Motives for Participation in Open-Source Software Projects: A Survey among R Package Authors

Mair, Patrick; Hofmann, Eva; Gruber, Kathrin; Hatzinger, Reinhold; Zeileis, Achim; Hornik, Kurt

DOI:
10.57938/96d90c73-39ac-479c-9813-403417bd361d

Published: 01/04/2014

Document Version:
Publisher's PDF, also known as Version of record

Document License:
Unspecified

Motives for Participation in Open-Source Software Projects: A Survey among R Package Authors

Patrick Mair, Eva Hofmann, Kathrin Gruber, Reinhold Hatzinger, Achim Zeileis, Kurt Hornik

Research Report Series
Report 126, Apri 2014

Institute for Statistics and Mathematics
http://statmath.wu.ac.at/

Motives for Participation in Open-Source Software Projects: A Survey among R Package Authors

Patrick Mair
Harvard University

Eva Hofmann
Universität Wien

Kathrin Gruber
WU Wirtschaftsuniversität Wien

Reinhold Hatzinger
WU Wirtschaftsuniversität Wien

Achim Zeileis
Universität Innsbruck

Kurt Hornik
WU Wirtschaftsuniversität Wien

Abstract

One of the cornerstones of the R system for statistical computing is the multitude of contributed packages making an extremely broad range of statistical techniques and other quantitative methods freely available. This study investigates which factors are the crucial determinants responsible for the participation of the package authors in the R project. For this purpose a survey was conducted among R package authors, collecting data on different types of participation in the R project, three psychometric scales (hybrid forms of motivation, work design characteristics, and values), as well as various specie-demographic factors. These data are analyzed using item response theory and generalized linear models, showing that the most important determinants for participation are a hybrid form of motivation and the knowledge characteristics of the work design. Other factors are found to have less impact or influence only specific aspects of participation.

Keywords: R Project for Statistical Computing, open-source software, CRAN, IRT, SIMEX.

1. Introduction

The story of the R environment for statistical computing (R Core Team 2014) has been one of tremendous success. Since it was first conceived by Ihaka and Gentleman (1996), R has been attracting more and more users and contributors from different fields where data analysis plays a major role. Fox (2009) conducted a series of interviews with members of the R Core Team in order to explore the social organization of R and to identify factors crucial to its success. The study presented here aims to examine *why* package authors participate in the R project.

We use scales on work design characteristics, personal values, and types of motivation – based on theories from a general open-source software (OSS) perspective – to learn about factors and incentives that drive authors to develop R packages as well as participate in R conferences and mailing lists.

The overwhelming majority of R packages are released under open-source licenses, thereby placing no restrictions on users and usages and guaranteeing that these packages can be-

come public goods (von Hippel and von Krogh 2003). While from a traditional economic point of view, it appears to make no sense to give away one's skills and efforts for free, thousands of highly skilled developers have organized into communities like the *Comprehensive R Archive Network* (CRAN; http://CRAN.R-project.org/), *Bioconductor* (Gentleman, Carey, Bates, Bolstad, Dettling, Dudoit, Ellis, Gautier, Ge, Gentry, Hornik, Hothorn, Huber, Iacus, Irizarry, Leisch, Li, Maechler, Rossini, Sawitzki, Smith, Smyth, Tierney, Yang, and Zhang 2004, http://www.Bioconductor.org/), and *R-Forge* (Theußl and Zeileis 2009, http://R-Forge.R-project.org/) to contribute code and documentation to open-source R packages distributed by these communities.

Studying software developer's motivations to contribute to OSS projects is not an easy or straightforward task. There are many external factors and hybrid forms of motivation that might potentially play a role and, hence, have to be taken into account when one wishes to explain OSS contributions. Empirical findings in this research area are rather limited and partially ambiguous (e.g., Roberts, Il-Horn, and Sandra 2006). In this study, we apply models from item response theory (IRT) and generalized linear models (GLM) to data collected in a survey, conveyed on the popular platforms CRAN, R-Forge, and Bioconductor, to examine the motivation of contributors to the R project.

This article is organized as follows. Section 2 gives an overview of studies examining motivation, work design, and value theory from a general OSS perspective and explains underlying theories with respect to the scales employed. Section 3 elaborates the scales used in the survey and presents research questions and hypotheses. The data are described in Section 4 before Section 5 presents some background on the statistical methodology and corresponding empirical results. Finally, Section 6 discusses our findings. The questionnaire for the assessment of participants' motivation is given in the appendix.

2. Empirical findings on participation in OSS projects

In this section we elaborate the substantive background relevant to the survey from a social sciences point of view. In general, contribution to OSS projects is influenced by an individual's professional background. The most popular perception of OSS is that participation is entirely voluntary and, thus, unpaid. However, since more and more commercial companies begin contributing to OSS projects, monetary rewards cannot be ignored as a possible source of motivation (Lakhani and Wolf 2005). In fact, Henkel (2006) finds that contributions to OSS do follow an economic rationale. The degree developers are willing to contribute to OSS is decisively influenced by their professional background (i.e., job category). While academics readily contribute all of their code, developers with a commercial background maximize their commercial benefits by employing a range of strategies to protect particularly valuable parts of code (Henkel 2006).

Referring to basic psychological concepts, the following paragraphs elaborate on motivational theory, work design theory, and value theory.

First, let us focus on *motivation* and its classical distinction between intrinsic and extrinsic motivation. Intrinsic motivation represents the enjoyment of an activity itself. It is strongly linked to an individual's perception of autonomy and competence (Deci, Koestner, and Ryan 1999). In the context of OSS, intrinsic motivation is what one would describe as "fun coding" (Lakhani and Wolf 2005). Extrinsic motivation refers to any scenario in which a person

is motivated by external control. Some of the most salient extrinsic motives are monetary rewards and peer pressure. In addition, it has been found that satisfying a personal need ("scratching a personal itch") (Raymond 1999; Shah 2006), further improvements by others (Raymond 1999; Henkel 2006), enhancing personal reputation (Wasko and Faraj 2005; Hertel, Niedner, and Hermann 2003; Roberts *et al.* 2006), reciprocity and general exchange (Shah 2006; Lakhani and von Hippel 2003), and social norms (Lakhani and Wolf 2005) are other extrinsic motives to be considered in OSS development.

Intrinsic motivation is the most pervasive motive for contributions to OSS (Lakhani and Wolf 2005; Shah 2006; Hertel *et al.* 2003). When speaking of programming, developers suggest that it is analogous to playing chess, rock climbing, and putting together a difficult jigsaw puzzle (Shah 2006).

However, most researchers agree that a simple model of purely intrinsic and extrinsic motives is insufficient to capture the motivational patterns in OSS (e.g. Roberts *et al.* 2006; Lakhani and Wolf 2005). Instead, motivation is to be more accurately understood as a highly complex continuum of intrinsic, extrinsic, and internalized extrinsic motives. Following Deci and Ryan (1987) we speak of internalized extrinsic motivation when individuals "get used" to extrinsically motivated behavior. Thereby, particular habits can be observed even when the initial external factor ceases. For example, a developer could get used to enhance his professional reputation by contributing to a prestigious OSS project at an early stage of his career. Once he has reached his initial career goals, he would quite likely continue contributing to OSS on a self-regulated basis. Thus, it would be misleading to define his motivation as, for instance, purely extrinsic (Roberts *et al.* 2006; Deci and Ryan 1987).

Motives evolve over time, as task characteristics are shifting from need-driven problem solving to mundane maintenance tasks within the community. Shah (2006) finds that most developers initially participate in OSS projects because of a need for a particular software improvement, and that the majority leave the OSS project after they have contributed only once. These one-time contributions are usually driven by extrinsic motives like reciprocity, source code commits, or career concerns. Long-term participation typically includes mundane but necessary tasks like providing support for new participants. These tasks are usually taken on by a relatively small group of committed individuals who are intrinsically motivated (Shah 2006; Lakhani and von Hippel 2003). To conclude, one-time contributors are characterized by goal-oriented problem-solving and extrinsic motivation whereas long-term contributors are characterized by their work "mundane but necessary" tasks and intrinsic motivation.

When examining the motivation to contribute to OSS projects, it has been suggested to take into account *work design* characteristics as well (Hertel 2007). Corresponding underlying traits refer to task complexity, significance of work, autonomy mastering the task, feedback from the task, etc. (Hackman and Oldham 1976; Morgeson and Humphrey 2006). The model for work design is a comprehensive approach to assess work-related factors that determine motivation to participation in a task. It allows organizations, which coordinate tasks like business organization as well as OSS project administrations, to assess the current state of specific task related characteristics and, afterwards, to change their design in a way that tasks become more motivating.

Lastly, personal *values* are another factor which need to be considered in order to understand the contributions to OSS projects. The Schwartz value theory distinguishes ten different values: benevolence, conformity, tradition, security, power, achievement, hedonism, stimulation,

self-direction, and universalism (Schwartz 1992). From these values only few were actually tested for their impact on OSS developers' productivity: in some studies the impact of values was theoretically assumed (Carillo and Okoli 2008; Millar, Choi, Russell, and Kim 2005) whereas in other studies they were assessed from a rather general perspective without precise values distinctions (Hemetsberger and Pieters 2001). Nevertheless, it has been shown that values, such as power (i.e., the desire to dominate and influence the behavior of others), self-direction (which expresses the need for independence regarding thoughts and actions), and universalism (which stands for the protection of welfare for people and nature through tolerance, understanding and appreciation) influence the different types of motivation for OSS project contribution (Oreg and Nov 2008).

3. Survey design and research questions

Based on the research results described above, three lines of possible incentives are pursued: (a) hybrid forms of motivation, (b) work design characteristics and (c) values. We investigate the hypothesis that these determine the degree of the participation in the R project. The following subsections describe the variables and constructs included in our study. Figure 1 summarizes the latent structure of the psychometric scales we use and their relation to the measures for participation in the R project.

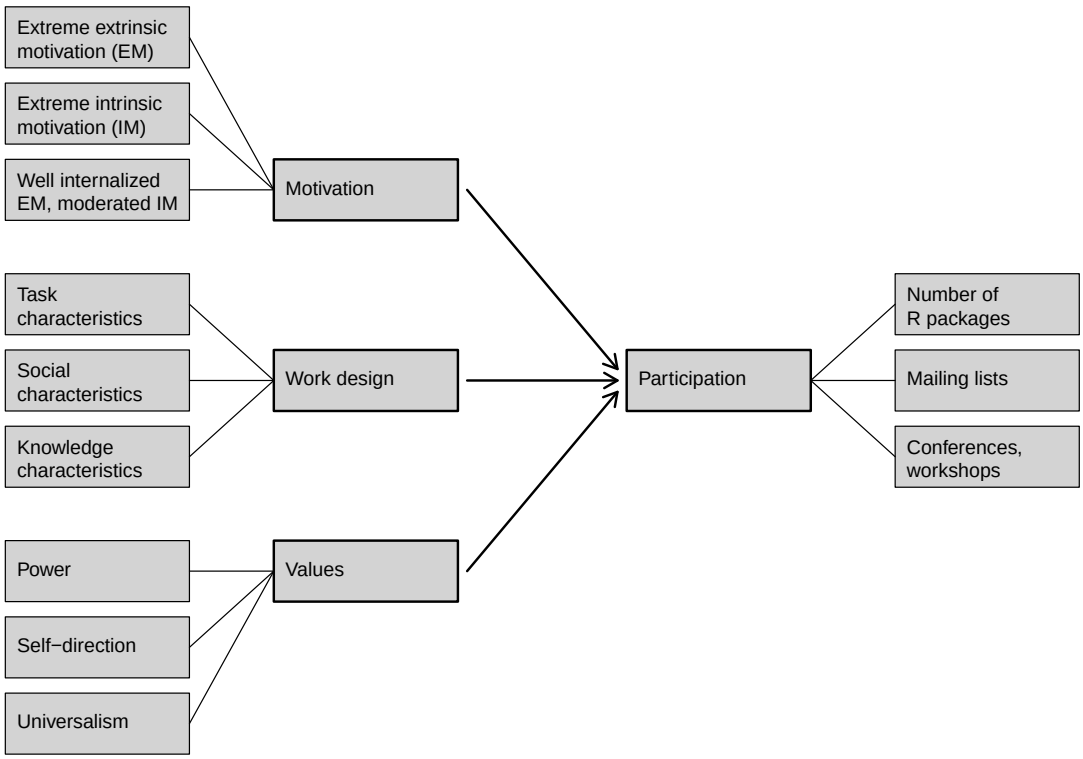

Figure 1: Hybrid forms of motivation (Reinholt 2006), work design characteristics (Morgeson and Humphrey 2006), and values (Schwartz 1992) determining participation in the R project.

3.1. Degree of participation

We investigate different forms of participation. Clearly, participation in open-source software projects will primarily manifest itself in the form of code contributions. In addition, previous studies have shown that this is just one part of an underlying learning and information process (Lakhani and von Hippel 2003). A prominent example of other forms of contribution is the active engagement in social media platforms such as mailing lists or blogs (Shah 2006).

In the context of the R project, contributed code is typically conveniently organized in packages and distributed via repositories such as CRAN or Bioconductor. This makes packages the primary vehicle for communicating conceptual and computational tools related to R. Hence, the number of R packages (co-)developed by an individual author can easily be interpreted as the main indicator of the extent of participation in the R project. Additionally, we use active participation in the official R project mailing lists (R-help, R-devel, special interest groups, ...) as an indicator for engagement in social media. A timely reanalysis might also include activities such as posting on Stack Exchange question and answer sites (e.g., Stack Overflow or Cross Validated), or blogging. Finally, we consider attending R conferences such as the annual useR! or the Directions in Statistical Computing (DSC) meetings as another form of participation.

The corresponding questionnaire items can be found in Appendix A.7.

3.2. Hybrid forms of motivation

As already elaborated, the classical distinction between intrinsic and extrinsic motivation is seen as too rigid within our context. Reinholt (2006) presents a concept that distinguishes between extreme intrinsic motivation, well internalized extrinsic motivation/moderated intrinsic motivation, and extreme extrinsic motivation. Well internalized extrinsic motivation and moderated intrinsic motivation comprise hybrid types of intrinsic and extrinsic motivation.

The corresponding scales built are based on this concept of motivation, because it provides a nuanced and coherent understanding of motivational types along a continuum of motivation. This framework also accounts for potential interaction effects between intrinsic and extrinsic types of motivation. For the intrinsic and extrinsic motivation sub-scales 36 items are included in our questionnaire (see Appendix A.5 and A.6). Each sub-scale (i.e., enjoyment based intrinsic motivation, self-reinforcement, obligation based motivation, integrated regulation, identification, introjection based regulation, external regulation) consists of four to eight items.

3.3. Work design characteristics

As suggested by previous studies (Shah 2006; Hertel *et al.* 2003), the Work Design Questionnaire (WDQ; Morgeson and Humphrey 2006) is a prominent tool to investigate work design characteristics. This work design model, based on the well-known job diagnostic survey by Hackman and Oldham (1976), captures, inter alia, the effects of task characteristics (autonomy, task variety, task significance, task identity, feedback from job), social characteristics (received and initiated interdependence, feedback from others), and knowledge characteristics (job complexity, information processing, problem solving, skill variety, specialization), on job satisfaction.

The WDQ in its original form comprises 77 items. In our questionnaire, the three sub-scales

mentioned above were used which leads to a total amount of 48 items (see Appendix A.2 and A.3). Although work characteristics and project governance have already been mentioned in previous research (Hertel 2007; Chung-Yan 2010; Johns 2011), to our knowledge we are the first to explicitly investigate work design in OSS projects. Note that WDQ items referring to the "job" were adapted to refer to the work on R packages. As an example, the item

> "The job allows me to make my own decisions about how to schedule my work"

was reformulated in terms of

> "The work on R packages allows me to make my own decisions about how to schedule my work".

3.4. Values

Schwartz's concept of values describes how an individual's experiences affect his/her action. The value scales consist of distinct motivational types which are largely universal across cultures (Schwartz 1992). Although the original scale includes 10 different motivational types, we only adopt three of them (self-direction, power, and universalism), as they seem most appropriate in the context of previous findings on OSS project participation discussed in Section 2. Although other values such as benevolence or achievement could correlate with the participation in the R project, we believe that the selected three dimensions are most suitable because they cover different kind of relevant values. For example, it is stated that universalism is a value close to benevolence and power close to achievement (Schwartz 1992).

Self-direction type values (e.g., creativity, choosing own goals, curiosity) are driven by independent thought and action. Thus, they are closely related to forms of intrinsic motivation. Power type values (e.g., social power, social recognition, authority) reflect abstract outcomes on an individual's achievements. These values do not refer to the direct outcomes of any particular action, but to the status in social structure an individual is able to derive from actions. Hence they relate directly to forms of internalized extrinsic motivation. Universalism type values (e.g., equality, wisdom, social justice) refer to action for the welfare of all people and are derived from people's awareness of the scarcity of resources. They imply that individuals will consciously protect their own survival needs through the acceptance and just treatment of anyone outside their group (Schwartz 1992). All 19 items pertaining to these value sub-scales are included in the questionnaire (see Appendix A.4).

3.5. Research questions

Based on the theoretical extension of the concept of intrinsic and extrinsic motivation (Reinholt 2006), we hypothesize that *extreme extrinsic motivation* (comprising external regulation and introjection-based regulation), *extreme intrinsic motivation* (stemming solely from enjoyment-based intrinsic motivation), and *well internalized extrinsic motivation/moderated intrinsic motivation* (identification, obligation-based intrinsic motivation, self-reinforcement, and integrated regulation), are positively related to the participation in the R project.

Additionally, it is expected that *task characteristics* (comprising autonomy, task variety, task significance, task identity and feedback from the job), *knowledge characteristics* (including job complexity, information processing, problem solving, skill variety and specialization), and

social characteristics (consisting of received and initiated interdependence and feedback from others), are positively related to participation. The more positive these characteristics are perceived, the more a package author should participate in R activities.

Finally, it is hypothesized that, in line with earlier studies, the values *self-direction* and *universalism* relate positively to participation, whereas *power* is expected to relate negatively.

4. Sample description

In total, 4274 R package authors were contacted via email in April 2010 and asked to fill out an online questionnaire before the end of May 2010. The survey was conducted using the online survey software Unipark (QuestBack AG 2013). The platforms we used for the acquisition of the email addresses were CRAN, R-Forge, and Bioconductor. In total we sent out 4274 emails of which approximately 200 could not successfully be delivered ("bounced").

A total of 1448 persons considered the questionnaire. 310 respondents quit immediately and 51 respondents scrolled through without answering. Altogether, a sample of 1087 persons remained, 45 of which quit the questionnaire at some point and 764 users completed the whole questionnaire without skipping any of the items described below. These subjects represent the sample considered for subsequent statistical analysis and modeling, with variables as follows.

The first variable measuring participation (see Section 3.1) is the number of packages (co-)developed by an individual author. It is summarized in Figure 2 and has a mean of 2.9, a median of 2, and a standard deviation of 3.45. The other two participation variables are binary, with 57.07% contributing to the R mailing lists and 31.02% attending R conferences.

The items pertaining to the motivation, work design and value scales are transformed to

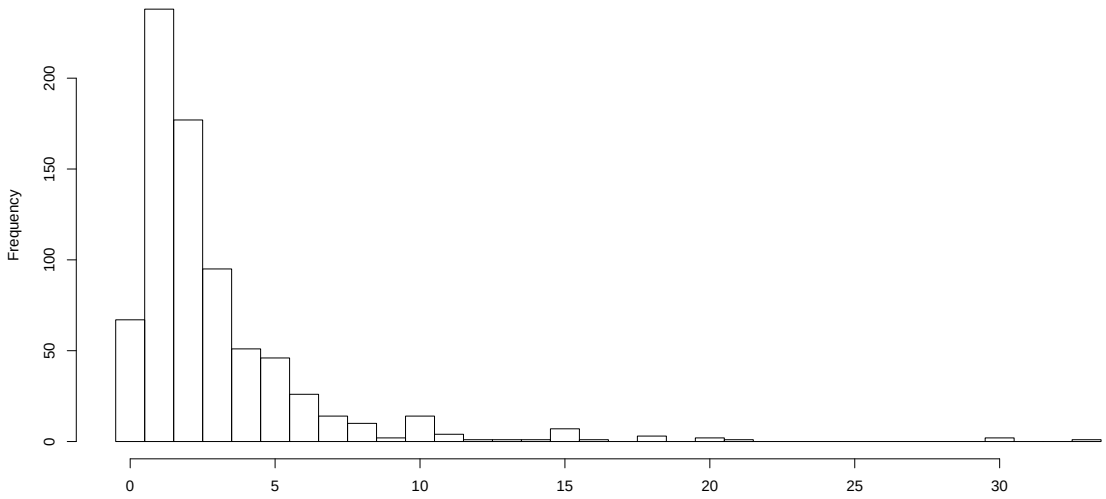

Figure 2: Histogram for the number of packages.

psychometric scores using IRT analysis as described in detail in Section 5.1.

Finally, to control for the effects of various socio-demographic factors when relating participation to psychometric scores, we derive several binary variables from the questionnaire items in Appendix A.8. Specifically, we record whether respondents have a PhD degree (*phd*, yes: 71.47%), an education in statistics (*statseduc*, yes: 63.09%), are employed full time (*fulltime*, yes: 85.21%), work in academia (*academia*, yes: 60.47%), and work as statisticians (*statswork*, yes: 63.22%).

5. Statistical analysis

The goal of our analysis is to determine the effect of hybrid forms of motivation, work design and values on participation in the R project, controlling for socio-demographic factors. We thus start by constructing the three times three psychometric scores as depicted in Figure 1 from the corresponding questionnaire items using the two-parameter logistic IRT model. Subsequently, these scores are used as explanatory variables in generalized linear models (GLMs) for each of the three variables measuring participation. For the count response *number of packages* we fit a negative-binomial (NB) GLM, while for the binary responses participation in *mailing lists* and *conferences*, binomial GLMs with logit links are employed. To account for potential measurement error (ME) when constructing the psychometric scores, estimation by the simulation-extrapolation (SIMEX) method is performed in addition to standard maximum likelihood (ML) estimation.

5.1. IRT analysis

The latent constructs occurring in the scales for hybrid forms of motivation, work design, and values are scored using models from item response theory (IRT). IRT is a family of latent variable models to score item responses of subjects on a single latent trait. The resulting parameter estimates are item and person parameters. Numerous IRT models have been proposed over the last decades (see, e.g., De Ayala 2008, for an overview). Note that it is not our aim in the current study to establish scales that fulfill highest standards with respect of measurement theory, as, for instance, in Rasch analysis (Rasch 1960). Rather, we take the scales as they are, with the exception that we remove items that produce a high misfit. Eventually, the package authors are scored on uni-dimensional latent traits (three times three traits in total, see Figure 1).

Based on these thoughts, our model of choice is the two-parameter logistic model (2-PL; Birnbaum 1968) defined as:

$$P(X_{vi}) = \frac{\exp\left(\alpha_i(\theta_v - \delta_i)\right)}{1 + \exp\left(\alpha_i(\theta_v - \delta_i)\right)}. \tag{1}$$

Given the responses X_{vi} by author v ($v = 1, \ldots, n$) on item i ($i = 1, \ldots, k$), we estimate two item parameters: an *item discrimination parameter* α_i ($\alpha_i > 0$), and an *item difficulty parameter* δ_i that locates the item on the latent trait. Subsequently, for each author v we estimate a *person ability parameter* θ_v that maps the author on the same latent trait.

Since the 2-PL is a uni-dimensional model, we perform our IRT analysis separately for each scale dimension using the R package **ltm** (Rizopoulos 2006). We eliminate misfitting items

using the $Q1$ fit statistic (Yen 1981). Once the final model is found, the corresponding person parameters are extracted.

With respect to the Reinholt (2006) scale for extrinsic and intrinsic motivation, only one item is eliminated:

> "I work on R packages because it is expected from me."

which is supposed to load on *extreme extrinsic motivation* dimension with subtrait "introjection". Pertaining to the WDQ, we need to eliminate two *task characteristics* items, namely

> "The work performed on R packages has a significant impact on a lot of subjects outside the R community."

and

> "The development of R packages provides me the chance to completely finish the pieces of work I begin."

from the "task significance" and "task identity" subtraits, respectively. With respect to the *social characteristics* in the WDQ we eliminate

> "My work on R packages affects the activity of other R developers."

and

> "The tasks of others depend directly on my task."

from the "initiated interdependence" subtrait. Finally, for the *social characteristics* the item

> "I receive feedback on my R package performance from other people in the R community."

from the "feedback from others" subtrait will not be taken into account. For the three value scales none of the items has to be eliminated.

All remaining items can be assigned to their corresponding dimensions, yielding person parameters (factor scores) for each of the nine traits and each R package author (along with the corresponding standard errors). In subsequent analyses and tables, the resulting variables are labeled *mextrinsic*, *mintrinsic*, and *mhybrid* for the motivation scales; *wtask*, *wsocial*, and *wknowledge* for the scales obtained from the WDQ; and *vpower*, *vselfdirection*, *vuniversalism* for the value scales.

5.2. Generalized linear models

The person parameters obtained in the IRT analysis are included as the main determinants of interest for the degree of participation in a subsequent GLM analysis (McCullagh and Nelder 1989). A general representation of our model is

$$g(\boldsymbol{\mu}) = \boldsymbol{\Theta}\boldsymbol{\beta} + \mathbf{X}\boldsymbol{\gamma}, \tag{2}$$

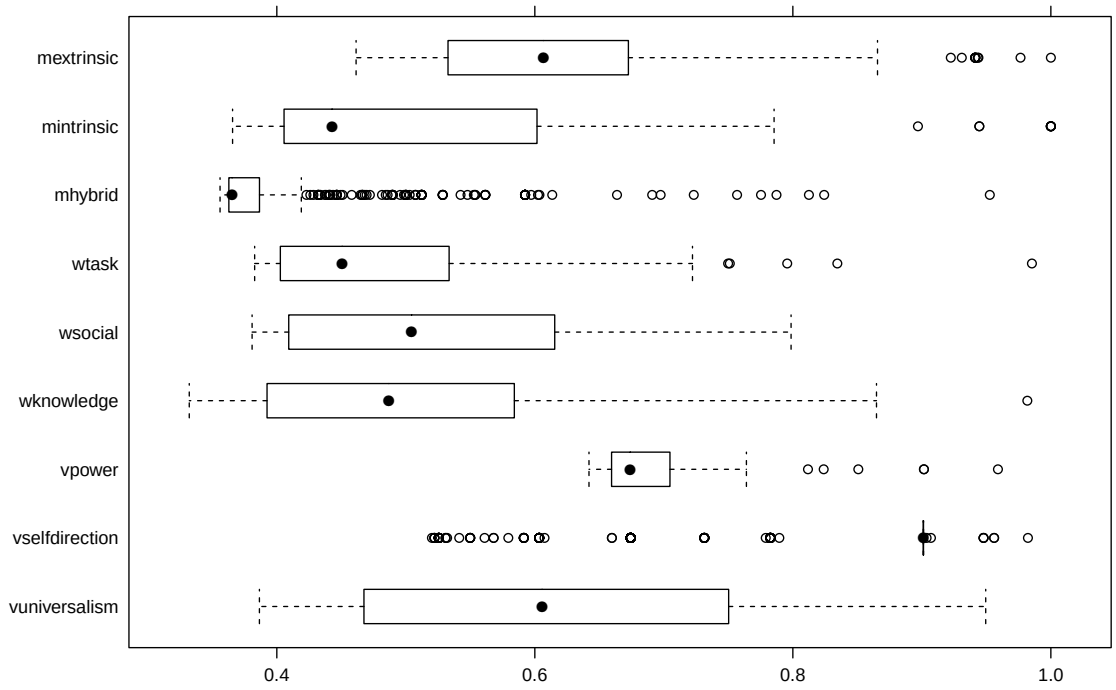

Figure 3: Box plots for measurement errors, i.e., standard errors of estimated person parameters from the 2PL models.

where $\boldsymbol{\mu}$ is the mean of the participation response variable, $g(\cdot)$ represents the corresponding link function, $\boldsymbol{\Theta}$ is the matrix containing of person parameters with corresponding regression coefficients $\boldsymbol{\beta}$, and finally $\mathbf{X}$ is the matrix with socio-demographic variables and corresponding regression coefficients $\boldsymbol{\gamma}$. Depending on the participation response considered, different models from the GLM class are employed: For the count response *number of packages* we fit a negative-binomial (NB) model which, as opposed to a regular Poisson regression, accounts for over-dispersion (for more details see Cameron and Trivedi 2013, Chapter 3). Binomial GLMs with logistic link are used for the binary responses capturing participation in *mailing lists*, and *conferences*, respectively.

Note that unlike the socio-demographic variables in $\mathbf{X}$ the person parameters $\boldsymbol{\Theta}$ are subject to measurement error (ME) as they are obtained from IRT analysis. Due to this ME the standard maximum likelihood (ML) estimation of the model is in general biased. To mitigate this problem, the heteroskedasticity of the MEs needs to be taken into account. See Figure 3 for the standard errors of the person parameters obtained from the 2-PL model.

More formally, let $\boldsymbol{\beta}$ be the true value of the parameter vector and $\hat{\boldsymbol{\beta}}$ the estimates of the ML model. Since we do not account for the MEs in this model, estimates are biased; that is, $E(\hat{\boldsymbol{\beta}}) \neq \boldsymbol{\beta}$. In order to get unbiased regression estimates for the psychometric predictors, we have to account for the MEs. This can be achieved by the use of the simulation-extrapolation method (SIMEX) by Cook and Stefanski (1994), after fitting the naive GLMs. This leads to corrected estimates in the presence of additive ME.

Note that we apply the jackknife variant of SIMEX (Stefanski and Cook 1995) which is based the following idea: The starting point is the standard error of the person parameters in construct c ($c = 1, \ldots, C$) which reflects the ME. This could be a single value for each construct c, or a vector of length n allowing for varying ME's across persons. In our analysis we allow for full ME heteroskedasticity (across constructs, across persons) which leads to the ME matrix $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$ of dimension $n \times C$ with column vectors $\boldsymbol{\sigma}_{\theta_{\cdot|c}}$.

Through ME-based jackknife resampling, the SIMEX approach simulates repeated measurements. By refitting the model in each step we get a new parameter vector $\hat{\boldsymbol{\beta}}(\boldsymbol{\Sigma}_{\boldsymbol{\Theta}})$. SIMEX theory states that the mean of the parameter distribution resulting from resampling, that is, $\bar{\hat{\boldsymbol{\beta}}}(\boldsymbol{\Sigma}_{\boldsymbol{\Theta}})$, is an unbiased estimator for $\boldsymbol{\beta}$. Technical details for SIMEX can be found in Cook and Stefanski (1994) and Stefanski and Cook (1995). A corresponding implementation is given in the R package **simex** (Lederer and Küchenhoff 2006).

5.3. Results

In this section we present the results of the regression models. For each of the participation responses (number of packages, and participation in mailing lists and conferences, respectively), a full GLM is considered first using all potential determinants considered: the three times three psychometric scores and all socio-demographic factors. Subsequently, a stepwise backward selection of the predictor variables in the GLM is carried out based on the Akaike information criterion (AIC) to highlight which determinants are most relevant. For both, the full model and the stepwise selected model, SIMEX estimates are presented in addition to the standard ML estimates to assess sensitivity of the results to measurement error in the psychometric scores.

First, we look at the negative binomial regression with the number of packages an author has (co-)authored as the response variable, see Table 1. It is positively influenced by all three forms of motivation with hybrid motivation playing the most important role (both in terms of significance and effect size). Work design is also an important determinant of the number of packages, with knowledge characteristics being positively associated and task characteristics negatively associated while social characteristics have a smaller positive albeit non-significant effects. Thus, package authors whose job implies a high amount of complexity, skill variety, etc., are involved in more packages. Conversely, the higher a package author scores on the task dimension, the lower the number of packages co-authored. In terms of the values scales, only power is found to be significantly associated with the number of packages showing a negative effect. On the socio-demographic side, the fact that a package author works full time and her field of work is statistics have a significant effect.

The results for the logistic regression model of participation in mailing lists are given in Table 2. Again, hybrid motivation significantly increases the probability of participation. However, extrinsic motivation has a similar absolute effect (both in terms of coefficient estimate and standard error) but the effect is negative. Regarding the WDQ, knowledge and task characteristics are again found to be more important than social characteristics with knowledge characteristics having a large positive impact and task characteristics a somewhat smaller negative impact. None of the value scale variables has a significant effect on the participation in mailing lists. For the socio-demographic predictor part, the fact that a package author works in the field of statistics leads to a significantly lower participation probability.

Finally, Table 3 presents the results of the logistic regression model for the binary response

	Full (ML)	Full (SIMEX)	Step (ML)	Step (SIMEX)
(Intercept)	0.596***	0.594***	0.653***	0.645***
	(0.126)	(0.124)	(0.120)	(0.117)
mextrinsic	0.097	0.172*	0.082	0.149*
	(0.053)	(0.077)	(0.052)	(0.073)
mintrinsic	0.026	−0.002		
	(0.063)	(0.092)		
mhybrid	0.158**	0.213**	0.172***	0.239***
	(0.057)	(0.080)	(0.049)	(0.062)
wtask	−0.172**	−0.288***	−0.175**	−0.305***
	(0.054)	(0.078)	(0.054)	(0.073)
wsocial	0.070	0.101	0.076	0.128
	(0.049)	(0.068)	(0.049)	(0.067)
wknowledge	0.322***	0.489***	0.327***	0.484***
	(0.055)	(0.079)	(0.055)	(0.079)
vpower	−0.162**	−0.297**	−0.166**	−0.305**
	(0.059)	(0.095)	(0.058)	(0.092)
vselfdirection	−0.008	0.000		
	(0.071)	(0.120)		
vuniversalism	−0.056	−0.077		
	(0.055)	(0.084)		
phdyes	0.129	0.141		
	(0.095)	(0.097)		
statseducyes	0.023	0.008		
	(0.100)	(0.099)		
fulltimeyes	0.326**	0.279*	0.363**	0.336**
	(0.120)	(0.119)	(0.117)	(0.116)
academiayes	−0.168*	−0.169*	−0.146	−0.159*
	(0.083)	(0.084)	(0.080)	(0.079)
statsworkyes	0.158	0.178	0.182*	0.192*
	(0.101)	(0.100)	(0.080)	(0.079)
Log-likelihood	−1608.328		−1610.386	
AIC	3246.657		3240.772	
BIC	3316.235		3287.157	

Table 1: Negative binomial GLM parameter estimates (and standard errors) for number of packages (with significance codes *** for the 0.001 level, ** for the 0.01 level, and * for the 0.05 level).

indicating participation of package authors in R conferences and workshops. Regarding the motivational dimension, hybrid motivation is again found to be the most important determinant. Its influence is again clearly positive with a p-value of about 6% in the full model and highly significant after AIC selection. In terms of work design, knowledge characteristics are again clearly significant with a positive impact on participation. Task characteristics again have a negative but for this response variable non-significant influence. All three value scales have only little impact with only the universalism scale being significant at 5% after step-

	Full (ML)	Full (SIMEX)	Step (ML)	Step (SIMEX)
(Intercept)	0.963***	1.072***	0.872***	0.878***
	(0.249)	(0.260)	(0.157)	(0.160)
mextrinsic	−0.351**	−0.629***	−0.334**	−0.553***
	(0.111)	(0.163)	(0.106)	(0.155)
mintrinsic	0.216	0.271	0.220	0.298
	(0.128)	(0.188)	(0.127)	(0.192)
mhybrid	0.437***	0.551***	0.432***	0.552***
	(0.119)	(0.165)	(0.114)	(0.156)
wtask	−0.254*	−0.415*	−0.236*	−0.401*
	(0.113)	(0.162)	(0.111)	(0.157)
wsocial	−0.115	−0.134		
	(0.101)	(0.139)		
wknowledge	0.427***	0.700***	0.420***	0.647***
	(0.114)	(0.161)	(0.113)	(0.157)
vpower	0.046	0.122		
	(0.120)	(0.186)		
vselfdirection	0.051	0.097		
	(0.149)	(0.247)		
vuniversalism	−0.057	−0.123		
	(0.116)	(0.172)		
phdyes	−0.145	−0.255		
	(0.194)	(0.205)		
statseducyes	−0.285	−0.329		
	(0.207)	(0.214)		
fulltimeyes	0.084	0.040		
	(0.234)	(0.241)		
academiayes	−0.206	−0.177	−0.243	−0.260
	(0.172)	(0.178)	(0.159)	(0.164)
statsworkyes	−0.436*	−0.379	−0.606***	−0.569***
	(0.208)	(0.214)	(0.164)	(0.169)
Log-likelihood	−481.398		−483.615	
AIC	992.795		983.229	
BIC	1062.374		1020.338	

Table 2: Logistic GLM parameter estimates (and standard errors) for participation in mailing lists (with significance codes *** for the 0.001 level, ** for the 0.01 level, and * for the 0.05 level).

wise selection (or SIMEX estimation). The only significant socio-demographic variable is the occupational status: A full-time employment of a package author is a strong determinant to participate in R conferences. None of other socio-demographic variables (except, to a certain degree, statistics as the field of work which has a minor influence) has any impact on the model.

In summary, the broad picture is very similar across all three participation responses (and

	Full (ML)	Full (SIMEX)	Step (ML)	Step (SIMEX)
(Intercept)	-1.593^{***}	-1.594^{***}	-1.613^{***}	-1.627^{***}
	(0.283)	(0.290)	(0.267)	(0.271)
mextrinsic	-0.022	-0.070		
	(0.114)	(0.170)		
mintrinsic	0.120	0.156		
	(0.137)	(0.208)		
mhybrid	0.224	0.276	0.276^{**}	0.325^{**}
	(0.124)	(0.174)	(0.098)	(0.125)
wtask	-0.075	-0.164		
	(0.118)	(0.166)		
wsocial	-0.025	-0.043		
	(0.106)	(0.150)		
wknowledge	0.455^{***}	0.719^{***}	0.433^{***}	0.633^{***}
	(0.120)	(0.162)	(0.116)	(0.162)
vpower	0.119	0.244		
	(0.126)	(0.195)		
vselfdirection	-0.053	-0.129		
	(0.155)	(0.255)		
vuniversalism	-0.219	-0.402^{*}	-0.238^{*}	-0.408^{*}
	(0.119)	(0.177)	(0.116)	(0.169)
phdyes	-0.069	-0.092		
	(0.204)	(0.214)		
statseducyes	0.127	0.088		
	(0.217)	(0.221)		
fulltimeyes	0.716^{**}	0.675^{*}	0.623^{*}	0.601^{*}
	(0.270)	(0.275)	(0.254)	(0.258)
academiayes	-0.169	-0.124		
	(0.180)	(0.184)		
statsworkyes	0.265	0.271	0.313	0.308
	(0.218)	(0.221)	(0.170)	(0.172)
Log-likelihood	-449.989		-451.852	
AIC	929.978		915.703	
BIC	999.556		943.535	

Table 3: Logistic GLM parameter estimates (and standard errors) for participation in conferences (with significance codes *** for the 0.001 level, ** for the 0.01 level, and * for the 0.05 level).

corresponding models), even if the details vary to a certain degree: Hybrid motivation and knowledge characteristics are the most important determinants for higher levels of participation in the R project. The picture for extrinsic and intrinsic motivation is less clear and varies over the particular type of participation. Authors that score highly on the task characteristics scale generally participate less while social characteristics do not play an important role. Similarly, values (in terms of the Schwartz scales) are not found to be important

drivers of participation as they rarely show up in the selected models. The influence of the socio-demographic variables also varies somewhat: Full-time employment generally increases participation while a job in academia somewhat lowers it. Working in statistics has a positive effect on the number of packages and participation in the conferences but a negative on participation in mailing lists. The remaining two variables (having a PhD and an education in statistics, respectively) cannot be shown to have an impact on participation in any of the models.

6. Discussion

This study has asked what motivates R package authors to participate in the R project for statistical computing. A survey was conducted and the data were analyzed using IRT models and, subsequently, GLMs. In what follows our findings are discussed in more detail and related to the literature on participation in open-source software projects.

6.1. Hybrid forms of motivation

In line with the literature as discussed in Section 2 – see especially Roberts *et al.* (2006), Lakhani and Wolf (2005), and Shah (2006) – hybrid motivation is crucial while purely intrinsic and purely extrinsic forms of motivation are less important. This is exactly reflected in our regression results and conforms well with the academic "life cycle". Various factors, including reputation, reciprocity, or social norms, can contribute to an internalization of extrinsic motives. Consequently, many academics do what they "have to do" but also select tasks they enjoy doing which can also encompass activities such as "fun coding" (see Lakhani and Wolf 2005).

The influence of purely extrinsic motivation which in particular includes monetary rewards (suggested as a potentially important factor by Lakhani and Wolf 2005) varies across the participation variables. In part this may be due to a strong rooting of the R project in various academic communities which only rather recently saw an increasing commercialization. In particular, this might explain the negative (and highly significant) coefficient of extrinsic motivation for mailing list contributions. While packages and conferences are by now regarded as scientific contributions, mailing list contributions have no (direct) impact on academic performance measures. This is somewhat substantiated by the positive (but not significant) influence of intrinsic motivation on contribution to mailing lists. We note that Wasko and Faraj (2005) find that contributions to "electronic networks of practice" are increased if the contributors perceive that this enhances their reputation (i.e., a typical extrinsic motive). Thus, participation in R mailing lists is apparently not perceived to do so. This might be different in the more recently established question and answer websites such as Stack Exchange which work differently from classical mailing lists and explicitly try to capture the reputation of its contributors.

6.2. Work design characteristics

To the best of our knowledge, this paper is the first empirical investigation of the influence of work design characteristics on participation in OSS projects. Our results confirm that in particular the knowledge characteristics of a "job-related perspective" are important drivers

for "a better scope for creating motivating conditions" as hypothesized by (Hertel 2007).

The positive and highly significant impact of knowledge characteristics on all three dimensions of participation (package development, mailing lists and conference participation) reflects that all of these involve a high amount of information processing and solving complex problems – comparable to putting together a difficult jigsaw puzzle (Shah 2006).

Task characteristics, on the other hand, are found to have a negative influence on participation. This could be explained as follows: If the work is organized around the development of an R package as the central task (from development of code, via writing of manuals and vignettes to maintenance and bug fixing), then R authors appear to do that but are less involved in the development of further packages or discussions on mailing lists. Or conversely, those authors who participate more and develop several packages do not appear to be driven be the task of R package development as such but by the underlying knowledge characteristics involved (see above).

6.3. Values

Generally speaking, we find that values have little impact on participation. which conforms with (Oreg and Nov 2008) who show that the values are correlated with the different types of motivation. Our results indicate that in the context of R packages there appears only little additional direct effect of the values – other than potential indirect effects through the types of motivation, see above. There are two notable exceptions, though, that are discussed in the following: power is shown to have a clear negative effect on the number of packages and universalism has a clear negative effect on conference participation.

The former reflects that package authors, for whom social power, wealth, social recognition and authority are important, produce fewer packages than their dimensional counterparts. The way the field of applied and computational statistics has developed over the last years, R package implementations have increased in scientific value. Thus, for a researcher, a corresponding implementation has become an academic status symbol to the effect that they refer to themselves as "R package author" even when involved in a single package only.

The latter shows that the higher a package author scores on the universalism dimension, the less likely she is to attend meetings. A closer look at what is meant by "universalism" provides an interesting interpretation of this result. According to Schwartz, attributes associated with universalism include: a world of beauty, unity with nature, protecting the environment, and inner harmony. These are derived from an awareness of the scarcity of resources. Thus, universalism implies a strong environmental attitude that is incompatible with carbon-intensive long distance travels to conferences.

6.4. Socio-demographic variables

Full-time employment always has a positive impact on participation; significantly so for the number of packages and conference participation. This suggests that many contributions to the R project are made as part of the job. For mailing lists the influence is weaker but, as already argued above, such participation is typically not part of the job description. Additionally, there may also be direct effects of full-time employment on conference participation, e.g., through reimbursement of expenses.

Working in statistics also has positive impact on the number of packages and conference

participations but clearly negative impact on mailing list participation. While the former is not surprising given that the R system is dedicated to statistics, the latter may not be obvious. However, statisticians will typically have other ways of asking questions related to R (e.g., colleagues within their department) and other ways of providing feedback about the corresponding statistical methods (e.g., in forms of papers, books, or lectures). However, for R authors and users coming from other domains (say, ecology, finance, or epidemiology) the R mailing lists may be a more crucial means of obtaining information related to R. This also conforms with the findings of Bohn, Feinerer, Hornik, and Mair (2011) who show that answers on the R mailing lists are mainly given by a few central players feeling responsible for certain topics.

Interestingly, an academic "background" – i.e., having a PhD or a job in academia – does not lead to more participation as hypothesized by Henkel (2006); but, in fact, has almost no impact on any of the three response variables.

6.5. Conclusions

Overall, our results are important for institutions and individuals that want to stimulate growth of OSS development: they must provide a work environment and corresponding incentives that foster a high amount of skill variety (data analysis, programming, technical skills), information processing (e.g., large amount of thinking, keeping track of more than one thing at a time, analyze a lot of information), job complexity (e.g., work on relatively complex tasks) and problem solving (e.g., problems with no obvious correct answer, requiring creativity, presenting unique ideas).

Furthermore, the results have shown that growth of R-related projects is positively influenced by hybrid motivation while purely intrinsic or extrinsic motives are less important. Hence, this suggests that extrinsic motives (such as monetary rewards or building reputation) can be important drivers but need to be balanced by possibilities of internalizing them. However, given the ongoing commercialization of the R ecosystem this aspect deserves re-investigation in the future.

Acknowledgments

We would like to commemorate with sorrow the death of Reinhold Hatzinger. His contributions to this paper were invaluable and he will be fondly remembered by his friends and colleagues.

References

Birnbaum A (1968). "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability." In FM Lord, MR Novick (eds.), *Statistical Theories of Mental Test Scores*, pp. 395–479. Addison-Wesley, Reading, MA.

Bohn A, Feinerer I, Hornik K, Mair P (2011). "Content-Based Social Network Analysis of Mailing Lists." *The R Journal*, **3**(1), 11–18.

Cameron AC, Trivedi PK (2013). *Regression Analysis of Count Data*. 2nd edition. Cambridge University Press, Cambridge, UK.

Carillo K, Okoli C (2008). "The Open Source Movement: A Revolution in Software Development." *Journal of Computer Information Systems*, **49**, 1–9.

Chung-Yan GA (2010). "The Nonlinear Effects of Job Complexity and Autonomy on Job Satisfaction, Turnover, and Psychological Well-Being." *Journal of Occupational Health Psychology*, **15**, 237–251.

Cook JR, Stefanski LA (1994). "Simulation-Extrapolation Estimation in Parametric Measurement Error Models." *Journal of the American Statistical Association*, **89**, 1314–1328.

De Ayala RJ (2008). *The Theory and Practice of Item Response Theory*. The Guilford Press, New York.

Deci EL, Koestner R, Ryan RM (1999). "A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation." *Psychological Bulletin*, **125**, 627.

Deci EL, Ryan RM (1987). "The Support of Autonomy and the Control of Behavior." *Journal of Personality and Social Psychology*, **53**, 1024–1024.

Fox J (2009). "Aspects of the Social Organization and Trajectory of the R Project." *The R Journal*, **1**(2), 5–13.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J (2004). "**Bioconductor**: Open Software Development for Computational Biology and Bioinformatics." *Genome Biology*, **5**, R80. URL http://genomebiology.com/2004/5/10/R80.

Hackman JR, Oldham GR (1976). "Motivation through the Design of Work: Test of a Theory." *Organizational Behavior and Human Performance*, **16**, 250–279.

Hemetsberger A, Pieters R (2001). "When Consumers Produce on the Internet: An Inquiry into Motivational Sources of Contribution to Joint-Innovation." Paper presented at the Fourth International Research Seminar on Marketing Communications and Consumer Behavior, La Londe.

Henkel J (2006). "Selective Revealing in Open Innovation Processes: The Case of Embedded Linux." *Research Policy*, **35**(7), 953–969.

Hertel G (2007). "Motivating Job Design as a Factor in Open Source Governance." *Journal of Management and Governance*, **11**, 129–137.

Hertel G, Niedner S, Hermann S (2003). "Motivation of Software Developers in Open Source Projects: An Internet-Based Survey of Contributors to the Linux Kernel." *Research Policy*, **32**, 1159–1177.

Ihaka R, Gentleman R (1996). "R: A Language for Data Analysis and Graphics." *Journal of Computational and Graphical Statistics*, **5**, 299–314.

Johns G (2011). "Attendance Dynamics at Work: The Antecedents and Correlates of Presenteeism, Absenteeism, and Productivity Loss." *Journal of Occupational Health Psychology*, **16**, 483–500.

Lakhani KR, von Hippel E (2003). "How Open Source Software Works: Free User-to-User Assistance." *Research Policy*, **32**(6), 923–943.

Lakhani KR, Wolf RG (2005). "Why Hackers Do What They Do: Understanding Motivation and Effort in Free/Open Source Software Projects." In J Feller, B Fitzgerald, S Hissam, KR Lakhani (eds.), *Perspectives on Free and Open Source Software*. MIT Press, Cambridge.

Lederer W, Küchenhoff H (2006). "A Short Introduction to the SIMEX and MCSIMEX." *R News*, **6**(4), 26–31.

McCullagh P, Nelder JA (1989). *Generalized Linear Models*. 2nd edition. Chapman & Hall, London.

Millar CCJM, Choi CJ, Russell ET, Kim JB (2005). "Open Source Communities: An Integrally Informed Approach." *Journal of Organizational Change Management*, **18**, 259–268.

Morgeson FP, Humphrey SE (2006). "The Work Design Questionnaire (WDQ): Developing and Validating a Comprehensive Measure for Assessing Job Design and the Nature of Work." *Journal of Applied Psychology*, **91**, 1321–1339.

Oreg S, Nov O (2008). "Exploring Motivations for Contributing to Open Source Initiatives: The Roles of Contribution Context and Personal Values." *Computers in Human Behavior*, **24**, 2055–2073.

QuestBack AG (2013). "Unipark: Online Survey Software for Universities and Students." http://www.unipark.com/, accessed 2013-09-25.

Rasch G (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen.

Raymond E (1999). "The Cathedral and the Bazaar." *Knowledge, Technology & Policy*, **12**, 23–49.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Reinholt M (2006). "No More Polarization, Please! Towards a More Nuanced Perspective on Motivation in Organizations." *Technical report*, Center for Strategic Management Working Paper Series, Copenhagen Business School, Copenhagen, Denmark.

Rizopoulos D (2006). "**ltm**: An R Package for Latent Variable Modelling and Item Response Theory Analyses." *Journal of Statistical Software*, **17**(5), 1–25. URL http://www.jstatsoft.org/v17/i05/.

Roberts JA, Il-Horn H, Sandra AS (2006). "Understanding the Motivations, Participations and Performance of Open Source Software Developers: A Longitudinal Study of the Apache Projects." *Management Science*, **52**, 984–999.

Schwartz SH (1992). "Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries." *Advances in Experimental Social Psychology*, **25**, 1–65.

Shah SK (2006). "Motivation, Governance, and the Viability of Hybrid Forms in Open Source Software Development." *Management Science*, **52**, 1000–1014.

Stefanski LA, Cook JR (1995). "Simulation-Extrapolation: The Measurement Error Jackknife." *Journal of the American Statistical Association*, **90**, 1247–1256.

Theußl S, Zeileis A (2009). "Collaborative Software Development Using R-Forge." *The R Journal*, **1**(1), 9–14.

von Hippel E, von Krogh G (2003). "Open Source Software and the Private-Collective Innovation Model: Issues for Organization Science." *Organization Science*, **14**, 209–223.

Wasko M, Faraj S (2005). "Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice." *MIS Quarterly*, **29**, 35–56.

Yen W (1981). "Using Simulation Results to Choose a Latent Trait Model." *Applied Psychological Measurement*, **5**, 245–262.

A. Questionnaire

A.1. Start

Dear R package author,

You have been selected as a potential participant in a survey about motivation for developing R packages and participating in the R community more generally. The study is being conducted by researchers from the Institute for Statistics and Mathematics and the Institute for International Marketing Management of (**WU** Vienna University of Economics and Business).

Filling in this questionnaire is **voluntary** and will take approximately **15 minutes** to complete.

Your answers are anonymous and confidential and it will not be possible to identify your individual responses when the data is analyzed and reported. The answers you provide serve the improvement of the Comprehensive R Archive Network (CRAN) to offer developers and maintainers of R an even more effective platform. They also are for research purposes and aim to examine what motivates persons to participate actively in the development and maintenance of R packages. You can withdraw your participation until you have completed the online questionnaire and pressed the send-button at the end of the questionnaire. After this point, it is not possible to withdraw your data as all responses are anonymous and individual responses cannot be identified. The study data will be stored securely in accordance with the Austrian Data Protection Act, and only the project researchers will have access to it. The overall results from the questionnaire will be used to undertake adaptations in CRAN and will be included in academic publications, conference presentations and for teaching purposes.

By filling in the questionnaire, you are providing your consent for your responses to be used in the ways previously described.

If you have any queries regarding the study or its results, please contact us!

Thank you for your time.

CRAN Motivation Survey Team
(Reinhold Hatzinger, Eva Hofmann, Kurt Hornik, Patrick Mair, Elfriede Penz, & Achim Zeileis)

In case of questions, please contact Eva Hofmann or Patrick Mair!
Eva.Hofmann@wu.ac.at (Tel: +43 1 31336 - 5126)
Patrick.Mair@wu.ac.at (Tel: +43 1 31336 - 4844)

A.2. Section 1.1

Below find a list of statements on your development of R packages. Please indicate whether you agree or disagree with the following statements! Choose the option that slightly better represents your position!

The work on R packages involves performing a variety of tasks.

My work on R packages affects the activity of other R developers.

The major work on R packages I undertake is the maintenance of R packages.

The work on R packages comprises relatively uncomplicated tasks.

The work on R packages requires the use of a number of skills.

The tasks of others depend directly on my task.

The development of R packages is arranged so that I can work on an entire package from beginning to end.

The work on R packages requires data analysis skills.

The major work on R packages I undertake is the development of code.

The work on R packages itself is very significant and important in the broader scheme of things.

I receive feedback on my R package performance from other people in the R community.

The work on R packages often involves dealing with problems that I have not encountered before.

The results of my work on R packages are likely to significantly affect the lives of other people.

The work on R packages is highly specialized in terms of purpose, tasks, or activities.

The work on R packages requires a depth of expertise.

The work on R packages requires that I only do one task or activity at a time.

Other people in the R community provide information about the effectiveness (e.g., quality and quantity) of my R package performance.

The development of R packages allows me to complete the work I start.

The work on R packages requires technical skills regarding package building and documentation.

The development of R packages provides me the chance to completely finish the pieces of work I begin.

Unless my work on the R package gets done, other tasks cannot be completed.

The work on R packages requires me to keep track of more than one thing at a time.

The work on R packages itself provides me with information about my performance.

The work performed on R packages has a significant impact on a lot of subjects outside the R community.

A.3. Section 1.2

Below find again a list of statements on your development of R packages. Please indicate whether you agree or disagree with the following statements! Choose the option that slightly better represents your position!

The work on R packages requires very specialized knowledge.

The work on R packages involves solving problems that have no obvious correct answer.

The activities while working on R packages are greatly affected by the work of other people. The development of R packages involves completing a piece of work that has an obvious beginning and end.

The work on R packages requires that I engage in a large amount of thinking.

The tools, procedures, materials, and so forth used to develop R packages are highly specialized in terms of purpose.

The development of R packages allows me to make decisions about what methods I use to complete my work.

My work on R packages cannot be done unless others do their work.

I receive a great deal of information from the R community about my R package performance.

The work on R packages requires me to analyze a lot of information.

The work on R packages involves doing a number of different things.

The work on R packages requires me to be creative.

The work on R packages provides me with significant autonomy in making decisions.

The work on R packages involves a great deal of task variety.

The work on R packages allows me to make my own decisions about how to schedule my work.

The work on R packages requires unique ideas or solutions to problems.

The work on R packages has a large impact on people outside the R community.

The work on R packages involves performing relatively simple tasks.

The work on R packages requires programming skills.

The work on R packages requires the performance of a wide range of tasks.

The work on R packages depends on the work of many different people for its completion.

The work on R packages itself provides feedback on my performance.

The major work on R packages I undertake is the packaging/documentation for CRAN.

The work activities themselves provide direct and clear information about the effectiveness (e.g., quality and quantity) of my performance.

A.4. Section 2

Find a list of values below. Please evaluate the importance *(unimportant vs. important)* **of each value as a guiding principle in your life! Choose the option that slightly better represents your beliefs!**

Equality (equal opportunity for all)
Social Power (control over others, dominance)
Freedom (freedom of action and thought)
Wealth (material possessions, money)
Self-Respect (belief in one's own worth)
Creativity (uniqueness, imagination)
A World at Peace (free of war and conflict)
Social Recognition (respect, approval by others)
Unity with Nature (fitting into nature)
Wisdom (a mature understanding of life)
Authority (the right to lead or command)
A World of Beauty (beauty of nature and the arts)
Social Justice (correcting injustice, care for the weak)
Independent (self-reliant, self-sufficient)
Broad-Minded (tolerant of different ideas and beliefs)
Protecting the Environment (preserving nature)
Choosing Own Goals (selecting own purposes)
Preserving My Public Image (protecting my "face")
Curious (interested in exploring everything)

A.5. Section 3.1

Find a list of statements on your development of R packages below. Please indicate whether you agree or disagree with the following statements! Choose the option that slightly better represents your position!

I develop R packages, because...

I can publish the packages in scientific journals.
it is in line with my personal values.
it reflects my responsibility towards the R community.
I believe that it is appropriate to do so.
it is an important task for me.
they are a byproduct of my empirical research. If I cannot find suitable existing software to analyze my data, I develop software components myself.
it is important for my personal goals but for no apparent rewards, such as money, career opportunities, etc.
I am committed to the R community.
I think that it is of importance.
I take pleasure in applying my skills.
it is expected from me.

it gives me satisfaction to produce something of high quality.
I believe that it is a necessity.
I can feel satisfied with my performance.
it is part of my identity.
it is an integral part of my personality.
I aim for social approval of my activities.
I get the feeling that I've accomplished something of great value.

A.6. Section 3.2

Again, find a list of statements on your development of R packages below. Please indicate whether you agree or disagree with the following statements! Choose the option that slightly better represents your position!

I develop R packages, because...

I feel an obligation towards the R community.
it means pure fun for me.
I enjoy undertaking the required tasks.
that's what my friends do.
I feel that R requires continuous enhancement.
it is a joyful activity.
I need them for teaching courses.
I believe it is vital to improve R.
it leaves me with a feeling of accomplishment.
they are part of my master / PhD thesis.
I feel that it is an interesting exercise.
that's what my work colleagues do.
I expect an enhancement of my career from it.
they are a byproduct of my methodological research. If I develop/extend methods, I develop accompanying software, e.g., for illustrations and simulations.
it comes more or less with my job.
my employer pays me to do so.
I develop them for clients who pay me.
it is part of my character to do so.

A.7. Section 4

Please give some details on your participation in the R community!
Where did you first get in touch with R?

- As student at a university
- As academic at a university
- At work outside of the university
- Media (Internet, Newspaper, etc.)
- Other
  If other, please specify! _____

For how long have you been participating in the R community (in years)?

_____

Do you plan to continue to participate in the R community? Please indicate the extent to which you think further participation probable on a scale from 1 to 5 (1 = Very unlikely, 5 = Very likely)!

Do you use other statistical software packages than R? Multiple answers are acceptable.

- IBM SPSS (former SPSS, PASW)
- Stata
- SAS
- S-PLUS
- Minitab
- Systat
- EViews
- MATLAB
- Other
  If other, please specify! _____

If you are working in a team coding R packages, how many people other than you work approximately in this team? In case you are working alone, please fill in 0!

_____

Until now, in the development of how many R packages have you been involved?

_____

Where do you distribute your R packages? Multiple answers are acceptable.

- CRAN
- Bioconductor
- R-Forge (http://R-Forge.R-project.org/)
- RForge (http://www.RForge.net/)
- Other
  If other, please specify! _____

In case you have published manuscripts on your R packages, in which media have you published them? Multiple answers are acceptable.

- Journal of Statistical Software
- The R Journal (or formerly R News)
- Journal of Computational & Graphical Statistics
- Computational Statistics and Data Analysis
- Computational Statistics
- Other
  If other, please specify! _____

Do you participate in other activities of the R community?

- R mailing lists (R-help, R-devel, R-SIGs, ...)
- R conferences (useR!, DSC, ...)

A.8. Section 5

Lastly, please fill in some details on your person!
How old are you (in years)?

_____

Are you ...?

- Male
- Female

What is your highest level of education?

- High school
- Vocational/technical qualification or apprenticeship
- University degree (BA, MSc., MBA, etc.)
- University degree (PhD)

In which fields have you been educated? Multiple answers are acceptable.

- Statistics
- Business & economics
- Social sciences
- Life sciences
- Engineering & computer technology
- Mathematics & natural sciences
- Other
  If other, please specify! _____

Which of the following describes your occupational status?

- Part time 1–20hr/wk
- Full time work
- Training/student
- Full time homemaker, carer or parent
- Temporary leave (e.g., maternity or sick leave)
- Retired
- Not working

Which of these categories best describes your job?

- Academic at a university (e.g., Researcher, Lecturer)
- Public official (e.g., Researcher at Governmental Body)
- Private research institute (e.g., Researcher)
- Private sector (e.g., Technician, Statistician)
- Student
- Not applicable (e.g., Not Working, Homeworker, Carer, Parent, Retired)
- Other
  If other, please specify! _____

In which field do you work? Multiple answers are acceptable.

- Statistics
- Business & economics
- Social sciences
- Life sciences
- Engineering & computer technology
- Mathematics & natural sciences
- Other
  If other, please specify! _____

Which country do you work in?

- Abkhazia
- Afghanistan
- Albania
- Algeria
- Andorra
- ...

Affiliation:

Patrick Mair
Department of Psychology
Harvard University
33 Kirkland Street
02138 Cambridge, MA, Unitest States of America
E-mail: mair@fas.harvard.edu