# Effects of Judicial Instructions and Juror Characteristics on Interpretations of Beyond Reasonable Doubt

Katrin Mueller-Johnson[1], Mandeep K Dhami[2], and Samantha Lundrigan[3]

[1]University of Cambridge

[2]Middlesex University

[3]Anglia Ruskin University

Send correspondence to:
Katrin Mueller-Johnson
Institute of Criminology
University of Cambridge
Sidgwick Avenue
Cambridge
CB3 9DA, UK
E-mail: kum20@cam.ac.uk
Tel: +44 1223 767184
Fax: +44 1223 335356

**Acknowledgements**

This research was funded by AHRC Early Career grant number: AH/E008607/1

(PI: KMJ, Co-I: MKD). We would like to thank NAPP Plc for assisting with data collection.

**Abstract**

Purpose and Methods: The standard of proof, beyond reasonable doubt (BRD), serves as a threshold for reaching verdicts in criminal cases. Past research has demonstrated that factors such as the wording of judicial instructions defining the standard can influence people's interpretation of it. In addition, there is some concern that instructions may not be effective for the wider jury-eligible population. In an experimental study involving members of the general public, we examined the effect of two commonly used judicial instructions (i.e., *sure* and *firmly convinced*) against a situation when BRD was undefined, on people's quantitative interpretations of BRD as well as on their self-reported understanding of the standard and confidence in applying it. We also explored the effect of juror characteristics (i.e., gender, age and education).

Results: Compared to when the standard was undefined, the *sure* instruction helped to reduce inter-individual variability in interpretations of BRD and the *firmly convinced* instruction increased people's understanding of the standard. However, neither instruction was effective in increasing confidence in applying the standard or in reducing observed individual differences.

Conclusion: These findings underscore the importance of developing evidence-based judicial instructions that can benefit the broad jury-eligible population equally and in a variety of ways.

**Keywords:** Standard of proof, beyond reasonable doubt, membership function, judicial instructions, jury decision-making.

# Effects of Judicial Instructions and Juror Characteristics on Interpretations of Beyond Reasonable Doubt

Proof 'beyond reasonable doubt' (BRD) is the standard of proof used in criminal trials in many adversarial justice systems such as in Anglo-American jurisdictions. The standard specifies the degree of belief in (or probability of) guilt required for conviction, and as a principle of due process it provides a stringent threshold so as to reduce the number of innocent defendants being wrongfully convicted. Indeed, BRD has been theorized to be equivalent to a level of certainty of .90 (or 90%; see Newman, 1993), although opinions differ as to whether the standard should be quantified (e.g., see Kagehiro, 1990 - pro quantification and Stoffelmayr & Diamond, 2000 - against quantification).

While there is empirical evidence showing that some judges and mock jurors interpret BRD at around .90 (e.g., McCauliff, 1982; Zander, 2000), there is also evidence demonstrating that interpretations may vary according to case characteristics (e.g., Martin & Schum, 1987), such as the severity of the offence, and individual differences in juror attitudes (e.g., Devine & Caughlin, 2014; Lundrigan, Dhami, & Mueller-Johnson, 2016). Interpretations may in fact be considerably lower than .90 (e.g., Horowitz & Kirkpatrick, 1996; Simon & Mahan, 1971).

Some of the differences in numerical interpretations of BRD reported across studies are likely to be due to differences in the methods used to elicit individuals' interpretations (such as direct rating, decision theory-based, and Membership Function method) (see Dane, 1985; Dhami, 2008). However, all methods demonstrate that there is considerable inter-individual variability in how BRD is interpreted (for a review see Hastie, 1993). For instance, Dhami (2008) found that average interpretations of BRD across individuals and methods varied from .53 to .96 (or 53% to 96%). Such inter-individual variability in how much

evidence jurors require before they are willing to convict can lead to disagreements among jurors, and in the most extreme case, to a hung jury.

A hung jury is followed by the possibility, or in some jurisdictions the requirement, of a retrial with all its associated financial costs to the justice system, as well as its psychological costs to victims, witnesses, and the defendant. Therefore, it is desirable that hung juries do not come about because of differences in the interpretation of BRD. Judicial instructions attempting to aid jurors' correct interpretation of BRD should be worded in a way that minimize inter-individual differences in interpretations and make it clear to jurors how BRD should be applied.

Both researchers and courts have expressed concern over the difficulty that people may have in understanding the concept of BRD (e.g., Elwork, Sales, & Alfini, 1982; Heffer, 2006; Power, 1999). In an effort to reduce inter-individual variability in interpretations of BRD as well as to bring interpretations closer to that intended, some judges have attempted to define the standard for jurors. However, judges may inadvertently influence jurors to overly increase or reduce the standard. For instance, the instruction used by the judge in Cage v. Louisiana (1990) i.e., "doubt as would give rise to grave uncertainty" and "actual substantial doubt" was held be unconstitutional because it was believed to require too high a degree of doubt for acquittal, and was thus rejected by the US Supreme Court. In order to avoid such situations, standardized instructions have been proposed and a considerable number of jurisdictions have adopted them. Currently, there are various different instructions (even within a jurisdiction), all of which use qualitative terms (e.g., "moral certainty") to define BRD (Heffer, 2006; Hemmens, Scarborough, & Del Carmen, 1997). Importantly, these are not evidence-based (see Dhami, Lundrigan & Mueller-Johnson, 2015).

## Effect of Judicial Instructions Defining Beyond Reasonable Doubt

### *Being 'Firmly Convinced' of the Defendant's Guilt*

One of the most common instructions in the US, advocated by the Federal Judicial Center, is the instruction "you must be *firmly convinced* of the defendant's guilt". This instruction is also used in at least eleven US states (Hemmens et al., 1997). Another common instruction, advocated, for instance, in the UK by the Judicial Studies Board for England and Wales, is "you must be *sure* of the defendant's guilt". The majority of English judges and magistrates have been found to use some variation of the *sure* instruction (Heffer, 2006) and it is also used in New Zealand (Young, 2003). Across the considerable body of studies investigating people's interpretations of BRD, a few have systematically studied the *firmly convinced* or *sure* instructions, and we review these below.

The evidence for the superiority of the *firmly convinced* instruction over the undefined standard (i.e., BRD) is mixed. As part of a study of five different instructions, Horowitz and Kirkpatrick (1996) compared interpretations of BRD under the *firmly convinced* instruction and when the standard was undefined. This was done in the context of two hypothetical murder cases where the strength of the evidence was manipulated to be weak or strong. A sample of the jury-eligible public was asked to provide their interpretations of BRD using a direct rating method at both the pre- and post-deliberation stages of six person juries (i.e., they were asked at each time to indicate on a 21-point scale what minimum probability of the defendant having committed the crime they required in order to convict). Under all conditions, the *firmly convinced* instruction led to higher numerical interpretations of BRD than the undefined standard. Participants also reported greater confidence in their verdicts under the *firmly convinced* instruction than when the standard was undefined.

Koch and Devine (1999) also compared the *firmly convinced* instruction against the undefined standard. A sample of students in mock juries of four to seven members were

asked to render verdicts on a hypothetical murder case that did or did not include a lesser charge of manslaughter. Guilty verdicts were used as an indirect measure of the standard of proof. Here, mock jurors were asked whether they considered the defendant guilty or not guilty at several points when reading the case transcript. The authors reported jurors' decisions after receiving the final instructions (and before deliberation). Given that there were more guilty verdicts under the *firmly convinced* instruction than when the standard was undefined, this study did not find that juries had more stringent interpretations of BRD under the *firmly convinced* instruction (i.e., it did not lead to a higher evidentiary threshold) than when the undefined standard.

***Being 'Sure' of the Defendant's Guilt***

Variants of the *sure* instruction have also been investigated. Montgomery (1998) examined the interpretations of BRD under the *sure* instruction in a sample of British adults who were the next of kin of university students as well as in a random sample drawn from the electoral register. Respondents were asked to judge a hypothetical murder case. Around three quarters of the whole sample who found the defendant not guilty said they needed 100% confidence of guilt to convict. Of those who gave a guilty verdict, around a third said they needed 100% confidence of guilt to convict. This suggests that BRD tended to be interpreted above 90%.

Zander (2000) surveyed samples of randomly selected members of the British general public and lay magistrates (judges) for their interpretation of BRD under the *sure* instruction. Here, half of the public and over a third of the lay judges interpreted BRD as 100% certainty. Around three-quarters of the public and lay judges interpreted BRD as 90% or higher. A small minority of both groups (around 4% to 5%) interpreted BRD to be lower than 70%.

Thus, the findings from the small body of past research investigating interpretations of BRD under the *firmly convinced* instruction are mixed as to whether the instruction leads

to a more or less stringent standard of proof compared to when the standard is undefined. The present study contributes to this body of research. The few studies that have examined the *sure* instruction demonstrate that it leads to an extremely stringent standard of proof, although these studies have not compared the *sure* instruction to when BRD is undefined, as we do in the present study. This is also the first study to directly compare the *firmly convinced* and *sure* instructions.

**Effects of Juror Characteristics on Beyond Reasonable Doubt**

As juries are expected to represent a cross-section of the general population, it is important to ascertain if different sub-sections of the population have systematically different interpretations of BRD, and to explore any differential effect of judicial instructions across these sub-sections. Nevertheless, few have investigated the effects of juror characteristics such as gender, age and education level on interpretations of BRD either when defined by instructions or when left undefined. We found only two relevant published studies, and both of these focus solely on gender differences. Zander (2000) found that a slightly greater proportion of female than male members of the public interpreted BRD as requiring 100% certainty of guilt under the *sure* instruction. Nagel (1979) found that male students had higher interpretations of BRD when undefined than their female counterparts in the context of a hypothetical rape case.

It is important to investigate individual differences in self-reported understanding of BRD and jurors' self-reported confidence in applying the standard because perceived difficulties in understanding the law have been shown to be related to juror stress (Bornstein, Miller, Nemeth, Page, & Musil, 2005). Having to understand the intricacies of legal and court procedures may be an additional, albeit under-researched, source of stress. Bornstein et al. (2005) found that although the burden of responsibility carried by jurors was seen as the greatest stressor, trial complexity (including an understanding of the law, and deciding on

guilt) was reported as the second great stressor. Findings from a UK based juror survey suggested that jurors most often reported understanding legal terminology as the most difficult aspect of a trial (Matthews, Hancock, & Briggs, 2004). Stress can have a negative effect on decision-making such as limiting the information that people attend to and process (Mather & Lighthall, 2012).

## The Present Study

The first aim of the study was to examine the effect of judicial instructions for BRD (i.e., *firmly convinced*, *sure*, and undefined) on people's interpretations of the standard of proof, including inter- and intra-individual variability in interpretations. The second aim was to examine the effects of instructions on people's reported ease/difficulty in understanding BRD. The third aim was to examine the effects of instructions on people's confidence in applying the standard. The final aim of the study was to examine the relationship between people's demographic characteristics (i.e., gender, age, and education level) and their interpretations of BRD as well as their confidence in applying the standard.

Based on past research it was hypothesized that the *sure* standard would lead to significantly more stringent interpretations of BRD than the undefined standard. As the evidence on the *firmly convinced* instruction is mixed and as there is no prior work comparing the *firmly convinced* and the *sure* instructions, we make no predictions as to whether *firmly convinced* leads to higher or lower interpretations than the undefined standard and as to which of the two, *firmly convinced* or *sure*, is interpreted more stringently.

Given that judicial instructions were introduced to bring people's interpretations of BRD closer to that intended by the law, it is reasonable to predict that the *sure* and *firmly convinced* instructions would lead to reduced inter- and intra-individual variability in

numerical interpretations of BRD compared to when the undefined standard. Finally, given the lack of research on the effect of juror characteristics on people's interpretations of BRD and their confidence in applying the standard we did not make a priori predictions about the effects of gender, age and education level on these variables.

## Method

### *Participants*

The initial sample consisted of 170 members of the jury-eligible British public who volunteered to participate in the study in return for a payment of £10. The data for three participants had to be excluded as their MF function ratings were not complete, thus leaving a final sample of n = 167. This sample size provides 80% power to detect a Cohen's $d_z$ (the effect size for dependent data) of .23, about halfway between a small (0.14) and a medium (0.35) $d_z$ effect size (Lankin, 2017).

Fifty-three percent of the sample was male, and 95.8% of the sample described themselves as white. On average, participants were 36.30 years old (*SD* = 10.57; ranging from 19 to 69 years). Secondary school (up to age 16) was the highest educational attainment for 27.1% of the sample; 30.1% had been to college (up to 18 years), and 42.2% had a university education. This is roughly in line with the distribution of education in the general public (ONS, 2013) Fifteen percent of the sample reported having served on a jury in the past. On average, participants rated the likelihood of them serving on a jury, if they were called for service, as being as 73.0%.

### ***Design***

We used a mixed quasi-experimental design. The within-subjects variable was judicial instruction which was manipulated and had three levels (i.e., BRD undefined, *sure*

instruction, and *firmly convinced* instruction). Participant gender, age and education were treated as between-subjects variables in the data analyses.

### *Stimuli and Measures*

Participants were asked to imagine that they were serving on a jury in a criminal trial. Each participant was presented with the standard of proof (i.e., BRD undefined, defined as *sure* and *firmly convinced*). In the undefined condition, the instruction read: "The defendant is presumed innocent unless the prosecution has proved guilt beyond reasonable doubt". In the *sure* condition, the instruction read: "The defendant is presumed innocent unless the prosecution has proved guilt beyond reasonable doubt. Proof beyond reasonable doubt is proof that makes you sure." Finally, in the *firmly convinced* condition, the instruction read: "The defendant is presumed innocent unless the prosecution has proved guilt beyond reasonable doubt. Proof beyond reasonable doubt is proof that leaves you firmly convinced."

Numerical interpretations of BRD were measured using the Membership Function method (MF; Dhami, 2008; Dhami et al., 2015; Lundrigan et al., 2016; 2017, Park, Seong, Kim & Kim, 2016) which has been shown to be a valid predictor of verdicts (Dhami, 2008; Lundrigan et al., 2016; also see Dhami & Wallsten, 2005 for more on the reliability and validity of this method). As is typically done in the Membership Function Method, participants were presented with 21 scales that each corresponded to one of 21 values, from 0% to 100% (in 5% intervals; see Appendix). Participants responded to the question "to what extent would each of these values substitute for BRD?" Each scale had 21-point points and was labeled at each from *not at all* to *absolutely*. Responses were provided by circling a point on each scale. The MF method provides measures of the 'peak' value that *absolutely* substitutes for BRD and the 'spread' of values that represent BRD to varying degrees (see Appendix).

Participants were also asked to report how well they understood each instruction and how confident they would be using each instruction in a real trial. Ratings of both understanding of BRD and confidence in applying the standard were provided on 7-point scales anchored at each end (i.e., understanding: 1 = "very easy" to 7 = "very difficult", and confidence: 1 = "not at all confident" and 7 = "extremely confident").

Gender was defined as male versus female. Education was measured as a two level variable consisting of "having obtained a university degree" or not. Age was measured as a continuous variable. For any ANOVA in which age was included, age was dichotomized using a median split (< 31 years v. 31 years and older), as is common in psychological research.

***Procedure***

Participants were recruited from a large UK company based in the East of England that employs over 700 people in a wide range of roles from manual labour through clerical staff to professionals (i.e., medics). Recruitment posters were placed across the multi-building site and emails asking for volunteers were sent out to all employees. The data was collected individually, on paper, at the recruitment site. The experiment took approximately 15-20 minutes. Participants first completed the MF method. The order of presentation of the three experimental conditions (i.e., BRD undefined, sure instruction, and firmly convinced instruction) was counter-balanced across participants. Participants then responded to the questions asking about understanding and confidence. Finally, participants provided their demographic details (i.e., gender, age, educational background, jury experience and willingness to serve on a jury).

No experimental conditions were excluded. All measures used in the study are reported. Participants were only excluded from this study if they failed to complete the study materials, i.e. did not give ratings for all three judicial instructions.

## Results

### *Descriptive Statistics for Dependent Measures*

The last column in Table 1 presents the mean and standard deviations for the peak and spread of interpretations of BRD as measured by the MF method as well as the self-reported understanding of BRD and confidence in using the standard.

TABLE 1 HERE

### *Effect of Judicial Instructions*

***Interpretations of BRD***. The first aim of the study was to examine the effect of judicial instructions for BRD (i.e., *firmly convinced*, *sure*, and undefined) on people's interpretations of the standard, including inter- and intra-individual variability in interpretations. Before conducting this analysis, we were interested in examining how close people's interpretations of the standard were to .9 (or 90%). Bonferroni adjusted one-sample two-tailed *t*-tests revealed that the mean peak interpretations of BRD were significantly greater than 90% for the *undefined* condition, $t\ (160) = 2.55$, $p = .012$, $d = -.20$ and the *sure* condition, $t\ (164) = 3.09$, $p = .001$, $d = -.24$. The *firmly convinced* condition was not significantly different from the 90% threshold, $t\ (164) = 1.89$, $p = .61$, , $d = -.14$.

The first row of Table 1 presents the means and standard deviations for the peak interpretations of BRD by judicial instruction. A one-way repeated measures ANOVA revealed no significant effect of instruction on peak interpretation, $F\ (2, 314) = .677$, $p = .509$, *partial eta*$^2 = .004$, *90% CI* [.000; .024]. Since a lack of statistical significance could be

due to a lack of statistical power, equivalence tests using the Two One-Sided Tests (TOST) procedure for dependent samples (Lakens, 2017) were calculated to test if the instructions were indeed not different from each other [1]. As our data had a power of 80% to detect an effect size of dz = .23, this was taken as the widest equivalence bounds, in line with the suggestion in Lakens (2017). Using Lakens' (2017) excel based calculator, the comparison of the *sure* versus *undefined* condition showed that the observed effect size for this difference (*dz* = -.003) showed equivalence given a lower boundary of *dz* = -.23 and upper boundary of *dz* =.23, i.e. a small effect. This means that if there is a possible difference between the two instructions despite the non-significant ANOVA, it must be smaller than a small effect size. An analysis for equivalence of the firmly convinced versus undefined conditions using the TOST procedure showed that the observed effect size (dz =- .08) was not significant with the equivalent bounds of *dz* = -.14 and *dz* = .14, t(166) =.79, p= .216; but it was significantly within the equivalent bounds of *dz* = .-.23 and *dz* = .23, t(166) = 1.95, p=.026. Similarly the comparison for the sure and the firmly convinced instruction (observed effect size dz = -.08) showed no equivalence at *dz* = -.14 and *dz* = .14, t(166) =.75, p= .227; but it was significantly within the equivalent bounds of *dz* = .-.23 and *dz* = .23, t(166) = 1.91, p=.029. Thus all three instructions showed equivalence at a level of a *dz* =.23 effect size (i.e., a small to medium effect).

***Intra-individual variability***. The second row in Table 1 presents the means and standard deviations for the spread of interpretations of BRD by judicial instruction. A one-

[1] The TOSTprocedure can be used to establish equivalence by testing whether an observed effect size is statistically different from a pre-specified effect size boundary (for instance *d* = ± .3) . The observed effect size is examined in a one-way t-test to investigate the null-hypothesis that it is statistically significantly smaller than this lower bound (e.g. smaller than *d* = -.3) and in a second one-way *t*-test to see if it is statistically significantly greater than the upper bound (e.g. bigger than *d* = .3). If both of these null hypotheses are rejected, then it is concluded that the observed effect falls within the pre-specified statistical equivalence bounds (in this example d = ± .3), which suggests that the two means tested are close enough to each other to be practically equivalent (Lakens, 2017).

way repeated measures ANOVA showed no significant effect of instruction on spread $F$ (2, 314) = .836, $p$ = .434, *partial eta*$^2$ = .005, *90% CI* [.000; .022].

***Inter-individual variability***. Finally, in order to examine the extent of inter-individual variability in the peak interpretations of BRD, we computed Pearson correlations followed by Pitman-Morgan tests. The correlation between the peak interpretation of BRD when the standard was undefined and when it was defined as *sure* was *r (158)* = .59, $p$ < .001. The correlation between the peak interpretation of BRD under the undefined standard and when it was defined as *firmly convinced* was *r (157)* = .63, $p$ < .001. The correlation between the peak interpretation of BRD when the standard was defined as *sure* and when it was defined as *firmly convinced* was *r (161)* = .56, $p$ < .001. The Pitman-Morgan test compares the variances for two paired-sample variables, taking into account the correlation between the two variables (Kenny, 1953).[2] We found a significant difference in the variance of the peak interpretations of BRD when the standard was undefined (*M* = 92.30, *variance* = 130.64) and when it was defined as *sure* (*M* = 92.33, *variance* = 94.09), $t_{two\text{-}tailed}$(156)= 2.55, $p$ = .012 (thus smaller than the Bonferroni adjusted p-threshold of $p$ = .017) . Here, the inter-individual variability was greater in the undefined condition. There was no significant difference in the variability of the peak interpretations of BRD when the standard was undefined and when it was defined as *firmly convinced* (*M* = 91.55, *variance* = 110.46), $t_{two\text{-}tailed}$ (156) = 1.34, $p$ = .178 or between the *sure* and the *firmly convinced* conditions, $t_{two\text{-}tailed}$ (156) = 1.22, $p$ = .221.

***Self-reported understanding.*** The second aim of the study was to examine the effect of judicial instructions on people's reported understanding of BRD. A repeated measures ANOVA was computed (the Greenhouse-Geisser correction was applied to the $F$-statistic). There was a significant main effect of instruction on people's understanding of BRD, $F$ (2,

[2] As the Pitman Morgan Test is not readily available in statistics software packages, the following online calculator was used: http://www.how2stats.net/2011/06/testing-difference-between-correlated.html

295) = 4.19, $p$ = .020, *partial eta*$^2$ = .025, . *90% CI* [.003; .061].. Post-hoc *t*-test comparisons using Bonferroni corrections revealed that understanding of BRD was significantly greater under the *firmly convinced* instruction than when the standard was *undefined, p* = .013 (thus smaller than the adjusted p-value .017). No other significant differences between instructions were observed, *p*s > .05.

***Confidence.*** The third aim was to examine the effect of judicial instructions on people's reported confidence in applying the standard. A repeated measures ANOVA using the Greenhouse-Geisser correction showed no significant main effect of instruction on confidence in applying BRD, $F$ (2, 311) = 1.20, $p$ = .302, *partial eta*$^2$ = .007, 90% CI [.000; .027]. Confidence in applying the standard was significantly negatively correlated with the spread of interpretations of BRD ( i.e. intra-individual variability) when it was undefined, *r (159)* = -.21, $p$ = .005 (and thus larger than the Bonferroni adjusted threshold of p=.017). There was no significant correlation for the spread of the interpretation of BRD and confidence for the *sure* instruction, *r (163)* = .17, $p$ = .168, or the *firmly convinced* instruction, *r (163)* = .08, $p$ = .329.

### *Relationship Between Juror Characteristics and Interpretations of BRD and their Confidence in Applying the Standard*

The final aim of the study was to examine if people's demographic characteristics (i.e., gender, age, and education level) are associated with their interpretations of BRD as well as their confidence in applying the standard. The findings are presented below.

***Interpretations of BRD***. Table 2 presents the means and standard deviations of peak interpretations of BRD and spread of interpretations by age, gender and education level. We computed mixed ANOVAs on the peak and spread. Judicial instruction was the within-subjects factor. Age, gender and education level were the between-subjects factors. We

performed a median split on age, which had two levels (i.e., < 31 years v. 31 and over). Gender had two levels (i.e., male, female), and education had two levels (i.e., university educated or not). No statistically significant main effects or interaction effects were observed, *p*s > .05.

TABLE 2 HERE

Since age had been was measured on a continuous scale, we also computed correlations between age and the peak and spread of interpretations of BRD when the standard was undefined, when it was defined as *sure* and when it was defined as *firmly convinced*. There were no significant correlations between age and peak interpretations of BRD when the standard was undefined, or when it was defined as *sure* and *firmly convinced*, *p*s > .05. Although there were no significant correlations between age and the spread of interpretations of BRD when it was defined as *sure* or *firmly convinced*, *p*s > .05, there was a significant positive correlation of $r\ (158) = .17$, $p = .030$, when the standard was undefined. However, after adjusting for multiple testing (with a resulting new p-threshold of p=.017), this correlation failed to reach significance.

***Confidence in Applying BRD.*** Finally, Table 3 presents the means and standard deviations of reported confidence in applying BRD by age, gender and education level. We computed mixed ANOVAs on reported confidence in applying BRD. Judicial instruction was the within-subjects factor, and age, gender and education level were the between-subjects factors. There were significant main effects of gender and education level, whereas the main effect of age was marginally significant (gender: $F(1,157) = 13.88$, $p < .001$, *partial eta*$^2$ = .081, *90% CI* [.026; .155].; education level: $F(1,157) = 4.50$, $p = .036$, *partial eta*$^2$ = .028, *90% CI* [.001; .082]. and age: $F(1,157) = 3.68$, $p = .057$, *partial eta*$^2$ = .023, *90% CI* [.000; .074]., see Table 3). Self-reported confidence in correctly applying the standard was

significantly greater for males than for females *(p* = .002), for those who were university educated than for those without university degree (*p* = .036) and marginally significantly greater for those aged 31 and over than for those who were younger (*p* = .057).

TABLE 3 HERE

In addition, there was a marginally significant two-way interaction effect of gender by education level, $F(1,157) = 3.71$, $p = .056$, *partial eta*$^2$ = .023, *90% CI* [.000; .075]. As Figure 1 illustrates, less educated females reported lower levels of confidence in applying BRD than their more educated female counterparts. No other significant interaction effects were observed.

FIGURE 1 HERE

Finally, we also computed correlations between age and confidence in applying BRD when it was undefined, defined as *sure* and defined as *firmly convinced*. No significant correlations were observed, *p*s > .05.

## Discussion

Given that the legal system confers great responsibility on jurors to make decisions that may have severe consequences for the defendant's liberty and for public safety, the system should be responsible for setting out clearly what it asks of jurors, so that jurors can, and are confident that they can, accomplish these tasks. Efforts at providing jurors with judicial instructions that define the standard of proof have followed from empirical evidence suggesting that jurors' interpretations of BRD are lower than that intended, and may vary across jurors, as well as from evidence suggesting that jurors report difficulty in understanding the standard. Although the perceived need for judicial instructions was derived from empirical evidence, the actual wording of the instructions has not been, to our knowledge, based on empirical evidence of how these instructions may be understood.

In the present study, we focused on the *firmly convinced* instruction used in several US jurisdictions and the *sure* instruction used in England and Wales and New Zealand. We compared the effect of these two instructions against the effect of leaving the standard undefined, on people's interpretations of BRD, including inter- and intra-individual variability in interpretations. We also investigated the effect of instructions on people's reported ease/difficulty in understanding the standard of proof and their confidence in applying it. In addition, we studied the relationship between people's demographic characteristics (i.e., gender, age, and education level) and their interpretations of BRD under the different instructions, as well as their confidence in applying the standard.

We found that on average, in each of the three conditions, people interpreted BRD at just over .90 (90%), a somewhat more stringent standard than intended by law, but consistent with past research (Montgomery, 1998; Zander, 2000). There was no significant effect of judicial instructions on people's interpretations of the standard: interpretations of BRD were similar under the *sure* and *firmly convinced* instructions, and no different from when the standard was undefined. Although Horowitz and Kirkpatrick (1996), using a different method, did find a difference in interpretations of BRD under the *firmly convinced* instruction and the undefined standard, our finding of no difference is compatible with research by Koch and Devine (1999) who also used a different method. The present findings are also consistent with Dhami (2008) who found no significant effect of judicial instructions on interpretations of BRD as measured by the MF peak. The present findings thus suggest that existing instructions may be unnecessary.

Although people's peak interpretations of BRD may be around and above 90%, there is often considerable *inter*- and *intra*-individual variability in interpretations (e.g., Dhami, 2008). In the present study, interpretations of BRD ranged from around .5 to 1 across people for each of the three conditions. Therefore, instructions may be necessary to reduce such

variability. We found that although the *sure* instruction was useful in reducing *inter*-individual variability in interpretations of BRD, there was no such effect for the *firmly convinced* instruction. In addition, we observed that both the *sure* and *firmly convinced* instructions were ineffective in reducing *intra*-individual variability in people's interpretations of BRD. Indeed, the spread of interpretations of BRD was around 40 percentage points in each of the three conditions. Our findings are compatible with research also demonstrating significant differences in the ability of instructions to reduce *inter*- and *intra*-individual variability in interpretations of BRD (Dhami, 2008; Dhami et al., 2015).

Beyond examining the effect of judicial instructions on people's interpretations of BRD, it is useful to investigate people's reported understanding of BRD because perceived difficulties in understanding the law may be associated with juror stress (Bornstein et al., 2005), and this can have a negative effect on decision-making (Mather & Lighthall, 2012). Judicial instructions may have a role in increasing jurors' sense of understanding of what is required and thus could reduce juror stress. We found that when people were asked to report how easy or difficult it was for them to understand the concept of BRD, there was no difference between the *sure* and undefined conditions. However, in the *firmly convinced* condition participants reported greater ease of understanding compared to the undefined condition. One possible explanation for these findings may be that 'being sure' refers to an internal mental state in the same way that 'beyond reasonable doubt' does, whereas 'being firmly convinced' refers to an external source providing certainty such as that of having been convinced by legal argumentation. This and other potential explanations need to be explored in future research.

Understanding and interpreting the concept of BRD are different from having confidence in applying the standard. Again, one might expect judicial instructions to increase jurors' confidence in applying the law. Jurors lacking in confidence may be more likely to

acquit or struggle to reach consensus on a verdict. In the only past study to examine confidence in applying the standard of proof, Horowitz and Kirkpatrick (1996) found that participants had greater confidence in their verdicts under the *firmly convinced* instruction than when the standard was undefined. Until now, no-one has examined the effect of the *sure* instruction on confidence. In the present study, we found no significant effect of judicial instructions (i.e., *sure*, *firmly convinced* and undefined) on people's reported confidence in applying BRD. We found that average levels of confidence were around 5 as measured on a 7-point scale. Upon further exploration of the data we found a significant negative correlation of -.21 between the spread of interpretations of BRD (i.e., intra-individual variability) and confidence in applying the undefined standard. This suggests that the more fuzzy the concept of BRD is in an individual's mind, the less confident he/she feels in applying the standard.

Judicial instructions are meant to be used across a broad spectrum of society (i.e., the jury-eligible public), and so should be equally effective across different segments of the population. This raises the importance of studying individual differences in interpretations of BRD as well as confidence in applying the standard. However, few researchers have done so. Nagel (1979) found that males had higher interpretations of BRD than females when the standard was undefined. Zander (2000) found that females had higher interpretations of BRD than males under the *sure* instruction. We did not observe any significant differences in peak and spread of interpretations of BRD according to people's demographic characteristics (i.e., age, gender and education level) across the three instruction conditions (i.e., *sure*, *firmly convinced* and undefined. However, we found that there was a tendency for younger people (i.e., under 31) to be less confident in applying the standard than their older counterparts. In addition, less educated females were significantly less confident in applying the standard than their more educated counterparts. These findings reinforce the importance of ensuring people

not only understand and interpret BRD as intended but that they also feel confident in applying it.

***Potential Implications***

Policy and practical implications emerge from these findings. First, instructions aimed at defining the standard of proof for jurors in terms of the *sure* and *firmly convinced* language may not be necessary, given that jurors' interpretations of BRD when it is undefined already reach the desirable threshold for conviction. Indeed, currently, some jurisdictions in the US such as Illinois, Mississippi, Texas and in Australia such as New South Wales leave the standard undefined. As Heffer (2006, p. 168) points out, in New South Wales, judges advise jurors that the words BRD are "ordinary everyday words" and thus should be self-explanatory.

However, where the judicial instructions could be of value is in efforts at reducing the *inter*-individual variability in interpretations of BRD, increasing people's reported understanding of the standard, and reducing any individual differences in people's interpretations of BRD, as well as their confidence in applying the standard.

The *sure* instruction, while maintaining the same desired threshold for conviction as the undefined standard, led to lower *inter*-individual variability in interpretations of BRD than when the standard was undefined. Similarly, the '*firmly convinced*' instruction also led to the same desired threshold as when the standard was undefined, but it also increased self-reported understanding of BRD compared to the undefined standard. However, neither the sure nor the firmly convinced instructions reduced *intra*-individual variability in interpretations of BRD compared to when the standard was undefined. In addition, neither of these two instructions increased people's confidence in applying the standard beyond the undefined standard.

Although the two instructions examined in the present study did not do so, some form of judicial instructions for BRD may be helpful to improve older people's interpretations of BRD so the concept is less fuzzy in their minds. Women with lower levels of education may also benefit from instructions that increase their confidence in applying the standard.

***Potential Limitations***

It could be argued that the external validity of the present findings is limited because mock jurors were used rather than real jurors, and that we studied the standard of proof outside the context of a legal case. It would be inappropriate to study real juries in real trial situations where the wording of the standard of proof was manipulated experimentally, as we did in the present study. We made a concerted effort to minimize these limitations in several ways.

First, unlike most past psychological research on jury decision-making that utilizes student samples as mock jurors (for a review see Devine, Clayton, Dunford, Seying & Pryce, 2001) we sampled participants from a large company with great diversity in employment roles and education levels among their staff. Although this is not equivalent to random sampling from the jury eligible population (which to our knowledge has never been done in past research on this topic), it did afford the opportunity to study a wide cross-section of people in a controlled data collection environment, and enabled examination of individual differences.

Second, it is possible that people's interpretations of BRD or their perceived understanding of the standard may differ if applied to an actual case or if studied in the absence of a case. However, theoretically speaking, the interpretation of BRD should not vary as a function of case, and Dhami (2008) found no significant difference in people's interpretations of BRD in and outside the context of a real manslaughter case (see also Lundrigan et al., 2017). If we had studied BRD in the context of a criminal case our findings

would have been potentially limited to that specific type of case. Future research, nevertheless, could examine the effects of the *sure* and *firmly convinced* instructions in the context of a variety of criminal cases.

## Conclusions

The present study provides some evidence on the effectiveness of two judicial instructions that are used in the Anglo-American criminal justice system. We have demonstrated that not all judicial instructions are equally useful. We have also highlighted the importance of using several outcome measures when evaluating the impact of judicial instructions, and demonstrate the usefulness of examining individual differences in people's understanding of the standard of proof under different instructions. Although a practical application of the present findings may want to await replication, this study demonstrated that efforts to improve legal language would benefit from concurrent empirical testing in order to create better instructions and further evidence-based law-making (see also Dhami et al., 2015).

**References**


Bornstein, B. H., Miller, M. K., Nemeth, R. J., Page, G. L., & Musil, S. (2005). Juror reactions to juror duty: perceptions of the system and potential stressors. *Behavioral Sciences and the Law, 23*, 321-346.

Dane, F. C. (1985). In search of reasonable doubt: A systematc examination of selected quantification approaches. *Law and Human Behaviour, 9*(2), 141-158. doi:10.1007/BF01067048

Devine, D.J., & Caughlin, D. E. (2014). Do they matter? A meta-analytic investigation of individual characteristics and guilt judgments. *Psychology, Public Policy and Law, 20*, 109–134. doi: 10.1037/law0000006

Devine, D. J., Clayton, L. D., Dunford, B. B., Seying, R., & Pryce, J. (2001). Jury decision making: 45 years of empirical research on deliberating groups. *Psychology, Public Policy and Law, 7*, 622–727. doi:10.1037//1076-8971.7.3.622

Dhami, M. K. (2008). On measuring quantitative interpretations of reasonable doubt. *Journal of Experimental Psychology: Applied, 14*(4), 353-363. doi:10.1037/a0013344

Dhami, M. K., Lundrigan, S., & Mueller-Johnson, K. (2015). Instructions on reasonable doubt: Defining the standard of proof and the juror's task. *Psychology, Public Policy and Law, 21,* 169-178. doi: 10.1037/law0000038

Dhami, M. K., & Wallsten, T. S. (2005). Interpersonal comparison of subjectiveprobabilities: toward translating linguistic probabilities. *Memory and Cognition, 33*,1057-68. doi: 10.1037/a0013344

Elwork, A., Sales, B. D., & Alfini, J. J. (1982). *Making jury instructions understandable*. Charlottesville, Virginia: The Michie Company.

Hastie, R. (1993). *Inside the juror: The psychology of juror decision making.* New York: Cambridge University Press.

Hastie, R., Penrod, S. D., & Pennington, N. (1983). *Inside the jury*. Cambridge, MA: Harvard University Press.

Heffer, C. (2006). Beyond reasonable doubt: the criminal standard of proof instruction as communicative act. *The International Journal of Speech, Language and the Law, 13*(2), 159-188.

Hemmens, C., Scarborough, K. E., & Del Carmen, R. V. (1997). grave doubts about reasonable doubt: confusion in state and federal courts. *Journal of Criminal Justice, 25*(3), 231-254. doi:10.1016/S0047-2352(97)00008-1

Horowitz, I. A., & Kirkpatrick.L.C. (1996). A concept in search of a definition: the effects of reasonable doubt instructions on certainty of guilt standards and jury verdicts. *Law and Human Behaviour, 20*(6), 655-670. doi:10.1007/BF01499236

Kagehiro, D. K. (1990). Defining the standard of proof in jury instructions. *Psychological Science, 1*, 194-200. doi:10.1111/j.1467-9280.1990.tb00197.x

Kenny, D.T. (1953). Testing of differences between variances based on correlated variates. *Canadian Journal of Psychology/Revue canadienne de psychologie,* 7(1), 25-28.

Koch, C. M., & Devine, D. J. (1999). Effects of reasonable doubt definition and inclusion of a lesser charge on jury verdicts. *Law and Human Behaviour, 23*(6), 653-674. doi:10.1023/ A:1022389305876

Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science.* DOI: 10.1177/1948550617697177

Lundrigan, S., Dhami, M. K., & Mueller-Johnson, K. (2017). A re-examination of the acquittal biasing effect of offence seriousness. *Manuscript submitted for publication.*

Lundrigan, S., Dhami, M.K. & Mueller-Johnson, K (2016). Predicting verdicts using pre-trial attitudes and standard of proof . *Legal and Criminological Psychology, 21*, 95–110 , DOI:10.1111/lcrp.12043

Mather, M., & Lighthall, N. R. (2012). Risk and reward are processed differently in decisions made under stress. *Current Directions in Psychological Science, 21,* 36-41. DOI: 10.1177/0963721411429452

Martin, A.W., & Schum, D.A. (1987). Quantifying burdens of proof: A likelihood ratio approach. *Jurimetrics Journal*, Summer, 383-402.

Matthews, R., Hancock, L., & Briggs, D. (2004). *Jurors' perceptions, understanding, confidence and satisfaction in the jury system: a study in six courts*. London: Home Office.

McCauliff, C. M. A. (1982). Burdens of proof:degrees of belief, quanta of evidence, or constitutional guarantees? *Vanderbilt Law Review, 35*, 1293-1335.

Montgomery, J. W. (1998). The criminal standard of proof. *New Law Journal, April*, 582-585.

Nagel, S. (1979). Bringing the values of jurors in line with the law. *Judicature, 63*(4), 189-195.

Newman, J. O. (1993). Beyond 'reasonable doubt'. *New York University Law Review, 68*, 979 -1002.

Ogloff, J. (1998). J*udicial instructions and the jury. A comparison of alternative strategies. Final report*. Vancouver, BC: British Columbia Law Foundation.

ONS (Office of National Statistics) (2013). Full Report - Graduates in the UK Labour Market 2013. Last retrieved on 09.09.217 from http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/dcp171776_337841.pdf

Park, K., Seong, Y., Kim, M. & Kim, J. (2016) Juror adjustments to the reasonable doubt standard of proof, *Psychology, Crime & Law,* 22:6, 599-618, DOI: 10.1080/1068316X.2016.1168427

Power, R. C. (1999). Reasonable and other doubts: the problem of jury instructions. *Tennessee Law Review, 67*, 45-123.

Simon, R. J., & Mahan, L. (1971). Quantifying burdens of proof: a view from the bench, the jury and the classroom. *Law and Society Review, Feb*, 319-330.

Stoffelmayr, E., & Diamond, S. S. (2000). The conflict between precision and flexibility in explaining beyond a reasonable doubt. *Psychology, Public Policy and Law, 6*(3), 769-787. doi:10.1037/1076-8971.6.3.769

Young, W. (2003). Summing-up to juries in criminal cases - what jury research says about current rules and practice. *Criminal Law Review*, 665 - 689.

Zander, M. (2000). The criminal standard of proof-how sure is sure? *New Law Journal, Oct*, 1517-1519.

**Legal Cases**

Cage v. Louisiana 498 U.S. 39 (1990).

*Table 1*.

Means and Standard Deviations of BRD Interpretations by Judicial Instructions

| | Undefined | | Sure | | Firmly convinced | | Total | |
|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| MF Peak | 92.30 | 11.43 | 92.33 | 9.70 | 91.55 | 10.51 | 92.08 | 9.09 |
| MF Spread | 41.80 | 23.80 | 40.36 | 25.15 | 41.00 | 25.84 | 40.33 | 21.53 |
| Understanding | 3.18 | 1.69 | 2.96 | 1.58 | 2.77 | 1.46 | 2.98 | 1.21 |
| Confidence | 4.92 | 1.73 | 4.86 | 1.65 | 5.08 | 1.53 | 4.95 | 1.23 |

***Note***. Self-reported understanding and confidence were measured on 7-point scales with higher values representing greater difficulty in understanding of BRD and greater confidence in applying the standard.

*Table 2.*

Means and Standard Deviations of BRD Interpretations by Judicial Instructions and Demographic Characteristics

| | | Undefined | | Sure | | Firmly convinced | |
|---|---|---|---|---|---|---|---|
| | | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| **MF Peak** | | | | | | | |
| Age: | <31 | 94.36 | 9.13 | 93.85 | 9.35 | 93.13 | 9.93 |
| | 31 and over | 91.14 | 12.38 | 91.54 | 9.86 | 90.79 | 10.79 |
| Gender: | Male | 92.53 | 10.83 | 91.82 | 10.29 | 90.74 | 10.84 |
| | Female | 92.02 | 12.16 | 92.92 | 9.01 | 92.47 | 10.12 |
| Education: | University | 94.55 | 9.28 | 92.75 | 10.45 | 92.86 | 10.27 |
| | No university | 90.59 | 12.57 | 91.95 | 9.18 | 90.48 | 10.65 |
| **MF Spread** | | | | | | | |
| Age: | <31 | 36.04 | 22.33 | 33.38 | 26.35 | 35.66 | 25.98 |
| | 31 and over | 44.09 | 22.65 | 41.20 | 23.99 | 42.50 | 25.21 |
| Gender: | Male | 37.94 | 22.62 | 36.12 | 23.05 | 38.70 | 25.26 |
| | Female | 45.55 | 24.56 | 43.01 | 26.40 | 41.99 | 25.87 |
| Education: | University | 39.10 | 25.20 | 36.79 | 26.25 | 39.18 | 27.44 |
| | No university | 43.61 | 22.40 | 41.44 | 23.65 | 41.28 | 24.11 |

*Table 3.*

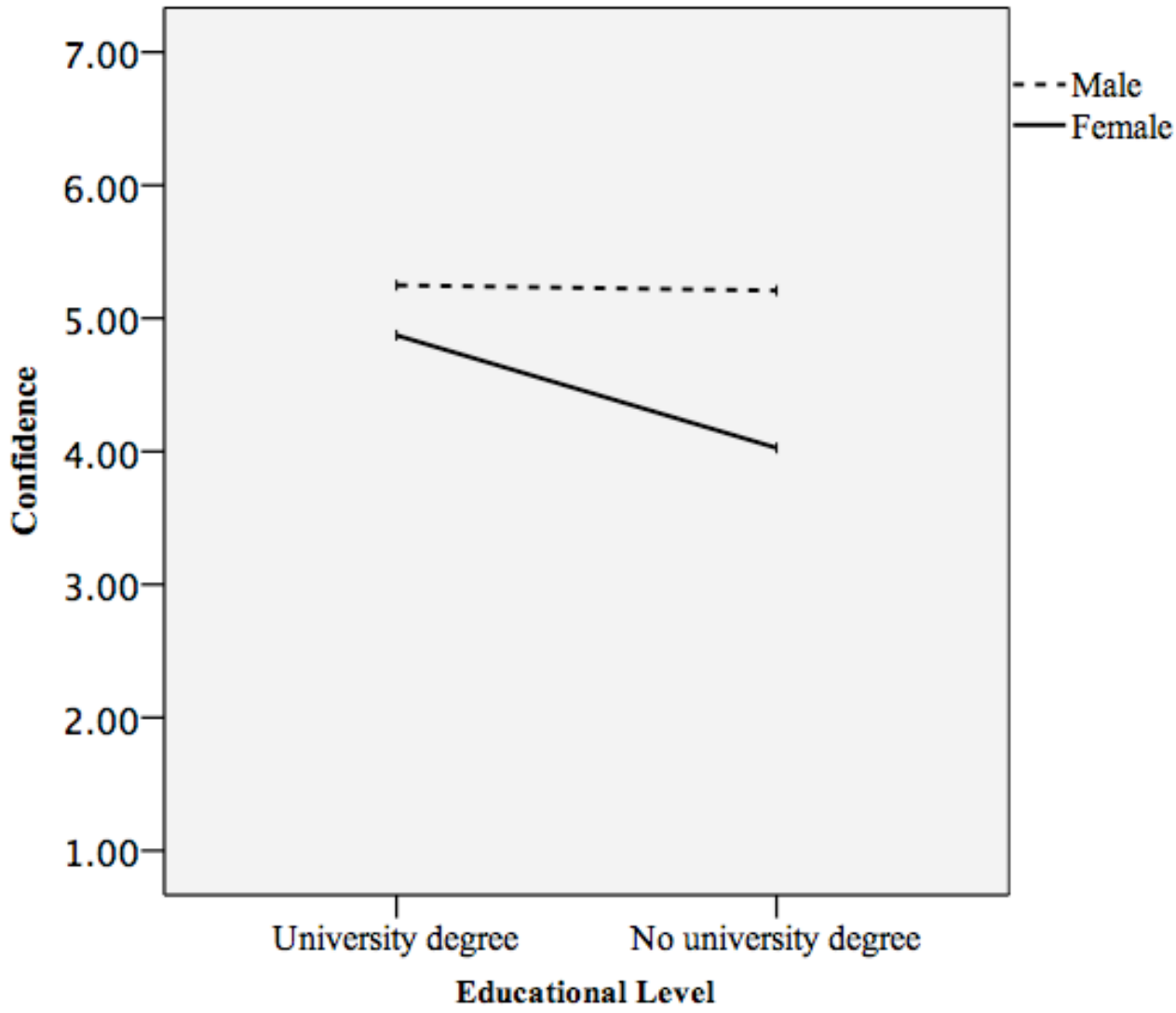
Means and Standard Deviations of Confidence in Applying BRD by Judicial Instructions and Demographic Characteristics

| | | Undefined | | Sure | | Firmly convinced | |
|---|---|---|---|---|---|---|---|
| | | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Age: | <31 | 4.35 | 1.87 | 4.93 | 1.59 | 5.02 | 1.43 |
| | 31 and over | 5.19 | 1.60 | 4.81 | 1.67 | 5.10 | 1.59 |
| Gender: | Male | 5.39 | 1.58 | 5.07 | 1.56 | 5.27 | 1.48 |
| | Female | 3.37 | 1.73 | 4.62 | 4.62 | 4.86 | 1.58 |
| Education: | University | 5.06 | 1.54 | 4.83 | 1.72 | 5.23 | 1.78 |
| | No university | 4.79 | 1.85 | 4.89 | 1.60 | 4.96 | 1.58 |

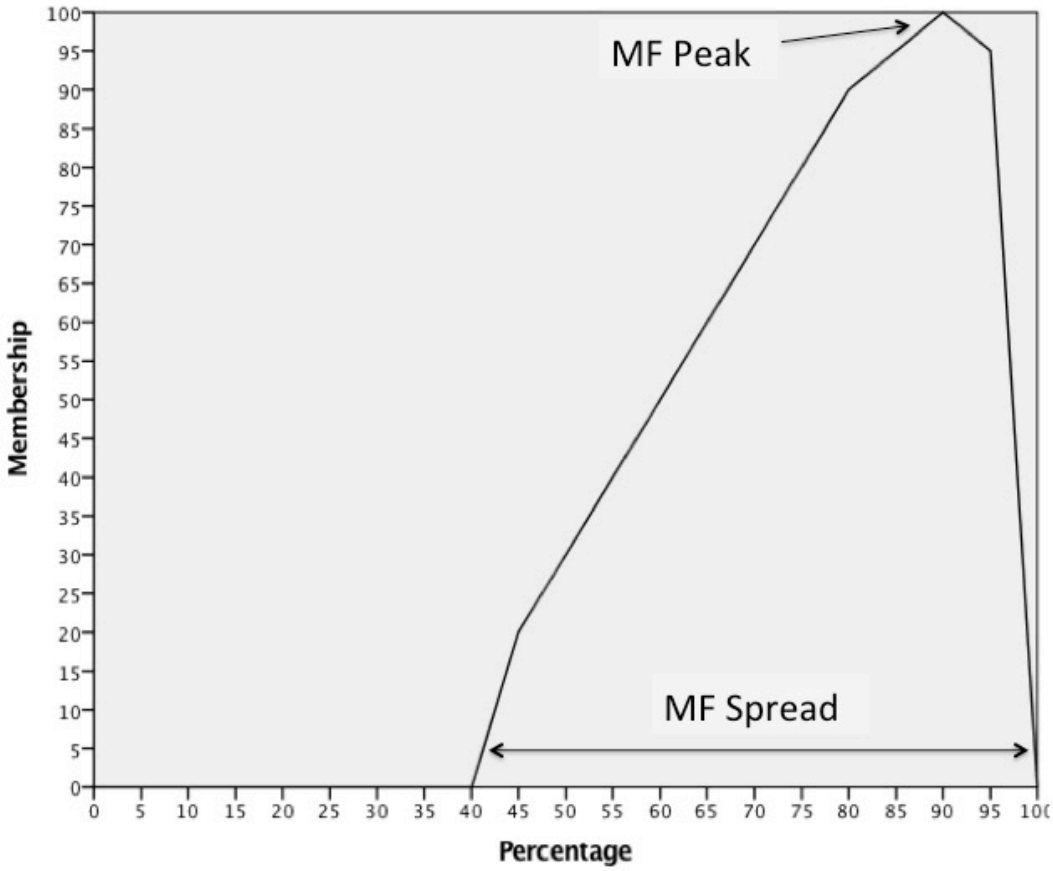
***Note***. Confidence was measured on 7-point scales with higher values representing greater confidence in applying the standard.

*Figure 1.*

Two-Way Interaction Effect of Gender by Education Level for Confidence in Applying BRD



**Appendix**


MF Peak
MF Spread
Membership
Percentage
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100