Vol. 13 No. 1 ISSN 2220-6140

Editorial

Journal of Economics and Behavioral Studies (JEBS) provides distinct avenue for quality research in the ever-changing fields of economics & behavioral studies and related disciplines. Research work submitted for publication consideration should not merely limited to conceptualization of economics and behavioral developments but comprise interdisciplinary and multi-facet approaches to economics and behavioral theories and practices as well as general transformations in the fields. Scope of the JEBS includes: subjects of managerial economics, financial economics, development economics, finance, economics, financial psychology, strategic management, organizational behavior, human behavior, marketing, human resource management and behavioral finance. Author(s) should declare that work submitted to the journal is original, not under consideration for publication by another journal and that all listed authors approve its submission to JEBS. Author (s) can submit: Research Paper, Conceptual Paper, Case Studies and Book Review. Journal received research submission related to all aspects of major themes and tracks. All submitted papers were first assessed by the editorial team for relevance and originality of the work and blindly peer-reviewed by the external reviewers depending on the subject matter of the paper. After the rigorous peer-review process, the submitted papers were selected based on originality, significance, and clarity of the purpose. The current issue of JEBS comprises papers of scholars from Namibia, USA, Germany and Zimbabwe. Analysis of the Money Demand Function for Zambia, Influence of Empathy on Moral Judgments, To Analyze Communication Data from Laboratory Experiments without Being a Machine Learning Specialist, Assessing Funding Mechanism Available for Mining Companies and Random Actions in Experimental Zero-Sum Games were some of the major practices and concepts examined in these studies. The current issue will therefore be a unique offer where scholars will be able to appreciate the latest results in their field of expertise and to acquire additional knowledge in other relevant fields.

Editor In Chief

Prof Dr Ijaz

TABLE OF CONTENTS

PAPERS

Journal of Economics and Behavioral Studies (JEBS)
Vol. 13, No. 1, February 2021 (ISSN 2220-6140)

An Analysis of the Money Demand Function for Zambia: A Gregory Hansen Cointegration Approach

Peter Nsokolo Mumba, Emmanuel Ziramba
Economics Department, University of Namibia, Pioneerspark, Windhoek, Namibia
peternsoxmumba@gmail.com

Abstract: The objective of this study was to analyze the money demand function for Zambia for the period 1978 – 2018 using annual time series data. The study employed the Gregory Hansen cointegration technique. The study also employed Hendry's General to Specific technique to estimate the error correction model by obtaining a parsimonious model. The results of the Gregory Hansen test confirmed the presence of a cointegrating relationship and selected the GH-2 model as the most plausible model with a level shift and a trend. The results also endogenously determined 1994 as the break year in the money demand function. Other interesting results obtained by the study suggest that inflation and interest rate are the robust determinants of real money demand both in the short and long run. Furthermore, unlike many other developing countries, the results show that money is a necessity in Zambia. The other interesting results suggested by the study are that the financial sector reforms of 1994 diminished the demand for real money; however, the positive time trend suggests that there has been an increase in real money holdings over time in Zambia. The low-interest elasticity of money demand also potentially compromises the effectiveness of money supply as a monetary policy tool for economic stabilization. The results of the CUSUM and CUSUMSQ confirm the stability of the money demand function in Zambia.

Keywords: *Gregory-Hansen, Money demand function, Structural break.*

1. Introduction

Monetary policies have been widely used by many developed and developing countries to achieve low levels of inflation and to stimulate economic growth. Arguably, the money demand function assumes a significant part in the detailing of monetary policy strategies. According to Cinar and Nur (2018), the money demand function relates money demand that is made by various motives to its determinants. A sound comprehension of the stability and determinants of the money demand function is key in the implementation of monetary policy. This enables the Central Bank to implementation of policy-driven changes in monetary aggregates to influences macroeconomic variables[1] (Subraram, 1999; Nachega, 2011; Halicioglu & Ugur, 2005). Over the past three decades, many countries have revised their economic structures. Starting from the deregulation policies of the 1980s, where industrial and labor policies in most developed countries moved from direct government intervention policies to market forces and competition[2]. Initially, market-based monetary policy instruments were not plausible. This was due to the perceived underdeveloped financial markets and the control of interest rates. Consequently, many developing countries later started adopting similar structural changes.

Hence, many countries later abandoned the use of direct monetary policy instruments (Omotor, 2011). Advocates of financial liberalization supported the shift to a more robust system in which monetary variables would be determined by market forces and competition. In Zambia, financial liberalization began in 1987 through to 1993, after the cancellation of the structural adjustment program and the 1989-1993 Economic Recovery Program (ERP) (Maimbika & Mumangeni, 2016). From 1964, Zambia has explored at least three monetary policy frameworks. This has entailed the shift in monetary policy instruments and their targets. It is opined that developments in the monetary system might affect the stability of the money demand function (Al Rasasi, 2016; Soto & Tapia, 2001). Generally, there is scant literature on the stability of money demand especially in Sub-Saharan Africa (SSA). Although some researchers have estimated the stability of money demand (Al Rasasi, 2016; Kjosevski, 2013; Soto & Tapia, 2001), a limited number of studies have explored the

[1]For more details on the importance of a stable money demand function, see Nduka (2014).
[2]The objective of this move was to enhance the resilience of economies, promote resource allocation and, to facilitate economic growth, see Pera (1989).

money demand function stability in light of structural breaks (Nyong, 2014; Omotor (2011); Nachega, 2011; Soto & Tapia, 2001).

Particularly, in Zambia, Zgambo and Chileshe (2014) empirically analyzed money demand stability in Zambia but their study did not account for structural breaks. Therefore, this study purposes to fill this literature gap. Furthermore, this study purposes to endogenously determine the break date in the cointegrated money demand function for Zambia. Additionally, it purposes to investigate the robust determinants and stability of the money demand function for Zambia in light of the endogenously determined break date. Such an empirical inquiry can be done by applying cointegration methods in models with the regime and trend shifts[3]. The rest of this paper is organized as follows; section 2 looks at the overview of the Zambian monetary policy frameworks, then the relevant literature reviewed is presented in section 3. Section 4 gives the theoretical framework and outlines the specific methodology that is employed. The results are presented in section 5. Section 6 presents the conclusion and recommendations from the study.

Review of the Regime shifts and Macroeconomic Performance in Zambia: The monetary policy framework in Zambia has undergone some changes in the recent past particularly following multiparty political independence in 1964, the Structural Adjustment Programme (SAP) of the late 1980s and financial liberalization of the early 1990s.

The Period 1964 – 1991: After Zambia's independence in 1964, all economic activities were controlled by the Government through nationalization and monetary policy in Zambia were performed using direct instruments and were guided by multiple objectives. Because of this, the economy plunged into long-term stagflation[4]. This to a large extent propagated the emergence of a new Government in 1991[5] (Maimbika & Mumangeni, 2016; Kalyalya, 2001; Zgambo & Chileshe, 2014).

The Period 1991 – 2011: The new government liberalized the economy in 1991[6]. After 1991, the Bank of Zambia (BoZ) adopted the Monetary Aggregate Targeting (MAT) framework. According to Zgambo and Chileshe (2014), the MAT framework was based on a strong and stable relationship between monetary aggregates and inflation which was the primary monetary policy target. This framework is believed to have been effective looking at the reduction in inflation rates from triple digits such as 165.7% in 1990 and 183.34% in 1991 to current single digits; 6.6% in 2015 and 7.5% in 2016 (World Development Indicators). In 1996, the Bank of Zambia was given autonomy to conduct monetary policy and began making use of indirect instruments (Kalyalya, 2001). These changes led to an improvement in Zambia's macroeconomic environment.

The Period After 2011: In 2012, inspired by a new modernized monetary policy framework, BoZ introduced the Policy Rate in retaliation to the MAT framework to achieve price stability. The Policy Rate provides a credible anchor in determining interest rates. Following the introduction of the Policy Rate, the operational target of monetary policy shifted to interest rates from reserve money (BoZ Monetary Policy Statement, December 2012). The introduction of the policy rate also poses several advantages for macroeconomic management in Zambia[7].

2. Literature Review

Although there is scant empirical work on the stability of money demand and its function. Some studies have made use of panel data techniques such as Naraya et al. (2009) who assessed the money demand function for five South Asian countries between 1974 and 2002. The study employed panel cointegration tests and

[3] See Gregory and Hansen (1996a; 1996b).
[4] This was as a result of the global oil price shocks of the 1970s and the plunge in the global prices of copper.
[5] Movement for Multi-party Democracy (MMD) which adopted economic and structural reforms
[6] This included the decontrolling of interest rates and the removal of exchange rate controls among others, see Kalyalya (2001).
[7] See Zgambo and Chileshe (2014) and BoZ Monetary Policy Statement (December, 2012)

established that money demand is cointegrated with its determinants. These include real exchange rates, real income, and both the short-term domestic and foreign interest rates. Hamdi et al. (2015) also made a similar inquiry in the Gulf Cooperation Council Countries between 1980Q1 and 2011Q4. The study applied panel cointegration tests. The results reveal that there is cointegration in the model. The results suggest a stable long-run money demand function.

These results are similar to the ones obtained by Narayan et al. (2009) who used a similar methodology. The Granger non-causality test procedure suggests bidirectional causality. Other studies have explored time series techniques making use of the error correction model. Vega (1998) estimated the money demand stability for Spain using structural stability tests in regressions with variables integrated of order one between 1979 and 1995. This study also used the error correction model. The results indicate that financial system openness affects the long-run stability of the money demand function. Lestano et al. (2011) estimated the stability of narrow money demand in Indonesia between 1980Q1 and 2004Q4 making use of an Autoregressive Distributed Lag (ARDL) model. The findings suggest that broad and narrow money demand equations are cointegrated. The results also reveal that the narrow money demand is stable, whereas the converse is true for broad money demand. Cziraky and Gillman (2006) used monthly data to estimate the money demand for Croatia from 1994 to 2002. A two-equation cointegrated system was used and evidence shows that there is a stable money demand that rapidly convergences back to equilibrium after-shocks. Other studies also used the unrestricted error correction model such as Al Rassai (2016) for Saudi Arabia, who assessed the stability of money demand between 1993Q1 and 2015Q3. The study applied the Johansen cointegration test and the findings suggest stability of money demand in the long run. Likewise, the results also suggest that the long-run estimates are consistent with theoretical expectations. As opposed to using the conventional CUSUM and CUSUMSQ stability tests, this study used Hansen (1992) stability tests and established that the money demand.

In the same way, Kjosevski (2013) investigated the determinants and stability of money demand in Macedonia. This study employed monthly data from January 2005 to October 2012. The results of the VECM provide evidence that exchange rate and interest rates explain most long-run variations of money. A few studies have used estimation techniques that allow for structural changes. Omotor (2011) used the Gregory and Hansen procedure to analyze the demand for money in Nigeria in light of structural breaks for the period 1960 - 2008. The study determined that 1994 was the endogenous break date. Like previous studies, the findings of this study also suggest a stable money demand function for Nigeria. Kumar, Webber and Fargher (2013) also made use of the same methodology to determine the level and stability of narrow money demand in Nigeria for the same period. However, unlike the results obtained by Omotor (2011), the findings of this paper suggest that the improved the scale economies of money demand to a less extent and money demand is stable. These results agree with those obtained by Nduka (2014) who also made use of the same methodology by analyzing the behavior of money demand in India between 1953 and 2008. The results of this study confirm the presence of cointegration. The results also suggest a break in the year 1965. Additionally, the study suggests a reduction in the demand for money by about 0.33% around the break year. The results also suggest that the demand for money is stable except between the years 1975 and 1998. These results are similar to those obtained by Omotor (2011) and; Kumar et al. (2013). Similarly, Nyong (2014) estimated the demand for money in the Gambia between 1986Q1 and 2012Q4 in light or regime shifts. The findings show that there exists a cointegrating relationship between money and its determinants namely income, inflation, exchange rate and interest rate. The results further suggest a structural break in 1995Q1.

The results also suggest the instability of money demand. However, the stability results are contrary to the findings for Omotor (2011) mainly due to the military coup in the Gambia and fall in foreign aid during the period. Very few studies on the stability of the money demand function have been done in Zambia. Zgambo and Chileshe (2014) modelled the money demand function in Zambia using the Autoregressive Distributed Lag (ARDL). The findings indicate that exchange rates, treasury bills rates and real income affect the money demand function in the long-run while inflation plays a similar role in the short-run. The findings also show that the money demand function stable and this iterates the relevance of monetary aggregates in the implementation of monetary policy in Zambia. Mutoti et al. (2012) in a similar study established that income, exchange rate and 90 days Treasury bill rate all affect money demand. The study also shows that the time trend which was used as a proxy for financial liberalization is positively related to money demand. The study also established that Zambia's demand for money function is stable. All these results confirm the finds of

Zgambo and Chileshe (2014). Another study by Adam (1999) analyzed monetary policy reforms in Zambia. The findings of the study suggest a stable money demand function with a break in the long run. These results are in agreement with the results of other studies like Zgambo and Chileshe (2014) and Mutoti et al. (2012). Also, the findings suggest that there is an increase in the variation of money demand around 1989, but it begins to reduce around 1994. From the literature reviewed, researchers have explored various methodologies on various data sets there is, however, a gap in empirical work on the stability of the money demand function allowing.

3. Theoretical Framework and Methodology

Money demand theory originated from the Keynesian Liquidity preference theory of holding money and the contributions from the monetarists such as Milton Friedman (1956). Additionally, the inventory theory[8] also contributed to the extensions of the Keynesian Liquidity preference theory of holding money (Nyong, 2014). Keynes Liquidity preference theory postulates that there are three motives for holding real money balances: transaction motives; precautionary motives; and speculative motives for money demand. Under Classical economics, the transactions and precautionary motives of money demand argue that the demand for real money balances depends on the interest rates (Zgambo & Chileshe, 2014). Under the speculative motive, Keynes argued that interest rates cause uncertainty about the future and this may influence the demand for money. For structural breaks especially in Zambia, no study has used the Gregory-Hansen cointegration approach in the presence of an endogenously determined structural break. Keynes believed that money does not earn any interest because it is a perfectly liquid asset. On the other hand, bonds pay interest on future income.

Several authors (Baumol, 1952; Tobin, 1956; Friedman, 1956) have contributed to theoretical literature by outlining theoretical distinctions between the transactions demand and the asset motive. Theoretically, real GDP positively affects the demand for money whilst interest rates negatively affect the price level as shown below (Zgambo & Chileshe, 2014). These relationships have been summarized in Keynes liquidity preferences theory equation below:

$$\frac{M^d}{P} = (Y^{(+)}, I^{(-)}) \qquad (4.1)$$

Where M is the demand for nominal demand for money; P is the nominal price level; therefore $\frac{M^d}{P}$ is the demand for real money; Y is real income; I is the interest rate. The interest rate is the rate of return on money and also the opportunity cost of holding money. The liquidity preferences theory equation assumes a unit elasticity of the nominal cash balances concerning the price level. On the other hand, an unstable money demand function implies that changes in money supply will not be closely related to prices and income hence it will be difficult to control inflation using adjustments in money supply (Zgambo & Chileshe, 2014). Based on the theoretical framework and empirical the suitable real money demand function is one in which real money demand a function of income, domestic interest rates and expected inflation. Furthermore, the Error Correction Model (ECM) is also observed to be the model of choice by many authors (Zgambo and Chileshe, 2014; Al Rassai (2016); Lestano et al. (2011) among others). This study adopts the model used by Lungu et al. (2012) and Zgambo and Chileshe (2014). This study also estimates money demand in the log-linear form, allowing the monetary aggregates and the scale variables to be expressed in logarithms (Akinlo, 2006; Omotor, 2011).

The model is therefore specified as follows:

$$In\frac{M^d}{P_t} = \beta_0 + \beta_1 lnY_t + \beta_2 E_t + \beta_3 R_t + \beta_4 INF_t + \varepsilon_t \qquad (4.2)$$

$$\beta_1, \beta_3 > 0\ ;\ \beta_2, \beta_4 < 0\ or > 0$$

Where: $In\frac{M^d}{P_t}$ is the natural logarithm of the real demand for money *(lnRM)* which is broad money divided by the consumer price index (*M3 / annual change in inflation rate*), lnY_t is the natural logarithm of *GDP*, E_t is the

[8] See Baumol (1952) and Tobin (1956, 1958).

real effective exchange rate, R_t is the real interest rate, INF_t is the annual *GDP* deflator and ε_t is the error, term. This study makes use of secondary annual data for the period 1978 to 2018. Real effective exchange rate data were obtained from Bruegel data sets and the rest of the data from the World Development Indicators. When dealing with time-series data, it is common practice to test for the presence of a unit root to avoid spurious regressions which offer misleading estimates. It is important to note that Zambia has experienced regime changes as a result of the reforms put forward by the IMF. Hence, the data is likely to have structural breaks. Therefore, it is not enough to rely on conventional unit root tests[9]. Therefore, this study further employs the Zivot and Andrews (ZA) unit root test in the presence of a structural break. Due to a potential structural break, standard cointegration tests are not appropriate.

Therefore, this study employs the Gregory-Hansen (G-H) (1996a; 1996b)[10] cointegration technique that allows the structural break to be endogenously determined by the model (Sadeghi & Ramakrishna, 2014). It has many advantages over conventional cointegration tests in light of structural breaks. It can test cointegration in the presence of structural changes (Gregory & Hansen, 1992). The G-H is cointegration tests on the residuals that propose against the alternative hypothesis that there may be one break in the cointegrating vector. The G-H makes use of three models to test for cointegration in light of a structural break in the cointegrating vector. According to Sadeghi and Ramakrishna (2014), these three models take into account the existence of a potentially unknown endogenous single break date. They allow for structural changes in either the intercept alone, in both trend and level shift and a full break. The G-H is based on the *ADF,* Z_α and Z tests for cointegration and they do not provide any information concerning the timing of the break (Gregory & Hansen, 1996). Considering cointegration with a trend and no structural change:

$$y_{1t} = \mu + \beta_t + \alpha^T y_{2t} + \varepsilon_t \qquad t = 1, \dots, n \qquad (4.3)$$

Where y_{2t} is *I* (1) and ε_t is *I* (0). The general structural change considered in The Gregory and Hansen (1996) considers a general structure change that only allows changes in the intercept μ and/or the slope α but not the trend β. To model the structural change, a structural dummy variable as defined below is used:

$$D_{tb} = \begin{matrix} 0, & if\ t \leq [Tb] \\ 1, & if\ t > [Tb] \end{matrix}$$

Where the unknown parameter $b \in (0, 1)$ is the change point, and [ ] is the integer part. The G-H, three models follow the pattern of the structural change as follows:

Model 1: Level Shift (C)

$$Y_{1t} = \alpha_0 + \alpha_1 D_{tb} + \beta_1^T y_{2t} + \varepsilon_t \qquad t = 1, \dots, T \qquad (4.4)$$

Model 2: Level Shift with Trend (C/T)

$$Y_{1t} = \alpha_0 + \alpha_1 D_{tb} + \gamma t + \beta_1^T y_{2t} + \varepsilon_t \qquad t = 1, \dots, T \qquad (4.5)$$

Model 3: Regime Shift (C/S)

$$Y_{1t} = \alpha_0 + \alpha_1 D_{tb} + \beta_1^T y_{2t} + \beta_{11}^T y_{2t} D_{tb} + \varepsilon_t \qquad t = 1, \dots, T \qquad (4.6)$$

Where; equation (4.4) is the level shift, equation (4.5) is the level shift with a trend and equation (4.6) is the regime shift (structural change) and D is the break dummy. The G-H cointegration test is based on three test statistics. The small values of the test statistics provide enough evidence against the null hypothesis. These test statistics are:

$$Z_\alpha^* = \inf_{b \epsilon T} Z_\alpha(b) \qquad (4.7)$$

$$Z_t^* = \inf_{b \epsilon T} Z_t(b) \qquad (4.8)$$

$$ADF^* = \inf_{b \epsilon T} ADF(b) \qquad (4.9)$$

[9] These tests might give misleading results if the series have structural breaks by failing to reject the hypothesis that series have unit root in the presence of structural break.

[10] The Gregory-Hansen methodology draws it's foundations from the Engle-Granger (1987) cointegration analysis (Omotor, 2011).

According to Gregory and Hansen (1996), simulation is used to approximate the limiting distribution of the test statistics (4.7) - (4.9) which are then calculated by fitting a response surface (MacKinnon, 1991). Therefore, to analyze the stability of the money demand function for Zambia, equation (4.2) is applied to equations (4.7) – (4.9).

GH-1: C: The crush model

$$In\frac{M^d}{P_t} = \alpha_0 + \alpha_1 D_{tb} + \beta_1 lnY_t + \beta_2 E_t + \beta_3 R_t + \beta_4 INF_t + \varepsilon_{1t} \quad (4.10)$$

GH-2: C/T: Changing growth model

$$In\frac{M^d}{P_t} = \alpha_0 + \alpha_1 D_{tb} + \gamma t + \beta_1 lnY_t + \beta_2 E_t + \beta_3 R_t + \beta_4 INF_t + \varepsilon_{1t} \quad (4.11)$$

GH-3: C/S: Regime Shift

$$In\frac{M^d}{P_t} = \alpha_0 + \alpha_1 D_{tb} + \beta_1 lnY_t + \beta_{11} lnY_t D_{tb} + \beta_2 E_t + \beta_{22} E_t D_{tb} + \beta_3 R_t + \beta_{33} R_t D_{tb} + \beta_4 INF_t + \beta_{44} INF_t D_{tb} + \varepsilon_{1t} \quad (4.12)$$

Where; D_{tb} is the shift in the slope, intercept or trend coefficient, b is the break date. The parameter α_0, is the intercept prior to the shift, α_1 is the change in the intercept at the time of the break.

In model GH-3; β_1, β_2, β_3 and β_4 are the slope coefficients before the regime change. β_{11}, β_{22}, β_{33} and β_{44} are the change in slope coefficients at the time of the break. Following the cointegration test, the residuals from the selected model of the canonical specifications are used to estimate the Error-Correction Model (ECM). Hendry's General to Specific (GETS) technique is used to estimate the ECM by obtaining a parsimonious model. To test the stability of the parsimonious model, the CUSUM and CUSUMSQ tests are used.

4. Empirical Results and Analysis

According to table 1, the unit root results show that all the variables are I (1). However, both the ADF and PP unit root tests might be misleading in the presence of structural breaks due to their shortcomings. Therefore, the ZA unit root test in the presence of a structural break is used to complement these tests and the results are presented in Table 2.

Table 1: ADF and PP Unit Root Tests

Variables		ADF		PP	
		Levels	First Difference	Levels	First Difference
lnRM	***Intercept***	-0.228328	-9.773299***	-0.868210	-9.694295***
	Trend & Int	-1.161654	-10.00791***	-1.738280	-9.910215***
lnY	***Intercept***	0.114075	-4.598012***	0.114075	-4.554694***
	Trend & Int	-2.218635	-4.688897***	-1.712389	-4.580111***
E	***Intercept***	-1.120869	-5.925513***	-1.047277	-6.138222***
	Trend & Int	-2.385512	-5.876143***	-2.211517	-6.506151***
R	***Intercept***	-1.951388	-7.251833***	-2.023902	-7.201539***
	Trend & Int	-2.174016	-7.161103***	-2.291245	-7.114582***
INF	***Intercept***	-1.869009	-5.430166***	-1.975870	-5.444447***
	Trend & Int	-2.484109	-5.380117***	-2.329027	-5.394862***

Source: Author's computations. **Note:** *, ** and *** denote rejection of the null hypothesis at 10%, 5% and 1% significance level respectively.

Table 2: Zivot and Andrews (1992) unit root tests in the presence of a structural break

	InRM	***InY***	***E***	***INF***	***R***
Break Date	1992	2005	2005	1995	1993
t-statistic	-10.23425***	-3.387936	-4.079244	-8.170064***	-5.904098***

Source: Author's computations. **Note:** *, ** and *** denote rejection of the null hypothesis at 10%, 5% and 1% significance level respectively.

The ZA unit root test results show that all variables; M, Y, E, INF and R have breaks in 1992, 2005, 2005, 1995 and 1993 respectively. Due to the presence of breaks, conventional cointegration tests will not be feasible; hence, the G-H cointegration test is performed[11]. The results of the G-H cointegration test in table 3 suggest a strong presence of cointegration in the ADF and the Z_t statistics. The results suggest that GH-2 with a break date in 1994 is the most plausible model[12]. Based on this rule of thumb[2], it is evident that the break was in 1994 and if it is ignored, the estimation will lead to wrong inferences which are not best for the model. The identified structural break date reflects the year in which several financial sector reforms were enacted[13].
According to the GH-2 cointegration estimates suggested in table 4, income is less than unity but has a positive and insignificant relationship with real money demand, implying that real money demand is income inelastic. This suggests that money is a necessity and supports the transactions motive for holding money. This is consistent with the findings of Omotor (2011) and Yesigat, Rao and Nagaraja (2018) who argue that financial sector reforms and technological improvements in payment systems among other factors decrease the income elasticity of money demand. However, this is inconsistent with the findings of Nyong (2014) and Simawu, Mlambo and Muriwirapachena (2018). Meanwhile, like the results obtained by Nyong (2014) and Kjosevski (2013), the exchange rate has a negative but insignificant relationship with real money demand.

This negative relationship suggests that economic players may exchange foreign currency assets for domestic currency assets following a depreciation of the Zambian Kwacha (Zgambo & Chileshe, 2014). However, this contradicts with the findings of Asongu, Folarin, and Biekpe (2019) who found a positive effect for some West African countries. On the other hand, real money demand negatively and significantly responds to variations in both the real interest rate and inflation rate. This suggests that money and financial assets are substitutes. It also suggests that interest rate is the appropriate opportunity cost variable in the long run. These findings conform with the findings of Nyumuah (2017) and Mansaray and Swaray (2012) but contradict with the findings of Narayan, Narayan and Mishra (2007) and Asongu, Folarin, and Biekpe (2019).

Table 3: The Gregory and Hansen (1996) Cointegration Test with Structural Breaks

Gregory Hansen Models	*ADF*		Z_t		Z_α	
	GH test statistic	Break-Point	GH test statistic	Break-Point	GH test statistic	Break-Point
GH-1	-6.76 ***	2012	-6.85***	2012	-44.03	2012
GH-2	-9.33***	1994	-9.45***	1990	-56.99	1990
GH-3	-7.17***	1998	-7.26***	1998	-45.99	1998

Source: Author's computations. **Note:** *** denotes rejection of the null hypothesis at the 1% significance level.

However, the low-interest elasticity of money demand compromises the effectiveness of money supply as a monetary policy tool for economic stabilization in Zambia. Furthermore, financial sector reforms diminished the demand for real money as seen from the negative and significant break dummy. This finding is not as expected because the financial sector reforms were targeted towards improving financial competitiveness and hence, increasing demand for real money. However, these findings conform with the findings of but contradict the findings of Nyong (2014) and Mansaray and Swaray (2012). Both the time trend and intercept have positive and significant coefficients. This also supports the selection of the GH-2 equation which implies that the endogenous change was indeed a level shift with a time trend. A positive and significant time trend suggests an increase in real money demand over time. This finding is consistent with the findings of Mansaray and Swaray (2012).

[11] The G-H cointegration test is performed with the null hypothesis of no cointegration tested against the alternative hypothesis of cointegration with breaks on unknown dates.
[12] The break date is determined at the point where the absolute value of the ADF test statistic is at its maximum.
[13] Towards the end of 1993, interest rates where decontrolled, exchange rates where then liberalized in 1994 following the cessation of the Exchange Control Act in 1994. Another notable event was the introduction of the Treasury bill tender system, the establishment of the Lusaka Stock Exchange (LuSE) in February 1994 and the strengthening of the financial sector by the enactment of the Banking and Financial Services Act in December 1994.

Table 4: Cointegrating Equation

Variables (Break date)	GH–1 (2012)	GH–2 (1994)	GH–3 (1998)
Intercept			
	12.10856***	18.23676***	28.73105***
	(0.0002)	(0.0000)	(0.0005)
Trend	-	0.052915***	0.023068
	-	(0.0000)	(0.1315)
LnY	0.310715**	0.019275	-0.486991
	(0.0355)	(0.9098)	(0.1757)
E	-0.003048	-0.002143	0.007844
	(0.5206)	(0.5966)	(0.2716)
R	-0.017990***	-0.012047***	-0.022052***
	(0.0000)	(0.0070)	(0.0004)
INF	-0.011072***	-0.013567***	-0.012500***
	(0.0000)	(0.0000)	(0.0000)
Dum_1994	-	-0.965599***	-
	-	(0.0000)	-
Dum_1998	-	-	-26.89540***
	-	-	(0.0075)
Dum_2012	0.341683**	-	-
	(0.0213)	-	-
LnY_1998	-	-	1.237588***
	-	-	(0.0073)
E_1998	-	-	-0.020916**
	-	-	(0.0312)
R_1998	-	-	0.021016
	-	-	(0.2034)
INF_1998	-	-	0.019541
	-	-	(0.2627)

Source: Author's computations. P-values are in parenthesis. **Note:** *, ** and *** denote rejection of the null hypothesis at the 10%, 5% and 1% significance level respectively.

The ECM is estimated using the GETS technique[14]. The results of the parsimonious short-run regression in table 5 reveal that real money balances are influenced by income, real interest rate, the dynamics of inflation and the past values of real money demand. The results also show that the ECT has a negative and highly significant coefficient. This implies that real money demand is cointegrated into its determinants in the pre and post-financial liberalization periods. The ECT implies that approximately 98.7 percent of the disequilibrium in real money demand from the previous period's shocks on the determinants will converge back to the long-run equilibrium in the current year. The diagnostic tests also confirm the robustness of the results obtained in this study[15]. The results of CUSUM and CUSUM Square tests in figures 1 and 2 respectively

[14] The ECM is performed on the preferred model; the GH-2. To achieve this, the optimal number of lags to be used in the estimation of the over-parameterised model are selected based on the AIC criterion. The AIC criterion selects a lag length of three. Using three lags, the differenced series of real money demand are regressed on its lags, the differenced and lagged terms of real GDP, real interest rate, exchange rate, inflation and the lagged residuals from GH-2 hence obtaining the over-parameterised model. Hendry's GETS technique is then applied to reduce the number of the regressors until the most parsimonious model is obtained.

[15] The DW statistic confirms that the model has no autocorrelation. The LM test confirms that there is no serial correlation among the variables. The Breusch-Pagan-Godfrey test confirms the presence of homoscedasticity and the Ramsey RESET test also confirms that the model is correctly specified at 1% significance level.

suggest that despite the presence of a policy-induced structural break, real money demand is stable over time. These findings conform to the findings of Zgambo and Chileshe (2014) and Doguwa et al. (2014) but contradict the findings of Nyong (2014).

5. Conclusion and Policy Recommendations

A stable money demand function is an important ingredient for the formulation of sound monetary policy and economic growth. A number of authors have argued the effect of financial sector reforms on real money demand. Owing to the financial sector reforms in Zambia during the early 1990s, this study investigated the money demand function for Zambia. The main departure of this study from other studies is that it incorporates a dummy variable for financial sector reforms and puts forward the Gregory Hansen cointegration approach which is the recommended cointegration test in light of structural breaks. After testing three models with different endogenously determined break dates, the study identified 1994 as the break date. This was the year in which exchange rates were liberalized following the cessation of the Exchange Control Act in 1994. Another notable event was the introduction of the Treasury bill tender system, the establishment of the Lusaka Stock Exchange (LuSE) in February 1994 and the strengthening of the financial sector by the enactment of the Banking and Financial Services Act in December 1994.

The results of the study suggest that both in the long run and short run, real interest rate and inflation are the significant determinants for real money demand in Zambia. However, the interest elasticity is very low hence this seems to compromise the effectiveness of the money supply in economic stabilization. The negative inflation elasticity supports the theoretical expectations of Milton Friedman and suggests that increases in inflation will lead to substitution between money and other financial assets. Additionally, although insignificant, the positive income inelasticity supports the transactions motive for holding money and suggests that money is a necessity in Zambia. The negative insignificant exchange rate suggests that agents may substitution between foreign currency assets and domestic currency assets following depreciation in the Kwacha. The other interesting results suggested by the study are that the introduction of financial sector reforms diminished the demand for real money in Zambia. Also, the positive and significant time trend suggests an increase in real money holdings over time. Additionally, after estimating a parsimonious ECM, it is evident that the model is stable and appropriate as the error correction term is highly significant. Finally, the stability tests suggest evidence of the stability of the money demand.

This finding conforms to the unitary elasticity of the demand for money. This implies that the Central Bank can effectively target interest rates using monetary policy. The findings of this study pose several policy implications. The stability of real money demand suggests that the Bank of Zambia can still use the money supply for monetary policy implementation. However, the low-interest elasticity would mean that the money supply would not be very effective. Therefore, even as the Bank of Zambia has opted to modernize monetary policy by introducing the policy rate as the key operating target, it would, therefore, be suggested that the central bank adopts a mixture of operating targets. Finally, the Bank of Zambia should also closely monitor the real interest rate and inflation rate. The low-interest elasticity could also be due to the underdeveloped financial sector in Zambia. Therefore, efforts to develop the financial sector could also increase the effectiveness of the money supply for monetary policy implementation. The central bank should also try to keep moderate and stable levels of inflation to minimize agents from substituting money for other assets. This is because increasing levels of inflation tend to exert downward pressure on the demand for real money by increasing the return on other assets.

Table 5: Parsimonious Dynamic Short –Run Money Demand Estimates, Dependent Variable: LNRM

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.034084	0.021331	1.597881	0.1257
D(LNRM(-2))	0.367667	0.121771	3.019340	0.0068***
D(LNRM(-3))	0.270466	0.122554	2.206911	0.0392**
D(LNY)	-0.315490	0.125427	-2.515320	0.0205**
D(LNY(-2))	-0.359191	0.234679	-1.530563	0.1415

D(LNY(-3))	0.312739	0.259568	1.204844	0.2423
D(E(-1))	-0.000519	0.002459	-0.211085	0.8350
D(E(-2))	0.003064	0.003649	0.839487	0.4111
D(E(-3))	-0.008593	0.004340	-1.979785	0.0617
D(R)	-0.015922	0.004120	-3.864189	0.0010***
D(R(-1))	0.003876	0.003818	1.015203	0.3221
D(R(-2))	-0.006583	0.004199	-1.567733	0.1326
D(INF)	-0.017174	0.002029	-8.462084	0.0000***
D(INF(-1))	0.005532	0.001829	3.025468	0.0067***
D(INF(-2))	-0.005842	0.002193	-2.663506	0.0149**
D(INF(-3))	0.004349	0.001311	3.317271	0.0034***
ECT(-1)	-0.986690	0.148404	-6.648676	0.0000***

Source: Author's computations. **Note:** *, ** and *** denote rejection of the null hypothesis at the 1%, 5% and 10% significance level respectively.

Figure 1: Stability Test for Short-Run Money Demand (CUSUM)

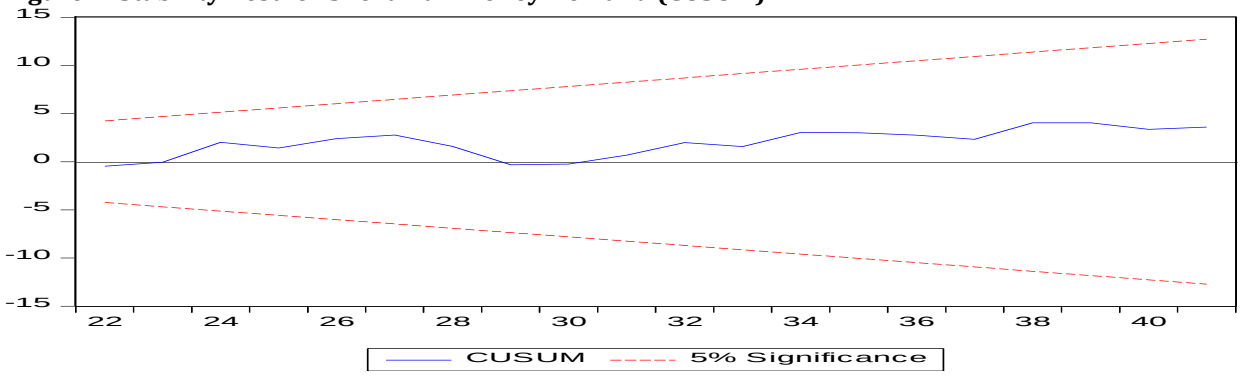
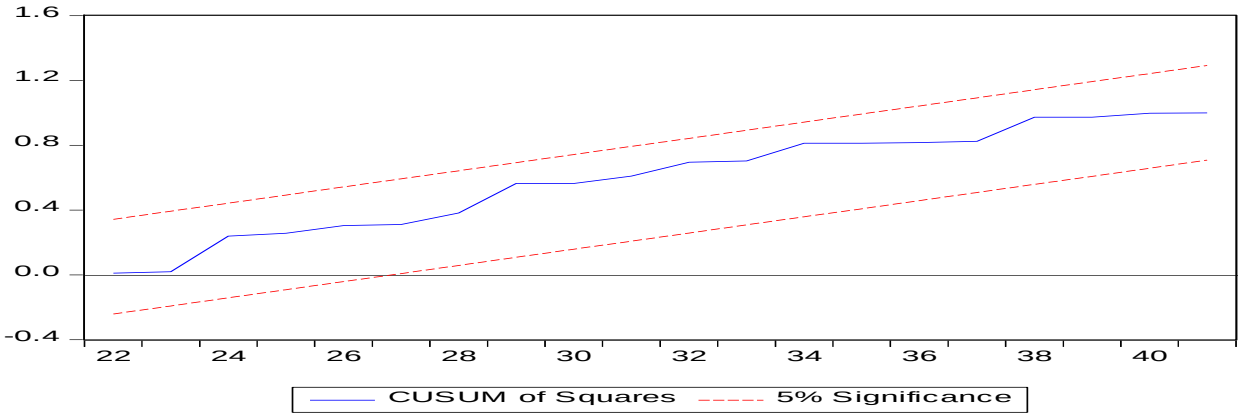


Figure 2: Stability Test for Short-Run Money Demand (CUSUMSQ)



References

Asongu, S., Folarin, O. & Biekpe, N. (2019). The long-run stability of money demand in the proposed West African monetary union. *Munich Personal RePEc Archive,* (92343), 1-28.

Adam, C. S. (1999). Financial liberalization and currency demand in Zambia. *Journal of African Economics,* 8(3), 268-306.

Al Rasasi, M. H. (2016). On the stability of money demand in Saudi Arabia. WP/16/6. Saudi Arabia Monetary Agency.

Bank of Zambia. (2012). Monetary Policy Statement.

Baumol, W. J. (1952). The Transactions demand for cash: An Inventory Theoretic Approach. *Quarterly Journal of Economics,* 66, 545-556.

Cinar, S. & Nur, H. B. (2018, July). Determinants and stability demand for money: A sample of BRIC-T countries. *The Empirical Economics Letters*, 17(7), 852-862.

Cziráky, D. & Gillman, M. (2006). Money demand in an EU Accession country: A VECM study of Croatia. *Bulletin of Economic Research,* 58(2), 105-127.

Doguwa, S. I., Olorunsola, O. E., Uyaebo, S. O., Adamu, I. & Bada, A. S. (2014). Structural breaks, cointegration and demand for money in Nigeria. *CBN Journal of Applied Statistics*, 5(1), 15-33.

Friedman, M. (1956). The quantity theory of money; A restatement. *Studies in Quantitive Theory of Money,* 5, 3-31.

Friedman, M. (1970). A theoretical framework for monetary analysis. *Journal of Political Economy,* 78(2), 193-238.

Gregory, A. W. & Hansen, B. E. (1992, 1996). Practitioners Corner: Tests for cointegration in models with regime and trend shifts. *Oxford Bulletin of Economics and Statistics,* 58(3), 555-565.

Halicioglu, F. & Ugur, M. (2005). On stability of the demand for money in a developing OECD country: The case of Turkey. *Global Business and Economics Review,* 5(2), 203-213.

Hamdi, H., Said, A. & Sbia, R. (2015). Empirical evidence on the long-run money demand function in the Gulf Cooperation Council countries. *International Journal of Economics and Financial Issues,* 5(2), 603-612.

Kalyalya, D. H. (2001). Monetary Policy framework and implementation in Zambia. South African Reserve Bank Conference on Monetary Policy Frameworks in Africa, 17-19.

Kjosevski, J. (2013). The determinants and stability of money demand in the Republic of Macedonia. *Zbornik radova Ekonomskog fakulteta u Rijeci: časopis za ekonomsku teoriju i praksu,* 31(1), 35-54.

Kumar, S., Webber, D. J. & Fargher, S. (2013). Money demand stability: A case study of Nigeria. *Journal of Policy Modeling,* 35(6), 978-991.

Lestano, L., Jacobs, J. & Kuper, G. H. (2011). Money demand stability in Indonesia. Working Paper.

Lungu, M., Simwaka, K., Chiumia, A., Palamuleni, A. & Jombo, W. (2012). Money demand function for Malawi: Implications for monetary policy conduct. *Banks and Bank Systems,* 7(1), 50-63.

MacKinnon, J. G. (1991). Critical values for cointegration tests. (J. G. MacKinnon, Ed.) Long-run Economic Relationships: Readings in cointegration relationships.

Maimbika, S. & Mumangeni, J. N. (2016). Analysis of the effects of financial liberalization on Zambia's economic growth. *International Review of Research in Emerging Markets and the Global Economy,* 2(2), 840-854.

Mansaray, M. & Swaray, S. (2012). Financial liberalization, monetary policy and money demand in Sierra Leone. *Journal of Monetary and Economic Integration,* 12(2), 62-90.

Mutoti, N., Kapambwe, C. M. & Zgambo, P. (2012). Inflation targeting in Zambia. Unpublished Bank of Zambia Working Paper.

Nachega, J. C. (2011). A cointegration analysis of broad money in Cameroon. International Monetary Fund, WP/01/26.

Narayan , P. K., Narayan, S. & Mi, V. (2009). Estimating money demand functions for South Asian countries. *Empirical Economics,* 36(3), 658-696.

Narayan, S., Narayan, P. K. & Mishra, V. (2009). Estimating money demand functions for South Asian countries. *Empirical Economics,* 36(3), 685-696.

Nduka, K. E. (2014). Structural breaks and the long-run stability of demand for real broad money function in Nigeria: A Gregory-Hansen Approach. *The Economics and Finance Letters,* 1(8), 79-89.

Nyong, M. O. (2014). The demand for money, structural breaks and monetary policy. *Developing Country Studies,* 4(19), 93-106.

Nyumuah, F. S. (2017). An investigation into the interest elasticity of demand for money in developing countries: A panel data approach. *International Journal of Economics and Finance,* 9(3), 69-80.

Omotor, D. G. (2011). Structural breaks, demand for money and monetary policy in Nigeria. *Ekonomski Pregled,* 62(9-10), 559-582.

Pera, A. (1989). Deregulation and privatization in an economy-wide context. *OECD Economic Studies,* 12, 159-204.

Pesaran, H. M., Shin, Y. & Richard, S. J. (2001). Bounds testing approaches to the analysis of level relationships. *Journal of Applied Econometrics,* 16(3), 289-326.

Sadeghi, A. & Ramakrishna, G. (2014). An empirical analysis of imports of Iran: A Gregory Hansen method of cointegration. *American Journal of Business, Economics and Management,* 2(4), 105-112.

Simawu, M., Mlambo, C. & Murwirapachena, G. (2018). An investigation into the demand for broad money in South Africa. *International Business & Economics Research Journal,* 13(6), 1419-1436.

Soto, R. & Tapia, M. (2001). Seasonal cointegration and the stability of the demand for money: A seasonal cointegration approach. Banco Central de Chile.

Subramanian, S. S. (1999). Survey of literature on demand for money: Theoretical and empirical work with special reference to error-correction models. International Monetary Fund.

Tobin, J. (1956). The interest elasticity of the transactions demand for cash. *Review of Economics and Statistics,* 38, 241-247.

Tobin, J. (1958). Liquidity preference as behavior towards risk. *The Review of Economic Studies,* 25(2), 65-86.

Vega, J. L. (1998). Money demand stability: Evidence from Spain. *Empirical Economics,* 23(3), 387-400.

Yesigat, A., Rao, K. T. & Nagaraja, G. (2018). Determinants of demand for money and its stability in Ethiopia. *International Journal of Research and Analytical Reviews,* 5(4), 349-360.

Zgambo, P. & Chileshe, P. M. (2014). Empirical analysis of the effectiveness of monetary policy in Zambia. COMESA Committee of Central Bankers (pp. 1-50). COMESA Monetary Institute.

Journal of Economics and Behavioral Studies (JEBS)
Vol. 13, No. 1, February 2021 (ISSN 2220-6140)

The Influence of Empathy on Moral Judgments

Rojhat Avsar, Rami Gabriel
Columbia College Chicago, Department of Humanities, History, and Social Sciences, USA
ravsar@colum.edu, rgabriel@colum.edu

Abstract: Empathy is expected to correlate with pro-social attitude s, but what effect does empathy have on judgments of distributive fairness? In our study, we found that participants with higher empathy scores on the Interpersonal Reactivity Index (IRI) were more likely to: (a) favor the use of egalitarian distribution when the joint effort is involved, and (b) deem overly self-interested or opportunistic behavior unfair. Female participants were more consistent in the exercise of moral judgments across diverse scenarios. Furthermore, empathy has several dimensions (e.g., perspective-taking or empathetic concern) and we observed that they interacted with gender and the nature of the hypothetical problem differently in some cases. Although the findings of the study are not counterintuitive, it has identified some avenues for further explorations and highlighted some potential methodological shortcomings of the IRI as a measure of empathetic traits.

Keywords: *Empathy, perspective taking, fairness, gender, distributive justice, social norms.*

1. Introduction: How does Empathy Inform Judgments of Fairness?

Do we decide what's fair based upon our reasoned judgments or our emotional reactions? This distinction is of paramount importance for our notions of responsibility, our appraisal of a person's character, and the functioning of our legal system. In fact, since social and communal interaction by a set of individuals is the very project of civilization, this question intrigued early political philosophers. During the Enlightenment, emotions were considered uncontrollable passions that distort reasoning or forces that weaken the flourishing of the will. Some philosophers focused on the moral and ethical nature of emotions, and how these forces played out in our forms of social organization. A clear understanding of empathy and its neural correlates bears upon larger questions of morality and social living (De Oliveira-Souza, Zhan & Moll, 2014). In this paper, we describe findings from an empirical investigation of how empathetic traits redound to notions of fair distribution and opportunism in economic exchange. Judging the fairness of any given act engages the emotional experience of empathy. It thus serves as a crucial locus of our interpretation of behavior and subsequent actions.

Yet, there may be a fundamental difference between our knowledge of external objects, our self-knowledge and our knowledge of others (Zahavi, 2014), and when we enter the realm of ethics, judgments of fairness may not be reducible empathy (Thompson, 2005). We will argue Empathy (see review, Cuff, Brown, Taylor & Howat, 2014) as manifested in personality traits has an effect on ethical decisions. Empathic processes consist of a complex set of interactions between cognitive and affective components (Heberlein & Saxe, 2005; Strayer, 1987)[16]. Empathy is best thought of as a set of processes, from evolutionarily earliest in affective states to the latest in linguistically-shaped descriptions. Perception of another person's emotional state or perspective-taking, i.e. putting one in someone else's shoes, generally activates one's feelings (though there are differences in brain activity, see Preston et al., 2007). Since it captures the range, function, and phenomenology of the phenomenon. We sponsor a view of empathy forged by animal ethologists like Preston & de Waal (2002) who considered the evolutionarily earliest form of empathy processes to be 'emotional contagion, a type of affective resonance.

[16] Cuff et al. (2014) put forward the following definition:
Empathy is an emotional response (affective), dependent upon the interaction between trait capacities and state influences. Empathic processes are automatically elicited but are also shaped by top-down control processes. The resulting emotion is similar to one's perception (directly experienced or imagined) and understanding (cognitive empathy) of the stimulus emotion, with recognition that the source of the emotion is not one's own (p. 7).

Where an individual (unconsciously, and possibly through subtle nonverbal cues) picks up the emotion of people around her (think for example of how infants take on the emotions of their caregiver). The next layer is the ability to cognitively (and usually consciously) empathize with another creature. The third, and most abstract layer, is the ability to consciously take the perspective of another creature (De Waal, 2007). A reasoned judgment is always based upon the facts of the case, and some cases include more emotional facts than others. One recent model suggests empathy consists of four components: (1) a type of affective sharing based on perception-action coupling, (2) a judgment of the distinction and relation between self and other, (3) a type of mental flexibility which allows the ability to adapt the perspective of another, and finally, (4) empathy as a form of regulating one's own emotions in a social scenario (Decety, 2007; Decety & Michalska, 2010). Decety et al. argue that there are at least three different types of empathy, cognitive empathy, affective sharing (i.e. emotional contagion), and prosocial motivation (Decety & Jackson, 2004; Decety & Cowell, 2014). Another model suggests empathy is a type of simulation, where the act of imagination of what it might be like to be in the situation another person is in is a cognitive simulation. Whether it is simply the affective resonance of matching emotional state, or a conscious act of trying to understand what it is like to be someone else, feeling empathy for someone changes the reasoning process. Reasoning about fairness in the case of in-group members is different, i.e. more empathetic, judgment than reasoning about non-group members.

The three-layered empathy complex we espouse probably evolved for allowing social animals to interact within-group conspecifics (Thompson, 2005). Do we experience empathy more, or only in, our interactions with members of our own in-group? This would suggest that making decisions regarding fairness feels different, i.e. generates empathy, only in certain cases. If this were true, then making a judgment about fairness concerning two individuals in the same situation, say your cousin who is being sentenced for drunk driving, and someone on the other side of town who is being sentenced for the same crime will feel different for you. The in-group-out-group bias reflects the empirically substantiated tenet that Individuals tend to treat members of their in-group in a more egalitarian fashion relative to members of out-groups (Eferson, Lalive & Fehr, 2008). Larger questions arise as well, like how large can one's in-group be? Is it more accurate to say there are concentric circles of in-groups, say from the brotherhood of man to ethnic groups, neighbors, and family? If so, is there a continuum of feelings of empathy, from a little blip for making a judgment of fairness for someone who went to the same high school to someone who likes the same hockey team to surges of empathy for intimate friends (Dunbar, 1998; de Waal, 1996)? Now that we have some clarity on the role of empathy, how can we characterize reasoned judgments, i.e. judgments based on critical comprehension and subsequent logical contemplation which result in decisions followed by actions? Such judgment must be based on the facts of the case.

For example, what are the reasons given for such action, what is the context, at what time did the action take place, etc. Beyond the facts upon which the judgment is made, there are the prospective implications of making the judgment. For example, being asked to gamble a hypothetical amount of 'money in a psychology experiment is very different from gambling with real rupees in a card game. Prospective implications often frame the reasoned judgment; it is a part of the set of pertinent facts albeit emotional and imaginative. Maybe a more accurate question is: are emotions, like empathy, the same kind of facts to be considered as other kinds of facts, like the causes and consequences of a given act? Does a hint of empathy affect the reasoned judgment as much as a simple fact about the case, such as whether the crime being judged took place during the day or the night? Therefore, a decision concerning fairness is always dependent upon the level and type of emotion that informs the reasoned judgment being made. It turns out our colloquial descriptions of reason have yet to take in the psychological evidence that emotion is one of the rivers that flow into the lake of Reason (Damasio, 1994). It is a river of swirling currents that transport the black soot down from the mountain and divulges an inordinate fertility to the contemplative field. We believe that economic behavior is not satisfactorily captured by the rationality paradigm and is in fact driven by a set of complex motivations, such as empathetic notions of fairness. Accordingly, this study investigates whether empathetic traits predict moral judgments in a set of hypothetical scenarios. In the next section, we discuss how economists have addressed this aspect of economic decisions.

2. Complexity of Economic Motivation

Economists describe individual differences through the concept of diversity in "preferences." They seek to model moral judgments that consider idiosyncratic variations across individuals. According to the approach popularized by Gary Becker (1976), differences in value judgments could simply be explained as differences in "taste." It is not inconceivable that certain "types" emerge when individuals are put in a choice situation in which economic incentives are modified by moral considerations. Modeling the possible heterogeneity of moral types attracted the scholarly attention of economists of experimental leanings. For instance, though there is always fluidity across these categories modified by the context, in Public Good games, it is possible to identify some individuals as "altruists" (i.e. those who contribute to the common pool generously) or "reciprocals" (i.e. those who contribute to the common pool as long as others are doing it) based on their reactions to the scenarios. That is to say, personality types have empirical effects when tested through scenarios. In this study, we are particularly interested in the notion of "fair" distribution and "socially appropriate" (or moral) behavior. In distributing resources, fairness or social propriety can take on several meanings. Sometimes, it is the equal split that resonates as fair; we call this "egalitarian." At other times, merit-based considerations dominate, and these results in a type of distribution we call "utilitarian[17]." Egalitarian distributions are not sensitive to either difference in skills or the productive contribution of the parties involved.

While utilitarian distributions favor an accounting method that prioritizes productivity above all other potential considerations. It is of course possible that any given individual's moral judgment consists of elements of both egalitarian and utilitarian types, depending on the case of in-group favoritism. Further, it is reasonable to expect fairness considerations to be sensitive to social context and display some (but not complete) convergence when additional information is provided about the circumstances of potential recipients in a resource-distribution scheme. Experimental findings from Frohlich, Oppenheimer, & Eavery (1987), Faravelli (2007), Cappelen et al. (2007), Côté, Piff, and Willer (2013) among others, appear to corroborate our expectation about a potential convergence in moral judgments when the participants are primed to feel more empathetic. Côté, Piff, and Willer (2013) showed that disinterested upper-class participants (recruited through MTurk as we did) whose judgments were originally geared toward maximizing the total gains irrespective of its effects on the least advantaged ("lose member") in a resource distribution game would be no more utilitarian than their lower-class counterparts once empathy was induced[18]. Frohlich and his colleagues (1987) found that a choice maximizing the average income with a floor that prevents extreme destitution in distribution is most likely. Particularly when participants are not given all the necessary information or asked to reach a consensus.

Faravelli's (2007) results are comparable: Rawlsian minimax (as opposed to egalitarian or utilitarian) in evaluating the fairness of various resource distribution outcomes is the prevalent choice of participants majoring in economics. More interestingly, if when the differences in productivity are explained in terms of circumstances beyond the control of the individual (e.g. in the case of an injury incurred), the Rawlsian choice becomes much more likely. In a clever experimental setting (a dictator game preceded with a production

[17] Utilitarianism, in the Benthamite sense of the term, not the later iterations, could result in equal distribution justified by the declining marginal utility of income. We use the term in the way in which it is commonly understood/interpreted in economic literature today that finds its clearest expression in the marginal productivity theory of distribution: allocating more resources to the higher productivity individuals maximize the total pie. A merit-blind distribution, on the other hand, creates disincentives, as the argument goes.

[18] The authors had the upper-class participants read the following instructions before they played the game: "As you make your decision, think about the feelings and the wellbeing of the 'lose member' of the group. Concentrate on trying to imagine how the 'lose member' feels and how your decision will influence him or her. Try to feel the impact of your decision on how the 'lose member' of your group will feel."

stage), Cappelen and his colleagues (2007) allowed participants to tease out how fairness judgments come about as a result of an interplay between considerations of equity and egalitarian instincts. Although they reported a significant plurality of fairness ideals among the participants, the "liberal egalitarian" position that, of the options available, is the most sensitive to factors out of an individual's control (e.g. luck) appears to be the choice of the majority participating in the experiment. How do we come to decide what we think is fair? Is it the context of the scenario or the individual's upbringing that determines their notion of fairness? What is the role of personality factors in judgments of fairness? We posit that the notion of what constitutes fair or socially appropriate behavior must depend as much on our personality traits. In particular (the degree of empathy we feel for others) as it does on the contextual information that potentially affects our sense of worthiness and desert/equity in fairness distribution schemes.

Our study fills in the picture of the interaction between personality factors and information upon which moral judgments of fairness distribution are made since extant studies have not focused exclusively on the impact of empathetic, traits on the sense of egalitarianism. One study conducted by Hoffman and Spitzer (1985) demonstrated the following result: participants could be manipulated to become "greedy", i.e. make distributions based on a utilitarian sense of fairness distribution and thus deviate from egalitarian distribution. This was accomplished by making the first movers in the game, the "controllers," (whose role is identical to a dictator in Dictator Games) believe that they "deserve" their status as the "controller" as opposed to the belief that they landed on this position by mere luck. Whereas in the absence of such a belief of moral authority, participants tend to opt for much more egalitarian distributions. Passing judgment on whether others deserve our generosity depends as much on the perceived worthiness of the potential recipient as it does on our moral authority.

Fong (2007) investigated how empathetic dispositions interact with the attributes of the object of empathy, like the recipient of welfare transfers. Defined as the desire to help those who are deserving, empathetic responsiveness predicts charitable-giving in a Dictator Game-like setting if the recipient is worthy of their support. In a similar study, Klimecki et al. (2016) found a strong positive relationship between empathetic (self-reported) feelings, when artificially induced by the experimenter, and the size of the offer in Dictator Game settings regardless of the empathetic pre-disposition of the participants. This is not surprising because empathy may be modulated by various factors such as selective attention, emotional regulation, personal life histories, social distance, etc (Kirman and Teschl, 2010). However, as Klimecki and her colleagues discovered, those who possess emphatic traits responded more strongly to the priming. These participants must have been predisposed "to simulate the internal state" (Singer and Fehr, 2005) of others (e.g., feeling pain) more vividly than their less empathetic counterparts. Forsyth (2019) found that empathetic individuals and tend to be "idealistic" in their moral judgments.

Idealism, as he defines it, represents a strong disposition toward minimizing harm to others. Of the two idealist moral types, empathy, he found, proved to be a strong predictor of "absolutists" who believe that people should act in ways that are consistent with (universal) moral standards. We are also interested in identifying, if available, the differences in the way in which the impact of empathy on moral judgments is mediated by gender as some findings are pointing toward this direction. For instance, if the females are indeed more indiscriminately empathetic (Christov-Moore et al., 2104), some aspects of the scenarios may become more salient to the male participants. In sum, previous studies suggest that an individual's native sense of distributive fairness ranges between egalitarian and utilitarian schemes and that these responses can be modified by manipulating the information upon which decisions are made. We build on some of these earlier findings and expose our participants to a diverse set of hypothetical cases in which they are "provoked" to reveal or express moral judgments. Further, we focus on how personality traits on empathy and in-group favoritism play a role in moral judgments of distribution.

3. Methods and Hypotheses

The study we conducted investigates whether being empathetic affects fairness judgments in problems involving, distribution of resources by disinterested individuals. Using Amazon Mechanical Turk (AMT), a crowdsourcing marketplace, we recruited 303 participants of at least high school graduates and asked them to complete an online questionnaire which took participants an average of 26 minutes.[19] AMT allowed us to reach larger (considering our limited research budget) and more diverse participants (in terms of age and upbringing) than college students who are customary participants in such studies. The first part of the questionnaire consists of demographic questions (see Table 1). The Interpersonal Reactivity Index (Davis, 1983), a commonly used survey that measures the level of empathy along four equally weighted dimensions: *perspective-taking, fantasy, empathic concern, and personal distress*.

These constitute our independent variables. Since folk wisdom suggests that moral values are shaped partly by an individual's upbringing, we included the following as part of our demographic questionnaire: Did you have a religious upbringing/childhood? Did either of your parents graduate from college? What kind of area in which you were raised? The second part of the questionnaire asks participants to respond to three scenarios that motivate the exercise of moral judgment and thus serve as our dependent variables (see questionnaire with three scenarios in the appendix). Three scenarios were presented; the first one concerns the fairness of distribution of rewards relative to the distribution of work put in on building a bike. The second scenario is an ultimatum game with the addition of a question about the participant's emotional reaction to the results of the game. The third scenario is about a moral judgment concerning raising prices at a hardware store. (The details of the scenarios are provided in the Appendix.) Participants' empathy scores on the Interpersonal Reactivity Index (IRI) were used to predict their responses to three hypothetical scenarios in a set of logistic regressions and decision trees to test the following hypotheses.

- **(Distributive Justice) Hypothesis #1:** Higher empathy scores will be associated with more egalitarian distribution preferences when some favorable information is provided about the circumstances of the disadvantaged individual in our hypothetical productive exchange setting.
- **(Social Norm Violations) Hypothesis #2:** Higher empathy scores will increase the likelihood of rejecting the $2 offer in the ultimatum game.
- **(Moral Limits to Profit) Hypothesis #3:** Higher empathy score will increase the likelihood of finding the decision to raise prices by the hardware store owner unfair.
- **(Consistency of Moral Judgments) Hypothesis #4:** Empathetic individuals make consistent choices across all the cases. We define "moral consistency" as maintaining a notion of fairness across scenarios; in this case, it refers specifically to (i) opting for egalitarian distribution, and (ii) calling a $2 offer and the decision to raise prices unfairly.

Table 1: Demographic Variables

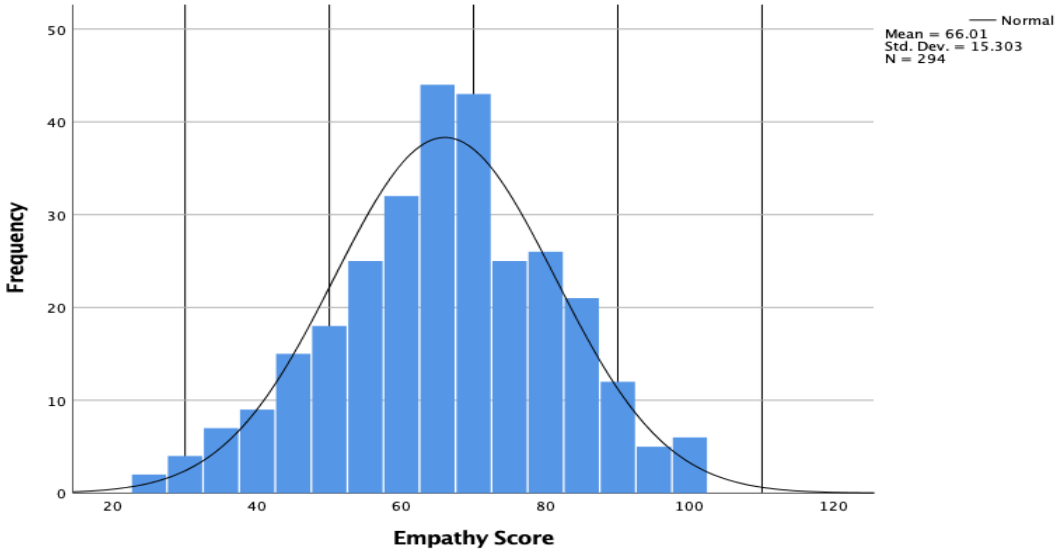
Race (n=294) ● White (86.4%) ● Non-white (13.6%) Gender (n=294) ● Male (44.4%) ● Female (55.6%)	Spiritual (n=294) ● Yes (49.2%) ● No (50.8%) Family income (n=294) ● Below average (34.6%) ● Average (53.6%) ● Above average (11.9%)

[19] Eight of the surveys have not been included in the final analysis. Two of them have been discarded since they did not pass the "attention" test where the responses were completely counterintuitive. The rest of the responses have been eliminated in the process of removing outliers. We excluded the empathy scores that lied on the both extremes, too low or too high, relying on the Stem-and-Leaf Plot tool provided by the SPSS statistical software package.

Age (n=294) ● 25 or younger (2.4%) ● Between 26 and 40 (57.6%) ● 41 and older (40%) Religious upbringing (n=294) ● Certainly (32.5%) ● Somewhat (34.9%) ● Not really (32.5%)	Number of parents who graduated from college (n=294) ● Both of them (21%) ● One of them (27.1%) ● Neither of them (51.9%) Location in which the person was raised (n=294) ● Rural (13.6%) ● Small Town (21%) ● Suburban (43.1%) ● Urban (22.4%)

The participants' sense of distributive justice in the real world. Moreover, there are drawbacks to capturing individual traits with aggregated scores because this approach could miss the nuances that may be present in the responses to 28 individual questions (7 under each category) in the IRI. For instance, when we reviewed the empathic concern scale, we noticed there are noticeable inconsistencies: of the seven different questions meant to measure the level of empathic concern, some participants scored very high on some questions while scoring very low on others. The latter strikes us as a more actionable feeling. Another possibility for the apparent inconsistency is that individual questions may measure aspects of personality traits more strongly than they do empathy.

Figure 1: Frequency Distribution - Empathy Scores (min 25; max=100)



4. Results

Result 1: In the first scenario, we investigated how empathy informs the sense of "fair" distribution. We asked participants to distribute collectively produced output between those who contributed to building the bicycle in the scenario. These choices could be described as "egalitarian (4)," "mild egalitarian (3)," "mild utilitarian (2)," and "utilitarian (1)," based on the gap between the respective shares ranging from 0 hours in the one extreme (egalitarian) to 84 hours in the other (utilitarian). Initially, we provided *no contextual information* about the individuals except that one party had contributed less than the other. The participants

made their selections. Then some contextual information is provided about the *causes* of the productivity differences. We used the Decision Tree with CHAID growing method[20] to trace if empathy and our demographic variables could explain the change of heart among our participants. We expected that empathetic individuals are likely to opt for a more equal distribution of bike hours once they are told that the productivity differences had been caused by factors beyond the individual's control. Indeed, the additional context changed the odds of opting for Egalitarian distribution, which increased in participants with higher empathy scores. However, the relationship between empathy and the shift toward the egalitarian distribution is true mostly for those who were Mild Egalitarians, to begin with, and not across the board (See Figure 2).

Of those who were mildly egalitarians, those with empathy scores exceeding 53 (which corresponds to the cutoff for the bottom quintile) are more likely to switch their position in favor of a fully egalitarian distribution once some context is provided. It also appears that those who are 41 and older are slightly overrepresented. This result supports our Hypothesis #1 that empathy increases sensitivity to other people's personal circumstances (in particular, those that are beyond their control) in exercising distributive fairness judgments when the joint effort is involved. This is a unique form of cooperation that Lawler and his colleagues called "productive exchange" (2008). It differs from other forms of exchange like reciprocal trades in its ability to foster a greater degree of group solidarity. That said, the empathy threshold, 53, may have been caused by the fact that our participants appeared already to be fairly egalitarian-minded, a hunch which Table 2 does appear to corroborate to some extent: Around 52 percent of the participants chose Mildly Egalitarian distribution (3) before any context was provided. Interestingly, this group's empathy scores are evenly distributed across each quintile with no apparent indication of empathy being the most significant driver of their decision in the first place. A word of caution is in order here: the fact that invoking empathy in the absence of emotional priming is rather challenging (considering its strong affective dimension) may have caused some underestimation in our study of the actual role that empathy plays in informing.

For instance, "Sometimes I feel very sorry for other people when they are having problems" expresses a more passive/reactive emotional reaction than "When I see someone being taken advantage of, I feel kind of protective towards them." What to make of this? One possibility is that the framing of questions is a determinative factor. This suspicion led us to conduct a two-stage cluster analysis. The results are noteworthy: based on responses to the questions meant to measure Empathic Concern, participants can be grouped into three clusters (See Figure 3). While it is beyond the scope of this paper to speculate if any of the clusters correspond to identifiable personality traits, we should note that compared to the other two Cluster 3 reliably predicted egalitarian choices in the allocation of bike hours with or without the context information. Moreover, its predictive ability is enhanced when it interacts with the "male" gender; i.e. being male and a member of Cluster 3 successfully predicted egalitarian choices.

Table 2: Frequency Distribution – Allocated Bike Hours With and Without Context

	Without Context		**With Context**	
Types	**frequency**	**percent**	**frequency**	**percent**
egalitarian	77 (12)	26.2 (15.5*)	177	60.2
mild-egalitarian	152 (32)	51.7 (21*)	95	32.3
mild-utilitarian	52 (11)	17.7 (21.1*)	18	6.1
utilitarian	13 (0)	4.4 (0*)	4	1.4

***** Percentage of those who fall in the top quintile (20%) of the empathy score distribution

[20] Minimum cases in Parent (the Child) Node is 50 (20).

Figure 2: Empathy and Egalitarian Attitudes

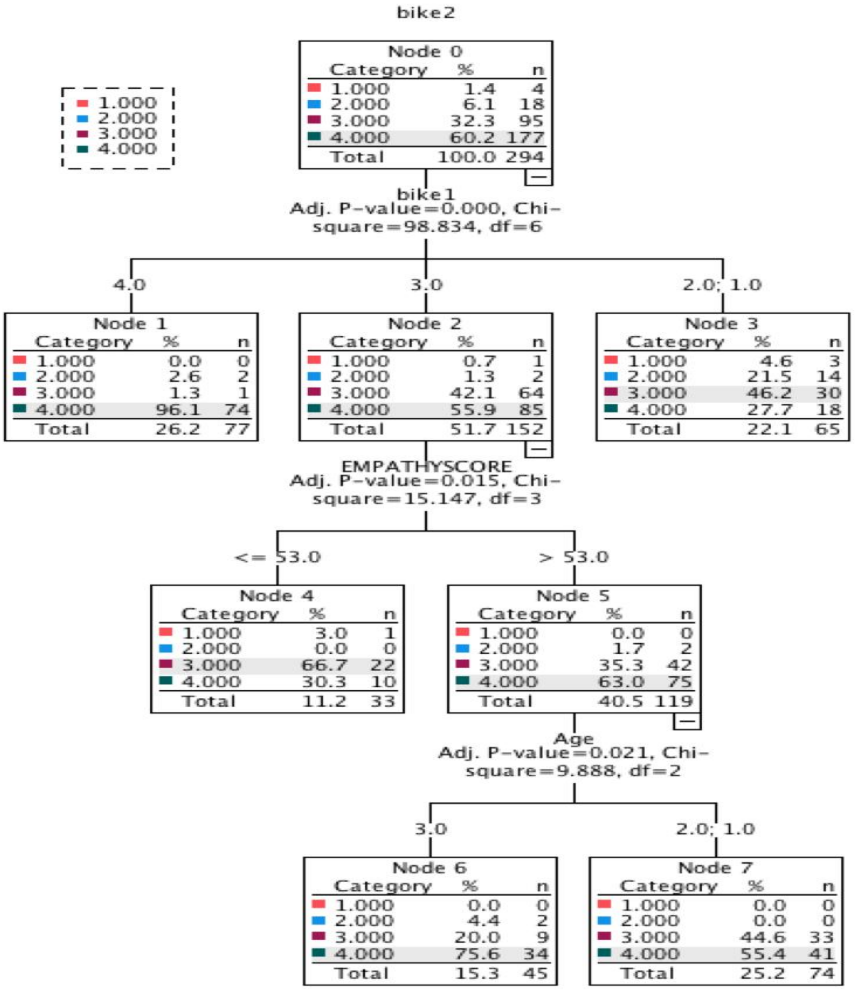


Figure 3: Clusters Based on Empathic Concern Scales Questions[21]

Clusters

Input (Predictor) Importance
1.0 0.8 0.6 0.4 0.2 0.0

Cluster	2	3	1
Label			
Description			
Size	44.0% (129)	34.5% (101)	21.5% (63)
Inputs	ec1 2.50	ec1 3.71	ec1 1.16
	ec2 3.25	ec2 3.92	ec2 1.81
	ec4 3.26	ec4 3.86	ec4 1.76
	ec3 1.30	ec3 0.26	ec3 2.41
	ec6 1.65	ec6 0.48	ec6 2.78
	ec7 1.60	ec7 0.43	ec7 2.81
	ec5 3.57	ec5 3.99	ec5 2.29

Result 2: In the second scenario, we investigated whether empathetic individuals are more sensitive to norm violations. We asked whether the participants would accept an offer of $2 out of $100 in a hypothetical ultimatum game—an offer that is clearly too low. The flat fee of $5 was offered to the subjects for their participation in the overall study and the fee was not conditioned on their responses. So, we recognize that this feature of our design is a major departure from the typical ultimatum games in the sense that the participants were not forced to "practice what they preach" and have no financial stakes in their decision to reject. However, this deliberate dissociation is consistent with our expectation that participants act like

[21] (EC1: I often have tender, concerned feelings for people less fortunate than me. EC2: Sometimes I don't feel very sorry for other people when they are having problems. EC3: When I see someone being taken advantage of, I feel kind of protective towards them. EC4: Other people's misfortunes do not usually disturb me a great deal. EC5: When I see someone being treated unfairly, I sometimes don't feel very much pity for them. EC6: I am often quite touched by things that I see happen. EC7: I would describe myself as a pretty soft-hearted person.)

"impartial spectators" even though this setup may have increased the rejection rate. Which we cannot independently verify considering there is no meaningful benchmark study for comparison except the receivers tend to reject offers less than 20% of the sum 50% of the time and the rejection rate increases as the share get ever smaller (Houser and McCabe, 2014). That said, we are much more interested in how participants *describe* such an offer and what their *emotional reactions* would be than whether they would reject the offer. After making their decision to accept or reject.

We asked the participants to select from a list of attributes describing the personality of the person making the lopsided. We also asked the participant how the offer would make them feel by selecting from a list of descriptive emotions. In the binomial logistic regression that we conducted, there is no statistically significant relationship between the individual's empathy score and the likelihood of rejecting the $2 offered (p=0.363). The results do not support our hypothesis (#2) that a higher empathy score is correlated with a higher likelihood of rejecting a very low offer. However, as Decety and Cowell (2014) argue, the relationship between morality and empathy is complex and nuanced; and the neural correlates of, say, empathetic concern differ in significant ways from those of, say, perspective-taking. Therefore, following their recommendations, we have decided to distinguish between the different facets of empathy as each facet might influence moral judgments somewhat uniquely. We disaggregated the empathy index into its constituent components. One of the four constituents of the aggregate empathy index on the IRI, the Perspective Taking (PT) score alone appears to have successfully predicted (p=0.008, Exp (B)=1.134) the odds of rejecting the offer at a 1% significance level. Only a one-point increase in one's PT score increases their odd of rejecting.

The offer by 1.134, a very substantive effect (See Table 3). This may be because the scenarios employed in the study are abstract and participants do not experience empathy with the vivacity of the more ecologically valid lived scenario. Our result supports our expectation that taking the perspective of the offender (to figure out their intention) as well as that of the offended would aid in detecting social norm violations and reinforcing reciprocal punishments. Moreover, we are curious whether the different facets of empathy interact with gender. Our expectation is that when modified by gender some of the subcategories of empathy categories can successfully predict the odds of rejecting the offer. We have rather intriguing results. Although Perspective Taking is explanatory regardless of the gender[22], Empathic Concern (p=0.043; Exp (B)=0.839) and Personal Distress (p=0.029; Exp(B)=0.880) scales are significant (and so in non-negligible ways) *only when* they interact with gender. Specifically, EC increases one's likelihood of rejecting the offer if they are male, while Personal Distress increases one's likelihood of rejecting the offer if they are female! (See Table 4) Christov-Moore et al. (2014) found such gender effects in empathy-related judgments to be common in the literature and suggest compelling explanations for such implicit differences. They discuss evolutionary considerations of caring instincts but also make room for developmental differences, neural differences, and socialized gender roles. In the second stage of the study, we allowed the participants to choose from a list of descriptions that characterize what they think of the person who offered $2.

We used the Decision Tree approach with the CHAID growing method as it is very conducive to revealing interaction among independent variables in a clearly interpretable visual fashion. Calling the proposer "Unwise" appeared to be the strongest predictor of rejecting the offer. For those who did not choose to select "Unwise," "Unfair" was the strongest predictor of rejection. Interestingly, the impact of "Unfair" has been modulated by the participants' PT score: among those who called the proper "Unfair," those with the PT score higher than 21 (out of possible 28) nearly unanimously (95.7%) rejected the offer. What about the emotional reactions? Of the emotional categories provided, "Angry" (87.6%) is the strongest predictor of whether someone penalized the proposer, by rejecting the meager offer even though it would have made them $2 better off than before. The role of anger in rejections of low offers in UGs was also corroborated by a smaller-scale study by Bosman, Sonnemans, and Zeelenberg (2001) who found that the intensity of "anger" that the participants said they felt to be negatively correlated with the level of offers. Moreover, they found that, for those who rejected, anger along with "irritation" and "contempt" were the most strongly felt emotions evoked

[22] Although not included in Table 4, when PT interacts with gender it comes out significant for both genders.

by the low offer. In our study, interestingly, anger also interacted with gender: among those who describe their primary emotion as "Angry" in reaction to the proposed amount, female participants rejected the offer to a greater degree (92.6% vs. 81.9% for males). Of those (female or male) who did not select "Angry," the emotion, "Frustrated," appeared to be a strong predictor of rejection (75.3%).

Table 3: Rejection of Low Offers and Perspective Taking

Dependent Variable: Rejection (=1)

	B	S.E.	Wald	df	Sig.	Exp(B)
Race (Non-white = 1)	.490	.428	1.307	1	.253	1.632
Gender (Female = 1)	.008	.321	.001	1	.981	1.008
Age (young)			3.736	2	.154	
Age(mid-aged)	-.332	.864	.148	1	.701	.717
Age(old)	.576	.331	3.039	1	.081	1.780
Religious (certainly)			1.200	2	.549	
Religious (somewhat)	-.425	.398	1.138	1	.286	.654
Religious (not really)	-.295	.372	.629	1	.428	.745
Spiritual (Yes =1)	-.003	.327	.000	1	.994	.997
Income (Below Avg)			1.294	2	.524	
Income (Average)	.553	.487	1.291	1	.256	1.739
Income (Above Avg)	.404	.457	.780	1	.377	1.498
Parents (No College)			.590	2	.745	
Parents (One College)	-.010	.376	.001	1	.980	.990
Parents (Both College)	.270	.433	.391	1	.532	1.310
Location (rural)			2.396	3	.494	
Location (s town)	.373	.546	.466	1	.495	1.452
Location (suburb)	-.329	.431	.582	1	.446	.720
Location(urban)	.168	.387	.187	1	.665	1.183
PERSONAL DISTRESS	.056	.044	1.644	1	.200	1.058
PESPECTIVE TAKING	**.126**	**.048**	**6.959**	**1**	**.008**	**1.134**
EMPATHIC CONCERN	.098	.071	1.916	1	.166	1.103
FANTASY	-.051	.049	1.078	1	.299	.950
Constant	-4.053	1.777	5.203	1	.023	.017

Table 4: Rejection of Low Offers when Different Empathy Categories Interact with Gender

Dependent variable: Rejection (=1)

	B	Std. Error	Wald	Df	Sig.	Exp(B)
Intercept	3.704	1.847	4.022	1	.045	
Age (young)	.134	.906	.022	1	.883	1.143
Age(mid-aged)	-.675	.339	3.962	1	.047	.509
Age(old)	0[b]	.	.	0	.	.
Religious (not really)	.489	.407	1.446	1	.229	1.631
Religious (somewhat)	.336	.376	.798	1	.372	1.400
Religious (certainly)	0[b]	.	.	0	.	.
[Spiritual = NO]	-.092	.334	.075	1	.784	.912
[Spiritual= YES]	0[b]	.	.	0	.	.
Income (Below Avg)	-.474	.496	.914	1	.339	.622
Income (Average)	-.348	.463	.567	1	.451	.706
Income (Above Avg)	0[b]	.	.	0	.	.
Location (rural)	-.460	.556	.684	1	.408	.632
Location (s town)	.307	.441	.486	1	.486	1.360
Location (suburb)	-.187	.392	.228	1	.633	.829
Location(urban)	0[b]	.	.	0	.	.
Parents (No College)	-.099	.385	.066	1	.797	.906
Parents (One College)	-.340	.437	.606	1	.436	.711
Parents (Both College)	0[b]	.	.	0	.	.
[Race = White]	-.547	.436	1.576	1	.209	.579
[Race = Non-white]	0[b]	.	.	0	.	.
PESPECTIVE TAKING	**-.127**	**.049**	**6.736**	**1**	**.009**	**.881**
[Gender =0] * EC	**-.175**	**.087**	**4.079**	**1**	**.043**	**.839**
[Gender =1] * EC	-.020	.093	.048	1	.827	.980
[Gender =0] * PD	.043	.068	.392	1	.531	1.044
[Gender =1] * PD	**-.128**	**.059**	**4.775**	**1**	**.029**	**.880**
[Gender =0] * FS	.038	.072	.272	1	.602	1.038
[Gender =1] * FS	.081	.061	1.743	1	.187	1.084

Figure 4: Describing the Person offering $2 and Likelihood of Rejecting

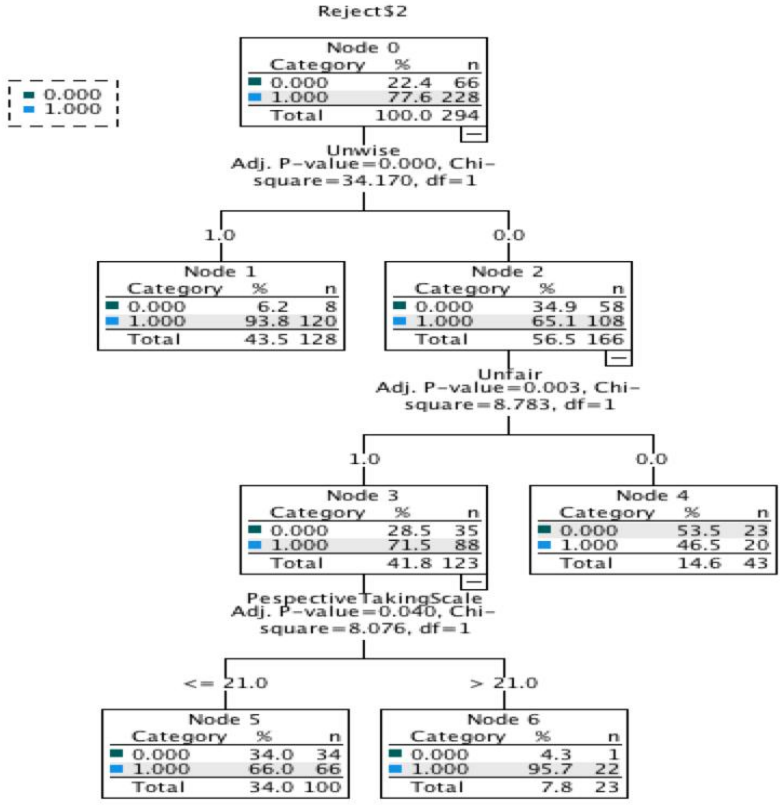
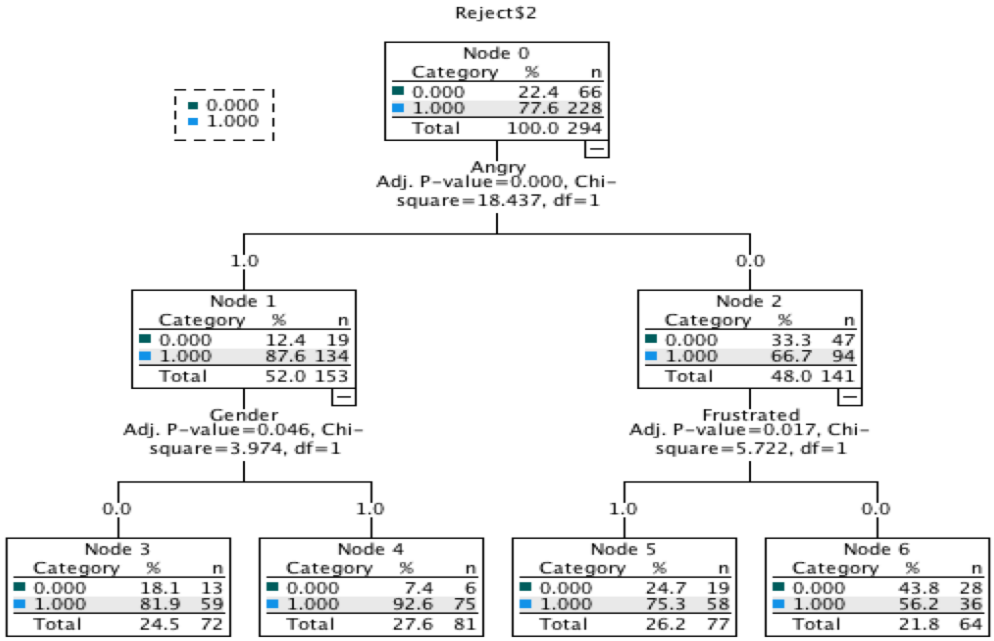


Figure 5: Emotional Reactions to the Low Offer and Likelihood of Rejection



Result 3: In Case 3, we investigated whether participants would consider opportunistic market behavior as unfair or just as another incidence of supply-demand interaction. Moral limits to markets like price gouging laws are commonplace, and we hoped this scenario would help capture our participants' attitude toward the acceptability of such constraints on profit-making. We expected that the higher the empathy level the lower the tolerance for opportunistic behavior of the type described here. In the multinomial logistic regression that we conducted, Empathy Score does appear to differentiate those who find this behavior "unfair" from those who consider it a fair game (p=0.024, Exp(B)=0.978). Therefore, a higher empathy score on the IRI increases the likelihood of finding the decision to raise prices unfairly[23]. Some demographic variables also appear significant for this question; for instance, male participants (p=0.002, Exp(B)=2.466) are nearly two and a half times more likely to find the decision to raise the prices fair (compared to unfair) compared to their female counterparts. Strikingly, male participants are more than four times more likely to opt for "does not apply" versus "unfair." Perhaps this reflects their inclination to see opportunism as all part of the game (p=0.003, Exp(B)=4.360). Having grown up in a rural area (compared to an urban area) along with those who described their upbringing environment as "Somewhat Religious" (compared to "Certainly Religious") do appear to increase the likelihood of finding the business practice of raising prices in the hardware store scenario unfair (p=0.008, Exp(B)=0.220).

Table 5: Hardware Store Case (The reference category: unfair)

Dependent Variable: Unfair to Raise Prices (0=no; 1=yes, 2=does not apply)[24]

0	B	Std. Error	Wald	df	Sig.	Exp(B)
Intercept	1.324	.964	1.887	1	.170	
EMPATHYSCORE	-.022	.010	5.096	1	.024	.978
[Gender = Male]	**.903**	**.297**	**9.221**	**1**	**.002**	**2.466**
[Gender = Female]	0[b]	.	.	0	.	.
[Religious = Not really]	-.288	.360	.642	1	.423	.750
[Religious = Somewhat]	**-.775**	**.342**	**5.128**	**1**	**.024**	**.461**
[Religious = Certainly]	0[b]	.	.	0	.	.
[Location = RURAL]	**-1.514**	**.573**	**6.991**	**1**	**.008**	**.220**
[Location = STOWN]	.393	.403	.947	1	.331	1.481
[Location = SUB]	.219	.359	.372	1	.542	1.245
[Location = URBAN]	0[b]	.	.	0	.	.
1	B	Std. Error	Wald	df	Sig.	Exp(B)
Intercept	-.981	1.486	.435	1	.509	
EMPATHYSCORE	-.005	.016	.101	1	.751	.995
[Gender = Male]	**1.473**	**.492**	**8.964**	**1**	**.003**	**4.360**

[23] However, empathy score on the IRI does not help predict the response of those who thought that the situation should not be evaluated within the fair vs. unfair dichotomy (i.e. those who selected "fairness does not apply")

[24] Of the control variables, only those that are significant at the 5% were reported.

[Gender = Female]	0[b]	.	.	0	.	.

Result 4: Finally, we were curious whether judgments exercised across these three relatively diverse cases have any common denominator. Accordingly, we created a variable to measure moral consistency in the local sense. By our definition, the variable of 'morally consistent' applies to individuals who (i) opted for the egalitarian distribution in our first case after the context is provided; (ii) found $2 offer unfair in the UG; and (iii) found the hardware store owner's decision to raise the price unfairly. Each aspect is equally weighted with the participant allotted one point if the condition is true. The highest possible score, the sign of moral consistency, is three (3) points. We should note that getting zero total points could be a form of consistency in the negative sense, i.e. to be morally consistent but non-egalitarian. However, the number of participants in this category is negligible (2.4%).[25] Based on the results of our multinomial regression we can make the following observation: female gender when coupled with higher empathy score is a reliable predictor of moral consistency in our study—a pattern that does not seem to carry over to those whose income is above average. Again, we refer the reader to Christov-Moore et al. (2014) for a multi-causal explanation of why empathy judgments consistently vary according to gender.

Table 6: Moral Consistency and its Determinants

Dependent Variable: Moral Consistency (3 = Morally Consistent is the Reference Category)[26]

		B	Std. Error	Wald	Df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
2.00	Intercept	1.222	1.018	1.443	1	.230			
	[Gender = MALE] *	-.008	.011	.445	1	.505	.992	.971	1.015
	[Gender = FMALE] *	**-.025**	**.010**	**6.538**	**1**	**.011**	**.975**	**.956**	**.994**
1.00	Intercept	2.928	1.214	5.813	1	.016			
	[Income=1]	-1.484	.610	5.917	1	.015	.227	.069	.750
	[Income=2]	-1.168	.558	4.389	1	.036	.311	.104	.927
	[Income=3]	0[b]	.	.	0	.	.	.	.
	[Gender = MALE] *	-.028	.014	3.874	1	.049	.972	.946	1.000
	[Gender = FMALE] * EMPATHYSCORE	**-.044**	**.013**	**12.145**	**1**	**.000**	**.957**	**.934**	**.981**

Discussion

Our results lend themselves to several interpretations. First, empathy tends to favor egalitarian distribution if the context provided suggests that circumstances in the scenario are depicted as being beyond the individual's control. Secondly, empathetic individuals display context-sensitivity in exercising fairness judgments. When empathy is measured as perspective-taking, our results indicate empathetic individuals are more sensitive to deviations from social norms and more likely to exercise negative reciprocity in the face of unfair (or anti-social) treatment as evidenced by their reactions in the ultimatum game. This may be because

[25] We merged this category, 0, with 1 in testing moral consistency.

[26] Of the control variables, only those that are significant at the 5% were reported

they are taking an explicit ethical position to an abstract scenario and then simulating the consequences, this may also be a form of virtue signaling. Why does one's ability to take the perspective of another increase their chances of rejecting the lopsided offer in the ultimatum game? Of the terms that participants chose to describe the lopsided split in the ultimatum game, three of them successfully predicted (at .01 significance) whether or not they would reject the offer: "acceptable," (p=0.005, Exp(B)=0.031) "rational," (p=0.001, Exp(B)=0.062) and "unwise" (p=0.001, Exp(B)=7.863)[27]. Needless to say, the first two are negatively, and the last one positively, correlated with the probability of rejecting the offer. Perspective-taking, an integral element of empathy, may have contributed to the description of the rejected offer as "unwise". This prediction is reinforced by the fact that "unwise" is strongly correlated with the emotions of frustration, sadness, and surprise. A set of emotional words the participants chose to describe the low offer.

Calling the low offer "unfair" was another strong predictor of rejection whose predictive power is even higher for those whose perspective-taking skills are fairly elevated. The low offer seems to be processed as a violation of social norms. Those who rejected the offer in the ultimatum game may have thought: "I would not have done that!" We asked participants to tell us how they *feel* about being offered so little in the ultimatum game scenario. Those who selected the following emotions were more likely to take the offer: "neutral" (p=0.001, Exp (B)=0.089), "pleased" (p=0.011, Exp(B)=0.048) and, interestingly, "jealous" (p=0.019, Exp(B)=0.242)[28]. Anger (p=0.035, Exp (B)=2.108), on the other hand, appears to have strongly motivated the participants to reject the offer: the odds of rejecting the offer are 2.1 times higher among those who characterized their feelings in this way. It is reasonable to expect that the low offer must have fed the sense of getting unfairly treated based on the high correlation between "anger" and "unfair" (p=0.001, co-efficient=0.389). This is not surprising as anger as an emotion is a potent tool to display discontent and the motivational force behind our push for fairness[29]. Apart from being inequity-averse by rejecting the low offer and, effectively, penalizing the behavior, empathetic individuals are inclined to see overtly self-interested behavior (e.g. raising prices) as unfair. That said, empathy is not the only variable that makes one sensitive. To the opportunistic behavior of pursuing one's interests at the expense of others. Gender and the population density of the location of upbringing also exert influence on one's moral judgments. The latter may be explained by the fact that in sparsely populated regions, social relationships tend to be much less impersonal. Our finding is in line with the findings of Sautter, Littvay, & Bearnes (2007) that empathy is more likely to result in cooperative behavior among those who were raised in sparsely populated localities due to the diminished, anonymity they experienced growing up. Lastly, we observed a formed consistency in exercising moral judgments among some of our participants. They proved to be Rawlsian compassionates in accommodating the least fortunate in the joint-production scenario, reciprocal in penalizing the norm-deviant in the UG, and, probably, followers of rule-based ethics in our last case (e.g., you shall not take advantage of the circumstances). Such consistency is much more pronounced among female participants with high empathy levels particularly when these two variables interact with income.

5. Conclusion

In this paper, we investigated the relation between empathetic personality characteristics measured by the IRI and decisions on a set of distribution scenarios. We found that an individual's Perspective Taking (PT)

[27] The terms "acceptable," and "rational" did not appear in our decision tree analysis because of its design. However, a logistic regression we ran separately showed that these two descriptions are positively correlated with accepting the offer.

[28] These terms did not appear in our decision tree analysis because of its design. However, a logistic regression we ran separately showed that these three descriptions are positively correlated with accepting the offer.

[29] The sense of fairness may have evolved as a way of dealing with the free-loader problem, some argue there is a cheater detection module (Cosmides and Tooby, 1992) while others emphasize the role of cooperation in building domestic and social organizations (Sterelny, 2016; Asma & Gabriel, 2019).

score on the IRI was a sole predictor of rejecting unfair offers in an ultimatum game. The active element of empathy in this scenario was cognitive empathy which allowed participants to imagine the perspective of the other as a frame to a fairness judgment on a distribution game. The effect of the location of upbringing is evident in our last case and may be explained by the fact that individuals from rural backgrounds have a different sense of responsibility towards in-group members, which then affects how they frame judgments of fairness. Providing context is crucial to adding an affective dimension to an abstract scenario as in case 1. Empathy and context-sensitivity are related psychological elements with the cognitive appraisal of fairness. To get back to our initial question of whether fairness is determined by reason or emotion, we suggest it may be that the effective tug of empathy modulates attentional processes to modify how reason weights contextual factors. We were not able to tease out more implicit forms of empathy, viz. emotional contagion, with these tests, in the future we will consider using a face-to-face dyadic paradigm so that the tasks take on a more embodied interactive tone which we predict will enable closer manipulation of implicit forms of empathy.

References

Asma, S. T. & Gabriel, R. (2019). The Emotional Mind: The Affective Roots of Culture and Cognition. Cambridge, MA: Harvard University Press.

Bosman, R., Sonnemans, J. & Zeelenberg, M. (2001). Emotions, rejections, and cooling off in the Ultimatum Game. Working paper, University of Amsterdam.

Cappelen, A. W., Astri D. H., Erik, Ø. S. & Bertil, T. (2007). The Pluralism of Fairness Ideals: An Experimental Approach. *American Economic Review,* 97(3), 818–27.

Christov-Moore, L., Simpson, E. A., Coudé, G., Grigaityte, K., Iacoboni, M. & Ferrari, P. F. (2014). Empathy: Gender effects in brain and behavior. *Neuroscience and biobehavioral reviews*, 46(4), 604–627.

Côté, S., Piff, P. K. & Willer, R. (2013). For whom do the ends justify the means? Social class and utilitarian moral judgment. *Journal of personality and social psychology*, 104(3), 490.

Decety, J. & Michalska, K. J. (2010). Neurodevelopmental changes in the circuits underlying empathy and sympathy from childhood to adulthood. *Developmental Science*, 13(6), 886–899.

Damasio, A. R. (1994). Descartes' error: emotion, reason, and the human brain. New York: G.P. Putnam,

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44, 113-126.

Decety, J. (2007). A social cognitive neuroscience model of human empathy, in Fundamentals of Social Neuroscience, ed. E. Harmon-Jones and P. Winkielman. New York: Guilford Press.

Decety, J. & Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews*, 3, 71-100.

Decety, J. & Cowell, J. M. (2014). The complex relation between morality and empathy. *Trends in Cognitive Sciences*, 18(7), 337-339.

De Oliveira-Souza, R., Zhan, R. & Moll, J. (2014). Neural Correlates of human morality: An overview. In Decety, J. & Wheatley, T. (Eds.), The Moral brain, a multidisciplinary perspective (pp. 183-195). Cambridge, MA.: MIT Press.

de Waal, F. B. M. (1996). Good Natured: The Origins of Right and Wrong in Humans and Other Animals (Cambridge, MA: Harvard University Press).

de Waal, F. B. M. (2007). The 'Russian doll' model of empathy and imitation. In S. Bråten (Ed.), Advances in consciousness research: Vol. 68. On being moved: From mirror neurons to empathy (pp. 49-69). Amsterdam, Netherlands: John Benjamins Publishing Company.

Dunbar, R. I. M. (1998). The social brain hypothesis. *Evol Anthropol*, 6, 178-190.

Efferson, C., Lalive, R. & Fehr, E. (2008). The Coevolution of Cultural Groups and Ingroup Favoritism. *Science*, 321(5897), 1844–1849.

Faravelli, M. (2007). How Context Matters: A Survey Based Experiment on Distributive Justice. *Journal of Public Economics,* 91(7–8), 1399–1422.

Fong, C. M. (2007). Evidence from an Experiment on Charity to Welfare Recipients: Reciprocity, Altruism and the Empathic Responsiveness Hypothesis. *The Economic Journal,* 117(522), 1008–24.

Forsyth, D. (2019). Making Moral Judgments: Psychological Perspectives on Morality, Ethics, and Decision-making. Routledge.

Frohlich, N., Joe A., Oppenheimer C, L. & Eavey. (1987). Choices of Principles of Distributive Justice in Experimental Groups. *American Journal of Political Science,* 31(3), 606.

Heberlein, A. S. & Saxe, R. R. (2005). Dissociation between emotion and personality judgments: Convergent evidence from functional neuroimaging. *Neuroimage*, 28, 770–777.

Hoffman, E., Matthew, L. & Spitzer. (1985). Entitlements, Rights, and Fairness: An Experimental Examination of Subjects' Concepts of Distributive Justice. *The Journal of Legal Studies,* 14(2), 259–97.

Houser, D. & McCabe, K. (2014). Experimental economics and experimental game theory. In *Neuroeconomics* (pp. 19-34). Academic Press.

Kandel, E. & Edward, P. (1992). Peer Pressure and Partnerships. *Journal of Political Economy,* 100(August 1992), 801–17.

Kirman, A. & Miriam, T. (2010). Selfish or Selfless? The Role of Empathy in Economics. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 365(1538), 303–17.

Klimecki, O. M., Sarah, V., Mayer, A. J., Jonathan, S. & Michael, S. (2016). Empathy Promotes Altruistic Behavior in Economic Interactions. *Scientific Reports,* 6(1), 1-5.

Lawler, E. J., Shane, R. & Thye, J. Y. (2008). Social Exchange and Micro Social Order. *American Sociological Review*, 73(4), 519–42.

Page, K. (2002). Empathy Leads to Fairness. *Bulletin of Mathematical Biology,* 64(6), 1101–16.

Preston, S. D. & de Waal, F. B. M. (2002). Empathy: Its ultimate and proximate bases. *Behavioral & Brain Sciences*, 25, 1–72.

Sautter, J. A., Levente, L. & Brennen, B. (2007). A Dual-Edged Sword: Empathy and Collective Action in the Prisoner's Dilemma. *The ANNALS of the American Academy of Political and Social Science,* 614(1), 154–71.

Singer, T. & Ernst, F. (2005). The Neuroeconomics of Mind Reading and Empathy. *American Economic Review,* 95(2), 340–45.

Stark, O. & Ita, F. (2000). Transfers, Empathy Formation, and Reverse Transfers. In *The Economics of Reciprocity, Giving and Altruism*, edited by L. A. Gérard-Varet, S. C. Kolm, and J. Mercier Ythier, 174–81. London: Palgrave Macmillan UK.

Sterelny, K. (2016). Cooperation, Culture, and Conflict, *British Journal for the Philosophy of Science*, 67(1), 31-58.

Strayer, J. (1987). Affective and cognitive processes in empathy. In N. Eisenberg & J. Strayer (Eds.), Empathy and its development (pp. 218– 244). New York, NY: Cambridge University Press.

Singer, P. (ed.) (1977). Animal Liberation. NY: Avon Books.

Cosmides, L. & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), The adapted mind: Evolutionary psychology and the generation of culture (p. 163–228). Oxford University Press.

Thompson, E. (2005). Empathy and human experience, in Science, Religion, and the Human Experience, ed. J.D. Proctor (New York: Oxford University Press).

Zahavi, D. (2014). Empathy and Other-Directed Intentionality, *Topoi,* 33, 129–142.

Appendix: **Table 5: Hypothetical Scenarios**

Case 1	**Case 2**	**Case 3**
Suppose John and Jane are building a bike together. Assume that bike parts are given to them for free. Once finished, they will have to decide how many hours per week each should keep the bike. Only other piece of information other information given is that Jane has put in more effort into the building of the bike than John has? (One week = 168 hours)	Imagine you participate in an ultimatum game. In the ultimatum game, there are two players who interact only once. There is no bargaining. The first player, the proposer, received a $100 which had not been earned by the proposer. The proposer now offers a split of this sum between himself/herself and the second player (e.g. 60-40). The second player, the receiver, is	Is it fair for a hardware store to raise the price of snow shovels in anticipation of a spring snowstorm when supplies of shovels are low? a. Yes b. No c. Fairness does not apply in this scenario

Which of the following distribution would be the fairest?

- 84 hours each
- Jane receives 126 hours and Jake receives 58 hours
- Jane receives 100 hours and Jake receives 84 hours
- Jane receives 92 hours and Jake receives 76 hours
- Jane receives 68 hours and Jake receives 100 hours

You were told that John's lower effort was caused by the familial obligations to which he had to attend. In the light of this new piece of information, which of the following distribution would be the fairest?

- 84 hours each
- Jane receives 126 hours and Jake receives 58 hours
- Jane receives 100 hours and Jake receives 84 hours
- Jane receives 92 hours and Jake receives 76 hours
- Jane receives 76 hours and Jake receives 92 hours

communicated the information of this proposal and has to decide to reject or accept this division offered by the proposer. You play the role of the receiver in this game.

If you reject the offer, neither party gets anything. If you accept the offer, the first player gets his/her demand and you get the amount you were offered. You have been offered $1. (This means the proposer demands $99.)

Would you take it?

- Yes
- No

How would you describe the offer made by the first player? You can choose multiple answers.

- Fair
- Acceptable
- Rational
- Mutually beneficial
- Opportunistic
- Unfair
- Unwise

How would you describe your feelings if you were to receive this offer? You can choose multiple answers.

- Neutral
- Pleased
- Angry
- Jealous
- Frustrated
- Sad
- Surprised
- None of the above

Journal of Economics and Behavioral Studies (JEBS)
Vol. 13, No. 1, February 2021 (ISSN 2220-6140)

How to Analyze Communication Data from Laboratory Experiments Without Being a Machine Learning Specialist

Benjamin Wegener
Ostwestfalen-Lippe University of Applied Sciences and Arts, Campusallee 12, Lemgo, Germany
benjamin.wegener@protonmail.com

Abstract: Recently, the analysis of communication has gained attention in experimental research. One important question is whether certain types of communication affect decisions differently than others. In this regard, Houser & Xiao (2011) present an approach for the classification of natural language messages. The primary limitation of their approach is its limited applicability to large message datasets. Therefore, Penczynski (2019) extends the methodological instruments by applying a machine learning classifier to experimental communication data. This is accompanied by the problem of a dearth of machine learning knowledge among experimenters. Hence, this paper presents an approach that employs a publicly available machine learning text analysis application. This makes it possible to analyze larger datasets based on small training datasets classified beforehand by human evaluators. As a first step, I use primary communication data reported by Charness and Dufwenberg (2006) to generate both training and test datasets. Following this approach, I am able to substantially replicate the original classification results obtained by Charness and Dufwenberg. The second step again involves messages from Charness and Dufwenberg as training data, while I take messages from a related trust game published by Deck et al. (2013) as a test, dataset. Promisingly, I am also able to replicate the classification results obtained by the external evaluators, as reported by Deck et al. The findings suggest that machine learning can be used to analyze large message datasets, both if the artificial intelligence is trained with data from the very same experiment and if it is trained with message data from a comparable experiment.

Keywords: *Laboratory Experiments; Communication; Classification of communication; Machine learning.*

1. Introduction

Experimental literature from economics and the social sciences in general provides rich data on the importance of natural language communication for decision-making in economic environments (see e.g. Isaac and Walker (1988)). In addition to demonstrating the importance of communication per se, the literature provides evidence that, in laboratory experiments, free-form communication results in better outcomes for the players as compared to restricted or pre-specified communication (for comparisons of free-form communication to pre-specified communication, see e.g. Lundquist et al. (2009); for an overview of related literature, see e.g. Agranov and Yariv (2018)). A major challenge associated with the use of free-form communication is that not only do the effects of the very existence of communication need to be analyzed, but also the content and intent of communication itself.[30] However, until now, very few methods have been available to gain deeper insight into free-form communication. Common approaches are, among others, the extraction of relevant keywords and the number of messages sent (see e.g. Huerta (2008); Moellers et al., (2017)). Besides analyzing keywords and message counts, economists have recently started to investigate whether certain types of communication (e.g. promises) might affect decisions differently than other types of communication (e.g. empty talk). Such investigations require the classification of natural language communication.

I am grateful for the helpful comments provided by participants of the 2018 annual meeting of the German Association for Experimental Economic Research (GfeW e.V.) and two anonymous referees. I would also like to thank Eva Tebbe, Korbinian von Blanckenburg, Christian Faupel, Philipp Russinger and Christoph Richert for their helpful comments. [30]The intent of communication refers to a category in which the communication can be classified, e.g. 'promise' or 'empty talk'.

In their well-cited paper, Charness and Dufwenberg (2006) present one of the first approaches of this kind. They classify messages from their experiment[31] into the two categories 'empty talk' and 'promise'. In response, Houser and Xiao (2011) criticize this approach due to its subjective nature and describe an objective procedure for the classification of natural language messages. Related to the ESP game[32] (Ahn & Dabbish, 2004), which was used to label images and enhance the accessibility of their contents, H&X reports a coordination game that is used to 'label' – that is, to classify – natural language messages. External evaluators read a corpus of messages and decide for each individual message whether it is to be classified as 'promise' or as 'empty talk'.[33] The participants of the game are incentivized financially so that they receive money if their evaluation matches the most common evaluation of the other participants.[34] Because the classification game of H&X is objective and easy to replicate, their game has been applied in many experiments (see e.g. Fischer and Normann (2019) and Huang and Xiao (2018)). Furthermore, H&X claims that its approach can clarify and extend the nature of conclusions reached, at a very small cost.

I argue that this is correct for a small dataset of messages, as one might see, for example, in the case of one-shot games with a small subject pool. On the other hand, their approach is no longer easy to replicate nor efficient for larger datasets. This might be why many experimenters analyze their data with fewer evaluators than H&X (see e.g. Ismayilov and Potters (2016) and Moellers et al., (2017)).[35] Even though this may save resources, the results potentially lack validity and robustness. I argue that human classification applied to a share of the messages from a laboratory experiment can be used as a training set for a machine learning approach. This is in line with the original intention of the ESP-Game – building a training dataset to label unlabeled pictures using a machine learning approach. In this way, larger datasets can be analyzed with only slightly more effort than small datasets. The most comparable publication to this study is Penczynski (2019). He presents a supervised machine learning approach to classifying messages according to their level in the level-k model of strategic reasoning. Research assistants classify messages from different laboratory experiments independently. Afterward, the classifications are reconciled and the research assistants have to agree upon one consistent classification.
The classified datasets are used to employ a Random Forest Classifier while common steps of natural language processing are utilized (all steps are implemented in R). Although Penczynski (2019) obtains promising results and argues that the implementation of a Random Forest Classifier can be accomplished without much effort, the task of classifying natural language messages is anything but trivial. Besides

[31] C&D implement a hidden-action trust game. Two participants are paired, one of whom is player A and the other player B. A has to choose 'in' or 'out'. Without knowing A's decision, B has to choose either 'roll (a die)' or 'don't roll (a die) '. If A has chosen 'out', A and B receive $5 each. If A has chosen 'in' and B has chosen 'don't roll', A receives $10 and B receives $14. If A has chosen 'in' and B has chosen 'roll', B receives $10 and rolls a six-sided die in order to determine the payoff of A. If the die yields a 1, A receives $0. If the die yields a number in the range from 2 through 6, A receives $12. The trust game played by C&D involves several treatments. Besides the explained approach, the authors vary in terms of whether preplay communication is allowed or not. If it is allowed, B can send a message to A or A can send a message to B. Additionally, the authors also change the payoff vector if A has chosen 'out' from (5, 5) to (7, 7).

[32] In the ESP game, two randomly paired participants label images without any pre-specified labels. Both are shown the same image. The participants do not know each other and are not allowed to communicate. The goal of the ESP game is to guess which image label the other participant has chosen. If the participants type in the same string while the image is on the screen, they move on to the next image. For every agreement, the participants get a certain number of points (Ahn & Dabbish, 2004).

[33] We do not discuss certain possible disadvantages of this approach. H&X point out that there are several aspects suitable for ongoing research, such as 'lay' evaluators. Furthermore, there is a discourse regarding the use of experts rather than non-experts, especially for deeper natural language tasks. Snow et al. (2008) analyze the use of non-experts recruited via Amazon Mechanical Turk as compared to experts. They find evidence that an average of four non-expert evaluators are required to emulate expert-level label quality.

[34] This is in contrast to the ESP game, which relies on the effect of entertainment and gamification as its incentives.

[35] H&X do not suggest a specific number of evaluators, but they argue that 'The average opinion of a large number of evaluators, (…), is a reasonable way to infer (…) the way the message was likely interpreted.'

determining a proper classification algorithm,[36] the preprocessing steps applied to the communication data require knowledge of stemming, lemmatization, part-of-speech tagging, feature engineering, and a multitude of other possible tasks. Therefore, I argue that an 'out-of-the-box approach can better fulfill the requirements of experimenters while reducing the effort and presupposed domain knowledge necessary to classify communication data. This study suggests a procedure utilizing such an 'out-of-the-box approach. This procedure involves the construction of a machine learning model based on IBM Watson Conversation[37]. The application is trained with a training dataset that encompasses human codings.

Based on these training messages, the machine learning classification algorithm learns the logical associations necessary to classify further messages. Consequently, it is able to classify messages in a test dataset. This study uses messages from C&D as well as Deck et al. (2013) (henceforth Deck et al.,) to test this approach. Using existing data allows me to compare the results generated by the machine learning classification algorithm to the original human coding results reported by C&D as well as Deck et al. To ensure the robustness of the approach, an estimation of the relative number of training set messages is needed. Therefore, software-generated results are cross-validated for various training set sizes with the original human classifications. In general, the results show that the approach is effective, both when training messages are taken from the very same experimental dataset as test messages, and when training messages are taken from an experimental dataset separate from that of the text messages. As a recommendation for use, this study provides rules of thumb regarding both the number of training messages necessary to generate good results and the distribution of messages per category.

2. Background

The approach presented in this paper focuses on adapting machine learning technology for unstructured data, i.e. text documents such as, in this case, chat messages. Although there will be no detailed discussion of the exhaustive literature on this topic, the essential approaches will be explained. Machine learning is the science of programming computers so that they can learn from data. Depending on whether the learning requires any type of supervision, machine learning systems can be classified as supervised or unsupervised.[38] In keeping with machine learning terminology, the classification of messages is to be assigned to the category of supervised machine learning, because the classification is supervised by the knowledge and intuition of human evaluators (Sebastiani (2002). The machine learning algorithm is capable of classifying natural language texts with a classification from a predefined set of classifications. This requires the availability of a corpus of manually pre-classified messages. These classifications must be provided by human evaluators. The pre-classified corpus of messages is the training dataset. The remaining corpus of unclassified messages is the test dataset. Both the training dataset and test dataset are subsets of the initial corpus of messages. There is no general rule on how to split the initial corpus (e.g., Hastie et al., (2009) suggest a split of 50/50).[39] Choosing too many messages for the training dataset could result in over fitting, i.e. the act of overgeneralizing in a training process by providing fewer training data for one set of cases.
The model interpretations become tighter and more specific the more messages are included in the training set so that the model performs poorly on a given random set. To preclude this possibility, the training dataset is split once more, so that there is an additional validation dataset to check for both over-fitting and under-fitting. In my approach, I adapt the machine learning techniques used to identify the intention of a message and to generate an appropriate response. To this end, this study uses the application IBM Watson

[36] Penczynski utilizes a Random Forest Classifier. In the area of classification algorithms, great progress has been made recently. Algorithms like XG Boost and Light Gradient Boosting Machine have outperformed Random Forest Classifier in several tasks (see, e.g., Chen & Guestrin (2016)).
[37] https://www.ibm.com/watson/
[38] Beside these categories there are some others, such as semi-supervised and reinforcement machine learning.
[39] It is obvious that these recommendations are dependent on the absolute number of available messages and that ongoing technical progress in the area of machine learning ensures ever better learning, also on the basis of small datasets.

Conversation.[40] I focus on the identification of the intention of a message. An intention is a purpose or goal that is expressed in natural language. In the following section, I explain the general methodological approach.

3. Methodology

The starting point of my approach is an initial corpus of messages (M) and a predefined set of categories (C):

(1) $M := \{m_1, \ldots, m_n \mid \text{all messages from the initial corpus of messages}\}$

(2) $C := \{c_1, \ldots, c_b \mid \text{all predefined categories}\}$

The number of categories, or intents (e.g. 'empty talk' and 'promise'), is only limited by the number of messages from the initial corpus, i.e. an appropriate number of messages is needed for every category c_i ($i \in \{1, \ldots, b\}$) in C. In this sense, the number of categories might be much higher than just two. For ease of presentation, I discuss the methodology as a binary classification task with only two intents, subsequently referred to as 'intent 1' (c_1) and 'intent 2' (c_2), so that b = 2 and $i \in \{1; 2\}$.

Based on the messages and categories, the procedure includes seven steps. As a first step, the initial corpus of messages is split into a training dataset (M_{Tr}) and a test dataset (M_{Te}).

The allocation of messages to the training dataset and test dataset is given by:

(3) $M_{Tr} := \{m_1, \ldots, m_k \mid \text{randomized subset of M, with } k < n \text{ messages}\}$

(4) $M_{Te} := \{M \setminus M_{Tr} \mid \text{remaining } n - k \text{ messages from M}\}$

As a second step, to follow a supervised machine learning approach, the randomly chosen messages included in the training dataset have to be classified manually and independently by two or more human coders.[41] To obtain unambiguously classified training data, as a third step, the interrater agreement of the evaluators is to be checked using Cohen's Kappa or Fleiss' Kappa[42], depending on the number of evaluators (Landis and Koch (1977)). Landis and Koch differentiate between 'moderate agreement' ($0.41 \leq \kappa \leq 0.60$), 'substantial agreement' ($0.60 < \kappa \leq 0.80$), and 'almost perfect agreement' ($0.80 < \kappa \leq 1.00$). The interrater agreement should at least be 'substantial'. Only messages which are classified unanimously (in the case of two evaluators) or at least by the majority of evaluators (in the case of more than two evaluators) should be used for the training dataset.[43] Ambiguously categorized messages, i.e. messages with the same wording.

One of which is categorized with 'intent 1' and the other with 'intent 2', should be excluded.[44] Accordingly, the classified corpus of messages, i.e. the training dataset, has to be revised as follows:

(5) $M_{R\,Tr} := \{M_{Tr}\} \setminus \{\text{all inconclusively classified messages}\}$

After classification and revision for all $j \in \{1, \ldots, k\}$ and for all $i \in \{1, 2\}$ it is known whether $m_j \in C_1$ or $m_j \in C_2$ holds so that $C_i := \{m_{i1}, \ldots, m_{ix} \mid \text{all messages classified with the intent } c_i\}$. The revised and classified training dataset is given by:

(6) $M_{R\,C\,Tr} := \{m_1, \ldots, m_h \mid \text{classified massages from } M_{Tr};\ h \leq k\}$

As a fourth step, the revised and classified training dataset ($M_{R\,C\,Tr}$) is split into two more subsets, the original training dataset ($M_{O\,Tr}$) and the validation dataset (M_{Va}). The different datasets can be summarized as follows:

---

[40] There are several other applications that could have been used, as will be discussed in the conclusion and outlook section. I have chosen IBM Watson Conversation (since renamed IBM Watson Assistant) because it is one of the available machine learning software applications with the highest degree of maturity in text analysis.

[41] According to the literature, there are various ways to do so. As already stated, H&X argue that a classification by the authors themselves is subjective. Instead, they recommend involving external evaluators.

[42] Cohen's Kappa and Fleiss' Kappa are statistical measures for assessing the reliability of observer agreement for nominal scales between two (Cohen's Kappa) or more than two observers (Fleiss' Kappa) (see Cohen (1960); Landis and Koch (1977)). 'It is directly interpretable as the proportion of joint judgments in which there is agreement, after chance agreement is excluded' (Cohen (1960), p. 46).

[43] The original ESP game used a 'good label threshold', i.e. before a label was attached to an image, it must have been agreed upon by at least a specified number of evaluator pairs. The specified number is the threshold, which can be lenient or strict.

[44] E.g. a message like 'okay' could be labeled as a 'promise' or 'empty talk', depending on which kind of question is answered.

(7) $M_{O\,Tr}$:= {m_1, …, m_g | randomized subset of $M_{R\,C\,Tr}$, with g < h messages}
(8) M_{Va} := {{$M_{R\,C\,Tr}$} \ {$M_{O\,Tr}$}| remaining h-g messages from $M_{R\,C\,Tr}$}

In order to specify the allocation of the messages (according to their categorization as a member of C_1 or C_2) in the original training dataset, the set can be defined by the cardinality of the set itself and its subsets C_1 and C_2 as $M_{O\,Tr} = M_{|O\,Tr|\,|C1|\,|C2|}$.[45]

The distribution of 'intent 1' and 'intent 2' in the revised and classified training dataset should be split proportionally between the original training dataset and the validation dataset. The validation dataset is needed to test the effectiveness of the original training dataset. As a fifth step, the original training dataset is uploaded on the IBM Watson Conversation workspace to train the machine learning model. After the machine learning model has been trained, the validation dataset is uploaded. The software then classifies validation messages as 'intent 1', 'intent 2', or the generic classification 'irrelevant'.[46] All messages that could not be assigned to one of the 'intent' classifications are classified as 'irrelevant.[47] In the sixth step, the whole revised and classified training dataset is checked for interrater agreement once more. The rater results are the classified messages yielded by human evaluators on the one hand and the classified messages of IBM Watson Conversation on the other. In addition to the interrater agreement for the whole revised and classified training dataset, the interrater agreement results for its subsets, the original training dataset, and the validation dataset also need to be analyzed.[48] As was the case in the third step, the target range for an interrater agreement should at least be 'substantial'. If the interrater agreement is outside the target range, over fitting or under fitting of the model may be the reason.
In such a case, the size of the original training dataset should be varied (once with more, once with fewer messages taken from the validation dataset M_{Va}). Depending on whether more or fewer messages result in a better interrater agreement, the final number of messages should be determined (as well as the proportional division between the particular classifications). Machine learning text classification models rely heavily on features such as words, numbers, and punctuation marks (among others). If there is not a sufficient number of features to separate the categories, a machine learning model will probably not work, although humans are still capable of classifying messages in such situations. Once an original training dataset that yields good results has been defined, the seventh step entails uploading the test dataset to classify all messages.

4. Results

In order to assess the approach, I apply it to available and classified message data from simple one-shot trust games. The starting point is the corpus of messages and classifications reported in the supplemental material

[45] In the previous step three, all messages with the same wording were excluded. Therefore, the cardinality of the sets is equal to the number of elements in the sets.
[46] The machine learning algorithms and the initial training data on which the artificial intelligence of IBM Watson Conversation is built are unknown. IBM has not published any information about the technologies and datasets used (Braun et al. (2017)). However, it seems likely that IBM validates the 'optimal' classification algorithm within the IBM Watson Conversation application, as IBM offers a stand-alone solution for the estimation of the best-performing classification algorithm with IBM Watson Auto AI.
[47] In the next section I will discuss further conclusions and applications of messages classified as 'irrelevant'.
[48] Depending on the employed machine learning algorithm(s), it is possible that the machine learning model is capable of classifying 100% of the messages correctly (as was the case for IBM Watson Conversation). In this case, the interrater agreement is an indicator for error detection in the context of training a model. For robustness checks, I implement a Multi-layer Perceptron (MLP) Classifier and a Random Forest (RF) Classifier in Python and test the same datasets on these classifiers. As a result, I observe 100% interrater agreement on the training data for the RF Classifier and 99.16% interrater agreement for the MLP Classifier on the training data (i.e. the MLP Classifier is capable of detecting 97.8% of all training datasets with 100% interrater agreement). The results can be found in Appendix B. The Python code will be provided upon request.

of C&D.[49] All blank messages are excluded. Therefore, the initial corpus of messages consists of 81 messages:[50]

(9) $M_{CD} = \{m_1, \ldots, m_{81} \mid \text{all messages reported by C\&D, except blank messages}\}$

C&D defines three categories. Deleting all blank messages, the category 'no message' is excluded from the analysis.[51] I employ the remaining two categories, 'promise' and 'empty talk':

(10) $C_{CD} = \{c_1, c_2 \mid \text{with } c_1 = \text{'promise' and } c_2 = \text{'empty talk'}\}$

As the messages reported by C&D are already coded and checked for interrater agreement beforehand, for the data used in this study, the training dataset is similar to the revised training dataset, as well as the revised and classified training dataset, i.e. $M_{Te} = M_{R\,Tr} = M_{R\,C\,Tr}$. However, as this study is exploratory in nature, before defining the revised and classified training dataset, I need to include one further step, which is not part of the procedure described above. In this step, I estimate the relative share of the entire message corpus needed for the original training dataset, as well as the relative share of messages classified as 'promise' and 'empty talk'. To do so, I use an iterative heuristic. That is, I vary the overall size of the training set.

As well as the share of messages that are associated with either c_1 or c_2, to assess how many messages are needed to yield good results while minimizing the risk of overfitting. As a minimum, k is set to $k=|M_{O\,Tr}|=8$, which is about 10% of the size of the initial corpus or messages. In terms of the minimum number of messages per category, all of the tested training sets include at least one message classified as 'promise' and 'empty talk' respectively. Specifically, this implies that there are seven combinations of 'promise' and 'empty talk' messages when the training set includes eight messages in total. For the first case (a total of eight messages from C&D with one randomly chosen message coded as 'empty talk' and seven randomly chosen messages coded as 'promise'), the original training dataset is:

(11) $M_{1\_|O\,Tr|\,|Ce|\,|Cp|} = M_{8\,1\,7} = \{m_{ijep} \mid 1 \le j \le 8 \wedge e = 1 \wedge 1 \le p \le 7\}$,
with $1 \le i_1 < i_2 < \ldots < i_8 \le 81$ and $i_j \in \mathbb{N}$, j=1; 2; …; 8.

For the second case, the original training dataset can be defined as:

(12) $M_{2\_|O\,Tr|\,|Ce|\,|Cp|} = M_{8\,2\,6} = \{m_{ijep} \mid 1 \le j \le 8 \wedge 1 \le e \le 2 \wedge 1 \le p \le 6\}$,
with $1 \le i_1 < i_2 < \ldots < i_8 \le 81$ and $i_j \in \mathbb{N}$, j=1; 2; …; 8. For the upper limit in terms of training set size, there is no standard recommendation in the literature.

In order to limit the effort associated with manual evaluation, the upper limit is set to 2/3 of the total number of messages (which is 54). Thus, all possible solutions in terms of the allocation to 'empty talk' and 'promise' are varied within the range of 8 to 54 messages, i.e. |O Tr|=8; 9;…;54. In this way, I test 683 different combinations of |O Tr|, |Ce|, and |Cp|.[52] To achieve robust results, I test five randomly assigned original training datasets within the above-described limits for every possible variation of |O Tr|, |Ce|, and |Cp|, which amounts to a total of 3,415 datasets.[53] After the training of the model on IBM Watson Conversation has been finished for a certain original training dataset, the initial corpus of messages (which, in the case of this study, is composed of the original training dataset and the validation dataset) is uploaded and classified. With regard to the category 'empty talk', as in all other experiments concerning 'empty talk' intent.

[49] Because of its similarity, I used all messages sent from B in the payoff vectors (5,5) and (7,7). Overall, C&D reported 91 messages in these payoff vectors, of which 10 were blank and without any text.

[50] As stated in the previous section, messages with the same wording, which were categorized with 'intent 1' and 'intent 2', should be excluded. I did not identify any such messages. There are two messages which could have fallen into this category: 'I'll choose to roll' and 'I will choose to roll'. Because they do not feature exactly the same spelling and were both assigned the intent 'promise', I decided to include both messages in the initial corpus of messages.

[51] This is due to the fact that this would distort the results, as a training of just one blank message with the intent 'no message' would always result in the 100% agreement of all blank messages with this intent.

[52] In the reported and classified data of C&D, there are no more than 33 messages evaluated as 'empty talk'. To ensure that at least a (randomly selected) third of all 'empty talk' messages were part of the training dataset, the upper limit for 'empty talk' messages is set at a total of 22. Accordingly, the maximum number of messages classified as 'promise' in the original training dataset is 32.

[53] I used Stata to create the files.

It is very broad and can vary in terms of subject matter, dealing with topics such as the weather, politics, sports, and many more. Therefore, IBM Watson Conversation uses the category 'irrelevant', which encompasses all messages that could not be assigned to one of the pre-specified categories. Still, if the other categories are unambiguously classified, I argue that it is legitimate to assume that all messages that have been classified by IBM Watson Conversation as 'irrelevant' can be reclassified as 'empty talk'.[54] Applying this principle, I get the following interrater agreement results for the initial corpus of messages.[55] Of the 3,415 original training datasets, 2,476 yields a Kappa score on the initial corpus of messages that is at least substantial ($\kappa > 0.60$). Correspondingly, for the remaining 939 datasets, the Kappa score is equal to or smaller than the threshold level of 0.6.

Result 1: For more than 70% of training datasets, the machine learning-generated classifications on the initial corpus of messages are substantially similar to the classifications of C&D. Among the 939 poor-performing datasets, 925 have either eight or fewer 'empty talk' messages or eight or fewer 'promise' messages. Accordingly, there are a minimum number of messages per category that is necessary to train a valid model.

Result 2: The minimum number of messages per category necessary to train a valid machine learning model is roughly 10% of the initial corpus of messages. Checking for interrater agreement only on the messages of the validation dataset, 1,259 datasets have a Kappa score larger than 0.6 and 2,156 datasets a Kappa score equal to or smaller than the threshold of 0.6. I also checked the Kappa score on all original training datasets. However, these results are just for purposes of error detection, as IBM Watson Conversation was capable of classifying 100% of the messages of the original training dataset correctly. Aside from the agreement results, I am also interested in whether certain messages are regularly coded differently by the machine learning application in comparison to human coding. Therefore, I code the classifications of IBM Watson Conversation, as well as those of C&D, to '1' for 'empty talk' and '2' for 'promise'. Based on these values, I determine the mean assessment overall training datasets and compare this mean value to the assessment of C&D. Overall, there are three individual messages where the mean deviates by more than 0.5 from the human coding.[56] For most of the messages, the deviation was less than 0.15.

Result 3: Over the full range of training datasets, the average assessment of the machine learning approach per message corresponds to the assessment of C&D for 78 out of 81 messages. Up to this point, the descriptive results have been discussed. An unambiguous training of this kind of category would be close to impossible, not to mention cost-prohibitive. For a deeper understanding, the models reported in Table 1a and Table 1b evaluate the impact of the total number of messages in the original training dataset (#Messages_J), the number of empty talk messages (#EmptyTalk_I), the number of promise messages (#Promise_K), and interaction terms of these three variables on Kappa.

[54] This is in line with the results. Among all 3,415 datasets, IBM Watson Conversation classified 17,114 messages under the category 'irrelevant'. 16,148 of them (i.e. 94.4%) were assigned by C&D to the category 'empty talk', while 966 messages (i.e. 5.6%) were classified as 'promise'. In the course of implementing both the MLP Classifier and the Random Forest Classifier in Python for robustness checks, I adopted a similar approach. Those messages with an estimated probability of 0.5 for both categories were assigned to category 1, i.e. 'empty talk'.

[55] The results were calculated with Stata. The Stata code, as well as the full classification of messages, can be supplied upon request.

[56] These messages are those from B to A in the (5, 5) treatment with the ID 10 in Session 1, with the ID 14 in Session 1, and in the (7, 7) treatment with the ID 6 in Session 1.

Table 1a: Tobit Regression Results, All Messages Included

	(a)	(b_e)	(b_p)	(c_e)	(c_p)
#Messages_ J	0.0107*** (0.0002)	0.0113*** (0.0003)	0.0094*** (0.0004)	0.0042*** (0.0006)	0.0166*** (0.0007)
#EmptyTalk_I	--	-0.0019*** (0.0005)	--	-0.0180*** (0.0012)	--
#Promise_K	--	--	0.0019*** (0.0005)	--	0.0128*** (0.0009)
#Messages_ J X #EmptyTalk_I	--	--	--	0.0006*** (0.0000)	--
#Messages_ J X #Promise_K	--	--	--	--	-0.0004*** (0.0000)
Constant	0.3590*** (0.0073)	0.3648*** (0.0074)	0.3648*** (0.0074)	0.5450*** (0.0144)	0.1960*** (0.0141)
Obs.	3415	3415	3415	3415	3415
LR chi²	1585.28[a]	1600.89[a]	1600.89[a]	1803.07[a]	1790.85[a]

Standard errors are reported in parentheses. * p<0.1, **p<0.05, ***p<0.01, [a] p < 0.001.

Table 2b: Tobit Regression Results, All Messages Included

	(d_e)	(d_p)	(e_e)	(e_p)
#Messages_ J	-0.0026** (0.0011)	0.0193*** (0.0012)	0.0154*** (0.0012)	0.0100*** (0.0019)
#EmptyTalk_I	-0.0614*** (0.0029)	--	0.0329*** (0.0024)	--
#Promise_K	--	0.0463*** (0.0027)	--	-0.0151*** (0.0024)
#Messages_ J X #EmptyTalk_I	0.0023*** (0.0001)	--	-0.0006*** (0.0001)	--
#Messages_ J X #Promise_K	--	-0.0016*** (0.0001)	--	0.0001** (0.0001)
Constant	0.7393*** (0.0243)	0.0633** (0.0250)	0.0734* (0.0428)	0.6140*** (0.0650)
Obs.	1930	1930	1485	1485
LR chi²	729.28[a]	605.33[a]	1192.30[a]	1105.51[a]

Standard errors are reported in parentheses. * p<0.1, **p<0.05, ***p<0.01, [a] p < 0.001.

Kappa as dependent variable (i.e. Kappa score overall messages, while irrelevant messages are coded as empty talk). Model (a) supports the finding that there is no substantial evidence of over-fitting effects, as there is a positive effect of the total number of messages constituting the original training dataset. Model (b) specifies the influence of the number of messages per category. The higher the number of promise messages included in a certain training dataset, the higher the Kappa score. By contrast, a higher number of empty talk messages go along with lower Kappa scores. This seems to confirm that the distinction between the two categories 'empty talk' and 'promise' relies more on an unambiguous training of the category 'promise'. Considering the interaction of the number of all messages and of messages per category in the training data as seen in the model (c), when the size of the training dataset increases, the positive effect of including more promise messages decreases. Again, the number of empty talk messages induces a contrary effect.

When the size of the training set increases, the negative effect of including more empty talk messages is smaller. Based on these findings, I split the total number of messages in the original training dataset into two sub-groups: (d) a group of sets with a small number of messages, so that J ≤ 30, and (e) a group of sets with a large number of messages, so that J > 30. Given a small training dataset size, the results discussed above are still valid. However, given a large training dataset size, contrasting results can be observed. In this case, a higher number of 'promise' messages go along with a reduction in Kappa, while an increasing number of 'empty talk' messages yield higher Kappa scores. This result might display over-fitting effects of including too many 'promise' messages. Whereas with small training set sizes, including more 'promise' messages seems to

be vital to yield high Kappa scores, when the number of messages included in the training set is high enough, a higher share of 'empty talk' messages is to be preferred. Thus, a higher share of 'promise' messages could lead to an overfitting effect as the marginal utility of an additional 'promise' diminishes or becomes negative. In contrast, including more 'empty talk' will not lead to over-fitting effects because of its broad range of contents. In order to be able to get a recommendation for choosing a training set size, I analyze the distribution of Kappa, depending on the training set, size in more detail.

Figure 1: True and Predicted Distribution of Kappa (Basis for Prediction: Model (c))

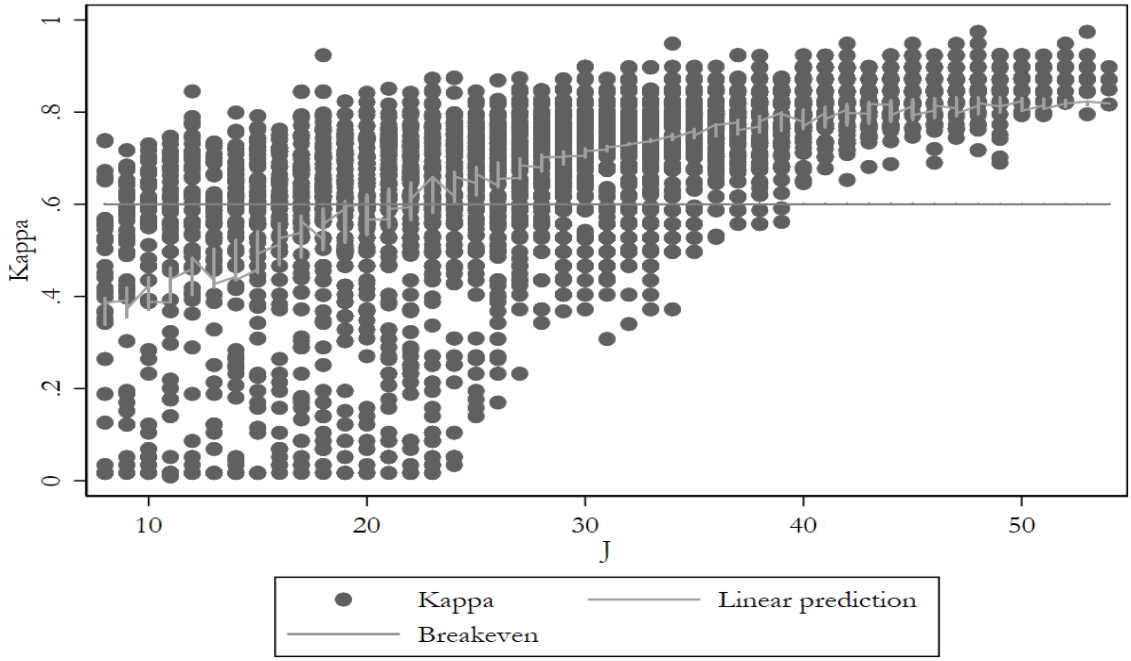


Figure 1 shows the observed Kappa scores for each training set size, as well as the Kappa scores predicted by the model (c). The predicted breakeven point, i.e. the point at which all predicted Kappa scores are higher than 0.6, is J = 23. The true breakeven point is at J = 40. The figure also suggests that there are three groups of training set sizes: (1) those that might result in very low Kappa scores, (2) those that allow for high Kappa scores regularly, but that also involve a certain risk of bad coding, (3) and those that generate Kappa scores that are consistently high. In other words, there seem to be three groups of training set sizes that differ in terms of the variance of Kappa. To determine whether this visual impression is verified analytically, I conduct a cluster analysis with three clusters, defined by the distinctive range of training set sizes. Table 2 presents the results. As expected from the analysis of Figure 1, each of the three clusters encompasses a range of J that represents one of the three groups discussed above. While there are Kappa scores > 0.9 in all three clusters, the minimum Kappa score per cluster is substantially higher in the third cluster as compared to the first and the second. In the same vein, and as expected from the analysis of Figure 1, the standard deviation of Kappa is smaller in the former cluster as compared to the latter two. I conclude that with training set sizes from the third cluster, the probability of accurate evaluations is very high.

With training set sizes from the second cluster, the mean standard deviation of Kappa is 1/3 smaller than with training sets from the first cluster. Consequently, I consider training sets from cluster 1 to be undesirable. I recommend using a training set size that is not smaller than a quarter of the whole set of messages, while training sets that include 1/2 of the total number of messages seem to be sufficiently informative to generate high Kappa scores, even if the number of messages per category cannot be influenced.

Result 4: Training sets that include 50% of the initial corpus of messages yield high Kappa scores even if the distribution of messages over categories is unknown. Yet it is also possible to yield high Kappa scores with far

less than 50% of the messages included in the initial corpus of messages. As already stated in the sixth step of the methodology, the size of the original training dataset, as well as the proportion of the messages per category in the training dataset, should be varied (once with more, once with fewer messages taken from the validation dataset M_{Va}). Depending on whether more or fewer messages result in a better interrater agreement, the final number of messages should be determined (as well as the proportion of messages drawn from the various classification categories).

Suggestion: If the distribution of messages over categories is unknown and the experimenter is unable to generate human codings for 50% of the initial corpus of messages, I recommend varying the number of messages in the training set so that interrater agreement is maximized. However, for small message datasets, I advise against using training set sizes smaller than 25% of the initial corpus of messages.

Table 3: Descriptive Statistics for Three Clusters of Training Sets, With Distinctive Ranges of Size_Training_Set

	# Training sets	Mean	SD	Min	Max
Cluster 1 – by J					
Kappa_All	1270	0.54	0.21	0.01	0.92
Number_ET	1270	8.74	5.42	1	22
Number_P	1270	8.83	5.42	1	23
Size_Training_Set	1270	17.57	4.62	8	24
Cluster 2 – by J					
Kappa_All	1380	0.70	0.12	0.14	0.95
Number_ET	1380	11.84	6.19	1	22
Number_P	1380	18.97	6.95	3	32
Size_Training_Set	1380	30.82	3.68	25	37
Cluster 3 – by J					
Kappa_All	765	0.82	0.06	0.56	0.97
Number_ET	765	16.67	4.11	6	22
Number_P	765	26.67	4.11	16	32
Size_Training_Set	765	43.33	4.11	38	54

Indeed, it must be considered that the same training data could have been used to classify far more messages. Depending on the experimental environment and the restricted instructions of the trust game reported by C&D, I argue that more 'promise' messages from more subjects would have been very similar in their structure and utilized phrases.[57] Therefore, I used the above-discussed training data, based on the C&D messages, to classify messages from a comparable experiment reported by Deck et al. In this way, I am also able to demonstrate the scalability of machine learning models for the classification of messages from laboratory experiments. The results are reported in Appendix A, while additional econometric analysis appears in Appendix B. If we now consider the messages reported from C&D along with those reported by Deck et al. as one initial corpus of messages, the results indicate that training sets that include around 1/3 of the total number of messages seem to be sufficiently informative to generate highly substantial classification results. Overall, comparing the results of the validation data derived from C&D messages with validation data from the Deck et al. messages, we see even better results in the case of the Deck et al. messages.

5. Conclusion and Outlook

This paper reports to the best of my knowledge, the first results of an 'out-of-the-box machine learning classification approach utilized to classify messages from a laboratory experiment. Using the knowledge and intuition of human evaluators who classify a corpus of messages, I am able to replicate the classification results presented by C&D and Deck et al. with far less than 50% of the messages from the initial corpus of messages used as training data. With training sets that encompass 50% of the initial corpus of messages, the

[57] It is logical to assume that this assumption is not valid for the 'empty talk' category.

machine learning algorithm robustly yields high agreement with the human coders, irrespective of the distribution of messages over categories. These results are especially interesting for datasets encompassing far more messages than reported by C&D. Using this approach, large message datasets can be analyzed with the same effort as small message datasets. Furthermore, the study provides evidence of the scalability of the approach. Using training data from one message corpus (reported by C&D) to classify another message corpus (reported by Deck et al.,), I achieve robust and highly substantial results. These findings rely heavily on the similarity of both studies and the instructions of their evaluators.

However, these findings are nevertheless interesting for experimenters who conduct replication studies or split their experimental sessions into several time slots. In the first case, already-labeled data from the original study can be used to train the machine learning model. In the latter case, the communication data from the first-run sessions can already be used to train the machine learning models. Thereby, the process of running experimental sessions and analyzing communication data can be parallelized and the time required to finish a publication can be shortened. Besides the potential to more efficiently analyze large datasets, another advantage of this approach is that the evaluation is history-independent. This means that there is no human bias in the process of evaluating a vast corpus of messages. Although many experimenters consider this in their instructions, it is unlikely that human evaluators are able to act consistently over time.[58] The machine learning algorithm instead evaluates all messages independently and with the same reliability for all messages, as the same algorithm is used for all messages.[59] In their experiment, Nielsen et al. (2019) classify messages into numerous categories.[60] In doing so, they try to achieve deeper insight into the rich chat content. Their results show relatively low agreement rates for most of the categories.

I acknowledge that my approach works well for a binary classification task with two categories, but problems might arise when trying to apply it to fields with a broader range of categories, especially if the amount of potential training data for each category is restricted. The agreement of the machine learning classifications with the human coding of the communication data, reported by C&D and Deck et al. relies on the separability of the categories 'empty talk' and 'promise'. The separability is inherent in the diversity of these two categories, as they are characterized by different features (i.e. the words of the messages). The less the categories are characterized by different, and therefore unique, features with regard to their category, the more difficult it is for a machine learning model to separate these categories.[61] Therefore, the application of such machine learning models to less separable categories should be handled with care. A further challenge is the analysis of multiparticipant chats (see, e.g., Uthus and Aha (2013)). Tracking the intention in synchronous discussions within a single message corpus is difficult even in the context of human evaluation, let alone machine analysis. While the classification in a one-shot game with preplay communication is based on the perception of a single isolated message, the classification of messages in a multi-participant chat rests on who is talking to whom, which question is being answered, and many other aspects.

In this paper, the approach was tested using an application developed by IBM: IBM Watson Conversation. In ongoing research, I evaluate other applications that might perform as well as this one or even better (see, e.g., Braun et al., (2017)). Besides this, an open question is whether different applications perform better or worse on different experimental tasks. In addition, the expansion of this approach to other natural language understanding functionalities, e.g. the identification of entities and sentiments, ought to be considered. I am

[58] See, e.g., coding rules of Deck et al.: 'The unit of observation is a single message'; 'Your job is to capture the content of the message (…). Think of yourself as a "coding machine".'

[59] Natural language understanding services, like IBM Watson Conversation, improve over time. The reproducibility of the results is therefore only conditionally possible. The data was collected in the period from October 2 to October 16, 2018. In this time, I worked on the IBM Watson Conversation version 2018-09-20 (the service has since been renamed IBM Watson Assistant). Therefore, the very same approach with the same messages, exactly as described in this paper, might lead to different (presumably better) results.

[60] Their experiment was a replication of an experiment by C&D, but using teams instead of individuals.

[61] Nielsen (2019), for example, employs the categories 'weak promise' and 'strong promise'.

also interested in the further possibilities of adaption and scalability, as well as the influence of external evaluators. Therefore, I want to see whether one could use already-trained models to classify the messages of further laboratory experiments, as already exemplified in the results section. In general, a machine learning model is most reliable if the training dataset follows a distribution similar to that of future datasets, which are still to be classified (Mitchell (1997)). Therefore, I will use already-trained models based on the classifications reported by C&D and H&X to classify the messages reported, e.g., by Ismayilov and Potters (2016).[62]

I do not focus on the impact of single messages on the performance of a machine learning model. However, I observe some messages which were, in the mean, evaluated differently by IBM Watson Conversation than by the human evaluators. This might be due to the fact that these messages entailed contradictory content. Messages with several sentences might be inconsistent if the first sentence of a message is classified as 'promise', whereas the second sentence is classified as 'empty talk'. Assigning one category to this message in the machine learning model is logically false. In further research, I will also clarify whether it is possible to identify in advance messages that will lower the performance of the machine learning model and how to handle such multi-class classified messages. This study has presented a novel approach to analyzing experimental communication data. The study showed several advantages, in particular, greater efficiency of machine learning message classification as compared to a human coding procedure. So far, this study has ignored the question of whether experimenters will trust such machine learning-generated results. Penczynski (2019), for example, analyzes the influence of particular features or tokens (i.e. words, numbers, and other components of natural language text) on the performance of a Random Forest Classifier.

By doing so, the human 'judges' of the classification algorithm could use their intuition to evaluate whether they would trust the machine. The paper of Ribeiro et al. (2016) is comparable in this regard. They use a model-agnostic approach to explain the outcomes of almost every classification algorithm, among them some so-called 'black box' classifiers. Inexperienced machine learning users are tasked with testing the results of their explanation approach. Thereby, Ribeiro et al. (2016) show that the users are more trusting in the machine learning models, and have the ability to tune hyperparameters of the models if they are guided by the model-agnostic approach. Besides these technical questions associated with the implementation of machine learning in the analysis of experimental data, I see an equally great challenge in convincing experimenters that results can be trustable, even if they derive from a black-box algorithm. Democratizing artificial intelligence is one approach to tackling this problem. IBM and other companies are working on tools to foster greater understandability of machine learning models and the interpretability of their results. However as shown in this paper, expectations regarding artificial intelligence must remain realistic.

References

Agranov, M. & Yariv, L. (2018). Collusion through communication in auctions, *Games and Economic Behavior*, 107, 93-108.

Ahn, L. v. & Dabbish, L. (2004). Labeling images with a computer game. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 319-326.

Braun, D., Hernandez-Mendez, A., Matthes, F. & Langen, M. (2017). Evaluating natural language understanding services for conversational question answering systems. Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, 174-185.

[62] Ismayilov and Potters (2016) utilized the trust game reported by C&D and employed external evaluators to classify the messages. Because of this, I want to apply the classifications of H&X, as Ismayilov and Potters (2016) used external evaluators as well. However, in contrast to H&X, Ismayilov and Potters (2016) incentivized the evaluators just for the completion of the task. There were neither additional nor variable earnings if the evaluation matched that of any other nor the majority of the other evaluators. Another essential difference is that the messages reported by Ismayilov and Potters (2016) are much longer. This might lead to ambiguous classification results, as one message could consist of sentences that were classified as 'empty talk' in an isolated evaluation as well as sentences classified as 'promise' in an isolated evaluation.

Charness, G. (2006). Promises and partnership. *Econometrica,* 74(6), 1579-1601.
Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 785-794.
Cohen, J. (1960). A Coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 20(1), 37-46.
Deck, C., Servátka, M. & Tucker, S. (2013). An examination of the effect of messages on cooperation under double-blind and single-blind payoff procedures. *Experimental Economics,* 16(4), 597-607.
Fischer, C. & Normann, H. T. (2019). Collusion and bargaining in asymmetric Cournot duopoly - An experiment. *European Economic Review,* 111, 360-379.
Hastie, T., Friedman, J. & Tibshirani, R. (2009). The elements of statistical learning: Data mining, inference, and prediction. New York: Springer.
Houser, D. & Xiao, E. (2011). Classification of natural language messages using a coordination game. *Experimental Economics,* 14(1), 1-14.
Huang, L. & Xiao, E. (2018). Peer effects in public support for Pigouvian taxation. Working Paper.
Huerta, J. (2008). Relative rank statistics for dialog analysis. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 965-972.
Isaac, R. & Walker, J. (1988). Communication and free-riding behavior: The voluntary contribution mechanism. *Economic Inquiry,* 26(4), 585-608.
Ismayilov, H. & Potters, J. (2016). Why do promises affect trustworthiness, or do they? *Experimental Economics,* 19(2), 382-393.
Khan, A., Baharudin, B., Lee, L. H. & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology,* 1(1), 4-20.
Landis, J. & Koch, G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 33(2), 363-374.
Lundquist, T., Ellingsen, T., Gribbe, E. & Johannesson, M. (2009). The aversion to lying. *Journal of Economic Behavior & Organization,* 70(1-2), 81-92.
Mitchell, T. (1997). Machine Learning. New York: McGraw-Hill.
Moellers, C., Normann, H. T. & Snyder, C. M. (2017). Communication in vertical markets: Experimental evidence. *International Journal of Industrial Organization,* 50, 214-258.
Nielsen, K., Bhattacharya, P., Kagel, J. & Sengupta, A. (2019). Teams promise but do not deliver. *Games and Economic Behavior,* 117, 420-432.
Penczynski, S. P. (2019). Using machine learning for communication classification. *Experimental Economics,* 22(4), 1002-1029.
Ribeiro, M., Singh, S. & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 1135-1144.
Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computer Surveys,* 34(1), 1-47.
Snow, R., O'Connor, B., Jurafsky, D. & Ng, A. (2008). Cheap and fast - But is it good?: Evaluating non-expert annotations for natural language tasks. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 254-263.
Uthus, D. & Aha, D. (2013). Multiparticipant chat analysis: A survey. *Artificial Intelligence,* (199-200), 106-121.

Appendix A: Additional Results on Deck et al. Message Corpus: Deck et al. implement a one-shot hidden action trust game of C&D in a single-blind and a double-blind payoff procedure. The single-blind payoff procedure is an exact replication of the trust game reported by C&D. However, both in the single-blind and in the double-blind payoff conditions, Deck et al. implement some slight modifications compared to C&D.[63] This is especially interesting, as I would like to know more about further scalability of a machine learning model trained by exogenous data on similar experimental results.[64] In the first stage, the procedure is the same as in the previous section, as data reported by C&D is used for the training data. Therefore, I consider a total number of 3415 training datasets. In contrast to the first series of datasets, the validation data set is not composed of the remaining messages. Instead, for all 3415 training datasets, the validation dataset is composed of the messages and classifications reported in the supplement material of Deck et al. (2013).[65] Again, all blank messages are excluded. Therefore, the validation dataset consists of 44 messages:[66]

(1) $M_{Va\,De} = \{m_1, \ldots, m_{44} \mid$ all messages reported by Deck et al., except blank messages$\}$

Using the messages reported by Deck et al., the question arises whether the human classification results are comparable to C&D. I argue that this is permitted, as Deck et al. also use a weak definition of promise. Furthermore, they employ three evaluators to evaluate their messages, as well as the messages of C&D.

Agreeing on 89% (i.e., $\kappa = 0.766$) of the reported messages by C&D, the evaluators seem to have the same understanding of a weak promise statement. As with the messages from C&D, the categories from Deck et al are not adopted one to one. Deleting all blank messages, the two categories 'promise' and 'empty talk' are employed:[67]

(2) $C_{De} = \{c_1, c_2 \mid$ with c_1 = 'promise' and c_2 = 'empty talk'$\}$

Applying the machine learning training data models from the C&D messages on the message corpus reported by Deck et al. I check for interrater agreement between the classifications given by Deck et al. and the classifications given by IBM Watson Conversation. I get the following interrater agreement results for the initial corpus of messages. Of the 3,415 training datasets, in 1,743 cases the Kappa score on the validation dataset ($M_{Va\,De}$) is at least substantial ($\kappa > 0.60$). Compared to the results discussed before, where 1,259 validation data sets[68] yield a Kappa score larger than 0.6, the machine learning model seems capable of performing and generalizing better on the test data. Correspondingly, for the remaining 1,672 data sets, the

[63] To their surprise, Deck et al. were not able to replicate the results given by C&D. Therefore, they point out several notable differences in the procedure (besides the payoff procedures): (1) Deck et al. conducted their experiment in a lab, not in a classroom; (2) the subject pool was different, as their experiments were conducted in the south eastern US, not in California; (3) Deck et al. used a curtain to separate As from Bs; (4) they did not include elicited subjects beliefs; (5) In C&D strict preplay communication from B to A was implemented. '(…) Messages were sent before As made their decisions, and Bs roll decision was made after. The decisions were made on separate forms.' In the setup of Deck et al., '(…), roll choices were made at the top and messages could be written at the bottom of the same form. (…) While there is no way to control the order in which subjects in the B role complete the response form, it is likely that many completed the top portion first.' This way, Bs had the opportunity to send messages about what they have done or what they plan to do, whereas the Bs in the setup of C&D can send messages exclusively based upon what they plan to do.

[64] I do not discuss how to identify 'similar' experimental setups. Although, from a theoretical point of view, it is possible to automatically classify the instructions of experimental designs with a machine learning model.

[65] I used all messages sent by B in the single-blind and in the double-blind condition. Deck et al. reported 74 messages in these conditions of which 30 were blank messages without any text.

[66] As stated in the main paper, messages with the same wording, which were categorized with 'intent 1' and 'intent 2', should be excluded. I did not identify any such messages.

[67] Deck et al. used the categories named 'promise', 'non-promise message' and 'blank'. They followed in their definition of the promise category C&D. As C&D define the category promise as '(…) any statement of intent (…)', Deck et al. define their 'promise category" as 'a promise or statement of intention (…)'. The 'empty talk category' from Deck et al. is defined as 'a message that is not blank, but does not contain a promise statement of intention (…)'.
As Deck et al. employed three evaluators, some messages had not been classified in the same category by all three evaluators. In these cases, the classification of the majority of the evaluators is employed.

[68] The validation data sets were composed of the remaining messages from C&D.

Kappa score is equal to or smaller than 0.6. Within these 1,672 datasets, 1,390 datasets have either eight or fewer 'empty talk' messages or eight or fewer 'promise' messages. This is in line with the previous results, indicating that a minimum number of messages per category are needed for a valid training of the model.

Table A1a: Tobit Regression Results, All Messages Included

	(a)	(b_e)	(b_p)	(c_e)	(c_p)
#Messages_ J	0.0133*** (0.0003)	0.0137*** (0.0004)	0.0124*** (0.0005)	0.0052*** (0.0008)	0.0213*** (0.0009)
#EmptyTalk_I	--	-0.0013** (0.0006)	--	-0.0205*** (0.0016)	--
#Promise_K	--	--	0.0013** (0.0006)	--	0.0147*** (0.0012)
#Messages_ J X #EmptyTalk_I	--	--	--	0.0007*** (0.0001)	--
#Messages_ J X #Promise_K	--	--	--	--	-0.0005*** (0.0000)
Constant	0.2073*** (0.0098)	0.2108*** (0.0010)	0.2108*** (0.0010)	0.4264*** (0.0195)	0.0026 (0.0192)
Obs.	3415	3415	3415	3415	3415
LR chi²	1394.76[a]	1398.66[a]	1398.66[a]	1556.87[a]	1554.59[a]

Standard errors are reported in parentheses. * p<0.1, **p<0.05, ***p<0.01, [a] p < 0.001.

Table A1b: Tobit Regression Results, All Messages Included

	(d_e)	(d_p)	(e_e)	(e_p)
#Messages_ J	0.0012 (0.0015)	0.0223*** (0.0016)	0.0153*** (0.0023)	0.0241*** (0.0034)
#EmptyTalk_I	-0.0614*** (0.0040)	--	0.0376*** (0.0046)	--
#Promise_K	--	0.0536*** (0.0036)	--	-0.0074* (0.0044)
#Messages_ J X #EmptyTalk_I	0.0022*** (0.0002)	--	-0.0006*** (0.0001)	--
#Messages_ J X #Promise_K	--	-0.0018*** (0.0001)	--	-0.0002 (0.0001)
Constant	0.5777*** (0.0329)	-0.1254*** (0.0338)	-0.0404 (0.0816)	0.1654 (0.1214)
Obs.	1930	1930	1485	1485
LR chi²	527.51[a]	506.49[a]	699.77[a]	676.56[a]

Standard errors are reported in parentheses. * p<0.1, **p<0.05, ***p<0.01, [a] p < 0.001.

Kappa as dependent variable (i.e. Kappa-score overall messages, while irrelevant messages are coded as empty talk). Regarding the Tobit regression results, as shown in Table A1a and Table A1b, I observe largely similar results compared to using the messages from C&D as validation data. Again, I observe that for small training sets, Kappa benefits from including more promise messages, while in bigger training sets, including more empty talk messages is to be preferred. The regression results are also in line with the cluster results shown in Table A2. Regarding Cluster 3, the mean Kappa score indicates an 'almost perfect agreement'. Within this Cluster 3 (as in Cluster 2) I also observe that a higher share of promise messages regularly goes along with higher Kappa scores. Overfitting on the Deck et al. messages driven by promise messages in larger training sets seems to be less strong compared to the results on the C&D messages. This is supported by the finding that the coefficient of the variable indicating the number of promise messages in the model (e_p) is very small. In fact, the reason might be that there is no subset from the original training dataset and the validation dataset with the corpora of messages utilized as test datasets.

Table A2: Descriptive Statistics for Three Clusters of Training Sets, With Distinctive Ranges of Size_Training_Set

	# Training sets	Mean	SD	Min	Max
Cluster 1 – by J					
Kappa_All	945	0.52	0.21	0.01	0.92
Number_ET	945	7.85	4.78	1	20
Number_P	945	7.85	4.78	1	20
Size_Training_Set	945	15.7	3.85	8	21
Cluster 2 – by J					
Kappa_All	1420	0.66	0.15	0.02	0.95
Number_ET	1420	11.5	6.31	1	22
Number_P	1420	16.5	7.26	1	32
Size_Training_Set	1420	28	3.72	22	34
Cluster 3 – by J					
Kappa_All	1050	0.8	0.07	0.5	0.97
Number_ET	1050	15.67	4.82	3	22
Number_P	1050	25.67	4.82	13	32
Size_Training_Set	1050	41.33	4.82	35	54

Appendix B: Additional Econometric Analyses: I estimate models similar to those in Table 1 with Kappa scores calculated based on the validation-messages only, i.e. messages that have been part of the training set are excluded when calculating the agreements between that specific training set and the original coding of C&D. The results are presented in Table B1 and show that the findings discussed above are also valid if I consider Kappa scores calculated on validation data only.

Table B1a: Tobit Regression Results on Test Data Only

	(a)	(b_e)	(b_p)	(c_e)	(c_p)
#Messages_ J	0.0063*** (0.0003)	0.0076*** (0.0003)	0.0034*** (0.0004)	-0.0003 (0.0006)	0.0097*** (0.0007)
#EmptyTalk_I	--	-0.0042*** (0.0005)	--	-0.0221*** (0.0013)	--
#Promise_K	--	--	0.0042*** (0.0005)	--	0.0137*** (0.0010)
#Messages_ J X #EmptyTalk_I	--	--	--	0.0006*** (0.0000)	--
#Messages_ J X #Promise_K	--	--	--	--	-0.0003*** (0.0000)
Constant	0.3453*** (0.0079)	0.3572*** (0.0079)	0.3572*** (0.0079)	0.5580*** (0.0155)	0.2095*** (0.0153)
Obs.	3415	3415	3415	3415	3415
LR chi²	544.95[a]	611.62[a]	611.62[a]	828.63[a]	735.55[a]

Standard errors are reported in parentheses. * p<0.1, **p<0.05, ***p<0.01. *Kappa_Test* as dependent variable (i.e. training messages are excluded and irrelevant messages are coded as empty talk). [a] p < 0.001.

Table B1b: Tobit Regression Results on Test Data Only

	(d_e)	(d_p)	(e_e)	(e_p)
#Messages_ J	-0.0054*** (0.0012)	0.0119*** (0.0012)	0.0118*** (0.0020)	0.0050* (0.0030)
#EmptyTalk_I	-0.0600*** (0.0030)	--	0.0297*** (0.0040)	--
#Promise_K	--	0.0483*** (0.0028)	--	-0.0139*** (0.0038)
#Messages_ J X #EmptyTalk_I	0.0021*** (0.0001)	--	-0.0006*** (0.0001)	--

#Messages_ J X #Promise_K	--	-0.0015*** (0.0001)	--	0.0002 (0.0001)
Constant	0.7166*** (0.0248)	0.0782** (0.0254)	0.0486 (0.0710)	0.5779*** (0.1057)
Obs.	1930	1930	1485	1485
LR chi²	585.29[a]	504.71[a]	313.10[a]	285.60[a]

Standard errors are reported in parentheses. * p<0.1, **p<0.05, ***p<0.01. *Kappa_Test* as dependent variable (i.e. training messages are excluded and irrelevant messages are coded as empty talk). [a] p < 0.001.

FE-Model C&D: The following model results in Table B2a and Table B2b depict fixed effects regressions. Within each model, fixed effects for training set size are included. The variable 'Messages_J_Rel' is defined as the absolute number of messages in the specific training set, divided by the total number of messages in the whole corpus of messages. Likewise, the variables 'Number_EmptyTalk_I_Rel' and 'Number_Promise_Rel' are defined as the absolute number of promise or empty talk messages in the specific training set, divided by the total number of promise or empty talk messages in the whole corpus of messages. The model results hold qualitatively if corresponding variables representing absolute numbers are included.

Table B2a: FE Regression Results C&D, All Messages Included

	Models Estimating the Influence of an Increasing Number of Promise Messages				
Limitations on J	--	--	--	J<31	J>30
	(a_p)	**(b_{p1})**	**(b_{p2})**	**(d_p)**	**(e_p)**
#Messages_J_Rel	0.9330*** (0.0222)	FE	--	FE	FE
#EmptyTalk_I_Rel	FE	--	FE	--	--
#Promise_K_Rel	--	0.0758*** (0.0232)	0.5529*** (0.0131)	0.4189*** (0.0353)	-0.4801*** (0.0169)
Constant	0.3357*** (0.0082)	0.6395*** (0.0086)	0.4712*** (0.0052)	0.4842*** (0.0092)	1.0174*** (0.0087)
Obs.	3415	3415	3415	1930	1485
R^2	0.3714	0.2759	0.2759	0.1712	0.0238

Standard errors are reported in parentheses. * p<0.1, **p<0.05, ***p<0.01. *Kappa* as dependent variable (i.e. Kappa-score overall messages, while irrelevant messages are coded as empty talk). FE: the fixed effects are estimated for that specific variable.

Table B2b: FE Regression Results C&D, All Messages Included

	Models Estimating the Influence of an Increasing Number of Empty Talk Messages				
Limitations on J	--	--	--	J<31	J>30
	(a_e)	**(b_{e1})**	**(b_{e2})**	**(d_e)**	**(e_e)**
#Messages_J_Rel	0.8865*** (0.0223)	FE	--	FE	FE
#EmptyTalk_I_Rel	--	-0.0521*** (0.0160)	0.361*** (0.0091)	-0.288*** (0.0243)	0.3301*** (0.0117)
#Promise_K_Rel	FE	--	FE	--	--
Constant	0.3521*** (0.0081)	0.685*** (0.0062)	0.5373*** (0.0037)	0.6672*** (0.0082)	0.6299*** (0.0054)
Obs.	3415	3415	3415	1930	1485
R^2	0.3714	0.0768	0.0768	0.0107	0.4795

Standard errors are reported in parentheses. * p<0.1, **p<0.05, ***p<0.01. *Kappa* as dependent variable (i.e. Kappa-score overall messages, while irrelevant messages are coded as empty talk). FE: the fixed effects are estimated for that specific variable.

The models indicate that…

(a) …overall training set sizes, Kappa increases when the number of promise messages increases, given a fixed number of empty-talk messages (Models a_p and b_{p2}),
(b) …overall training set sizes, Kappa increases when the number of empty talk messages increases, given a certain number of promise messages (Models a_e and b_{e2}),
(c) …overall training set sizes, given a certain training-set-size, substituting empty talk messages by promise-messages yields higher Kappa scores than the other way around (Models b_{p1} and b_{e1}),
(d) …with small training set sizes, result (c) holds (Model d_p and d_e), yet, with training set sizes >30, substituting empty-talk-messages by promise-messages yields lower Kappa scores than the other way around (Model e_p and e_e). The results hold when Kappa is calculated from test data only (see Table B3a and Table B3b).

Table B3a: FE Regression Results C&D on Test Data Only

	Models Estimating the Influence of an Increasing Number of Promise Messages				
Limitations on J	--	--	--	J<31	J>30
	(a_p)	**(b_{p1})**	**(b_{p2})**	**(d_p)**	**(e_p)**
#Messages_J_Rel	0.6221*** (0.0235)	FE	--	FE	FE
#EmptyTalk_I_Rel	FE	--	FE	--	--
#Promise_K_Rel	--	0.1855*** (0.0247)	0.3686*** (0.0139)	0.5440*** (0.0346)	-0.3953*** (0.0277)
Constant	0.3050*** (0.0087)	0.4599*** (0.0091)	0.3954*** (0.0056)	0.3481*** (0.0090)	0.7893*** (0.0143)
Obs.	3415	3415	3415	1930	1485
R^2	0.1475	0.1480	0.1480	0.1509	0.0244

Standard errors are reported in parentheses. * p<0.1, **p<0.05, ***p<0.01. *Kappa_Test* as dependent variable (i.e. training messages are excluded and irrelevant messages are coded as empty talk). FE: The fixed effects are estimated for that specific variable.

Table B3b: FE Regression Results C&D on Test Data Only

	Models Estimating the Influence of an Increasing Number of Empty Talk Messages				
Limitations on J	--	--	--	J<31	J>30
	(a_e)	**(b_{e1})**	**(b_{e2})**	**(d_e)**	**(e_e)**
#Messages_J_Rel	0.3899*** (0.0269)	FE	--	FE	FE
#EmptyTalk_I_Rel	--	-0.1275*** (0.0170)	0.1589*** (0.0110)	-0.374*** (0.0238)	0.2717*** (0.0190)
#Promise_K_Rel	FE	--	FE	--	--
Constant	0.3872*** (0.0097)	0.5708*** (0.0066)	0.4687*** (0.0044)	0.5857*** (0.0080)	0.4703*** (0.0088)
Obs.	3415	3415	3415	1930	1485
R^2	0.1475	0.0094	0.0094	0.0586	0.1711

Standard errors are reported in parentheses. * p<0.1, **p<0.05, ***p<0.01. *Kappa_Test* as dependent variable (i.e. training messages are excluded and irrelevant messages are coded as empty talk). FE: The fixed effects are estimated for that specific variable.

Table B4a: FE Regression Results Deck et al. All Messages Included

	Models Estimating the Influence of an Increasing Number of Promise Messages				
Limitations on J	--	--	--	J<31	J>30
	(a_p)	**(b_{p1})**	**(b_{p2})**	**(d_p)**	**(e_p)**
#Messages_J_Rel	1.0872*** (0.0291)	FE	--	FE	FE
#EmptyTalk_I_Rel	FE	--	FE	--	--
#Promise_K_Rel	--	0.0169 (0.0301)	0.6443*** (0.0172)	0.4626*** (0.0429)	-0.7053*** (0.0315)
Constant	0.2053*** (0.0108)	0.5845*** (0.0111)	0.3633*** (0.00689)	0.3785*** (0.0112)	1.0811*** (0.0163)
Obs.	3415	3415	3415	1930	1485
R^2	0.3433	0.2342	0.2342	0.1465	0.0321

Standard errors are reported in parentheses. * p<0.1, **p<0.05, ***p<0.01. *Kappa* as dependent variable (i.e. Kappa-score overall messages, while irrelevant messages are coded as empty talk). FE: the fixed effects are estimated for that specific variable.

Table B4b: FE Regression Results Deck et al. All Messages Included

	Models Estimating the Influence of an Increasing Number of Empty Talk Messages				
Limitations on J	--	--	--	J<31	J>30
	(a_e)	**(b_{e1})**	**(b_{e2})**	**(d_e)**	**(e_e)**
#Messages_J_Rel	1.1362*** (0.0355)	FE	--	FE	FE
#EmptyTalk_I_Rel	--	-0.0116 (0.0207)	0.4629*** (0.0145)	-0.318*** (0.0295)	0.4849*** (0.0217)
#Promise_K_Rel	FE	--	FE	--	--
Constant	0.1880*** (0.0129)	0.5946*** (0.0081)	0.4254*** (0.0058)	0.5806*** (0.0099)	0.5119*** (0.0100)
Obs.	3415	3415	3415	1930	1485
R^2	0.3433	0.0882	0.0882	0.0076	0.3508

Standard errors are reported in parentheses. * p<0.1, **p<0.05, ***p<0.01. *Kappa* as dependent variable (i.e. Kappa-score overall messages, while irrelevant messages are coded as empty talk). FE: the fixed effects are estimated for that specific variable.

The models indicate that…

(a) …overall training set sizes, Kappa increases when the number of promise messages increases, given a fixed number of empty-talk messages (Models a_p and b_{p2}),

(b) …overall training set sizes, Kappa increases when the number of empty talk messages increases, given a certain number of promise messages (Models a_e and b_{e2}),

(c) …overall training set sizes, Kappa neither significantly increases when empty talk messages are substituted by promise messages, nor when messages are substituted the other way around (Models a_e and b_{e2}),

(d) …with small training set sizes, given a certain training set size, substituting empty talk messages with promise messages yields higher Kappa-scores than the other way around (Model d_p and d_e), yet, with training set sizes >30, substituting empty talk messages by promise messages yields lower Kappa-scores than the other way around (Model e_p and e_e). That is, most of the qualitative results derived from FE models on C&D messages also hold when considering Deck et al. messages. The only difference is that over the whole range of training sets, there is no clear preference for including either more empty talk or more promise messages, given a certain size of the training set. Yet, there is a clear preference for either of the two categories when considering different ranges of training set sizes.

Appendix C: Additional Results on MLP and Random Forest Classifier: Table C1 shows the mean Kappa score (over five datasets) on the MLP Classifier for the whole message corpus. I employed the same five training data sets for each stratification of the training data corpus of the categories 'empty talk' and 'promise' as in the main section of this paper while employing IBM Watson Conversation.

Table C1: Mean Kappa-Scores on Whole Corpus C&D Using MLP Classifier

No. messages promise in training data set \ No. messages empty talk in the training data set	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
					0.4	0.3	0.4	0.3	0.3	0.3	0.3	0.3	0.2	0.5	0.3	0.3	0.2	0.3	0.2	0.1	0.2
2	-	-	-	-	4	9	5	4	6	5	6	9	8	1	7	5	6	6	7	6	1
				0.5	0.6	0.5	0.4	0.4	0.4	0.5	0.5	0.4	0.3	0.3	0.3	0.4	0.2	0.3	0.2	0.3	0.2
3	-	-	-	8	1	6	8	8	8	4	2	4	0	6	5	6	6	3	9	9	9
			0.5	0.4	0.5	0.6	0.5	0.4	0.5	0.5	0.6	0.6	0.4	0.3	0.5	0.5	0.3	0.4	0.4	0.5	0.4
4	-	-	8	7	8	4	7	9	5	1	3	4	8	9	6	1	7	7	2	2	9
		0.4	0.4	0.4	0.6	0.5	0.5	0.6	0.5	0.5	0.4	0.6	0.5	0.5	0.5	0.6	0.6	0.5	0.6	0.4	0.4
5	-	1	4	9	5	6	9	0	8	1	6	4	5	7	5	5	0	2	2	7	9
	0.4	0.4	0.5	0.6	0.6	0.6	0.6	0.5	0.6	0.6	0.4	0.6	0.5	0.5	0.4	0.5	0.6	0.6	0.6	0.4	0.6
6	0	1	7	1	5	9	6	7	3	6	4	7	7	5	7	2	0	4	1	8	2
	0.2	0.4	0.4	0.6	0.6	0.5	0.6	0.5	0.6	0.6	0.6	0.7	0.6	0.6	0.5	0.6	0.5	0.6	0.6	0.5	0.6
7	6	9	4	6	5	9	4	7	1	5	6	2	8	3	6	6	9	5	6	1	7
	0.1	0.4	0.6	0.5	0.7	0.6	0.6	0.7	0.6	0.6	0.6	0.7	0.7	0.6	0.7	0.6	0.6	0.6	0.6	0.6	0.6
8	0	9	4	0	1	5	8	1	7	7	6	1	1	1	2	5	3	4	0	0	1
	0.2	0.5	0.5	0.6	0.6	0.5	0.6	0.6	0.7	0.6	0.7	0.7	0.6	0.7	0.6	0.7	0.6	0.6	0.5	0.7	0.6
9	4	6	0	7	6	4	8	9	7	8	1	2	7	3	4	1	8	9	9	3	7
1	0.1	0.6	0.4	0.5	0.6	0.6	0.6	0.6	0.6	0.7	0.6	0.6	0.6	0.7	0.6	0.6	0.7	0.7	0.7	0.6	0.6
0	7	1	9	4	3	8	9	9	7	6	3	7	9	7	5	5	1	3	1	4	8
1	0.3	0.5	0.5	0.5	0.6	0.7	0.6	0.6	0.7	0.7	0.7	0.6	0.6	0.6	0.7	0.6	0.7	0.7	0.7	0.6	0.7
1	9	3	5	2	7	0	2	6	2	7	2	7	7	5	7	9	2	8	5	8	5
1	0.2	0.3	0.4	0.6	0.5	0.6	0.6	0.6	0.6	0.7	0.7	0.6	0.7	0.5	0.7	0.7	0.6	0.6	0.7	0.7	0.6
2	9	4	4	2	5	6	8	7	9	2	1	5	8	9	5	6	8	9	0	6	5
1	0.1	0.4	0.3	0.5	0.6	0.5	0.7	0.7	0.7	0.7	0.6	0.7	0.7	0.7	0.7	0.7	0.7	0.6	0.7	0.7	0.7
3	9	3	4	1	1	0	0	3	3	5	8	9	3	8	6	3	6	9	6	5	3
1	0.2	0.5	0.5	0.5	0.5	0.6	0.6	0.7	0.6	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7
4	7	4	3	7	3	7	8	4	7	2	4	5	5	2	3	7	3	8	2	6	3
1	0.1	0.1	0.3	0.6	0.6	0.5	0.6	0.7	0.7	0.7	0.7	0.8	0.7	0.7	0.8	0.7	0.7	0.7	0.7	0.7	0.7
5	1	5	7	5	6	9	7	4	1	0	6	0	4	3	0	9	7	9	7	8	8
1	0.3	0.3	0.2	0.6	0.5	0.6	0.6	0.6	0.7	0.6	0.6	0.5	0.7	0.6	0.7	0.7	0.8	0.7	0.7	0.8	0.8
6	3	3	8	4	5	5	9	6	1	9	1	6	8	8	8	5	0	2	8	1	1
1	0.1	0.3	0.5	0.6	0.5	0.7	0.7	0.6	0.5	0.7	0.7	0.7	0.7	0.8	0.7	0.7	0.8	0.7	0.7	0.7	0.7
7	2	0	8	3	5	6	2	1	5	2	8	5	4	3	9	9	1	8	8	8	9
1	0.1	0.4	0.5	0.5	0.5	0.5	0.6	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.8	0.8	0.8	0.8
8	8	2	1	6	5	9	7	6	0	4	6	9	4	2	5	4	9	0	0	5	0
1	0.2	0.2	0.3	0.6	0.5	0.6	0.7	0.7	0.6	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.8	0.7	0.7	0.8	0.8
9	6	7	7	8	7	6	2	1	5	7	8	6	3	9	8	9	1	9	7	3	1
2	0.1	0.3	0.5	0.4	0.6	0.6	0.6	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.8	0.8	0.7	0.7	0.8	0.8
0	0	7	4	7	8	0	8	3	4	4	9	9	0	9	9	1	1	8	5	0	2
2	0.3	0.3	0.5	0.5	0.5	0.6	0.6	0.7	0.7	0.7	0.6	0.7	0.7	0.8	0.7	0.8	0.7	0.8	0.8	0.8	0.8
1	2	5	6	6	8	9	8	2	7	1	7	6	9	1	6	1	8	3	4	4	0
2	0.1	0.3	0.4	0.5	0.5	0.6	0.7	0.7	0.7	0.7	0.6	0.8	0.7	0.7	0.8	0.8	0.8	0.7	0.8	0.8	0.6
2	5	2	4	6	4	4	2	6	1	4	8	1	1	7	0	2	0	9	0	0	5
2	0.1	0.3	0.4	0.5	0.6	0.5	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.8	0.7	0.8	0.8	0.8	0.7	0.8
3	9	7	1	7	3	7	2	0	3	1	9	9	6	8	3	9	3	0	2	9	4
2	0.2	0.3	0.4	0.5	0.5	0.6	0.5	0.6	0.7	0.7	0.6	0.7	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
4	6	9	6	8	4	3	7	1	5	3	8	5	8	0	0	1	0	0	0	7	1

2	0.2	0.3	0.5	0.5	0.5	0.6	0.6	0.7	0.7	0.7	0.7	0.7	0.8	0.8	0.7	0.8	0.8	0.8	0.7	0.8	0.7
5	8	0	2	2	5	5	5	1	2	0	3	0	0	3	9	0	1	2	9	1	9
2	0.3	0.2	0.4	0.6	0.6	0.5	0.6	0.6	0.7	0.7	0.7	0.7	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
6	0	5	2	3	4	8	5	1	2	5	3	7	8	3	4	2	2	1	5	5	1
2	0.1	0.4	0.5	0.5	0.5	0.5	0.6	0.6	0.7	0.7	0.7	0.7	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
7	7	2	0	4	4	0	2	3	9	1	2	5	5	1	6	0	1	3	4	7	1
2	0.1	0.2	0.4	0.5	0.4	0.6	0.6	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8
8	2	7	3	3	9	2	7	5	4	2	6	6	7	7	3	6	3	0	4	6	3
2	0.1	0.3	0.4	0.4	0.5	0.5	0.6	0.7	0.7	0.7	0.8	0.7	0.7	0.7	0.8	0.7	0.8	0.8	0.8	0.8	0.7
9	1	6	0	8	8	5	1	0	2	0	1	9	6	2	2	8	0	0	5	7	8
3	0.1	0.3	0.3	0.5	0.6	0.6	0.6	0.6	0.7	0.5	0.7	0.7	0.8	0.7	0.8	0.8	0.7	0.7	0.8	0.8	0.8
0	6	0	9	8	0	1	2	4	1	7	6	3	1	9	1	1	7	9	4	9	7
3	0.2	0.4	0.4	0.5	0.4	0.6	0.6	0.6	0.6	0.7	0.6	0.6	0.7	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8
1	0	1	7	0	5	9	5	7	9	6	8	1	8	4	4	5	3	4	1	7	8
3	0.1	0.2	0.4	0.5	0.5	0.6	0.6	0.7	0.7	0.6	0.8	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
2	7	6	1	2	7	5	2	3	1	9	0	3	3	1	1	1	2	7	7	6	6

Table C2 shows the mean Kappa score (over five datasets) on the MLP Classifier for the training message corpus. I employed the same five training data sets for each stratification of the training data corpus of the categories 'empty talk' and 'promise' as in the main section of this paper while employing IBM Watson Conversation.

Table C2: Mean Kappa-Scores on Training Data C&D Using MLP Classifier

No. messages promise in training data set	No. messages empty talk in the training data set																				
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
					1.0	0.8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.8	1.0
2	-	-	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
				1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	0.8	0.9	1.0	1.0	1.0	1.0	1.0	1.0
3	-	-	-	0	0	0	0	0	6	0	0	0	0	0	3	0	0	0	0	0	0
			1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
4	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.8	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
5	-	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0
	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.8	1.0	1.0	1.0	0.8	1.0	0.9	0.9	0.9	1.0	1.0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	8	6	0	0
	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.8	1.0
7	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0
	0.8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0
1	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.8	1.0	1.0	0.9	1.0	1.0	0.9	1.0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	9	0
1	1.0	1.0	0.8	1.0	1.0	0.8	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
3	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1.0	0.6	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
5	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1.0	1.0	0.8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.8	0.8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0

6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.8	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
7	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0	0.9	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0
8	0	0	0	0	0	0	0	0	8	0	0	0	7	0	9	0	0	0	0	0	0
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	0.9
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	9
2	0.7	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
0	3	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	0.9
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	9
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.8
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0	0.9	1.0	1.0	0.9	1.0	1.0	1.0
3	0	0	0	0	0	0	0	0	0	0	7	0	0	0	9	0	0	9	0	0	0
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9
4	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	9
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0
8	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	9	0	0	0	0	0
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0
9	0	0	0	0	0	0	0	9	0	0	0	0	0	0	9	0	0	0	0	0	0
3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0
3	0.9	1.0	1.0	1.0	0.8	1.0	1.0	1.0	1.0	0.9	1.0	0.8	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	0.9
1	6	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	8	0	0	9
3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0

Table C3 shows the mean Kappa score (over five datasets) on the Random Forest Classifier for the whole message corpus. I employed the same five training data sets for each stratification of the training data corpus of the categories 'empty talk' and 'promise' as in the main section of this paper while employing IBM Watson Conversation.

Table C3: Mean Kappa-Scores on Whole Corpus C&D Using Random Forest Classifier

No. messages promise in training data set	No. messages empty talk in the training data set																				
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
					0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.1	0.1	0.0	0.0	0.0
2	-	-	-	-	0	5	5	5	8	5	6	1	3	3	5	8	1	0	7	4	4
				0.4	0.2	0.4	0.1	0.2	0.2	0.1	0.1	0.2	0.0	0.1	0.1	0.2	0.1	0.1	0.1	0.1	0.0
3	-	-	-	4	9	3	5	5	2	6	6	5	9	7	5	0	2	3	5	7	9
			0.4	0.4	0.4	0.3	0.4	0.2	0.2	0.3	0.3	0.2	0.3	0.2	0.2	0.3	0.1	0.1	0.1	0.1	0.2
4	-	-	9	7	6	5	9	5	7	0	2	6	2	8	0	4	2	3	5	6	2
		0.4	0.5	0.6	0.7	0.6	0.4	0.3	0.4	0.4	0.4	0.4	0.4	0.1	0.2	0.2	0.2	0.3	0.1	0.2	0.2
5	-	0	8	4	5	2	0	4	7	3	1	7	0	3	9	6	2	2	7	4	5
	0.2	0.4	0.5	0.6	0.7	0.5	0.6	0.5	0.5	0.6	0.5	0.2	0.4	0.3	0.3	0.2	0.3	0.3	0.1	0.3	0.4
6	1	7	8	5	5	9	2	3	7	2	1	9	2	2	5	6	4	8	8	8	7
	0.2	0.4	0.4	0.7	0.5	0.7	0.6	0.6	0.4	0.6	0.6	0.6	0.5	0.5	0.3	0.5	0.3	0.3	0.4	0.4	0.4
7	0	9	9	1	8	4	7	2	6	1	8	8	0	0	6	9	9	9	4	2	7

	0.1	0.4	0.3	0.6	0.7	0.7	0.7	0.7	0.7	0.4	0.6	0.7	0.5	0.5	0.6	0.5	0.5	0.6	0.5	0.4	0.4
8	9	3	1	9	5	4	6	2	2	8	7	3	8	0	5	9	4	0	3	3	3
	0.2	0.1	0.6	0.6	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.5	0.5	0.6	0.5	0.5	0.4
9	4	3	7	2	2	5	6	6	3	0	4	4	9	3	5	7	6	3	9	7	8
1	0.0	0.2	0.2	0.5	0.7	0.8	0.7	0.7	0.7	0.8	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.7	0.6	0.6
0	8	7	0	9	2	0	3	8	7	1	4	4	1	5	9	5	6	9	3	1	3
1	0.0	0.3	0.3	0.5	0.7	0.7	0.7	0.7	0.8	0.8	0.7	0.6	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6
1	8	2	9	9	0	5	8	4	2	3	5	8	8	2	8	9	4	8	3	6	9
1	0.1	0.2	0.3	0.6	0.6	0.6	0.7	0.7	0.7	0.7	0.8	0.7	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.7	0.6
2	1	5	6	4	1	6	7	7	6	9	0	2	2	7	5	8	0	6	9	2	2
1	0.0	0.1	0.4	0.5	0.6	0.7	0.7	0.8	0.7	0.8	0.7	0.8	0.8	0.8	0.7	0.7	0.7	0.7	0.7	0.7	0.7
3	9	3	2	7	5	5	4	0	9	2	8	0	0	3	5	7	8	8	6	0	0
1	0.0	0.2	0.3	0.5	0.6	0.7	0.7	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.7	0.7	0.7	0.8	0.7	0.7	0.7
4	7	2	6	8	2	3	4	7	1	1	0	2	4	0	8	8	6	0	2	7	5
1	0.2	0.1	0.3	0.3	0.7	0.7	0.7	0.7	0.8	0.8	0.8	0.8	0.7	0.7	0.8	0.8	0.8	0.8	0.7	0.8	0.8
5	8	3	3	5	0	4	6	8	3	1	1	5	9	9	1	6	3	4	7	1	0
1	0.0	0.2	0.3	0.3	0.5	0.7	0.7	0.7	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.7	0.8	0.7	0.8	0.8
6	8	0	5	9	6	7	6	7	8	2	1	3	7	1	6	2	5	0	8	6	7
1	0.0	0.1	0.4	0.4	0.6	0.7	0.7	0.7	0.8	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
7	9	3	1	2	1	1	9	5	2	8	0	1	3	4	5	4	6	4	4	0	4
1	0.0	0.3	0.2	0.4	0.4	0.6	0.8	0.8	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
8	8	7	5	1	7	8	0	0	7	2	2	4	2	3	7	3	4	5	4	5	6
1	0.0	0.1	0.2	0.4	0.5	0.6	0.7	0.7	0.7	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
9	7	1	7	7	4	7	5	7	7	9	5	3	4	6	7	4	5	6	4	4	5
2	0.0	0.3	0.3	0.6	0.6	0.5	0.7	0.7	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
0	7	0	1	2	3	5	4	3	9	1	4	2	1	5	4	4	4	4	6	6	6
2	0.0	0.2	0.3	0.3	0.3	0.6	0.7	0.7	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.9	0.8	0.8
1	9	3	1	4	7	3	9	3	7	1	3	0	6	4	5	5	5	7	0	8	6
2	0.0	0.1	0.1	0.4	0.5	0.6	0.7	0.7	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
2	8	3	9	8	8	6	5	5	9	2	0	5	3	4	7	5	9	7	8	8	9
2	0.0	0.1	0.2	0.3	0.5	0.7	0.7	0.6	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.9	0.8	0.9
3	7	6	4	4	5	2	3	9	6	0	4	7	5	5	7	5	9	6	0	7	0
2	0.1	0.2	0.1	0.4	0.4	0.6	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.9	0.8
4	6	4	4	5	3	2	2	1	0	1	4	3	4	6	6	8	9	7	7	1	8
2	0.0	0.1	0.2	0.2	0.4	0.6	0.6	0.7	0.7	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.9	0.9	0.9	0.8
5	9	1	1	6	6	4	7	1	7	8	2	2	4	6	9	8	6	0	2	0	6
2	0.1	0.1	0.1	0.3	0.6	0.5	0.6	0.7	0.7	0.7	0.8	0.8	0.8	0.8	0.8	0.9	0.8	0.8	0.9	0.8	0.9
6	6	7	5	3	4	1	7	7	9	2	1	5	5	6	8	1	9	8	1	9	0
2	0.0	0.1	0.1	0.2	0.4	0.5	0.7	0.6	0.7	0.8	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.9	0.8	0.8	0.8
7	8	5	6	8	3	9	7	4	8	0	8	3	3	6	9	5	7	3	9	8	6
2	0.1	0.1	0.2	0.4	0.5	0.6	0.7	0.7	0.7	0.8	0.7	0.8	0.8	0.8	0.9	0.9	0.8	0.8	0.9	0.9	0.8
8	0	1	1	1	2	5	0	7	6	1	4	0	6	5	0	0	6	8	3	3	9
2	0.0	0.1	0.1	0.6	0.3	0.5	0.6	0.7	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.9	0.8
9	7	3	5	0	7	5	1	0	9	6	3	7	7	0	6	5	7	6	8	2	8
3	0.0	0.1	0.1	0.2	0.6	0.6	0.6	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.9	0.9	0.9	0.9
0	7	7	6	2	1	7	0	2	0	1	2	4	5	5	5	7	5	0	1	3	0
3	0.0	0.1	0.2	0.4	0.5	0.6	0.7	0.7	0.7	0.7	0.8	0.8	0.8	0.8	0.8	0.9	0.9	0.8	0.8	0.9	0.9
1	8	8	5	1	0	0	5	0	9	9	0	3	8	3	6	0	0	9	9	2	1
3	0.0	0.1	0.1	0.4	0.4	0.6	0.6	0.6	0.7	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.9	0.9	0.9
2	9	3	6	8	2	4	1	5	8	7	4	5	6	8	9	7	9	8	0	1	0

Table C4 shows the mean Kappa score (over five datasets) on the Random Forest Classifier for the training message corpus. I employed the same five training data sets for each stratification of the training data corpus of the categories 'empty talk' and 'promise' as in the main section of this paper while employing IBM Watson Conversation.

Table C4: Mean Kappa-Scores on Training Data C&D Using Random Forest Classifier

No. messages empty talk in the training data set

No. messages promise in training data set	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
					1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	-	-	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
				1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
3	-	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
			1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
4	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
5	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Journal of Economics and Behavioral Studies (JEBS)
Vol. 13, No. 1, February 2021 (ISSN 2220-6140)

Assessing Funding Mechanism Available for Mining Companies in Zimbabwe

Nyasha Kaseke, Gift Mapakame
University of Zimbabwe Business School Student, Freda Rebecca Mine, Zimbabwe
nykaseke@gmail.com, gmapakame@fredarebecca.co.zw

Abstract: The mining sector has been the cornerstone of economic growth in Zimbabwe hence funding becomes very crucial to resuscitate the economy. The study aimed to assess the funding mechanisms for the mining sector of Zimbabwe and the effect of these on the performance of the sector. A quantitative study was carried out in the mining sector. The research findings showed that respondents pointed out that the funding mechanisms used in the mining sector of Zimbabwe are project finance, finance by private equity, public bonds and loans from banks and other financial institutions. It was also revealed that over and above available mechanisms, investment in the mining sector is being influenced by Interest rate, Business economic empowerment policies, bank lending criteria and Technical information, simultaneously. Furthermore, the study established that the mining sector needs skilled and technical staff, Technical information, banks' lending criteria and Capital markets to get funding from investors. It was derived that, investment in the mining sector will increase production, product quality and profitability which in turn lead to infrastructural development. In addition, it is envisaged that funding will result in mining exports increase at the same time that new technologies are being introduced and the GDP is rising. Owing to the focus being exclusively on the funding of the mining sector, the study also recommended further studies on other factors, besides funding, that are affecting the performance of the sector.

Keywords: *Funding mining Sector, Banks Lending Criteria, Capital Markets, Mining Sector Investment.*

1. Introduction

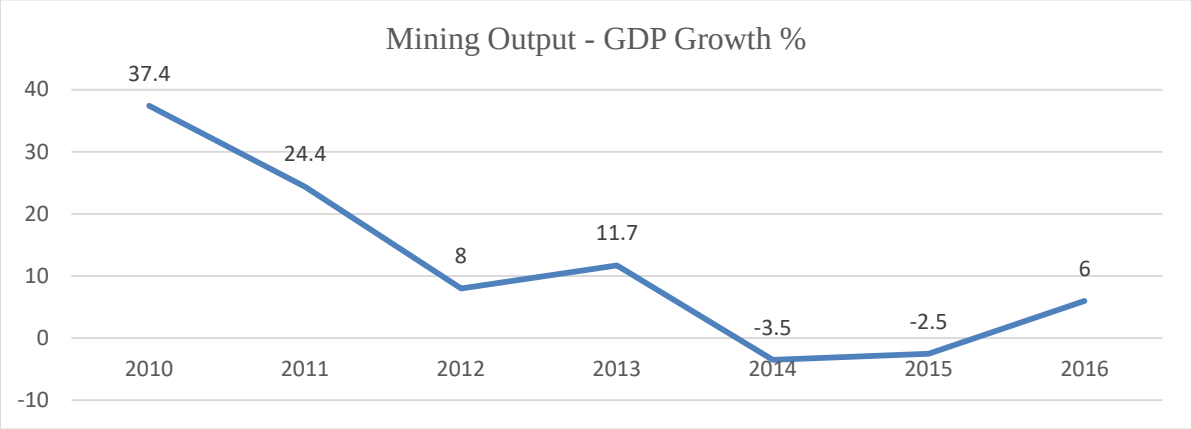
Minerals and minerals exploration and extractions have gained momentum over the past decade in Southern African. This saw investment by domestic, regional and international investors reaching high levels. The investment is mainly driven by booming commodity prices, high demand for commodities with China emerging as one of the largest consumers, and the need to revive economies, with some economies basing on natural mineral resources, for example, Zimbabwe targeting a mining economy of US$12billion by the year 2030. Governments have also supported the growth in investment through facilitating explorations, liberalizations of mining investment and privatization of state-owned mining companies. Zimbabwe with its vast mineral deposits, being a high ferrochrome producer, high concentration of platinum deposits, and one of the African gold high producers, have to increase production to achieve economic growth as targeted. These mineral deposits are supported by relatively good infrastructure, a large pool of skills and relatively favorable policies which seem to be advancing mineral extraction in Zimbabwe. However, in Zimbabwe, a lot of mining companies are failing to re-start operations or revamp key projects due to funding constraints that have impaired the sector and economy as a whole. The complete annihilation of output in this key sector has shown significant improvement. The mining sector is capital-intensive and is associated with high risk (both technical and economic).

A few resuscitated gold operations, improved platinum operations (new and existing mines) and diamond operations resulted in the country thriving in commodity markets. These characteristics present obstacles for many mining companies in terms of raising capital, especially for start-ups as there are a limited number of financiers who are willing to finance such Greenfield investments. Therefore, these challenges limit the sources of finance for mining investors in Zimbabwe and even in the region. In the past, the mining sector used to be financed through shareholder equity, debt financing, structured financing and syndicated loans but due to Zimbabwe's status as a struggling economy, characterized by limited Gross Domestic Product, budget deficits, negative current account, an underperforming banking sector and frail policy framework, the size of traditional investment in the mining sector has severely contracted (Zimbabwe Mining Development Company, 2015). There are conflicting expectations for the mining sector in Zimbabwe; on one hand, the mining sector is associated with high risk, limited production, low economic growth in the country, capital intensity of the sector and changing policies; and on the other hand, there is need to boost mining sector

production, extensively exploit the mineral resources for growth and development. To achieve the latter, there is a need to increase investment by both the private sector and government in mineral extraction.

Background: Since the stimulation of the Zimbabwean economy by introducing multi-currency trading & liberalization of minerals marketing, frontline investment poured into the mining sector boosting minerals output and mining gross domestic product contribution. The aim of the study is therefore to assess the forms of financing mechanisms for the mining sector to improve production. Gold, platinum, ferrochrome and diamond mining companies received extensive capital to either resuscitate or ramp up output leading to overall sector performance. However, the trend of growth of mining output reveals poor sector performance, declining tremendously from 37.4% in 2010 to as low as -3.5% in 2014 owing to a variety of factors (Figure 1 below). These include under-capacitation of small-scale mining, prolonged mine closures, inadequate funding to develop new mines, payment currency for minerals (for example gold) and lack of capital to fund mineral output beneficiation assets. Revenues have also declined with $1.86Bn reported in 2015 against $1.9Bn in 2014 (Muganyi, 2016). It is undisputed that the sectors' performance declined due to financing constraints thereby contradicting the thrust to catalyze economic growth as a whole.

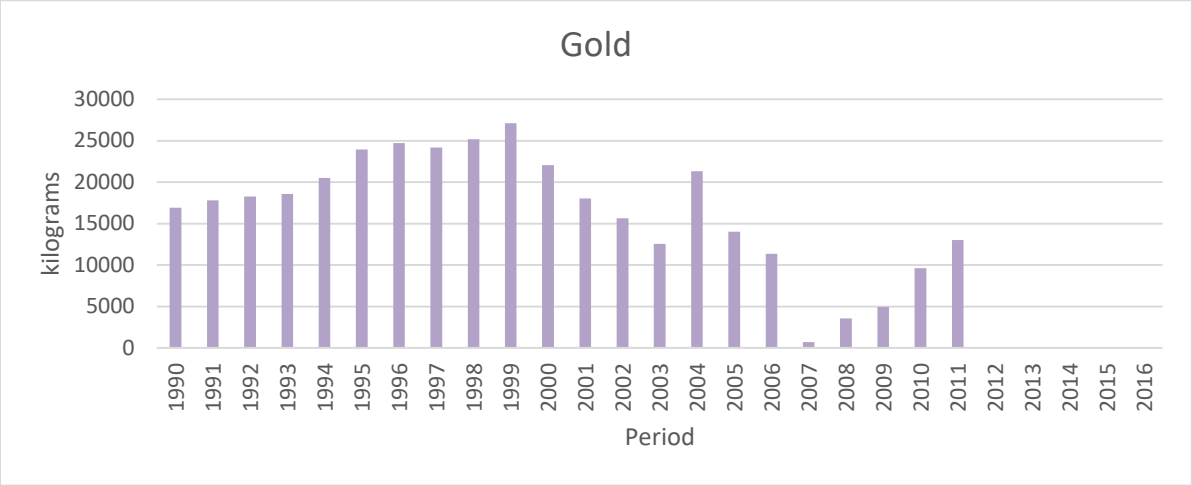
Figure 1: Mining GDP Growth



Source: MoF (2016)

Mine Closures and Idle Operations: The performance of the sector is critically inhibited by idle capacity due to capital shortages to bail out dormant assets. A snapshot of the mining sector output by the commodity of the pre-hyperinflation period and current period reveals a systematic gap that points out to a downward shift in operating entities. This signifies that it is critical to resuscitate existing mining operations to improve the performance of the sector. Figure 2 indicates how output in the key commodity facets has remained subdued even after the economy received stimulus from reforms that took place after hyperinflation in 2009. This infers that a substantial number of operations contributing to the sectors which recorded improved performance in the late 90s remained under care and maintenance or on shutdown due to failure to secure funding for resuscitation despite sitting on vast exploitable mineral reserves. As a result, the overall performance of the sector continues to perform below projections.

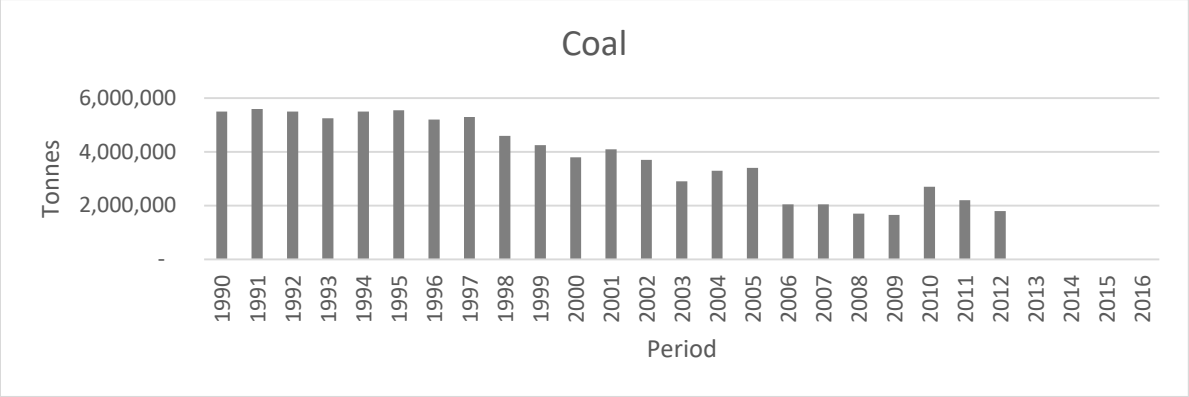
Figure 2: Gold Output Highlights



Source: Chamber of Mines (2015)

Gold output originates from both large-scale operators and rudimentary small-scale miners who have had no access to formal lines of credit to finance sustainable operations. Mining houses such as Metallon continue to lack expected contribution to output as they are running only 1 out of 4 large-scale operations.

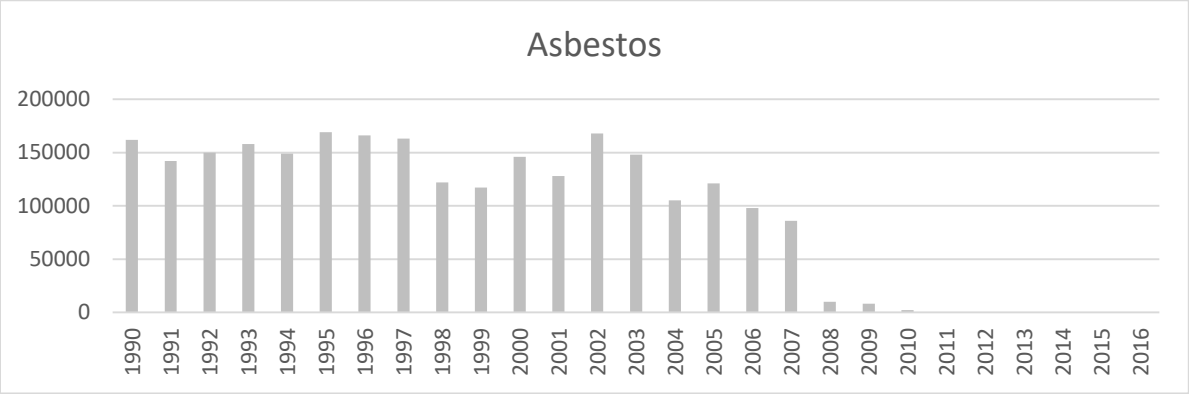
Figure 3: Coal Output Highlights (from the Zambezi basin)



Source: Chamber of Mines (2015)

State-owned enterprise Hwange Colliery is struggling to secure funding for the underground operations which generates 75% of the revenue through exporting high-grade coking coal (Chamber of Mines, 2015). As a result, the commodity has low contribution to export earning since the output is dominated by power coal mined and sold locally to Zimbabwe Power Company for power generation.

Figure 4: Asbestos Output from AA Mines



Source: Chamber of Mines (2015)

The once symbolic AA-Mines which exploited the vast fibrous deposits of Mashava for white asbestos have been reduced to shell infrastructure, with no reasonable prospects to obtain a suitable funding mechanism to re-capitalize owing to a prolonged shutdown which essentially depreciated all fixed assets and downgraded the investment grade to almost Greenfield.

Exploration and Mine Development: Though there is a demand to increase mineral output, explorations as the major starting point for mining have remained behind with reliance on past explorations. There was little investment in exploration over the last decade for Zimbabwe. The Department of Geological Survey failed to carry out its key role of geological mapping for the country over the past 10 years due to economic challenges (Chamber of Mines, 2015).

The Study Problem: The output from the mining sector is generally inconsistent with expectations substantiated by the reforms made by stakeholders in addressing policies and funding to capacitate the sector to achieve economic growth. To date, the performance of the mining sector remains subdued and exhibits a declining trend as several existing mines have failed to secure capital. Besides, crucial growth of the sector continues to spiral down commensurate with the lack of exploration spending and Greenfield project development. Ultimately, on the backdrop of an alarming funding deficit, miners are being driven out of the industry resulting in a poor performance outcome of the sector. Therefore, this research focuses on assessing funding mechanisms and options available to the mining sector in Zimbabwe.

2. Theoretical Review

This section presents literature on funding mechanisms, the factors affecting investment and its contribution to the performance of the mining sector.

Theories of Funding: Theories that explain funding are derived from the theories of investment (Nghifemwa, 2009). Three theories of investment highlighted the relationship between investment and industry performance. Agénor and Montiel (2009) note that funding of businesses is critical for any economy to grow. Little or no funding to businesses will result in little or no growth in the economy.

Marginal Efficiency Theory of Capital: This theory is based on the notion that investors look at the expected rate of return when considering investing in certain projects or institutions. In this case, the minimum expectation of the return is that rate that equates the cost of capital to the return on investment, simply put where the net present value is equal to or greater than zero (NPV ≥ 0). Thus, any return above the cost of capital will give a positive marginal return. The economic environment also matters when making investment decisions. Chidhakwa (2016) highlighted that Zimbabwean banks have not been up-coming in lending mining companies due to economic challenges. Therefore, marginal efficiency is the discount rate that equates the present value of the expected revenue from an investment in the capital with the current supply cost of capital goods (Keynes, 1936).

Tobin's Q-theory of Investment: The q-theory is based on the relationship between investment spending ratio and market value ratio determined by the additional unit of capital replacement cost (Tobin, 1969). This relationship will give a value often referred to as the q-ratio with a value of one (1) being the deciding level. A q-value above 1 means investment rises and below 1 means investment declines (Ferderer, 2009).

Accelerator Theory of Investment: This theory is based on the view that there is a relationship between investment and the performance of the firm or industry. The more the investment in the sector or industry, the more the growth. However, it is critical to note that there is the desired capital stock for a given level of output and interest rate. Beyond that, expected growth might not be realized. Also, an increase in demand for the product may increase demand for funding, as firms may be adjusting to meet the current demand.

Funding Mechanisms for the Mining Companies: Many funding mechanisms can impact the mining sector directly or indirectly. Vallerrie (2010) also added that there are several financing options available for mining projects. A study carried out by Leeman (2006) revealed that an effective funding mechanism can directly impact the performance of the mining sector. In addition, the Bankers Association of Zimbabwe (BAZ, 2014) highlighted that Zimbabwean mining company's need alternative funding mechanisms to perform in a depressed economy. Literature has shown various ways of financing mining companies. According to Mcmann (2010); Meyer (2014); Zhu (2011) and Vale (2012), there are various funding mechanisms which include; Project finance, Financing by IPO or Private Placement, Financing from Private Equity Funds, Royalty agreements, Escrow account arrangement, Streaming arrangements, Convertible loans, Public bonds, Loans from Banks and Other Financial Institutions, Joint venture partnerships, Off-take finance or contract financing, Corporate Bonds, Trusts, Equipment financing, Government Bonds, Export Credit Agencies and Financial Leasing. Meyer (2014) argued that the mining sector must be able to come up with suitable funding mechanisms.

Factors Affecting Funding: Eita and Du Toit (2007) found that the factors affecting funding are categorized according to three players in the mining sector; investor sphere, government sphere, and mine sphere.

Investor Sphere: As investment and funding largely emanate from different mechanisms all underpinned by investors looking for a return, it is prudent to note that, in the more standard Capital Asset Pricing Model (CAPM), an investment's riskiness is assessed against market return rather than consumption. Investors determine the country's risk. A country's risks in investment refer to the degree of uncertainty that exists about the occurrence of a future planned event. The capital budgeting decisions are also considered by the investor**.** Gawlik, (2008) and Pandey (2008) pointed out that when making capital budgeting decisions and investing in the mining projects, the following factors have to be considered: economic life of the mining projects; availability of the initial capital outlay; amount and timing of the cash flows; the need for additional capital requirements; the impact of the mining investment project on the entire firm; and how the initial capital outlay will be phased out.

Government Sphere: Chidhakwa (2016) affirmed that robust institutionalization of mining laws and competitive mining legislative framework is critical for local and international investors. These provide a conducive investment environment, simultaneously curtailing prejudice to national budget demands for investment funds. What investors want is a conducive environment and it can act as an incentive for investors even where there is no government support or incentives. The government must endeavor to promote efficient transactions by clearly stipulating, in black and white, what is acceptable and what is not. The government is a key player in motivating investment in the mining sector. Mothomogolo (2012) states that the government motivates investment through fast-tracking mining rights application process, providing investor conducive business economic empowerment policies, offering the security of tenure and transfer of title, refining royal tax and offering venture capital investments to allow the investor to claim funds.

Mine Operator Sphere: McMann (2010) indicated that the mine operator must ensure that there is technical information that is required to support funding available to mineral developers and financiers. The venture capital firms are also important to the investor through available companies or investors to enter into an agreement with. If partners are not willing to engage in contracts, then investors are affected negatively. Investors are also motivated by the banks' lending criteria. Skilled and technical staff must be availed. Mutiwa and Fondo (2015) argued that some projects need highly skilled and technical staff. In addition, there is a

need for a clear framework for managing labor. Labour issues are critical in determining profit or losses to investors. The Labor movement creates a new center of power and employees may listen more to the leaders of the labor movement than their employers thereby creating confusion for investors.

Mining Business Performance: The classification of business performance is very subjective. Business performance in the financial sense as being characterized by-product sales performance, profitability and return on capital employed (ROCE) (Selvarajan et al., 2007; Hsu et al., 2007). Besides, return on investment (ROI other measures of financial performance that can be used are Net Income after tax (NIAT) and earnings per share (EPS) (Grossman, 2012). Qualitatively, performance can be measured based on subjective performance measures using the perceived performance approach (PPA). This measures performance using the Likert-like scaling approach based on top management perspectives (Selvarajan et al., 2007). However, this is always criticized as it is based on individual opinions without specific financial information even though top management is privileged to the organisation's financial information.

3. Methodology

A descriptive research design was used for this study. The research used quantitative data which was drawn from a sample of the staff and management of the mining sector. A survey was carried out in the mining sector. This study expects the staff and management in the mining sector to respond to a questionnaire on the funding mechanism on mining performance. A self-administered structured questionnaire was used as a research instrument for this research. The population of this study consisted of all mining companies affiliated with the Chamber of Mines of Zimbabwe. Approximately, there are 33 mining companies affiliated with the Chamber of Mines of Zimbabwe. The sample selection for this study was guided by Leedy and Omron (2001).

4. Findings and Discussion

Findings from statistical analysis by SPSS reflect the Cronbach's Alpha of 0.820 which reflects the consistency of the questionnaire to be able to be used in researches measuring the same variable.

Appropriate Funding Mechanisms for the Zimbabwean Mining Sector

Table 1: Statistics on Funding Mechanisms

Funding Mechanism	Symbol	N	Minimum	Maximum	Mean	Std. Deviation
Project Finance	B1	20	2.00	5.00	4.3000	.97872
Financing by IPO	B2	19	1.00	5.00	2.6842	1.37649
Financing from private equity funds	B3	20	1.00	5.00	3.5000	1.31789
Royalty agreement	B4	20	1.00	5.00	2.7000	1.38031
ESCROW accounts	B5	19	1.00	4.00	1.6842	1.10818
Streaming arrangements	B6	18	1.00	5.00	2.5000	1.38267
Convertible loans	B7	20	1.00	5.00	2.9000	1.44732
Public bonds	B8	20	1.00	5.00	2.4000	1.35336
Loans from banks	B9	20	2.00	5.00	4.4500	.82558
Joint venture partnerships	B10	20	3.00	5.00	3.7000	.86450
Off-take finances	B11	20	1.00	5.00	3.1000	1.02084
Corporate bonds	B12	20	1.00	5.00	3.4500	1.19097
Trusts	B13	20	1.00	4.00	2.4500	.88704
Equipment financing	B14	20	3.00	5.00	4.3500	.74516
Export credit agencies	B15	20	1.00	4.00	2.6500	1.13671
Development finance institution	B16	20	1.00	5.00	3.4000	1.35336
Government bonds	B17	20	1.00	5.00	2.9500	1.50350
Financial leasing	B18	20	1.00	5.00	3.2000	1.28145

Table 1 revealed that whilst there was no least important funding mechanism, the less important funding mechanisms which were identified by this study were ESCROW accounts, Public bonds and Trusts. These were either uncommon or not preferred. On the other hand, the mechanisms which recorded high mean scores were Project Finance, Loans from banks and Equipment financing and are categorized as the more important funding mechanism. It is noted that none of the funding mechanisms involved qualified as most important implying that the sector adopts the funding mechanisms and complements each other without a single dominating mechanism.

Normality Tests: To ensure that the appropriate choice of data analysis techniques, normality tests were done to establish whether the distribution of the data availed suit to be analyzed by parametric or non-parametric analyses.

Table 2: Normality Tests – External Challenges

Funding Mechanisms	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	Df	Sig.
B1	.317	18	.000	.722	18	.000
B2	.222	18	.019	.860	18	.012
B3	.197	18	.064	.849	18	.008
B4	.206	18	.042	.889	18	.037
B5	.434	18	.000	.610	18	.000
B6	.197	18	.064	.855	18	.010
B7	.211	18	.034	.865	18	.015
B8	.270	18	.001	.832	18	.004
B9	.320	18	.000	.726	18	.000
B10	.342	18	.000	.728	18	.000
B11	.237	18	.009	.895	18	.048
B12	.239	18	.008	.909	18	.084
B13	.220	18	.021	.891	18	.040
B14	.276	18	.001	.788	18	.001
B15	.287	18	.000	.833	18	.005
B16	.237	18	.009	.875	18	.021
B17	.201	18	.054	.867	18	.016
B18	.289	18	.000	.858	18	.012

From the above analyses, all the mechanisms except B12 violate the normality assumption which parametric tests assume the sample under test is from a population with normal distribution. It follows then that non-parametric analyses were the best analyses to use for data of this nature. An assessment of the forms of funding mechanisms available for the mining sector using Factor Analysis revealed the following results.

Table 3: Total Variance Explained – External Factors

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.978	27.657	27.657	4.978	27.657	27.657
2	3.015	16.751	44.408	3.015	16.751	44.408
3	2.052	11.397	55.805	2.052	11.397	55.805
4	1.706	9.478	65.283	1.706	9.478	65.283
5	1.370	7.612	72.895	1.370	7.612	72.895
6	1.215	6.751	79.647	1.215	6.751	79.647
7	.939	5.217	84.864			
8	.850	4.720	89.583			
9	.620	3.444	93.028			
10	.421	2.340	95.368			
11	.329	1.829	97.197			
12	.236	1.312	98.510			
13	.154	.857	99.367			

14	.079	.440	99.806
15	.024	.136	99.942
16	.008	.043	99.985
17	.003	.015	100.000
18	-2.446E-016	-1.359E-015	100.000

Extraction Method: Principal Component Analysis.

From the foregoing, 6 components with eigenvalues greater than 1.0 were extracted, and all explained 79.647% of the total variation. The first component/factor had a variation contribution of 27.657%, while the second component had a variation contribution of 16.751%. The third component had a variance contribution of 11.397%, and the fourth component a contribution of 9.478%, the fifth had a variance of 7.612% and the sixth had a variance of 6.751%. The corresponding scree plot is presented below.

Figure 5: Scree Plot – External Factors

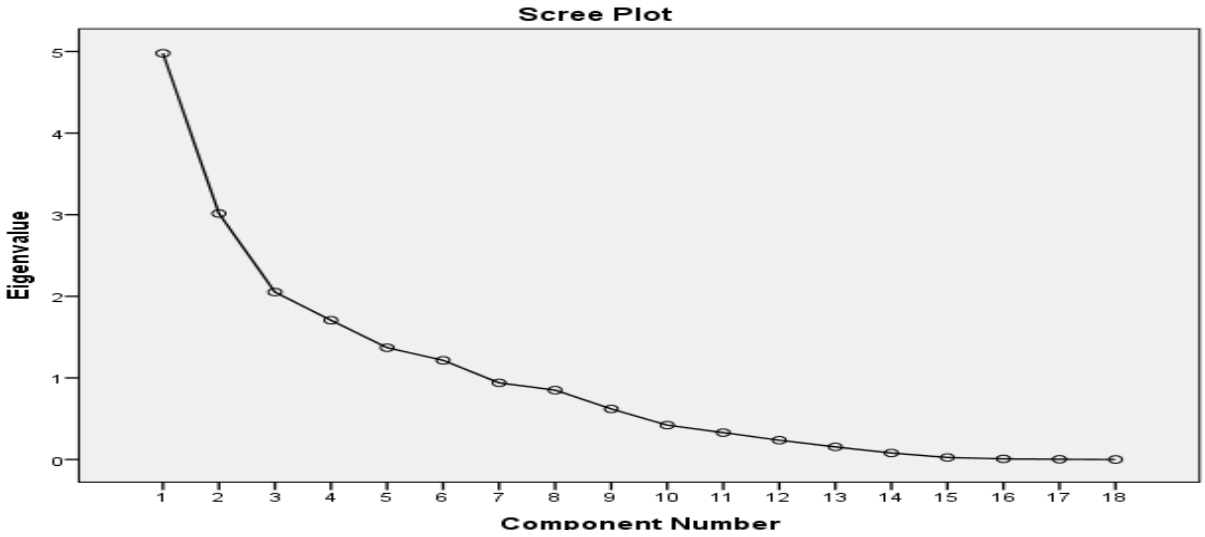


Table 4: Rotated Component Matrix

	Component					
	1	**2**	**3**	**4**	**5**	**6**
B2	.540					
B3	.785					
B5	.463					
B6	.499					
B7	.726					
B8	.750					
B12	.720					
B13	.470					
B15	.587					
B1		.828				
B10			.536			
B16				.645		
B17				.742		
B4	.				.673	

Extraction Method: Principal Component Analysis.
a. 6 components extracted.

From the analysis, it can be concluded that there were essentially 6 factors that were behind the current funding in the mining sector of Zimbabwe and these mechanisms are: Project finance; Financing from private equity funds; Convertible loans; Public bonds; Corporates bonds and Government bonds.

Analysis of Factors: Factors influencing investment in the mining sector.

Table 5: Factors Influencing Investment in the Mining Sector

Factors influencing Investment	Symbol	N	Minimum	Maximum	Mean	Std. Deviation
Politics	C1	20	4.00	5.00	4.9000	.30779
Interest rate	C2	20	3.00	5.00	4.2500	.71635
Royalty tax	C3	20	1.00	5.00	3.4000	1.18766
Business economic empowerment policies	C4	20	2.00	5.00	4.1500	.87509
Capital markets	C5	20	1.00	5.00	3.7000	1.08094
Venture capital incentives	C6	20	1.00	5.00	2.9000	1.11921
Bank lending criteria	C7	20	4.00	5.00	4.3000	.47016
Asset exit clauses	C8	20	1.00	4.00	3.1500	.98809
Venture capital incentive	C9	20	1.00	4.00	2.8500	1.08942
Technical information	C10	20	2.00	5.00	3.7500	1.16416
Skilled and technical staff	C11	20	2.00	5.00	3.4000	.88258
Labor issues	C12	20	1.00	5.00	3.3000	1.08094

The analysis from the table above was done using the key given below.

1	2	3	4	5
Not at all influential	Slightly influential	Somewhat influential	Very influential	Extremely influential

Table 5 shows that respondents agreed that the following factors are very influential to the investment in the mining sector; Interest rate, Business economic empowerment policies, bank lending criteria, Technical information whereas Politics was agreed to be extremely influential. The factors which were somewhat influential were Royalty tax, Venture capital incentives, Asset exit clauses, Venture capital incentive, Skilled and technical staff and Labor issues.

The Factors Influencing Funding Mechanisms in Mining: This section provides an analysis of the factors influencing funding of mining organizations and also how these factors affected anticipated funding either positively or negatively.

Table 6: Factors Influencing Funding Mechanisms

Factors influencing funding	N	Minimum	Maximum	Mean	Std. Deviation
C1A	6	10,000,000	35,000,000	23,000,000	8,717,797.97
C2A	3	4,000,000	10000000	8,000,000	3,464,101.62
C3A	1	30,000,000	30,000,000	30,000,000	.
C4A	3	15,000,000	30,000,000	24,000,000	7,937,253.93
C5A	1	30,000,000	30,000,000	30,000,000	.
C6A	0				
C7A	5	7,000,000	20,000,000	11,400,000	5,176,871.64
C8A	0				
C9A	0				
C10A	5	10,000,000	30,000,000	7,080,000	1,282,727.67
C11A	3	3,500,000	10,000,000	7,833,333	3,752,776.75
C12A	1	10,000,000	10,000,000	10,000,000	.

Table 6 shows that the factors influencing funding options in the funding options in the mining sector were C1A, C2A, C3A, C4A, C5A, C7A, C10A, C11A and C12A with each respective factor attributable to an amount of

23 million, 8 million, 30 million, 24 million, 30 million, 11.4 million, 7 million, 7.8million and 10 million of anticipated funding. The mid-tier range of 7 million – 30 million of anticipated funding also points out that the factors are generally revolving around funding meant for recapitalization and sustenance which is not excessively high when compared to new project development in the sector. Table 6 exposes that the factors with the highest frequency thus influence were found to be C1A, C7a and C10A namely, politics, bank lending criterion and technical information. The history of political unrest in Zimbabwe weighs in on the high level of country risk which impacts available funding.

Table 7: Positive Effect

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
C1PC	20	0	0	.00	.000	.	.	.	.
C2P	20	0	1	.15	.366	2.123	.512	2.776	.992
C3P	20	0	1	.10	.308	2.888	.512	7.037	.992
C4P	20	0	1	.15	.366	2.123	.512	2.776	.992
C5P	20	0	1	.20	.410	1.624	.512	.699	.992
C6P	20	0	1	.10	.308	2.888	.512	7.037	.992
C7P	20	0	1	.25	.444	1.251	.512	-.497	.992
C8P	20	0	1	.05	.224	4.472	.512	20.000	.992
C9P	20	0	1	.10	.308	2.888	.512	7.037	.992
C10P	20	0	1	.30	.470	.945	.512	-1.242	.992
C11P	20	0	1	.30	.470	.945	.512	-1.242	.992
C12P	20	0	1	.15	.366	2.123	.512	2.776	.992

Scale 1=Yes, 0=No

The table above shows that the factors with the highest scores were C11P (0.30), C10P (0.30), C7P (0.25) and C5P (0.20). The factors which recorded a positive impact were skilled and technical staff, Technical information, bank lending criteria and Capital markets.

Table 8: Negative Effect

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
C1N	20	0	1	.75	.444	-1.251	.512	-.497	.992
C2N	20	0	1	.45	.510	.218	.512	-2.183	.992
C3N	20	0	1	.15	.366	2.123	.512	2.776	.992
C4N	20	0	1	.35	.489	.681	.512	-1.719	.992
C5N	20	0	1	.10	.308	2.888	.512	7.037	.992
C6N	20	0	0	.00	.000	.	.	.	.
C7N	20	0	1	.40	.503	.442	.512	-2.018	.992
C8N	20	0	1	.15	.366	2.123	.512	2.776	.992
C9N	20	0	0	.00	.000	.	.	.	.
C10N	20	0	1	.30	.470	.945	.512	-1.242	.992
C11N	20	0	1	.05	.224	4.472	.512	20.000	.992
C12N	20	0	1	.05	.224	4.472	.512	20.000	.992
Valid (listwise)	N 20								

Scale 1=Yes, 0=No. The table above shows that the factors which recorded the highest negative effects were C1N (0.75), C2N (0.45), C4N (0.35), C7N (0.40) and C10N (0.30). According to the key, these factors are Politics, Interest rate, Business economic empowerment policies, bank lending criteria and Technical information respectively.

5. Conclusion and Recommendations

Conclusion: The study concludes that the funding mechanisms being used in the mining sector of Zimbabwe are project finance, finance by private equity, public bonds, corporate bonds and bank loans. The study concluded that funding will lead to an increase in mining production improvement in product quality, increase in profitability, Infrastructural development, and increase in mining exports, increase in new technologies and increase in GDP.

Recommendations

Royalty Agreements: It is recommended that the mining sector should do some Royalty agreements. This is financing based on the output produced. It is facilitated through upfront payment of funding required with repayment later based on units produced or percentage of products value or profits realized or revenues generated from the mining business. Royalty agreements are a very attractive source of finance, which offers benefits non-dilutive capital or shareholding, non-controlling on company share capital, and deferred repayment based on mining production and revenues. Royalty arrangements can be ideal for struggling mining companies and the findings from their study proved that royalty arrangements cause mining companies to improve their productivity.

Escrow Accounts: Escrow account arrangement can work in high-risk and high depressed economies. Operating an escrow account with an international guarantee may reduce the fear of high risk by investors or funders. This will provide the necessary working capital requirements for mining operations. This would work positively for Zimbabwe as the country is poorly rated in terms of risk as it removes the fear of financiers.

Streaming Arrangements: Streaming arrangements can be used by this sector. It involves financing through the supply of mineral outputs. Financiers would sign contracts to fund the mining operations with repayment based on the supply of minerals at an agreed price for a certain period. Streaming arrangements are effective funding mechanisms for mining companies in a depressed economy.

Joint Ventures: The sector can use a Joint venture which is a contractual agreement between partners for mutual benefits through sharing funding costs, risks or rewards. Mining companies should identify partners with a strong financial base to fund the operations during the time of need in return for an agreed share of revenues or profits for a specified period. Therefore, the agreements should specify the exit strategy to the joint venture to avoid disputes and conflicts.

Equipment Financing: The sector can also use Equipment financing. Assets can be used to raise finance in a mining company that is seeking bank loans based on critical assets available in the mining company. These assets act as security to the financiers.

References

African Development Bank. (2014). 2005-2009 Country strategy paper. Windhoek: Government of the Republic of Namibia.

Agénor, P. R. & Montiel, P. J. (2009). Development macroeconomics. Princeton, New Jersey: Princeton University Press.

Bankers Association of Zimbabwe. (2014). Accessing finance in a depressed economy. Bulawayo: Mine Entra 2014.

Chamber of Mines. (2015). State of the Mining Industry Survey Report, Harare, Zimbabwe.

Chamber of Mines. (2015). Proceedings of the 77th Chamber of Mines 2016 Annual General Meeting, Victoria Falls, May 2016

Chidhakwa, W. (Hon) (2016). Ministers Address. Proceedings of the 77th Chamber of Mines 2016 Annual General Meeting, Victoria Falls, May 2016.

Eita, J. H. & Du Toit, C. B. (2007). Explaining investment behavior of the Namibian economy. A conference paper. Pretoria: University of Pretoria.

Ferderer, J. P. (2009). The impacts of uncertainty on aggregate investment spending: An empirical analysis. *Journal of Money, Credit and Banking*, 25(1)30-48.

Gawlik, L. (2008). Construction and verification of an econometric model to determine linear relationship between coal acquisition cost and production volume. *Mineral Raw Materials Management*, 24(1/1), 27-44.

Grossman, T. T. (2012). A Thoroughly Modern Resource Curse? The New Natural Resource Policy Agenda and the Mining Revival in Peru, IDS Working Paper 300, Brighton, Institute of Development Studies.

Hsu, I. C., Lin, C. Y. Y., Lawler, J. J. & Wu, S. H. (2007). Towards a model of organizational human capital development: Preliminary evidence from Taiwana. *Asia Pacific Business Review*, 32(2), 251-275.

Keynes, J. M. (1936). The general theory of employment, interest, and money. United Kingdom: Macmillan Cambridge University Press.

Leedy, P. D. & Ormrod, J. E. (2001). Practical Research: Planning and design (7th Ed,). Pearson Educational International and Prentice Hall: New Jersey.

Leeman, A. (2006). Foreign investment in Russia: Reform and Regulation. New York: McGraw-m Hill Irwin.

McCann, M. (2014). Funding mechanisms in the African region. *Accountancy Journal*, 1(2), 20-35.

McMann, A. (2010). Innovative Funding Mechanisms. *Southern African Institute of Mining and Metallurgy*, 2(4)18.

Meyer, P. (2014) Impact and development in local economies based on mining; the case of Chilean II Region, *Resources policy*, 27,119 -134.

Mothomogolo, J. (2012). Development of innovative funding mechanisms for mining start-pus: A South African Case. The Southern African Institute of Mining and Metallurgy, Platinum.

Muganyi, T. (2016). Chamber of Mines President Address. Proceedings of the 77th Chamber of Mines 2016 Annual General Meeting Speech presentation, Victoria Falls, May 2016, Zimbabwe.

Mutwiwa, m. & Fondo, K. (2015). Factors Influencing Investment in the Mining Sector in Kenya: A Case Study of Base Titanium in Kwale, *County International Journal of Science and Research*, 10(4), 12-20.

Nghifenwa, N. (2009). Factors influencing investment: A case study of the Namibian economy: Windhoek

Pandey, I. M. (2008). Financial Management. (8th Edition), French forest NSW: Pearson Education Australia / Practice Hall. *The Review of Economic Studies*, 61(2), 197-222.

Selvarajan, T. T., Ramamoorthy, N., Flood, P. C, Guthrie, J. P., MacCurtin, S. J. & Liu, W. (2007). The role of human capital philosophy in promoting firm innovativeness and performance: Test of a causal model. *International Journal of Human Resources Management,* 18(8), 1456-1470.

The Global Mining Finance Guide. (2012).

Tobin, J. (1969). A general equilibrium approach to monetary theory. *Journal of Money, Credit, and Banking*, 1(1), 15-29.

Vale, I. (2012). Classical, Keynes' and neoclassical investment theory – a synthesis. United Kingdom: *Oxford Economic Papers*, 38(2), 305-316.

Vallerrie, D. N. (2010). Basic econometrics. New York: McGraw-Hill/Irwin.

Wang, D. W. (2012). Tax policy and investment behavior. *The American Economic Review*, 57(3), 391-414.

Zhu, D. (2011). Private investment in developing countries: An empirical analysis. *Washington, IMF staff papers*, 38(1).

Journal of Economics and Behavioral Studies (JEBS)
Vol. 13, No. 1, February 2021 (ISSN 2220-6140)

Random Actions in Experimental Zero-Sum Games

Jung S. You
Department of Economics, California State University-East Bay, Hayward, CA, USA
jung.you@csueastbay.edu

Abstract: A mixed strategy, a strategy of unpredictable actions, is applicable to business, politics, and sports. Playing mixed strategies, however, poses a challenge, as the game theory involves calculating probabilities and executing random actions. I test i.i.d. hypotheses of the mixed strategy Nash equilibrium with the simplest experiments in which student participants play zero-sum games in multiple iterations and possibly figure out the optimal mixed strategy (equilibrium) through the games. My results confirm that most players behave differently from the Nash equilibrium prediction for the simplest 2x2 zero-sum game (matching-pennies) and 3x3 zero-sum game (e.g., the rock-paper-scissors game). The results indicate the need to further develop theoretical models that explain a non-Nash equilibrium behavior.

Keywords: *Zero-sum game, strategy, randomness, experiment, statistical tests*.

1. Introduction

A mixed strategy is a strategy in which a player randomly takes actions from a set of available actions, based on a set of calculated probabilities (Pindyck & Rubinfeld, 2014). Using a mixed strategy, the player would benefit from being unpredictable; thus, the other players would not be able to predict which action is going to be played (Bernheim & Whinston, 2013; McCain, 2014). The application of mixed strategies extends to dealing with terrorism, tax evasion, playing poker, beating the stock market, and winning a new product market against competitors.

Since O'Neill (1987) and Brown & Rosenthal (1990), there has been mixed evidence regarding the empirical validity of the Nash equilibrium in mixed strategies. Negative evidence suggests that mixing does not occur as predicted, particularly when the data is analyzed at the individual level.[69] Positive evidence suggests that professional players play mixed strategies according to equilibrium predictions in competitive sports, such as soccer, tennis, and baseball, within which there are a winner and a loser[70]. Several articles show both positive and negative results. Walker & Wooders (2001) find that the expected payoff of a pure strategy is almost the same as that of another pure strategy in each 2x2 stage game, which is consistent with the equilibrium predictions. However, they note that the players' choices exhibit serial correlation. Van Essen & Wooders (2015) show that professional players' behaviors are closer to equilibrium than novices'. Emara et al. (2017) examined strategic decisions in the National Football League and investigated whether the chosen sequence exhibits serial correlation. The authors find that the choices of professional players exhibit serial correlation. And, more recently, using data from the rock-paper-scissors games played on a Facebook application, Batzilis et al. (2019) show that players deviate from the Nash equilibrium in response to the information of their opponent's historical matches with other players[71].

[69] This literature includes Brown & Rosenthal (1990), O'Neill (1991), Rapoport & Boebel (1992), Mookherjee & Sopher (1994), Ochs (1995), Mookherjee & Sopher (1997), Rapoport & Amaldoss (2000), Shachat (2002), Camerer (2011), Levitt et al. (2010), and Duffy et al. (2021).
[70] This literature includes Binmore et al. (2001), Chiappori et al. (2002), Palacios-Huerta (2003), Coloma (2007), Palacios-Huerta & Volij (2008), Azar & Bar-Eli (2011), and Buzzacchi & Pedrini (2014).
[71] The analysis in Batzilis et al. (2019) focuses on the first round in each match as non-equilibrium behavior on subsequent rounds involves a player's response to the prior rounds in the match in addition to the opponent's history with other players. Their games are effectively one-shot simultaneous-move games with the opponent's history before the match.

In this paper, I investigate how closely the mixed strategy Nash equilibrium theory can predict individual behavior in finitely repeated zero-sum games: matching-pennies and rock-paper-scissors[72]. The unique Nash equilibrium in a finitely repeated zero-sum game is playing the same one-shot game Nash equilibrium in each round. Repeated two-person zero-sum games produce multiple observations of individual behavior for testing Nash equilibrium assumptions of both best responses and backward induction.

I use two-person matching-pennies and rock-paper-scissors since they are the simplest among all the games that have been tested in the literature (most literature use games with which participants are unfamiliar and in which it may not be possible for them to become proficient in the limited timeframe of experiments). The payoff in matching-pennies and rock-paper-scissors includes only two outcomes of wins and losses; the games are symmetric with respect to the mixed strategies of the row and column players and gameplay is face-to-face. Due to their simplicity in structure and equilibrium solution, players face a cognitively less demanding task than those in other experiments. One can expect that nonprofessional player quickly learn both games and might approximate the behavior in the mixed strategy Nash equilibrium.

In my experiment, 36 pairs of subjects played a matching-pennies game or the rock-paper-scissor game 20 times in succession, face to face. The players were told in advance the exact number of repetitions. The matching-pennies game has a unique equilibrium: both players randomize with probabilities .5 and .5, respectively. The rock-paper-scissor game's unique equilibrium is that both players mix according to probabilities .33, .33, and .33, respectively. As in the typical experimental setting, the game has a precisely defined set of rules, only a few strategies are available, outcomes are decided immediately after strategies are chosen, and all relevant information is observable.

With the data collected from the experiment, I test two hypotheses that equilibrium theory of best responses and backward induction yields the behavior of players in repeated zero-sum games. The first hypothesis is that at every stage, players choose pure strategies with equal mixture probabilities (according to the equilibrium strategy of the symmetric one-shot game). The second hypothesis is that equilibrium strategies in finitely repeated zero-sum games are independent of the time lags between the stages of the game.

The results of the tests do not support equilibrium play for most participants. The first null hypothesis that the mixed probabilities for individual players are identical across pure strategies in each repetition of symmetric zero-sum games is rejected at a 10% significance level for all 72 participants except four players under the matching-pennies game. The second hypothesis that players generate serially independent sequences in repeated games is rejected at a 10% significance level for more than half of a total of 72 participants. The hypothesis rejection occurs for 83% of the players under the rock-paper-scissor games whereas it is rejected for 31% of the players under the matching-pennies games.

The results in this paper support the need for developing theoretical models of an individual's non-equilibrium plays. The results also support my observation and feedback from students and teachers that mixed strategies are difficult to understand and to implement. The difficulty of using mixed strategies resides in calculating mixture probabilities and using a random device according to chosen probabilities. In fact, people are known to have difficulty generating random numbers.[73]

The remainder of the paper is organized as follows. Section 2 describes the structure and setting of the play. Section 3 is devoted to the empirical analysis. Finally, Section 4 concludes.

[72] The rock-paper-scissor game has championship competitions around the world (for instance, see the World Rock Paper Scissors Association: https://www.wrpsa.com/rock-paper-scissors-world-championship/).

[73] Research has shown that people have difficulty with detecting and producing random sequences of the sort required for the execution of mixed strategies (Bar-Hillel & Wagenaar, 1991; Oskarsson et al., 2009; Rabin, 2002; Rapoport & Budescu, 1992; Willem A. Wagenaar, 1972).

2. Zero-Sum Game Experiments

A total of 72 students with various majors from a public university in California participated in the experiments from fall 2018 to fall 2019.

The Matching-Pennies Games: At the beginning of fall 2018, a group of 36 students played the simplest zero-sum game the matching-pennies games.[74]

The matching-pennies game involves two players. All players were informed in advance that the game would be played 20 times against the same opponent. One player in each pair is referred to as a row player and the other as a column player. Each player holds a penny and displays either head (H) or tails (T) simultaneously in each round. A row player gains one point, and a column player loses one point if the coins show the same side, and a row player loses one point and a column player gains one point if the coins show different sides, in each round. Table 1 shows the bimatrix form of the game where the left number in each cell is the payoff of a row player. Participants were not shown the bimatrix form. Since each player wins 1 or loses 1 depending on the opponent for the same side, the best strategy, in theory, is to play a mixed strategy in which each player chooses H or T with equal chance, that is, unbiased coin toss.

Table 1: Matching-Pennies

		Column Player H	Column Player T
Row Player	H	1, -1	-1, 1
	T	-1, 1	1, -1

At the time when students played the zero-sum game, they had not learned the concept of mixed strategies. On the instructional handout, students had to write their strategies to win the game and had to calculate payoffs. The following box provides the class handout for the game (the rounds after the first round are omitted due to repetition in the table):

Matching-Pennies

The instructor has you, group, into pairs. Decide who is a row player or column player.
Row Player: Column Player:
Each of you shows a coin simultaneously. Row Player's payoff is 1 (or -1) and Column Player's is -1 (or +1), when the two coins show the same side (or different sides), respectively. For each round, write down your strategy in the "Your action" column to play Head (H) or Tail (T) before you and your opponent take simultaneous actions. At the end of each round, write down your payoff for that round and compute your cumulative and average payoffs in the following table.

	Your action (H or T)	Opponent's action	Your payoff	Your cumulative payoff	Your average payoff
Round 1					

To incentivize the students to play the game seriously, the author announced that each student's cumulative payoffs after 20 rounds would be proportionally converted into participation credits that they needed to collect for their final grades.

The Rock-Paper-Scissors Games: At the beginning of fall 2019, another group of 36 students played the well-known zero-sum game RPS.[75] The RPS game also involves two players. All players were informed in

[74] You (2019) analyzed the same experiments of the matching-pennies games.

advance that the game would be played 20 times against the same opponent. In RPS, each player simultaneously forms one of three shapes, "rock" (R, a closed fist), "paper" (P, a flat hand), or "scissors" (S, a fist with the index finger and middle finger extended, forming a V), with an outstretched hand. A player who plays R will beat another player who has chosen S but will lose to one who has played P. A play of P will lose to a play of S. If both players choose the same shape, the game is tied.
Table 2 shows the bimatrix form of the game where the left number in each cell is the payoff of the row player. Participants were not shown the bimatrix form. The best strategy, in theory, is to play a mixed strategy in which each player chooses R, P, or S with equal chance, that is, an unbiased dice toss where number 1 or 2 indicates a play of R, 3 or 4 a play of P, and 5 or 6 a play of S.

Table 2: Rock-Paper-Scissors

		Column Player		
		R	P	S
Row Player	R	0, 0	-1, 1	1, -1
	P	1, -1	0, 0	-1, 1
	S	-1, 1	1, -1	0, 0

Students who participated in the experiment had not learned the concept of mixed strategies. On the instructional handout, students had to write their strategies to win the game and had to calculate payoffs. The following box provides the class handout for the game (The rounds after the first round are omitted due to repetition in the table):

Rock-Paper-Scissors

The instructor has you, group, into pairs. Decide who is a row player or column player.
Row Player: Column Player:
You play an RPS zero-sum game. Each of you shows your hand sign simultaneously. Each player simultaneously forms one of three shapes with an outstretched hand. These shapes are "rock" (a closed fist), "paper" (a flat hand), and "scissors" (a fist with the index finger and middle finger extended, forming a V). You know the rule: rock crushes scissors, paper covers rock, scissors cuts paper. The winner scores 1 and the loser scores -1. If the game is tied, each earns 0. For each round, write down your strategy in the "Your action" column to play Rock (R), Paper (P) or Scissors (S) before you and your opponent take simultaneous actions. At the end of each round, write down your payoff for that round and compute your cumulative and average payoffs in the following table.

	Your action (R, P, or S)	Opponent's action	Your payoff	Your cumulative payoff	Your average payoff
Round 1					

To incentivize the students to play the game seriously, the author announced that each student's cumulative payoffs after 20 rounds would be proportionally converted into participation credits that they needed to collect for their final grades.

3. Empirical Analysis

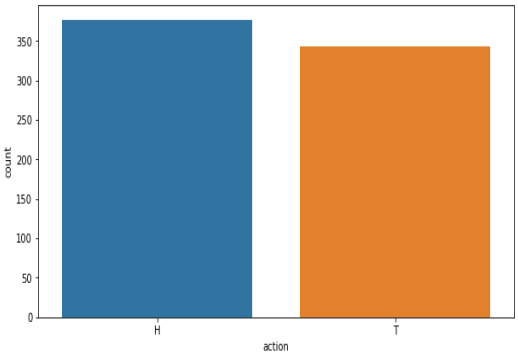
The marginal distributions of the aggregate data from both games seemingly exhibit uniformly distributed choices. For the matching-pennies game, my data consist of records of actions and payoffs that 36 students in 18 pairs submitted after a total of 20 rounds of the game. I cross-checked the actions and payoff calculations the students reported and did not find any mistakes in them. Heads were chosen 377 times, and tails were

[75] Instances of use of RPS in real-life scenarios include an American court case (Avista Management v. Wausau Underwriters in 2016), an auction house selection between Christie's and Sotheby's for the collection of Impressionist paintings in 2005, and numerous RPS games around the world including World Rock Paper Scissors Society that started in 2002.

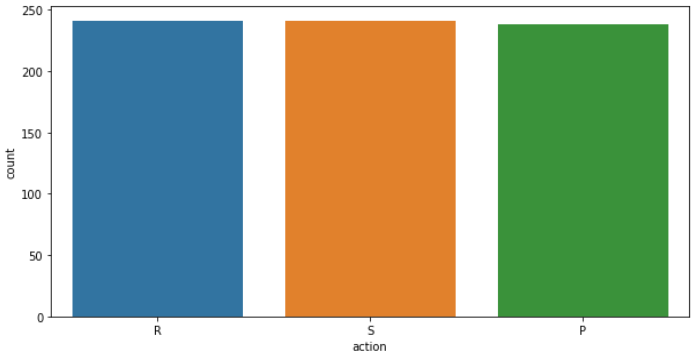
chosen 343 times out of a total of 720 times. The frequency of heads was 0.52 and that of tails was 0.48, close to .5 and .5. Figure 1(a) is the aggregate histogram of heads and tails after 20 rounds of the game with 36 students, and its appearance is close to that of a uniform distribution.

Figure 1

(a) "Matching-Pennies" Class Histogram



(b) "Rock-Paper-Scissors" Class Histogram



For the RPS game, my data consist of records of actions and payoffs that another 36 students in 18 pairs submitted after a total of 20 rounds of the game. I cross-checked the actions and payoff calculations the students reported and did not find any mistakes in them. In total, R appeared 241 times, P appeared 238 times, and S appeared 241 times; thus, the frequency of each shape makes .33, .33, and .33, respectively. Figure 1(b) is the aggregate histogram of R, P, and S after 20 rounds of the game with 36 students. Its appearance is close to that of a uniform distribution.

Game theory models individuals' rational behavior. Since aggregate data average out individual differences, I ask whether individual players' observed choices match the theoretical predictions.[76] The unique Nash equilibrium for both games predicts that in each round (i) players choose each action with equal probability and (ii) their choices are independent of their previous actions and their opponents' choices.

Individual Tests of Equal Mixture Probabilities: The tests of the null hypothesis that the mixture probabilities for individual players are identical across pure strategies in each repetition can be implemented with the Kolmogorov—Smirnov test of discrete uniform distributions. For the matching-pennies game, I apply a one-sample Kolmogorov—Smirnov test for each of a total of 36 individual players where I record tails as 0 and heads as 1.[77] The discrete Kolmogorov—Smirnov goodness-of-fit test (KS test) is an alternative to the Chi-square test, which does not achieve high statistical power for small sample sizes for discrete null distributions (Horn, 1977; Slakter, 1965). For binomial distributions, the p-values are known to be exact for the KS test (Arnold & Emerson, 2011). Using the dg of R package for discrete null distribution, I estimate the p-value via a Monte Carlo simulation with 10,000 replicates. Table 3 shows the observed frequency or mixture of choices, the Kolmogorov—Smirnov test statistic, and its p-value for Player 1 to Player 36.

[76] You (2019) performs the McNemar test on the joint distribution of paired plays. The test is based on aggregate averaged behavior that does not explain individual differences.
[77] The one-sample Kolmogorov—Smirnov test is a nonparametric test of the equality of (continuous or discontinuous) one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution.

Table 3: Matching-Pennies Game

Tests for equality of mixture probabilities

		Mixture		Kolmogorov—Smirnov		
Player	**#Obs**	**H**	**T**	**Statistic**	**P-Value**	
1	20	0.6	0.4	0.967	0.000	***
2	20	0.15	0.85	0.500	0.000	***
3	20	0.7	0.3	0.350	0.003	***
4	20	0.5	0.5	0.450	0.000	***
5	20	0.55	0.45	0.300	0.012	**
6	20	0.4	0.6	0.300	0.011	**
7	20	0.4	0.6	0.300	0.012	**
8	20	0.6	0.4	0.350	0.002	***
9	20	0.5	0.5	0.350	0.003	***
10	20	0.45	0.55	0.400	0.000	***
11	20	0.65	0.35	0.200	0.120	
12	20	0.55	0.45	0.350	0.003	***
13	20	0.7	0.3	0.300	0.013	**
14	20	0.45	0.55	0.300	0.010	**
15	20	0.55	0.45	0.300	0.012	**
16	20	0.5	0.5	0.200	0.117	
17	20	0.6	0.4	0.250	0.041	*
18	20	0.5	0.5	0.350	0.003	***
19	20	0.4	0.6	0.300	0.012	**
20	20	0.7	0.3	0.350	0.002	***
21	20	0.5	0.5	0.500	0.000	***
22	20	0.55	0.45	0.200	0.117	
23	20	0.65	0.35	0.350	0.003	***
24	20	0.45	0.55	0.200	0.040	*
25	20	0.45	0.55	0.333	0.000	***
26	20	0.55	0.45	0.133	0.198	
27	20	0.25	0.75	0.433	0.000	***
28	20	0.75	0.25	0.300	0.001	***
29	20	0.45	0.55	0.333	0.000	***
30	20	0.5	0.5	0.300	0.001	***
31	20	0.65	0.35	0.333	0.000	***
32	20	0.35	0.65	0.567	0.000	***
33	20	0.65	0.35	0.567	0.000	***
34	20	0.5	0.5	0.300	0.001	***
35	20	0.55	0.45	0.267	0.005	***
36	20	0.6	0.4	0.200	0.042	*

Note: *** Indicates rejected at 1% level, ** indicates rejected at 5% level, and * indicates rejected at 10% level.

The results in Table 3 show that the null hypothesis that mixing probabilities are identical across strategies is rejected for most players. Of the 36 players in the sample, the hypothesis is rejected for 22 players at the 1% significance level, 29 players at the 5% significance level, and 32 players at the 10% significance level. The hypothesis cannot be rejected for the remaining four players (Players 11, 16, 22, and 26) since the corresponding p-values range from 0.117 to 0.198. Given that the power of the KS test must be low for the small sample size for each player, the hypothesis could have been rejected with more observations.

Table 4: Rock-Paper-Scissors Game

Tests for equality of mixture probabilities

		Mixture			Kolmogorov—Smirnov		
Player	**#Obs**	**R**	**P**	**S**	**Statistic**	**P-Value**	
37	20	0.3	0.2	0.5	0.333	0.003	***
38	20	0.45	0.2	0.35	0.333	0.003	***
39	20	0.05	0.65	0.3	0.367	0.002	***
40	20	0.45	0.25	0.3	0.367	0.001	***
41	20	0.4	0.3	0.3	0.367	0.002	***
42	20	0.35	0.35	0.3	0.367	0.002	***
43	20	0.2	0.45	0.35	0.333	0.002	***
44	20	0.25	0.4	0.35	0.333	0.002	***
45	20	0.3	0.3	0.4	0.333	0.002	***
46	20	0.3	0.5	0.2	0.467	0.000	***
47	20	0.4	0.25	0.35	0.333	0.003	***
48	20	0.4	0.3	0.3	0.367	0.003	***
49	20	0.2	0.5	0.3	0.367	0.001	***
50	20	0.5	0.2	0.3	0.367	0.002	***
51	20	0.7	0.1	0.2	0.467	0.000	***
52	20	0.45	0.3	0.25	0.417	0.000	***
53	20	0.3	0.35	0.35	0.333	0.002	***
54	20	0.4	0.3	0.3	0.367	0.001	***
55	20	0.3	0.35	0.35	0.333	0.002	***
56	20	0.25	0.25	0.5	0.333	0.004	***
57	20	0.3	0.55	0.15	0.517	0.000	***
58	20	0.25	0.4	0.35	0.333	0.002	***
59	20	0.35	0.45	0.2	0.467	0.000	***
60	20	0.5	0.15	0.35	0.333	0.003	***
61	20	0.35	0.45	0.2	0.467	0.000	***
62	20	0.25	0.25	0.5	0.333	0.002	***
63	20	0.3	0.3	0.4	0.333	0.002	***
64	20	0.35	0.25	0.4	0.333	0.002	***
65	20	0.3	0.35	0.35	0.333	0.002	***
66	20	0.35	0.35	0.3	0.367	0.002	***
67	20	0.2	0.25	0.55	0.333	0.002	***
68	20	0.2	0.25	0.55	0.333	0.003	***
69	20	0.4	0.35	0.25	0.417	0.000	***
70	20	0.35	0.35	0.3	0.367	0.002	***
71	20	0.35	0.45	0.2	0.467	0.000	***
72	20	0.3	0.25	0.45	0.333	0.002	***

Note: *** Indicates rejected at 1% level, ** indicates rejected at 5% level, and * indicates rejected at 10% level.

For the RPS game, the unique Nash equilibrium is for every player to choose R, P, and S with an equal probability (i.e., .33), in each round. To check whether this theoretically optimal strategy was implemented, I applied the Kolmogorov—Smirnov test for each player where I record R as 0, P as 1, and S as 2. Using the dgof R package for discrete null distribution, I estimate the p-value via a Monte Carlo simulation with 10,000 replicates. Table 4 shows the observed frequency of choices, the KS test statistic and its p-value for Player 37 to Player 72.

The results in Table 4 show that the null hypothesis is rejected for all 36 players at the 1% significance level. These estimates suggest that at the individual level, the hypothesis that mixing probabilities are identical across strategies is rejected for all players at a conventional significance level.

Individual Tests of Serial Independence

In this section, I ask whether players' observed choices can be modeled as i.i.d. drawings from pair-specific stationary distributions.

Individual Tests of Serial Independence for the Matching-Pennies Game: For the matching-pennies game, I ask whether players' observed choices can be modeled as i.i.d. drawings from pair-specific stationary logit distributions. I investigate the possibility that both one's own and one's opponent's past plays are used to condition current plays. I here test whether this condition holds, and when it does not hold, I identify the sources of the failure.

To confirm that past choices have no role in determining current choices, I estimate logistic regressions for each player, applying the analysis of Palacios-Huerta (2003). My dependent variable takes a value of 1 if the play heads and a 0 otherwise. The independent variables in these equations were: first lagged indicators for both players' past choices and an indicator for the opponent's current choice. The latter is included to allow for the possibility that a player might be able to "read the face" of his or her opponent. I have experimented with the inclusion of second lags and the lagged interaction term of the two players' choice indicators in the equations underlying Table 5. However, I have found these terms to be statistically unimportant in explaining current choices for all players.

I then performed the likelihood-ratio tests of significance for the joint influence of lagged own choices, lagged opponent's choices, and contemporaneous opponent's choices. The results of the five hypothesis tests are summarized in Table 5. The first test measures the joint significance of all explanatory variables in accounting for a player's choice. According to the mixed strategy Nash equilibrium model, all the explanatory variables should be extraneous, and so we should be unlikely to find many cases in which these variables appear to be important in explaining players' choices. Only for three of the 18 pairs, at least one player's behavior is significantly determined at the 10% significance level by the set of explanatory variables included in my equations.

Table 5: Matching-Pennies Game

Results of significance tests from a logit equation for the choice of 'head'

Estimating Equation: $\Pr(C) = \frac{\exp[a_0 + a_1 \text{ lag}(C) + b_0 C^* + b_1 \text{lag}(C^*)]}{1+\exp[a_0 + a_1 \text{ lag}(C) + b_0 C^* + b_1 \text{lag}(C^*)]}$

			Players whose behavior allows rejection of the null hypothesis at the:			
Null hypothesis:			0.001 level	0.01 level	0.05 level	0.10 level
1	$a_1 = b_0 = b_1 = 0$	Row:			11,17	11,17
		Column:		18	18	12,18,32
2	$b_0 = b_1 = 0$	Row:			11,17	11,17,27
		Column:		18	12,18	12,18,30
3	$b_1 = 0$	Row:			7	7
		Column:			18,30	18,30
4	$b_0 = 0$	Row:		11,17	11,17	11,17,27
		Column:		12,18	12,18	12,18,28
5	$a_1 = 0$	Row:			17	17,31
		Column:			20,32	20,32

Notes: Tails is a pivot outcome.
C and C* denote the choice of a player and his or her opponent, respectively.
The terms 'lag' refers to the strategies previously followed in the ordered sequence of games.
Rejections are based on likelihood-ratio tests.

The observable correlations found in players' choices could have been exploited by their opponents. I, therefore, look for evidence of the influence of opponents' choices. The second test summarized in Table 5 measures the significance of terms involving the opponent's current and lagged choices in determining a

player's current choice. If players intended to play the mixed strategy Nash equilibrium and if players believed their opponents to be playing the mixed strategy Nash equilibrium, the various terms involving the opponent's current and lagged plays should not influence that player's choices. For four of the 18 pairs, at least one player's behavior is significantly influenced at the 10% significance level by the set of explanatory variables included in my equations.

The third test reported in Table 5 concerns the significance of the linear terms involving opponents' lagged play. If players attempted to predict their opponents' current choices partly based on the opponents' past choices, one would expect this term to influence observed choices. For three of the 18 pairs, at least one player's behavior is significantly determined at the 10% significance level by the set of explanatory variables included in my equations. If we cannot reject the unimportance of opponents' plays in determining players' choices, then we may accept the notion that the players themselves believed their opponents to be playing the mixed strategy Nash equilibrium. As indicated in Table 5, the data seem consistent with this notion for many players.

The fourth test reported in Table 5 concerns the ability of one player to discern the current choice of his or her opponent, even after controlling for the influence of past choices. My results indicate that "face reading" may have occurred for three pairs at the 10% significance level.

The fifth test focuses on the explanatory importance of a player's own past choices in determining his or her current choice. For three of the 18 pairs, at least one player's behavior is significantly determined at the 10% significance level by the set of explanatory variables included in my equations.

The results in Table 5 for the finitely repeated matching-pennies game might imply that nonprofessional players can make serially independent choices for each stage game as if they apply backward induction or try to play unpredictable moves. The reason why most players failed to play heads and tails with equal probabilities might be in the difficulty of generating random actions.[78]

Individual Tests of Serial Independence for the RPS Game: For the RPS game, I ask whether players' observed choices can be modeled as i.i.d. drawings from pair-specific stationary multinomial logit distributions. I estimate multinomial logistic regressions for each player, applying the analysis of Brown & Rosenthal (1990). The dependent variable is a trichotomous indicator of the choice of gesture from R, P, and S. My dependent variable takes a value of 0 if the play is R, 1 if P, and 2 otherwise. The independent variables are a first lagged indicator for both players' past choices and an indicator for the opponent's current choices. I then performed the likelihood-ratio tests of significance for the joint influence of lagged own choices, lagged opponent's choices, and contemporaneous opponent's choices. The results of the five hypothesis tests are summarized in Table 6.

Table 6: Rock-Paper-Scissors Game

Results of significance tests from multinomial logit equation for the choice of handshape

Estimating Equation:

$$\Pr(C_k) = \frac{\exp\left[a_{0,k}+a_{1,k}\text{lag}(C_k)+b_{0,k}C^*+b_{1,k}\text{ lag}(C^*)\right]}{\left[1+\exp\left[a_{0,k}+a_{1,k}\text{lag}(C_k)+ b_{0,k}\,C^*+b_{1,k}\text{ lag}(C^*)\right]+\exp\left[a_{0,k'}+ a_{1,k'}\text{lag}(C_{k'})+ b_{0,k'}C^* + b_{1,k'}\text{lag}(C^*)\right]\right]} \text{ for k, k'=1, 2, k}\neq\text{k'}$$

Null hypothesis:		Players whose behavior allows rejection of the null hypothesis at the: 0.001 level	0.01 level
1	$a_1 = b_0 = b_1 = 0$	Row: 47	37,41,43,47,61,63,65,67,69
		Column:	38,42,52

[78] Some literature find that people have difficulty with producing independent, random sequences. For instance, see W. A. Wagenaar, (1970), W. A. Wagenaar (1972).

2	$b_0 = b_1 = 0$	Row:	37,41	37,41,43,53,63,67,69
		Column:	42	38,42,50
3	$b_1 = 0$	Row:		41,43
		Column:		38,42,50
4	$b_0 = 0$	Row:	37,41	37,41,69
		Column:	42	38,42,48,50
5	$a_1 = 0$	Row:	47	37,41,47,61,63,71
		Column:		58
Null hypothesis:			0.05 level	
1	$a_1 = b_0 = b_1 = 0$	Row:	37,41,43,47,49,53,59,61,63,65,67,69,71	
		Column:	38,42,44,48,50,52,58,66,68,70,72	
2	$b_0 = b_1 = 0$	Row:	37,41,43,47,49,53,61,63,67,69	
		Column:	38,42,48,50,52,58,64,68,70,72	
3	$b_1 = 0$	Row:	41,43,53,61,63,67	
		Column:	38,42,48,50,58,62,64,72	
4	$b_0 = 0$	Row:	37,41,49,65,67,69,71	
		Column:	38,42,48,50,52,64,68,70,72	
5	$a_1 = 0$	Row:	37,41,43,47,49,53,59,61,63,65,71	
		Column:	42,44,50,52,56,58	
Null hypothesis:			0.10 level	
1	$a_1 = b_0 = b_1 = 0$	Row:	37,41,43,47,49,51,53,59,61,63,65,67,69,71	
		Column:	38,42,44,48,50,52,54,58,64,66,68,70,72	
2	$b_0 = b_1 = 0$	Row:	37,41,43,47,49,51,53,59,61,63,65,67,69,71	
		Column:	38,42,44,48,50,52,58,62,64,66,68,70,72	
3	$b_1 = 0$	Row:	37,41,43,53,59,61,63,67,69	
		Column:	38,42,44,48,50,58,62,64,72	
4	$b_0 = 0$	Row:	37,41,49,51,59,65,67,69,71	
		Column:	38,42,48,50,52,60,62,64,66,68,70,72	
5	$a_1 = 0$	Row:	37,41,43,47,49,53,59,61,63,65,71	
		Column:	42,44,50,52,56,58	

Notes: R is a pivot outcome, $C_1 = P$, and $C_2 = S$.
C and C* denote the choice of a player and his or her opponent, respectively.
The terms 'lag' refers to the strategies previously followed in the ordered sequence of games.
Rejections are based on likelihood-ratio tests

The first test measures the joint significance of all explanatory variables in accounting for a player's choice. For 15 of the 18 pairs, at least one player's behavior is significantly determined at the 10% significance level by the set of explanatory variables included in my equations. The statistical significance of these variables in explaining players' choices is the rule rather than the exception. I take this finding as strong evidence against the mixed strategy Nash equilibrium model.

The second test summarized in Table 6 measures the significance of terms involving the opponent's current and lagged choices in determining a player's current choice. If players intended to play the mixed strategy Nash equilibrium and if players believed their opponents to be playing the mixed strategy Nash equilibrium, the various terms involving the opponent's current and lagged plays should not influence that player's choices. However, for 15 of the 18 pairs, at least one player's behavior is significantly determined at the 10% significance level by the set of explanatory variables included in my equations.

The third test reported in Table 6 concerns the significance of the linear terms involving opponents' lagged play. The results of the third test are that for nine-row players and nine-column players, the set of terms involving opponents' lagged plays are statistically significant at the .10 level in determining current choices. In other words, for 13 of the 18 pairs, at least one player's behavior is significantly influenced by this term. The fourth test reported in Table 6 concerns the ability of one player to discern the current choice of his or her opponent, even after controlling for the influence of past choices. For 12 of the 18 pairs, at least one player's behavior is significantly influenced by this "face reading" term. Lastly, the results of the fifth test indicate that a substantial number of players made choices that were significantly related to their own previous choices. For 11 row players and six-column players, one can reject at the .10 level the null hypothesis that one's own past choices are uncorrelated with current choices. Only for two-row players (Players 40 and 46) and four column players (Players 39, 45, 55, and 57), any of the five null hypotheses cannot be rejected at the .10 level. In other words, 83% of the players behaved differently from the theoretical prediction of serial independence.

The findings in Table 6 indicate that the choices of at least three-quarters of players are related to their own previous choices and opponents' current and previous choices. These rejections occur despite the low power associated with pair-by-pair tests and even though some forms of behavior not in mixed strategy Nash equilibrium can generate a mixture. The most informative rejections come from the interdependent choices shown by so many player pairs. The results show little support for the mixed strategy model, even when this paper assumes stationarity in the process of generating players' choices.

4. Conclusion

My results confirm that most players behave differently from the Nash equilibrium prediction for the simplest matching-pennies and rock-paper-scissors games. The individual KS tests show that the first null hypothesis that the mixed probabilities for individual players are identical across pure strategies in each repetition of symmetric zero-sum games is rejected at a 10% significance level for all 72 participants except four players under the matching-pennies game. In addition, the individual likelihood ratio tests show that the second hypothesis that players generate serially independent sequences in repeated games is rejected at a 10% significance level for more than half of a total of 72 participants. The hypothesis rejection occurs for 83% of the players under the rock-paper-scissor games whereas it is rejected for 31% of the players under the matching-pennies games. The results of the tests do not support Nash equilibrium play for most participants.

The effects of the computational difficulties of mixed strategies could explain the results of deviations from Nash equilibrium prediction, especially for the finitely repeated RPS game.[79] Another possible explanation is that players have difficulties concealing hand gestures to their opponents even though they try to play random actions. Once a player reads the oncoming shape of the opponent's hand that is seemingly a non-equilibrium play. The player's best response is to shift from the mixed strategy Nash equilibrium and to play the corresponding move.[80] In addition to this, as people have difficulty producing independent, random sequences, serial independent plays might not only be irrational but also impossible for human subjects.[81] The results in this paper could be improved by greater sample sizes and higher incentives for players' competitive behavior. Nevertheless, given that the games this paper investigates are the simplest among all two-player games studied in the literature, the results support evidence against the validity of mixed strategy Nash equilibrium to describe most human behavior.

[79] For an example of computational difficulties, see Halpern & Pass (2015).
[80] A robot can beat the best human players in the RPS games, by using a high-speed camera and electronic reflexes to identify the oncoming shape of the opponent's hand and play the corresponding move to beat it.
[81] Randomness is easier to disprove than to prove since for disproving randomness, it is sufficient to show one type of systematic trend (W. A. Wagenaar, 1970; W. A. Wagenaar, 1972). This applies to testing on mixed strategy equilibrium.

Studying evolutionary game theory to model an individual's non-equilibrium plays is a potentially fruitful direction for future research. Recent studies show that evolutionary game theory outperforms Nash equilibrium in predicting average mixed strategies in asynchronous RPS-like games (Cason et al., 2014; Hoffman et al., 2015).[82]

References

Arnold, T. B. & Emerson, J. W. (2011). Nonparametric goodness-of-fit tests for discrete null distributions. *R Journal*, 3(2).

Azar, O. H. & Bar-Eli, M. (2011). Do soccer players play the mixed-strategy Nash equilibrium? *Applied Economics*, 43(25), 3591–3601.

Bar-Hillel, M. & Wagenaar, W. A. (1991). The perception of randomness. *Advances in Applied Mathematics*, 12(4), 428–454.

Batzilis, D., Jaffe, S., Levitt, S., List, J. A. & Picel, J. (2019). Behavior in Strategic Settings: Evidence from a Million Rock-Paper-Scissors Games. *Games*, 10(2), 18.

Benaïm, M., Hofbauer, J. & Hopkins, E. (2009). Learning in games with unstable equilibria. *Journal of Economic Theory*, 144(4), 1694–1709.

Bernheim, D. & Whinston, M. (2013). Microeconomics.

Binmore, K., Swierzbinski, J. & Proulx, C. (2001). Does minimax work? An experimental study. Economic Journal, 445–464.

Brown, J. N. & Rosenthal, R. W. (1990). Testing the minimax hypothesis: A re-examination of O'Neill's game experiment. Econometrica: Journal of the Econometric Society, 1065–1081.

Buzzacchi, L. & Pedrini, S. (2014). Does player specialization predict player actions? Evidence from penalty kicks at FIFA World Cup and UEFA Euro Cup. *Applied Economics*, 46(10), 1067–1080.

Camerer, C. F. (2011). Behavioral game theory: Experiments in strategic interaction. Princeton university press.

Cason, T. N., Friedman, D. & Hopkins, E. (2014). Cycles and instability in a rock–paper–scissors population game: A continuous-time experiment. *Review of Economic Studies*, 81(1), 112–136.

Chiappori, P. A., Levitt, S. & Groseclose, T. (2002). Testing mixed-strategy equilibria when players are heterogeneous: The case of penalty kicks in soccer. *American Economic Review*, 92(4), 1138–1151.

Coloma, G. (2007). Penalty kicks in soccer: An alternative methodology for testing mixed-strategy equilibria. *Journal of Sports Economics*, 8(5), 530–545.

Duffy, S., Naddeo, J., Owens, D. M. & Smith, J. (2021). Cognitive load and mixed strategies: On brains and minimax. Available at SSRN 3770824.

Emara, N., Owens, D., Smith, J. & Wilmer, L. (2017). Serial correlation in National Football League plays calling and its effects on outcomes. *Journal of Behavioral and Experimental Economics*, 69, 125–132.

Halpern, J. Y. & Pass, R. (2015). Algorithmic rationality: Game theory with costly computation. *Journal of Economic Theory*, 156, 246–268.

Hoffman, M., Suetens, S., Gneezy, U. & Nowak, M. A. (2015). An experimental investigation of evolutionary dynamics in the Rock-Paper-Scissors game. *Scientific Reports*, 5(1), 1–7.

Horn, S. D. (1977). Goodness-of-fit tests for discrete data: A review and an application to a health impairment scale. Biometrics, 237–247.

Levitt, S. D., List, J. A. & Reiley, D. H. (2010). What happens in the field stays in the field: Exploring whether professionals play minimax in laboratory experiments. *Econometrica*, 78(4), 1413–1434.

McCain, R. A. (2014). Game theory: A non-technical introduction to the analysis of strategy. World Scientific Publishing Company.

Mookherjee, D. & Sopher, B. (1994). Learning behavior in an experimental matching pennies game. *Games and Economic Behavior*, 7(1), 62–91.

[82] Cason et al. (2014) show that Time Average of the Shapley Polygon (TASP) outperforms Nash equilibrium in predicting average mixed strategies in asynchronous RPS-like games. Benaïm et al. (2009) propose TASP as an alternative model of non-equilibrium behavior in games whose Nash equilibria are mixed strategies.

Mookherjee, D. & Sopher, B. (1997). Learning and decision costs in experimental constant sum games. *Games and Economic Behavior*, 19(1), 97–132.

Ochs, J. (1995). Games with unique, mixed strategy equilibria: An experimental study. *Games and Economic Behavior*, 10(1), 202–217.

O'Neill, B. (1987). Nonmetric test of the minimax theory of two-person zero-sum games. *Proceedings of the National Academy of Sciences*, 84(7), 2106–2109.

O'Neill, B. (1991). Comments on Brown and Rosenthal's reexamination. Econometrica: Journal of the Econometric Society, 503–507.

Oskarsson, A. T., Van Boven, L., McClelland, G. H. & Hastie, R. (2009). What's next? Judging sequences of binary events. *Psychological Bulletin*, 135(2), 262.

Palacios-Huerta, I. (2003). Professionals play minimax. *The Review of Economic Studies*, 70(2), 395–415.

Palacios-Huerta, I. & Volij, O. (2008). Experientia docet: Professionals play minimax in laboratory experiments. *Econometrica*, 76(1), 71–115.

Pindyck, R. S. & Rubinfeld, D. L. (2014). Microeconomics. Pearson Education.

Rabin, M. (2002). Inference by believers in the law of small numbers. *The Quarterly Journal of Economics*, 117(3), 775–816.

Rapoport, A. & Amaldoss, W. (2000). Mixed strategies and iterative elimination of strongly dominated strategies: An experimental investigation of states of knowledge. *Journal of Economic Behavior & Organization*, 42(4), 483–521.

Rapoport, A. & Boebel, R. B. (1992). Mixed strategies in strictly competitive games: A further test of the minimax hypothesis. *Games and Economic Behavior*, 4(2), 261–283.

Rapoport, A. & Budescu, D. V. (1992). Generation of random series in two-person strictly competitive games. *Journal of Experimental Psychology: General*, 121(3), 352.

Shachat, J. M. (2002). Mixed strategy play and the minimax hypothesis. *Journal of Economic Theory*, 104(1), 189–226.

Slakter, M. J. (1965). A comparison of the Pearson chi-square and Kolmogorov goodness-of-fit tests with respect to validity. *Journal of the American Statistical Association*, 60(311), 854–858.

Van Essen, M. & Wooders, J. (2015). Blind stealing: Experience and expertise in a mixed-strategy poker experiment. *Games and Economic Behavior*, 91, 186–206.

Wagenaar, W. A. (1970). Appreciation of conditional probabilities in binary sequences. *Acta Psychologica*, 34, 348–356.

Wagenaar, Willem A. (1972). Generation of random sequences by human subjects: A critical survey of literature. *Psychological Bulletin*, 77(1), 65.

Walker, M. & Wooders, J. (2001). Minimax play at Wimbledon. *American Economic Review*, 91(5), 1521–1538.

You, J. S. (2019). Teaching Mixed Strategy Equilibrium through a Classroom Experiment. *Business Education Innovation Journal*, 11(2), 65.