.

- ● Very masculine
- ○ Somewhat masculine
- ○ Masculine and feminine/masculine nor feminine
- ○ Somewhat feminine
- ○ Very feminine

Figure 18. Example of three double responses to Q3 for which the Likert-type ones coincide while the Fuzzy-type ones clearly differ

Secondly, statistical conclusions were somewhat different depending on the scale. With the Likert scale, a Levene's test, a One-way ANOVA, a Kruskal-Wallis test, a Pearson X² test (Fisher exact test), and a bootstrapping approach were applied. For respondents AMAB versus AFAB, no significant difference in gender identity scores could be found (see Table 8 for an overview of results). A bootstrapping approach predicted mean and confidence intervals for the true population mean (with a 95% significance level). With the Fuzzy scale, a Fuzzy linear regression with a Fuzzy least squares method, a function to test differences between means and medians, as well as a One-Way ANOVA test were applied. Also here, no significant differences in gender identity scores were found between respondents AMAB and AFAB (see Table 8). For cisgender and transgender respondents, significant differences in gender identity scores were found on both scales (see Table 9 for an overview of results). The hypothesis that if there were differences between groups in terms of gender identity, they would be noticed by a Fuzzy scale rather than a Likert scale, was not confirmed. When comparing respondents, a high score was conceptualized as identifying strongly with the SAAB, a low score was conceptualized as identifying strongly with the gender opposite to the SAAB. The comparisons between AMAB and AFAB respondents showed that there are no significant differences between the two groups in terms of their gender identity. Both respondents AMAB and AFAB showed a similar pattern of responses, hence the non-significant results. Differences between cisgender and transgender respondents were significant, but these differences were detected by both scales. For a concept such as gender identity, it does not seem necessary to use a Fuzzy rating scale. To provide additional support for this claim, two questions were added to the questionnaire based on the pilot study.

Table 8. Summary of the results from the different tests used to compare respondents AMAB versus AFAB

LIKERT	*Test statistic* (*p*-value)
Levene's Test	$F(1,124) = .26$, $p = .610$
Analysis of Variance	$F(1,124) = .17$, $p = .683$
Kruskal-Wallis test	$X^2(1) = .39$, $p = .532$
Pearson's Chi-squared test	$X^2(2) = .68$, $p = .710$
Fisher's Exact Test	$p = .748$
FUZZY	
Function mean	$D(P>S) = 0.99$
Function median	$D(P>S) = 0.95$
Analysis of Variance	$F_o(1,125) = .24$, $F_l(1,125)= 0.36$, $p = .590$

Table 9. Summary of the results from the different tests used to compare cisgender versus transgender respondents

LIKERT	*Test statistic* (*p*-value)
Levene's Test	$F(1,142) = .11$, $p = .745$
Analysis of Variance	$F(1,142) = 144.83$, $p < .001$***
Kruskal-Wallis test	$X^2(1) = 45.53$, $p < .001$***
Pearson's Chi-squared test	$X^2(2) = 77.34$, $p < .001$***
Fisher's Exact Test	$p < .001$***
FUZZY	
Function mean	$D(S > P) = 1$***
Function median	$D(S > P) = 1$***
Analysis of Variance	$F_o(1,143) = 80.41$, $F_l(1,143)= 6.63$, $p = .006$**

*Note: * p < .05, ** p < .01, *** p < .001.*

The first question inquired about the difficulty of the gender identity question measured with a Fuzzy scale. Respondents could choose between three answer options: 'Not difficult at all', 'Somewhat difficult', and 'Very difficult'. 26.90% (*n* = 39) answered this question with 'Very difficult', as well as a 51.72% (*n* = 75) who found this question 'Somewhat difficult'. The qualitative results confirmed these findings: They highlighted that respondents did not understand the difference between the two 1-level points (the core) and the two 0-level points (the support), as well as some respondents stating that they had to read the question five times. The increase in variability when using a Fuzzy scale is not as important as the ability of respondents to answer questions correctly. Problems with understanding a specific type of scale increases the chance of bias, which can be a possible explanation for the increase in variability as well.

Another observation from the results, and in line with previous research, shows that the binary approach falls short when we want to map a concept like gender identity. When describing cisgender respondents AMAB and AFAB, 8.70% and 5.00% indicated to identify with both a masculine and feminine gender identity, or neither a masculine nor a feminine gender identity. This group, often referred to as gender non-binary persons, do not fit in a binary gender model. When using a binary question, those respondents would be wrongly placed within the category of a female/male gender identity, or would be treated as missing because they leave the gender identity question unanswered. The distribution of percentages also contrasts with transgender studies in this field, where persons AFAB identify as non-binary more often, compared to persons AMAB (Burgwal et al., 2019; Dierckx et al., 2017; Rosser, Oakes, Bockting, & Miner, 2007). The results also showed that transgender respondents more often identify as non-binary (16.67%) in comparison to cisgender respondents (6.35%). Previous research has proved that individuals with a transgender binary gender identity score very differently on health and well-being, compared to non-binary individuals (Burgwal et al., 2019; Fundamental Rights Agency, 2020; Harrison et al., 2012; Warren et al., 2016). Since this group seems to be an important group, with different outcome characteristics, future research should focus on this group in particular, in combination with a continuous approach on gender identity. In this way, we would eventually be able to predict outcome measures based on the position of a person on the scale.

The main implication from this paper: statistical conclusions can somewhat differ depending on the scale. However, in the case of gender identity, a 5-point Likert scale seems to make similar conclusions as to whether differences between groups exist as a Fuzzy scale. As a summary implication, this paper does not seem to fully corroborate what has been stated from other statistical perspectives (see e.g. de la Rosa de Sáa, Gil, García, and Lubiano (2013) and Lubiano, Montenegro, Sinova, de la Rosa de Sáa, and Gil (2016)): Responses on Fuzzy rating scales do not fully coincide with those based on responses from either Likert scales. However, when comparing groups on both scales, the same conclusions can be made.

To improve the analysis of this paper, some suggestions can be made. The sample size of the transgender group was small (*n* = 18). Therefore, this group might not be very representative of the transgender population in general. Thus, future research should focus on gathering more transgender respondents to obtain a sample size large enough to be representative. The data was collected through convenience sampling. The online survey has been distributed via Facebook, which may again introduce some bias. In order to obtain the most representative sample possible, it would be better, for example, to select random respondents from the national register. Obtaining a representative sample for each of the groups (in a cost-effective manner) can be done in several ways (see e.g. Scott and Wild (1986) and Anderssen and Malterud (2017)). At last, the two types of scales were compared descriptively, not analytically. Differences between groups were studied in order to provide an answer to the research question. However, analysis could be designed to compare the Likert-type responses with the Fuzzy rating scales. How this could be done, is outside the scope of this paper. There already has been some research about encoding Likert scales to a Fuzzy encoding (Calcagnì & Lombardi, 2014; Villacorta, Masegosa, Castellanos, & Lamata, 2014; Wang, Liu, & Zhang, 2014), but future research should expand the existing literature and subsequently apply it to the study of gender identity.

This paper has explained in detail an approach to descriptively analyze data obtained from the use of a Fuzzy rating scale-based questionnaire. It should be remarked that there are many other studies to be developed, although they are beyond the extend and length of this paper and will also depend in practice on the real interests users can have. Among them, there are still many statistical methods to be developed for both descriptive and inferential fuzzy data analysis, and this is a clear future direction to consider. Also, by optimizing knowledge of researchers about the use of a Fuzzy scale and the design of a Fuzzy scale within a questionnaire, this scale may become common practice when appropriate. A manual on how to set up a Fuzzy scale, for example, would make its use more popular. However, when studying gender identity, the 5-point Likert scale appears to be a good way to map gender identity. Further future research should therefore focus on developing an appropriate continuous scale for gender identity rather than categorizing respondents into groups. How gender identity can be represented best (with a 5-point or 7-point Likert scale, as a continuum from male to female or as two continuums, one for male and one for female) remains to be thoroughly investigated.

5. References

Åhs, J. W., Dhejne, C., Magnusson, C., Dal, H., Lundin, A., Arver, S., . . . Kosidou, K. (2018). Proportion of adults in the general population of Stockholm County who want gender-affirming medical treatment. *PLOS ONE, 13*(10), e0204606. doi:10.1371/journal.pone.0204606

Allen, I. E., & Seaman, C. A. (2007). Likert Scales and Data Analyses. *Quality Progress, 1*(1).

Anderssen, N., & Malterud, K. (2017). Oversampling as a methodological strategy for the study of self-reported health among lesbian, gay and bisexual populations. *Scandinavian Journal of Public Health, 45*(6), 637-646. doi:10.1177/1403494817717407

Angus, J. (2012). Gender, Sex, and Health Research: Developments and Challenges. *Canadian Journal of Nursing Research, 44*(3), 3-5.

Australian Bureau of Statistics. (2016). Standard for sex and gender variables. Retrieved from https://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/1200.0.55.012Main%20Features242016?opendocument&tabname=Summary&prodno=1200.0.55.012&issue=2016&num=&view=).

Bakker, F., & Vanwesenbeeck, I. (2007). *Seksuele gezondheid in Nederland 2006* (Eburon Ed.). Delft: Rutgers Nisso Group.

Balarajan, M., Gray, M., & Mitchell, M. (2011). *Monitoring equality: Developing a gender identity question (Equality and Human Rights Commission Research Report 75)*. Retrieved from Manchester: https://www.equalityhumanrights.com/sites/default/files/rr75_final.pdf

Bertoluzza, C., Corral, N., & Salas, A. (1995). On a new class of distances between fuzzy numbers. *Mathware & Soft Computing, 2*, 71–84. Retrieved from https://upcommons.upc.edu/bitstream/handle/2099/2462/berto.pdf

Blackless, M., Charuvastra, A., Derryck, A., Fausto-Sterling, A., Lauzanne, K., & Lee, E. (2000). How sexually dimorphic are we? review and synthesis. *American Journal of Human Biology, 12*(2), 151-166. doi:10.1002/(sici)1520-6300(200003/04)12:2<151::aid-ajhb1>3.0.co;2-f

Blalock, H. M. J. (1997). *Social statistics*. New York: Mc-Graw-Hill.

Bockting, W. O. (1999). From construction to context: Gender through the eye of transgendered. *Siecus Report, 28*(1), 3-7. Retrieved from http://www.siecus.org/_data/global/images/SIECUS%20Report%202/28-1.pdf

Bockting, W. O. (2008). Psychotherapy and the real-life experience: From gender dichotomy to gender diversity. *Sexologies, 17*(4), 211-224. doi:10.1016/j.sexol.2008.08.001

Bockting, W. O., Benner, A., & Coleman, E. (2009). Gay and bisexual identity development among female-to-male transsexuals in North America: Emergence of a transgender sexuality. *Archives of Sexual Behavior, 38*(5), 688-701. doi:10.1007/s10508-009-9489-3

Bryer, J. (2016). *Package 'likert'*. Retrieved from

Burgwal, A., Motmans, J., Vidic, J., Nieto, I. G., Gvianishvili, N., Kata, J., . . . Köhler, R. (2019). Health disparities between binary and non binary trans people: A community-driven survey. *International Journal of Transgenderism, 20*(2-3), 218-229. doi:10.1080/15532739.2019.1629370

Calcagnì, A., & Lombardi, L. (2014). Dynamic Fuzzy Rating Tracker (DYFRAT): a novel methodology for modeling real-time dynamic cognitive processes in rating scales. *Applied Soft Computing, 24*, 948-961. doi:10.1016/j.asoc.2014.08.049

Callens, N., Longman, C., & Motmans, J. (2017). *Het eerste sociaalwetenschappelijk onderzoek naar de zorgen sociale situatie van personen en ouders van kinderen met intersekse/DSD in Vlaanderen*. Retrieved from Gent:

Castleberry, J. (2019). Addressing the Gender Continuum: A Concept Analysis. *Journal of Transcultural Nursing, 30*(4), 403-409. doi:10.1177/1043659618818722

Collin, L., Reisner, S. L., Tangpricha, V., & Goodman, M. (2016). Prevalence of transgender depends on the "case" definition: a systematic review. *Journal of Sexual Medicine, 39*(6), 613-626. doi:10.1016/j.jsxm.2016.02.001

Colman, A. M., Norris, C. E., & Preston, C. C. (1997). Comparing Rating Scales of Different Lengths: Equivalence of Scores from 5-Point and 7-Point Scales. *Psychological Reports, 80*(2), 355-362. doi:10.2466/pr0.1997.80.2.355

Colubi, A., Coppi, R., D'Urso, P., & Gil, M. A. (2007). Statistics with fuzzy random variables. *International Journal of Statistics, 65*(3), 277-303. Retrieved from https://pdfs.semanticscholar.org/cb39/05e36b1c29d9ea4f882809d77a9de1db23bd.pdf

Colubi, A., Domınguez-Menchero, J. S., Lopez-Dıaz, M., & Ralescu, D. A. (2001). On the formalization of fuzzy random variables. *Information Sciences, 133*(1-2), 3–6. doi:10.1016/s0020-0255(01)00073-1

Conron, K. J., Scott, G., Stowell, G. S., & Landers, S. J. (2012). Transgender health in Massachusetts: results from a household probability sample of adults. *American Journal of Public Health, 102*(1), 188-122. doi:10.2105/ajph.2011.300315

Coppi, R., D'Urso, P., & Giordani, P. (2006). Component models for fuzzy data. *Psychometrika, 71*(4), 733-761. doi:10.1007/s11336-003-1105-1

Cox, E. P. (1980). The Optimal Number of Response Alternatives for a Scale: A Review. *Journal of Marketing Research, 17*(4), 407-422. doi:10.1177/002224378001700401

Crissman, H. P., Berger, M. B., Graham, L. F., & Dalton, V. K. (2017). Transgender Demographics: A Household Probability Sample of US Adults, 2014. *American Journal of Public Health, 107*(2), 213-215. doi:10.2105/ajph.2016.303571

Davis, G. (2014). The power in a name: diagnostic terminology and diverse experiences. *Psychology & Sexuality, 5*(1), 15-27. doi:10.1080/19419899.2013.831212

de la Rosa de Sáa, S., Gil, M. A., García, M. T. L., & Lubiano, M. A. (2013). Fuzzy Rating vs. Fuzzy Conversion Scales: An Empirical Comparison through the MSE. In *Synergies of Soft Computing and Statistics for Intelligent Data Analysis* (pp. 135-143).

Diamond, P., & Kloeden, P. (1994). Metric Spaces of Fuzzy Sets: Theory and Applications [Monograph]. doi:10.1142/2326

Dierckx, M., Meier, P., & Motmans, J. (2017). “Beyond the Box”: A Comprehensive Study of Sexist, Homophobic, and Transphobic Attitudes Among the Belgian Population. *Journal of Diversity and Gender Studies, 4*(1), 5-34. doi:10.11116/digest.4.1.1

Equality Act 2010. (2010). Equality Act 2010,. Retrieved from http://www.legislation.gov.uk/ukpga/2010/15/contents

Fausto-Sterling, A. (2000). *Sexing the body: gender politics and the construction of sexuality*. New York: Basic Books.

Finstad, K. (2010). Response interpolation and scale sensitivity: Evidence against 5-point scales. *Journal of Usability Studies, 5*(3), 104-110. doi:10.5555/2835434.2835437

Flores, A. R., Brown, T. N. T., & Herman, J. L. (2016). *How many adults identify as transgender in the United States*. Retrieved from Los Angeles, CA: https://williamsinstitute.law.ucla.edu/wp-content/uploads/Race-Ethnicity-Trans-Adults-US-Oct-2016.pdf

Fundamental Rights Agency. (2020). *A long way to go for LGBTI equality*. Retrieved from Luxembourg: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2020-lgbti-equality-1_en.pdf

Gagolewski, M., & Caha, J. (2019). *Package ‘FuzzyNumbers’*. Retrieved from http://www.gagolewski.com/software/

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin, 87*(3), 564-567. doi:10.1037/0033-2909.87.3.564

Gates, G. J. (2011). *How many people are lesbian, gay, bisexual and transgender?* Retrieved from Los Angeles, CA: https://williamsinstitute.law.ucla.edu/wp-content/uploads/Gates-How-Many-People-LGBT-Apr-2011.pdf

Gil, M. A., & González-Rodríguez, G. (2012). Fuzzy vs. Likert Scale in Statistics. In *Combining Experimentation and Theory* (pp. 407-420).

Goffman, E. (1963). *Stigma: Notes on the Management of Spoiled Identity*. London: Penguin.

Goodman, M., Adams, N., Cornell, T., Kreukels, B., Motmans, J., & Coleman, E. (2019). Size and Distribution of Transgender and Gender Nonconforming Population. A Narrative Review. *Endocrinology & Metabolism Clinics of North America, 48*(1), 303-321. doi:10.1016/j.ecl.2019.01.001

Grant, J. M., Mottet, L. A., Tanis, J., Harrison, J., Herman, J. L., & Keisling, M. (2011). *Injustice at every turn: A report of the national transgender discrimination survey*. Retrieved from National Gay and Lesbian Task Force website: http://www.thetaskforce.org/static_html/downloads/reports/reports/ntds_full.pdf

Harrison, J., Grant, J., & Herman, J. L. (2012). *A gender not listed here: Genderqueers, gender rebels, and otherwise in the National Transgender Discrimination Survey*. Retrieved from Los Angeles, CA:

Herdt, G. (1996). *Third sex third gender*. New York: Zone.

Hesketh, T., Pryor, R., & Hesketh, B. (1988). An application of a computerized fuzzy graphic rating scale to the psychological measurement of individual differences. *International Journal of Man-Machine Studies, 29*(1), 21-35. doi:10.1016/s0020-7373(88)80029-4

Homans, L. (2014). *Beleidsnota Gelijke Kansen 2014-2019*. Brussel: Vlaams Parlement Retrieved from http://docs.vlaamsparlement.be/pfile?id=1052026

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

Keuzenkamp, S. (2012). *Worden wie je bent. Het leven van transgenders in Nederland*. Retrieved from Den Haag: https://www.scp.nl/Publicaties/Alle_publicaties/Publicaties_2012/Worden_wie_je_bent

Klement, E. P., L., P. M., & Ralescu, D. A. (1986). Limit theorems for fuzzy random variables. *Proceedings of the Royal Society, 407*, 171–182. doi:10.1098/rspa.1986.0091

Komorita, S. S., & Graham, W. K. (1965). Number of Scale Points and the Reliability of Scales. *Educational and Psychological Measurement, 25*(4), 987-995. doi:10.1177/001316446502500404

Körner, R. (1997). On the variance of fuzzy random variables. *Fuzzy Sets and Systems, 92*(1), 83-93. doi:10.1016/s0165-0114(96)00169-8

Kruse, K., & Meyer, K. D. (1987). Descriptive statistics with vague data. In *Statistics with vague data* (pp. 71-130).

Kuyper, L., & Wijsen, C. (2014). Gender identities and Gender Dysphoria in the Netherlands. *Archives of Sexual Behavior, 43*(2), 377-385. doi:10.1007/s10508-013-0140-y

Kwakernaak, H. (1978). Fuzzy random variables I: Definitions and theorems. *Information Sciences, 15*(1), 1-29. doi:10.1016/0020-0255(78)90019-1

Kwakernaak, H. (1979). Fuzzy random variables II: Algorithms and examples for the discrete case. *Information Sciences, 17*(3), 253-278. doi:10.1016/0020-0255(79)90020-3

Leroy, E. (2019). *De seksuele gezondheid van cisgender- en transgender personen in Vlaanderen: Een kwantitatieve studie naar de socio-demografische verschillen en predictoren van seksueel gedrag, seksueel plezier en seksuele tevredenheid.* (Master of Arts in Gender and Diversity). Universiteit Gent, Gent.

Likert, R. (1932). *A technique for the measurement of attitudes.* New York: The Science Press.

Lin, P. C., Arbaiy, N., & Hamid, I. R. A. (2017). One-Way ANOVA Model with Fuzzy Data for Consumer Demand. In *Advances in Intelligent Systems and Computing* (pp. 111-121).

Lorber, J. (2006). Shifting Paradigms and Challenging Categories. *Social Problems, 53*(4). Retrieved from 10.1525/sp.2006.53.4.448

Lubiano, M. A., de la Rosa de Sáa, S., Montenegro, M., Sinova, B., & Gil, M. A. (2016). Descriptive analysis of responses to items in questionnaires. Why not using a fuzzy rating scale? *Information Sciences, 360*, 131-148. doi:10.1016/j.ins.2016.04.029

Lubiano, M. A., Gil, M. A., Lopez-Dıaz, M., & Lopez-Garcıa, M. T. (2000). The λ-mean squared dispersion associated with a fuzzy random variable. *Fuzzy Sets and Systems, 111*(3), 307–317. doi:10.1016/s0165-0114(97)00389-8

Lubiano, M. A., Montenegro, M., Sinova, B., de la Rosa de Sáa, S., & Gil, M. A. (2016). Hypothesis testing for means in connection with fuzzy rating scale-based data: algorithms and applications. *European Journal of Operational Research, 251*(3), 918-929. doi:10.1016/j.ejor.2015.11.016

Meier, P., & Motmans, J. (2020). Trans laws and constitutional rulings in Belgium: on the ambiguous relations between sex and gender. *Politics and Governance [accepted for publication]*.

Monro, S. (2020). Non-binary and genderqueer: An overview of the field. In *Non-binary and Genderqueer Genders* (pp. 8-13).

Motmans, J., Burgwal, A., & Dierckx, M. (2020). *Het meten van genderidentiteit in kwantitatief onderzoek*. Retrieved from Ghent, Belgium: http://transgenderinfo.be/wp-content/uploads/Adviesnota_Motmans_Burgwal_Dierckx_2020.pdf

Motmans, J., & Longman, C. (2017). Wat maakt het verschil? Een genderkritisch perspectief op het thema intersekse. *Tijdschrift voor seksuologie, 41*(2), 68-77. Retrieved from https://biblio.ugent.be/publication/8521133/file/8521135.pdf

Motmans, J., Wyverkens, E., & Defreyne, J. (2017). *Being transgender in Belgium: ten years later*. Retrieved from Brussels: https://igvm-iefh.belgium.be/sites/default/files/118_-_being_transgender_in_belgium.pdf?fbclid=IwAR0eUHpQpztT2FxYwL5x043Grw5TUqGfJJ97M-6m_QK2inpUxZNPJxagMng

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

Parchami, A., Taheri, S. M., & Mashinchi, M. (2010). Fuzzy p-value in testing fuzzy hypotheses with crisp data. *Statistical Papers, 51*(1), 209-226. doi:10.1007/s00362-008-0133-4

Parchami, A., Taheri, S. M., & Mashinchi, M. (2012). Testing fuzzy hypotheses based on vague observations: a p-value approach. *Statistical Papers, 53*(2), 469-484. doi:10.1007/s00362-010-0353-2

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*(1), 1-15. doi:10.1016/s0001-6918(99)00050-5

Puri, M. L., & Ralescu, D. A. (1986). Fuzzy random variables. *Journal of Mathematical Analysis and Applications, 114*(2), 409-422. doi:10.1016/0022-247x(86)90093-4

Reis, E. (2007). Divergence or disorder?: the politics of naming intersex. *Perspectives in Biology and Medicine, 50*(4), 535-543. doi:10.1353/pbm.2007.0054

Richards, C., Bouman, W. P., Seal, L., Barker, M. J., Nieder, T. O., & T'Sjoen, G. (2016). Non-binary or genderqueer genders. *International Review of Psychiatry, 28*(1), 95-102. doi:10.3109/09540261.2015.1106446

Ritz, S. A., Antle, D. M., Côté, J., Deroy, K., Fraleigh, N., Messing, K., . . . Mergler, D. (2017). First steps for integrating sex and gender considerations into basic experimental biomedical research. *the FASEB Journal, 28*(1), 4-13. doi:10.1096/fj.13-233395

Rosser, B. R. S., Oakes, J. M., Bockting, W. O., & Miner, M. (2007). Capturing the social demographics of hidden sexual minorities: An internet study of the transgender population in the United States. *Sexuality Research and Social Policy, 4*(2), 50-64. doi:10.1525/srsp.2007.4.2.50

RStudio Team. (2020). RStudio: Integrated Development for R. Boston, MA: RStudio, Inc. Retrieved from http://www.rstudio.com/

Schmidt, M. J. (1979). *Understanding and using statistics: Basic concepts*. Lexington, Mass: Heath.

Schoonacker, M., Dumon, E., & Louckx, F. L. (2009). *WELEBI, onderzoek naar het mentaal en sociaal welbevinden van lesbische en biseksuele meisjes.* Retrieved from Brussel: https://cavaria.be/sites/default/files/2009welebi_eindrapport.pdf

Scott, A. J., & Wild, C. J. (1986). Fitting Logistic Models Under Case-Control or Choice Based Sampling. *Journal of the Royal Statistical Society, 48*(2), 170-182. doi:10.1111/j.2517-6161.1986.tb01400.x

Sinova, B., Gil, M. A., Colubi, A., & Van Aelst, S. (2012). The median of a random fuzzy number. The 1-norm distance approach. *Fuzzy Sets and Systems, 200*, 99-115. doi:10.1016/j.fss.2011.11.004

Skrabanek, P., & Martinkova, N. (2018). *fuzzyreg: An R Package for Fuzzy Linear Regression.* Paper presented at the ENBIK2018, Prague.

Skrabanek, P., & Martinkova, N. (2019). Getting Started with Fitting Fuzzy Linear Regression Models in R. Retrieved from https://rdrr.io/cran/fuzzyreg/f/inst/doc/GettingStarted.pdf

Statistics New Zealand. (2015). *Statistical standard for gender identity*. Retrieved from https://unstats.un.org/unsd/classifications/expertgroup/egm2017/ac340-22.PDF

Stoller, R. (1968). *Sex and Gender: On the Development of Masculinity and Femininity*. New York: Science House.

Civil Rights Act 2020, (2020).

Trutschnig, W., Lubiano, M. A., & Lastra, J. (2013). SAFD — An R Package for Statistical Analysis of Fuzzy Data. In *Towards Advanced Data Analysis by Combining Soft Computing and Statistics* (pp. 107 to 118).

Van Caenegem, E., Wierckx, K., Elaut, E., Buysse, A., Dewaele, A., Van Nieuwerburgh, F., . . . T'Sjoen, G. (2015). Prevalence of Gender Nonconformity in Flanders, Belgium. *Archives of Sexual Behavior, 44*(5), 1281-1287. doi:10.1007/s10508-014-0452-6

Villacorta, P. J., Masegosa, A. D., Castellanos, D., & Lamata, M. T. (2014). A new fuzzy linguistic approach to qualitative Cross Impact Analysis. *Applied Soft Computing, 24*, 19-30. doi:10.1016/j.asoc.2014.06.025

Vlaamse Overheid. (2009). *De Vlaamse Regering 2009-2014. Een daadkrachtig Vlaanderen in beslissende tijden. Voor een vernieuwende, duurzame en warme samenleving.* Retrieved from Vlaams Parlement, Brussels, Belgium:

Wang, X., Liu, X., & Zhang, L. (2014). A rapid fuzzy rule clustering method based on granular computing. *Applied Soft Computing, 24*, 534-542. doi:10.1016/j.asoc.2014.08.004

Warren, J. C., Smalley, K. B., & Barefoot, K. N. (2016). Psychological well-being among transgender and genderqueer individuals. *International Journal of Transgenderism, 17*(3-4), 114-123.

Westat, A. E.-O. R. C. (2017). *Population Assessment of Tobacco and Health (PATH) Study [United States] Restricted-Use Files. Wave 3: Adult Questionnaire Data (English Version)*. Retrieved from Ann Arbor, Michigan: https://www.icpsr.umich.edu/icpsrweb/content/NAHDAP/path-study-faq.html#headingFive

Westbrook, L., & Saperstein, A. (2015). New categories are not enough: rethinking the measurement of sex and gender in social surveys. *Gender & Society, 29*(4), 534-560. doi:10.1177/0891243215584758

Zadeh, L. A. (1995). Discussion: Probability Theory and Fuzzy Logic are complementary rather than competitive. *Technometrics, 37*(3), 271-276. doi:10.1080/00401706.1995.10484330

Appendix

Appendix A

Q1. Welcome to this short anonymous survey that I conduct in the light of my Master thesis in Statistical Data Analysis at Ghent University. The aim of this survey is mainly statistical, we want to study the best way to measure the concept gender identity. Gender identity refers to the intrinsic feeling of being male, female, or an alternative gender. Herefor I would like to ask you a few questions about your gender identity. No other personal questions will be asked. At the end you have the possibility to give me feedback on these questions. Everyone aged 18+ can participate. It will only take you 5 minutes to complete the survey. Thank you for your participation. If you need further information, please contact me at aisa.burgwal@ugent.be. By clicking on YES, I declare I am at least 18 years of age and I agree to participate in the study:

- ○ Yes
- ○ No

IF Q1. = No THEN

Q2_b. Thank you very much for your interest in my survey. Unfortunately, without your consent, you cannot participate in this survey. If there is anything you want to tell me, you can write it here below:

____

END OF SURVEY

IF Q1. = Yes THEN

Q2_a. What was your sex assigned at birth, meaning on your original birth certificate? (We understand this question might not be pleasant for some to answer. We need to ask this question to be able to analyze the data correctly.)

- ○ Female
- ○ Male

IF Q1. = Yes & IF Q2_a. = Female THEN

Q3_a. How feminine / masculine do you feel? Please choose the option that fits you best.

- ○ Very masculine
- ○ Somewhat masculine

- ○ Not masculine nor feminine / masculine and feminine
- ○ Somewhat feminine
- ○ Very feminine

Q4_a. In the following question we want to ask you again to describe your gender, but in a different way, by using four scales. How feminine / masculine do you feel (on a range from masculine = 0 to feminine = 10)? We would like you to give us four numbers between 1 and 10, for example 7 - 8 - 9 - 10, which indicates that your gender identity falls between 7 and 10, and DEFINITELY between 8 and 9.

0 -- 10

IF Q1. = Yes & IF Q2_a. = male THEN

Q3_b. How masculine / feminine do you feel? Please choose the option that fits you best.

- ○ Very feminine
- ○ Somewhat feminine
- ○ Not feminine nor masculine / feminine and masculine
- ○ Somewhat masculine
- ○ Very masculine

Q4_b. In the following question we want to ask you again to describe your gender, but in a different way, by using four scales. How feminine / masculine do you feel (on a range from feminine = 0 to masculine = 10)? We would like you to give us four numbers between 1 and 10, for example 7 - 8 - 9 - 10, which indicates that your gender identity falls between 7 and 10, and DEFINITELY between 8 and 9.

0 -- 10

IF Q1. = Yes THEN

Q5. How difficult did you find the previous question?

- ○ Not difficult at all
- ○ Somewhat difficult
- ○ Very difficult

Q6. Is there anything you want to say about these questions?

____

Q7. Lastly, we would like to ask if you currently identify/have identified in the past as transgender? A transgender person can be conceptualized as someone who does not (fully) identify with their sex assigned at birth (either identifies opposite to their sex assigned at birth, or does not identify with the male/female labels).

- ○ Yes
- ○ No

Appendix B

The dataset used for this paper is represented in the table below. Some features were re-coded during the preprocessing phase, and other features were removed from the dataset because of their irrelevance for further processing. For each variable the feature name is shown, a short description, a label with possible recoding and the specific feature type.

Feature name	Description of the feature	Label	Type
IC	Informed consent	1 = Yes, 2 = No	Nominal
SAAB	Sex assigned at birth	1 = Male, 2 = Female	Nominal
Likert	Gender identity based on a five-point Likert scale	1 = Very masculine, 2= Somewhat masculine, 3 = Masculine and feminine / masculine nor feminine, 4 = Somewhat feminine, 5 = Very feminine	Numeric
Min	Gender identity based on a Fuzzy set scale: smallest number (support minimum).	[0.00-10.00]	Numeric
2nd	2nd number of the Fuzzy set scale (core minimum)	[0.00-10.00]	Numeric
3rd	3rd number of the Fuzzy set scale (core maximum)	[0.00-10.00]	Numeric
Max	Largest number of the Fuzzy set scale (support maximum)	[0.00-10.00]	Numeric
Diff	Difficulty of the Fuzzy set scale	1 = Not difficult at all, 2 = Somewhat difficult, 3 = Very difficult	Nominal
TRANS	Transgender identity	1 = Yes, 2 = No	Nominal
Open	Option to add something		String

GHENT
UNIVERSITY

FACULTY
OF SCIENCES

Measuring gender identity: A comparison between the Likert scale and the Fuzzy scale.

Aisa Burgwal

Master dissertation submitted to

obtain the degree of

Master of Statistical Data Analysis

Promotor: Prof. Dr. Christophe Ley

Co-Promotor: Prof. Dr. Joz Motmans

Department of Applied Mathematics, Computer Science and Statistics

Academic year 2019-2020