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

What Drives Package Authors to Participate in the R Project for Statistical Computing? Exploring Motivation, Values, and Work Design

Article (Submitted)

What Drives Package Authors to Participate in the R Project for Statistical Computing? Exploring Motivation, Values, and Work Design

Patrick Mair [*], **Eva Hofmann** [†], **Kathrin Gruber** [‡], **Reinhold Hatzinger** [‡], **Achim Zeileis** [§], **and Kurt Hornik** [‡]

[*]Harvard University, Cambridge, MA,[†]University of Vienna, Vienna, Austria,[‡]WU Vienna University of Economics and Business, Vienna, Austria, and [§]University of Innsbruck, Innsbruck, Austria

Submitted to Proceedings of the National Academy of Sciences of the United States of America

One of the cornerstones of the R system for statistical computing is the multitude of packages contributed by numerous package authors. This makes an extremely broad range of statistical techniques and other quantitative methods freely available. So far no empirical study has investigated psychological factors that drive authors to participate in the R project. This article presents a study of R package authors, collecting data on different types of participation (number of packages, participation in mailing lists, participation in conferences), three psychological scales (types of motivation, psychological values, and work design characteristics), as well as various socio-demographic factors. The data are analyzed using item response models and subsequent generalized linear models, showing that the most important determinants for participation are a hybrid form of motivation and the social characteristics of the work design. Other factors are found to have less impact or influence only specific aspects of participation.

R project for statistical computing | Schwartz values | motivation | work design | item response theory | generalized linear models | simulation-extrapolation

Abbreviations: CRAN, Comprehensive R Archive Network

Significance

Over the last years the open-source environment R has become the most popular environment for statistical computing and data analysis across many fields of research. The developer community is highly active: around 7200 packages are available in the official CRAN repository, and a few more on developer platforms like GitHub or R-Forge. One question that has not been studied yet is: WHY do people contribute to the R environment? What are the key motives that drive package authors? Do these developers have some specific personal value structures? Are some work environments more conducive to productivity than others? This is the first empirical study performed within the R package author community that finds answers to these questions.

The story of the R environment for statistical computing [1] has been one of tremendous success. Since it was first conceived by [2], R has been attracting more and more users and contributors from different fields where data analysis plays a major role. [3] 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 become public goods [4]. 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* [5] (`http://www.Bioconductor.org/`), *R-Forge* [6] (`http://R-Forge.R-project.org/`), and *GitHub* (`https://github.com/`) to contribute code and documentation to open-source R packages distributed by these communities.

Studying software developer's motivations and determinants for participating in OSS projects is not a straightforward task. There are many internal and external factors that might potentially play a role and, hence, have to be taken into account when one wishes to explain OSS participation. Empirical findings in this research area are rather limited and partially ambiguous [7]. 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.

Psychological Findings on Participation in OSS Projects

In terms of internal factors that influence participation in OSS projects, psychological literature suggests to consider motivational theory, work design theory, and value theory. Motivational theory distinguishes between intrinsic and extrinsic motivation. Intrinsic motivation is the most pervasive motive for contributions to OSS [8, 9, 10, 11]. It represents the enjoyment of an activity itself and is strongly linked to an individual's perception of autonomy and competence [12]. 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") [13, 9], further improvements by others [13, 14], enhancing personal reputation [16, 10, 7, 17], reciprocity and general exchange [9, 15], and social norms [8] are other extrinsic motives to be considered in OSS development. Most researchers agree that a simple model of purely

Reserved for Publication Footnotes

intrinsic and extrinsic motives is insufficient to capture the motivational patterns in OSS [7, 8]. Instead, motivation is to be more accurately understood as a complex continuum of intrinsic, extrinsic, and internalized extrinsic motives. Motives evolve over time, as task characteristics are shifting from need-driven problem solving to mundane maintenance tasks within the community.

The second potential influential factor for OSS contribution are *work design* characteristics [18, 19]. Corresponding underlying traits refer to task complexity, significance of work, autonomy mastering the task, feedback from the task, etc. [20, 21]. The model for work design allows organizations to assess the current state of specific task related characteristics and, afterwards, to change their design in a way that tasks become more motivating.

Third, personal *values* can be important for understanding contributions to OSS projects. The classic value theory by [22] distinguishes 10 different values: benevolence, conformity, tradition, security, power, achievement, hedonism, stimulation, self-direction, and universalism. [23] determine the following three values to be relevant for OSS developments: self-direction, power, and universalism [24]. 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 [22].

Survey Design and Research Questions

Our population consists of package authors who contributed to R packages on *CRAN*, *Bioconductor*, and *R-forge*. This includes package maintainers as well as people that received credit for contributing code and, therefore, appear in the package author list. We need to distinguish package authors clearly from users, i.e. people who are just using packages or providing code snippets without being "officially" involved in a package development. Our study does not aim to generalize the results to the whole R community.

The online questionnaire for the package authors, provided as Supporting Information (SI), included standard socio-demographic variables as well as more specific dichotomous work related variables such as whether respondents have a PhD degree, an education in statistics, are employed full time, work in academia, and work as statisticians.

Based on the research results described above, three lines of possible psychometric incentives are pursued: (a) hybrid forms of motivation, (b) work design characteristics and (c) values. We investigate to which extend these factors determine the degree of the authors' participation in the R project. The following subsections describe these 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.

[INSERT FIGURE 1 (cranpnas-dia.eps) HERE]

Degree of Participation. Participation in OSS projects will primarily manifest itself in the form of code contributions. As previous studies have shown, however, this is just one part of an underlying learning and information process [15]. A prominent example of other forms of contribution is the active engagement in social media platforms such as mailing lists or blogs [9].

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 first, main indicator of the extent of participation in the R project. As a second indicator we use active participation in R project mailing lists (R-help, R-devel, special interest groups, ...) as an indicator for engagement in social media. Finally, as third participation indicator we consider attending R conferences such as the annual useR! or the Directions in Statistical Computing (DSC) meetings.

Psychometric Constructs. As elaborated above, the classic distinction between intrinsic and extrinsic motivation is seen as too rigid within our context. [25] 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 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. 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.

As suggested by previous studies [9, 10] the Work Design Questionnaire (WDQ) [21] is a prominent tool to investigate work design characteristics. This work design model captures, among others, the following three subscales: 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). In its original form the WDQ comprises 77 items. Using the three sub-scales above reduces the questionnaire to 48 items. Note that WDQ items referring to work tasks in general were adapted to the work on R packages.

Regarding personal values, we consider three out of 10 values of the Schwartz value scale (self-direction, power, and universalism). All 19 items pertaining to these value sub-scales are included in the questionnaire.

Research Questions. Based on the theoretical extension of the concept of intrinsic and extrinsic motivation [25], 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.

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

Statistical Analysis and Results

Statistical Analysis Work Flow. Our sample consists of 1087 package authors. The statistical analysis work flow is the following: We scale each psychometric construct using a two-parameter logistic (2-PL) item response theory (IRT) model [29]. Unidimensionality is checked using categorical principal component analysis [30] and itemfit is tested using the $Q1$ fit statistic [31]. For the set of fitting items, the latent trait (person) parameters are estimated which then act as predictors, in addition to demographic variables, in the subsequent generalized linear models (GLM). For the first degree of participation response "number of packages" we fit a negative-binomial regression, for "participation in mailing lists" and "attending conferences" we fit two logistic regressions. For each of these regression models, first a full model is considered using all potential determinants: 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 the full model and the final model from stepwise selection, in order to account for the measurement error of latent trait scores as predictors in the GLM, we apply the simulation-extrapolation (SIMEX) approach [32]. Methodological details about each statistical analysis step is given in the SI as well as the outputs in terms of regression tables and effects plots.

Results. First, we look at the negative binomial regression with the number of packages an author has (co-)authored as the response variable, see Table S1. The effect plots for the final model are given in Figure S3.

The number of packages are positively influenced by hybrid and extrinsic motivation. Work design is also an important determinant of the number of packages, with social characteristics being positively associated and task characteristics negatively associated. Thus, the higher the intiated/received interdependence of an author and the more feeback he/she gets from the community, the more packages he/she is involved in.

Conversely, the higher a package author scores on the task dimension, the lower the number of packages (co-)authored. In terms of the value 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 his/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 S2 and the effect plots are shown in Figure S4.

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, social characteristics have 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 S3 presents the results of the logistic regression model for the binary response indicating participation of package authors in R conferences and workshops. The corresponding effect plot are given in Figure S5.

Regarding the motivational dimension, hybrid motivation is again found to be the most important determinant. Its influence is again positive. In terms of work design, social characteristics are significant with a positive impact on participation. Regarding values, universalism is significant at 5% after stepwise selection. 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.

To summarize, the broad picture is very similar across all three participation responses (and corresponding models), even if the details vary to a certain degree: Hybrid motivation and social 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 knowledge characteristics do not play an important role. Similarly, values 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 varies across the models: 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.

Discussion

This study has asked why R package authors participate in the R project for statistical computing. A survey was conducted and the data were analyzed using IRT models and, subsequently, GLMs (with SIMEX correction). In what follows, our findings are discussed in more detail and related to the literature on participation in OSS projects.

Hybrid Forms of Motivation. In line with the literature – see especially [7], [8], and [9] – 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. On the one hand, many academics "do what they have to do". One the other hand they select tasks they enjoy doing which can also encompass activities such as "fun coding" [8].

The influence of purely extrinsic motivation which, in particular, includes monetary rewards [8] varies across the participation variables. In part, this may be due to a strong rooting of the R project in various academic communities. While packages and conferences are by now regarded as scientific contri-

butions, 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 [16] 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.

Work Design Characteristics. Social work design characteristics reflect the fact that work is performed within a broader social environment [21] where single individuals highly depend on each other. Our results show that OSS projects provide high degrees of social dependency and feedback as theoretically hypothesized by [18]. That social characteristics are such an important factor in our models is not too surprising, given that we are interacting in a social media dominated environment and social coding platforms are widely used [33]. Psychological explanations for our results are the following: First, interaction with important others leads to reputation (self-esteem, future job opportunities, etc.). Second, interaction with alike minded individuals (i.e. interested in solving statistical problems) might be a possibility to express oneself and enjoy social inclusion.

From a broader perspective, social aspects include social recognition and identification. The R community seems to offer the opportunity for R developers to identify with this highly valued group and feel a sense of belonging. It can be assumed that they receive parts of their self-esteem by belonging to such a valued group [34] and are especially motivated to contribute to this group. It would be interesting to study such general social aspects of reputation gaining in a follow-up study.

Task characteristics 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), 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.

Values. 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. There are two notable exceptions: 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 trait 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 he or 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 may be incompatible with carbon-intensive long distance travels to conferences.

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 the field of 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 overlaps with the findings of [35] 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 [14]. In fact, it has almost no impact on any of the three response variables.

Conclusions. Our 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.

In conclusion, 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 interdependence and feedback from others. Such collaborative research strategies also include the encouragement to work on projects with researchers outside the institution and the engagement in social coding platforms.

1. R Core Team, *R: A Language and Environment for Statistical Computing*, 2015.
2. Ihaka R. and Gentleman, R. C. *R: A language for data analysis and graphics.* Journal of Computational and Graphical Statistics, 5 (1996), 299–314.
3. Fox, J. *Aspects of the social organization and trajectory of the R project.* The R Journal, 1 (2009), 5–13.
4. von Hippel E. and von Krogh G., *Open source software and the private-collective innovation model: Issues for organization science.* Organization Science, 14 (2003), 209–223.
5. Gentleman, R. C., Carey, V. J., Bates, D. M., 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, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J., *Bioconductor: Open software development for computational biology and bioinformatics.* Genome Biology, 5 (2004), R80.
6. Theussl S. and Zeileis, A. *Collaborative software development using R-Forge.* The R Journal, 1 (2009), 9–14.
7. Roberts, J. A., Il-Horn, H., and Sandra, A. S. *Understanding the motivations, participations and performance of open source software developers: A longitudinal study of the Apache projects.* Management Science, 52 (2006), 984–999.
8. Lakhani, K. R. and Wolf, R. G. Why hackers do what they do: Understanding motivation and effort in free/open source software projects. *Perspectives on Free and Open Source Software*, eds. Feller, J., Fitzgerald, B., Hissam, S., and Lakhani, K. R. (MIT Press, Cambridge), 2005.
9. Shah, S. K. (2006). *Motivation, governance, and the viability of hybrid forms in open source software development.* Management Science, 52 (2006), 1000–1014.
10. Hertel, G., Niedner, S., and Hermann, S. *Motivation of software developers in open source projects: An internet-based survey of contributors to the Linux kernel.* Research Policy, 32 (2003), 1159–1177.
11. Li, Y., Tan, C. H., and Teo, H. H. *Leadership characteristics and developers' motivation in open source software development.* Information & Management, 49 (2012), 257–267.
12. Deci, E. L., Koestner, R., and Ryan, R. M. *A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation.* Psychological Bulletin, 125 (1999), 627–668.
13. Raymond, E. *The cathedral and the bazaar.* Knowledge, Technology & Policy, 12 (1999), 23–49.
14. Henkel, J. *Selective revealing in open innovation processes: The case of embedded Linux.* Research Policy, 35 (2006), 953–969.
15. Lakhani, K. R. and von Hippel, E. *How open source software works: Free user-to-user assistance.* Research Policy, 32 (2003), 923–943.
16. Wasko, M. and Faraj, S. *Why should I share? Examining social capital and knowledge contribution in electronic networks of practice.* MIS Quarterly, 29 (2005), 35–56.
17. Bianchi, A. J., Kang, S. M., and Stewart, D. *The organizational selection of status characteristics: Status evaluations in a open source community.* Organization Science, 23 (2014), 341–354.
18. Hertel, G. *Motivating job design as a factor in open source governance.* Journal of Management and Governance, 11 (2007), 129–137.
19. Hemetsberger, A. and Reinhardt, C. *Collective development in open-source communities: An activity theoretical perspective on successful online collaboration.* Organization Studies, 30 (2009), 987–1008.
20. Hackman, J. R. and Oldham, G. R. *Motivation through the design of work: Test of a theory.* Organizational Behavior and Human Performance, 16 (1976), 250–279.
21. Morgeson, F. P. and Humphrey, S. E. *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 (2006), 1321–1339.
22. Schwartz, S. H. *Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries.* Advances in Experimental Social Psychology, 25 (1992), 1–65.
23. Oreg, S. and Nov, O. (2008). *Exploring motivations for contributing to open source initiatives: The roles of contribution context and personal values.* Computers in Human Behavior, 24 (2008), 2055–2073.
24. Engelhardt, S. and Freytag, A. *Institutions, culture, and open source.* Journal of Economic Behavior & Organization, 95 (2013), 90–110.
25. Reinholt, M. *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.
26. Wu, C. G., Gerlach, J. H., and Young, C. E. *An empirical analysis of open source software developer' motivations and continuance intentions.* Information and Management, 44 (2007), 253–262.
27. Kish, L. *Survey sampling*, 1965.
28. Armstrong, J. S. and Overton, T. *Estimating nonresponse bias in mail surveys.* Journal of Marketing Research, 14 (1977), 396–402.
29. Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. *Statistical Theories of Mental Test Scores* eds. Lord, F. M. and Novick, M. R. (Addison-Wesley, Reading, MA), 1968, pp. 395–479.
30. De Leeuw, J., and Mair, P. *Gifi methods for optimal scaling in R: The package homals.* Journal of Statistical Software, 31(4) (2009), 1–21.
31. Yen, W. *Using simulation results to choose a latent trait model.* Applied Psychological Measurement, 5 (1981), 245–262.
32. Cook, J. R. and Stefanski, L. A. *Simulation-extrapolation estimation in parametric measurement error models.* Journal of the American Statistical Association, 89 (1994), 1314–1328.
33. Dabbish, L., Stuart, C., Tsay, J., and Herbsleb, J. *Social coding in GitHub: transparency and collaboration in an open software repository.* Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, (2012), 1277–1288.
34. Tajfel, H. and Turner, J. C. The social identity theory of intergroup behavior *Psychology of intergroup relations* eds. Worchel, S. and Austian, W. G. (Nelson-Hall, Chicago, IL), 1986, pp. 7–24.
35. Bohn, A., Feinerer, I., Hornik, K., and Mair, P. *Content-based social network analysis of mailing lists.* The R Journal, 3 (2011), 11–18.
36. Rizopoulos, D. ltm: An R package for latent variable modeling and item response theory analyses. Journal of Statistical Software, 17(5) (2006), 1–25.
37. Stefanski, L. A. and Cook, J. R. *Simulation-extrapolation: The measurement error jackknife.* Journal of the American Statistical Association, 90 (1995) 1247–1256.
38. Lederer, W. and Küchenhoff, H. *A short introduction to the SIMEX and MCSIMEX.* R News, 6 (2006) 26–31.

Materials and Methods

Sample. In total, we had 4274 email addresses of R package authors. They were asked to fill out an online questionnaire within the following three weeks. The survey was conducted in May 2010 using the online survey software Unipark. 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"). Note that if packages had multiple authors, emails were sent out to those who provided an email address in the package description file. In addition, in the email list we used some package authors had multiple email addresses. Therefore, the response rate below reflects a lower bound.

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 which leads to a response rate of at least 27%. This is in line with related OSS studies such as [15], [10], and [26]. 764 package authors completed the whole questionnaire without skipping any of the items. From a statistical power point of view this sample size is sufficiently large to carry out all of our statistical analyses. The issue of non-response bias is addressed and analyzed in detail in the SI. In turns out that our results are representative for R package authors who contributed to more than one package.

Reproducibility Materials. The following materials were submitted in order to fully reproduce the analysis in the article. The raw data are submitted in the file `RMotivationRaw.csv` along with the variable descriptions (`RMotivationRawLabels.txt`). The code file `CRANdatprep.R` contains the R code for data preparation including the IRT analysis, resulting in the file `RMotivation.rda`. This file is also checked-in separately since the IRT itemfit analysis takes a considerable amount of time. The file `CRANreplication.R` performs all the computations (GLM, examining non-response bias) including regression tables and effect plots presented in the SI.

Fig. 1. Psychometric Constructs. Hybrid forms of motivation [25], work design characteristics [21], and values [22] determining participation in the R project.

Supporting Information

Non-response Bias

Non-response bias is an issue that often occurs in email surveys, especially when the response rate is not particularly large. The bias that arises if the answers of survey respondents (the sample) differ from the potential answers of those not in the sample. In this case results can not be generalized to the whole population. There are several strategies that address non-response bias as elaborated in classical texts such as [27] and [28]. The strategy we use to address potential biases in our survey is that of comparing sample values of a variable with known values from the population. The key variable within this context that we use is number of packages to which each author contributed. Using this variable, we can determine the population values by extracting the author names from the package description files and then computing the corresponding frequencies.

In our sample, 31.15% of the authors contributed to one package only, whereas in the population we have 67.86%. This indicates that one-package authors are underrepresented in our sample. This is not surprising, since people who contributed only one package are likely to have a lower committment to the R project and, therefore, are less likely to fill out such a questionnaire. Let us examine the (conditional) relative frequencies of authors of two or more packages. Note that we merge authors with 10 or more packages into a single category. Figure S1 shows the percentages across the number of packages for the sample and the population. We see that two-package authors are only slightly underrepresented in the sample, and for the remaining ones the sample and population proportions match closely. Therefore, our results are representative for the subpopulation "package authors who contributed to at least two packages".

[INSERT FIGURE S1 (cranpnas-pop-plot1.eps) HERE]

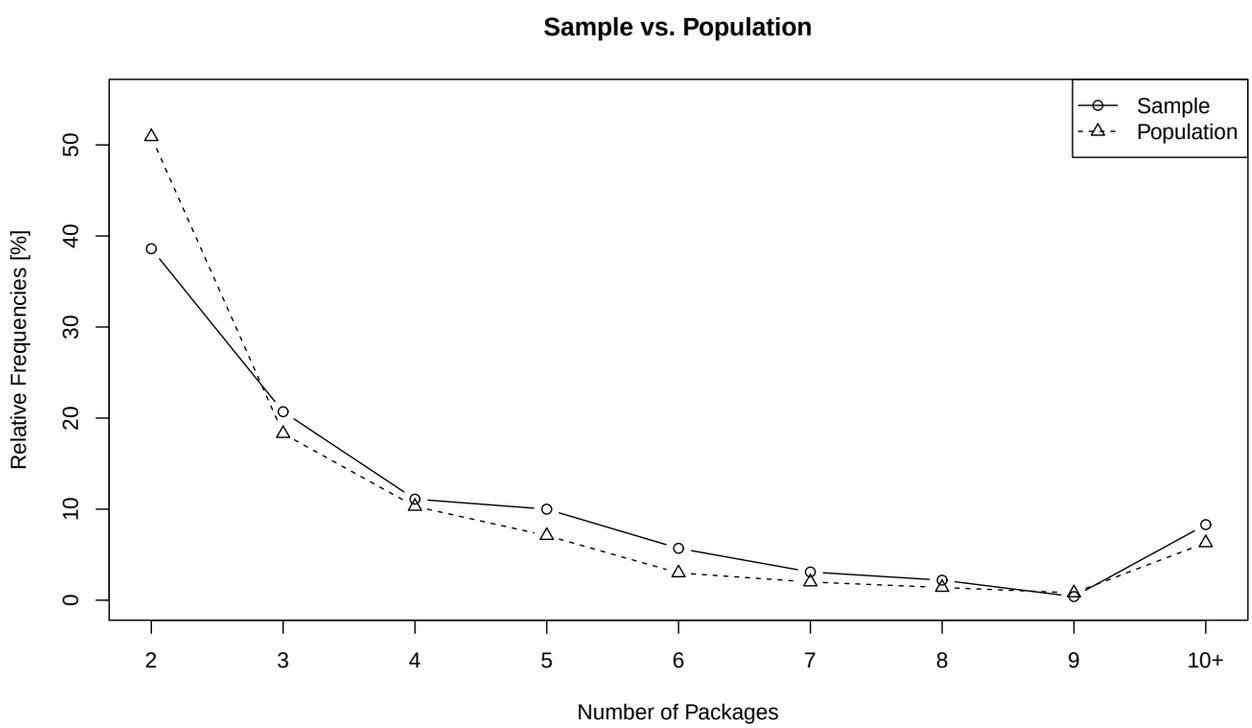

Fig. S2. Proportions of persons in the sample vs. the population for R package authors who contributed more than one package.

Statistical Methodology

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/work-related factors. We thus start by computing the psychometric scores as depicted in Figure 1 from the corresponding questionnaire items using the two-parameter logistic model. Subsequently, these scores are used as explanatory variables in GLM analysis for each of the three variables measuring participation. In order to account for potential measurement errors when psychometric scores enter the regression models, the estimators are corrected by the simulation-extrapolation (SIMEX). The following subsections give a brief methodological background.

IRT Analysis. The latent constructs occurring in the scales for hybrid forms of motivation, work design, and values are scored using IRT models. IRT is a family of latent variable models to score items and persons on a single latent trait. Our IRT model of choice is the two-parameter logistic model [29] defined as:

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

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 location parameter* δ_i that locates the item on the latent trait. Subsequently, for each subject v we estimate a *person parameter* θ_v that maps the package author on the latent trait.

We perform our IRT analysis separately for each scale dimension (with the items assigned to the dimensions correspondingly) using the R package **ltm** [36]. Before fitting an IRT model, we examined unidimensionality of each subscale using categorical PCA, implemented in the **homals** package [30]. The following items showed a strong deviation from unidimensionality: "Packages are a byproduct of my empirical research. If I cannot find suitable existing software to analyze my data, I develop software components myself" and "Packages are a byproduct of my methodological research. If I develop/extend methods, I develop accompanying software, e.g., for illustrations and simulations" from the motivation scale (extreme extrinsic motivation construct). In addition, "The work on R packages requires that I only do one task or activity at a time" had to be removed from knowledge characteristics.

All subsequent $Q1$ fit statistics were not significant and, therefore, no additional items were eliminated. Note that since we have a multiple testing problem, the alpha level was corrected by dividing 0.05 by the number of items per subscale. For our final item subsets we compute the person parameters for each of the nine traits. For subsequent analyses and tables, the resulting new 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.

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. A general representation of our model is

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

where $\boldsymbol{\mu}$ is the mean of the participation response variable, $g(\cdot)$ represents the corresponding link function, and $\boldsymbol{\Theta}$ is the matrix containing of person parameters with corresponding regression coefficients $\boldsymbol{\beta}$. $\mathbf{X}$ is the matrix of socio-demographic variables with corresponding regression coefficients $\boldsymbol{\gamma}$. 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. Binomial GLMs with logistic link function are used for the binary responses capturing participation in *mailing lists*, and *conferences*, respectively.

SIMEX Correction. Note that, unlike the socio-demographic variables in $\mathbf{X}$, the person parameters in $\boldsymbol{\Theta}$ are subject to measurement error (ME) as they are obtained from IRT analysis. Due to this ME the ordinary GLM estimates are, in general, biased. To mitigate this problem, the heteroskedasticity of the MEs needs to be taken into account.

Let $\boldsymbol{\beta}$ be the true value of the parameter vector and $\hat{\boldsymbol{\beta}}$ the estimated regression coefficients. In order to get unbiased estimates in the presence of additive MEs, we apply the simulation-extrapolation method (SIMEX) proposed by [32] after fitting the basic ("naive") GLMs.

For our specific problem we apply the jackknife variant of SIMEX [37] 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 MEs across persons. In our analysis we allow for full ME heteroskedasticity (across constructs, across persons) which leads to the ME matrix $\boldsymbol{\Sigma}_\Theta$ 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 $\hat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}_\Theta}$. SIMEX theory states that the mean of the parameter distribution resulting from resampling, that is, $\bar{\hat{\boldsymbol{\beta}}}_{\boldsymbol{\Sigma}_\Theta}$, is an unbiased estimator for $\boldsymbol{\beta}$ [37]. A corresponding R implementation is provided in the **simex** package [38].

Results

Descriptive Data Analysis. The first dependent variable measuring participation is the number of packages (co-)developed by an individual author. Its distribution is right-skewed, has a mean of 2.9, a median of 2, and maximum of 33, and a standard deviation of 3.45.

[INSERT FIGURE S2 (cranpnas-npkgs-plot1.eps) HERE]

Figure S2 shows the distribution of the number of packages. A few package authors stated that they have been involved in zero packages. The reason for this could be that they contributed code to a particular package, appear in the author list, but do not consider themselves being involved in the development of this particular R package (e.g. authors that are active on R-forge only). The other two dependent 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 psychometric scores using IRT analysis as described above. Our dichotomous work-related variables give the following descriptive results : PhD degree (*phd*, yes: 71.47%), education in statistics (*statseduc*, yes: 63.09%), employed full time (*fulltime*, yes: 85.21%), work in academia (*academia*, yes: 60.47%), work as statisticians (*statswork*, yes: 63.22%).

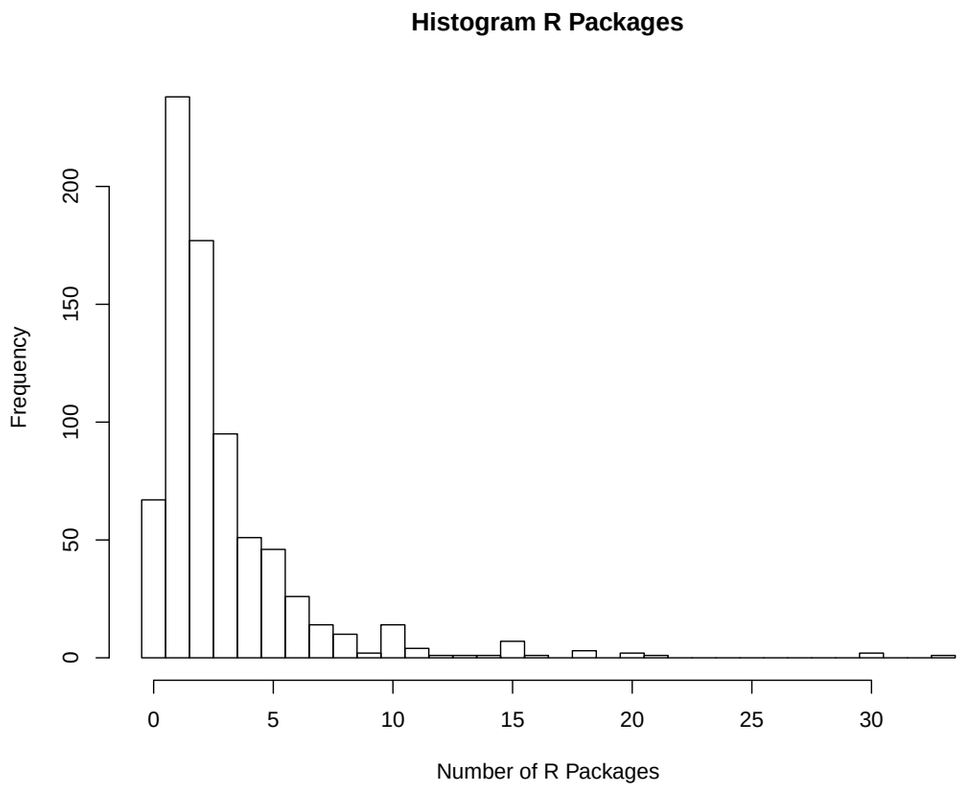

Fig. S3. Distribution of the number of packages the authors are involved in.

Regression Tables and Effects Plots. The following tables and plots show the results of the three GLMs. The first table refers to the negative binomial regression with "number of packages" as response, the second table to the logistic regression with "participation in mailing lists" as response, and the third table to the logistic regression with "participation in conferences" as response. The effects plots depict the effect structure based on the regression parameters for the for the predictors selected by stepwise regression.

[INSERT TABLE S1 (table1.tex) HERE]

[INSERT FIGURE S3 (cranpnas-npkgsglm-plot1.eps) HERE]

[INSERT TABLE S2 (table2.tex) HERE]

[INSERT FIGURE S4 [cranpnas-listsglm-plot1.eps] HERE]

[INSERT TABLE S3 (table3.tex) HERE]

[INSERT FIGURE S5 [cranpnas-meetglm-plot1.eps] HERE]

Table S1. Negative Binomial GLM Parameter Estimates for "Number of Packages" (standard errors in brackets; significance codes *** for the 0.001 level, ** for the 0.01 level, and * for the 0.05 level).

	Full (ML)	Full (SIMEX)	Step (ML)	Step (SIMEX)
(Intercept)	0.597^{***}	0.607^{***}	0.654^{***}	0.661^{***}
	(0.126)	(0.124)	(0.120)	(0.117)
wtask	-0.169^{**}	-0.293^{***}	-0.172^{**}	-0.299^{***}
	(0.054)	(0.072)	(0.054)	(0.073)
wsocial	0.323^{***}	0.490^{***}	0.328^{***}	0.505^{***}
	(0.055)	(0.078)	(0.055)	(0.077)
wknowledge	−0.068	−0.100	−0.074	−0.109
	(0.049)	(0.066)	(0.049)	(0.067)
mextrinsic	0.079	0.132	0.062	0.114
	(0.052)	(0.075)	(0.051)	(0.074)
mhybrid	0.159^{**}	0.221^{**}	0.174^{***}	0.234^{***}
	(0.058)	(0.078)	(0.049)	(0.062)
mintrinsic	0.026	0.017		
	(0.063)	(0.093)		
vuniversalism	−0.057	−0.088		
	(0.055)	(0.082)		
vpower	-0.159^{**}	-0.301^{**}	-0.163^{**}	-0.306^{***}
	(0.059)	(0.093)	(0.058)	(0.089)
vselfdirection	−0.008	−0.010		
	(0.072)	(0.119)		
phdyes	0.127	0.134		
	(0.095)	(0.097)		
statseducyes	0.023	0.003		
	(0.101)	(0.100)		
fulltimeyes	0.328^{**}	0.279^{*}	0.364^{**}	0.313^{**}
	(0.120)	(0.118)	(0.117)	(0.114)
academiayes	-0.165^{*}	-0.173^{*}	−0.144	−0.146
	(0.083)	(0.084)	(0.080)	(0.078)
statsworkyes	0.157	0.181	0.182^{*}	0.192^{*}
	(0.101)	(0.101)	(0.080)	(0.079)

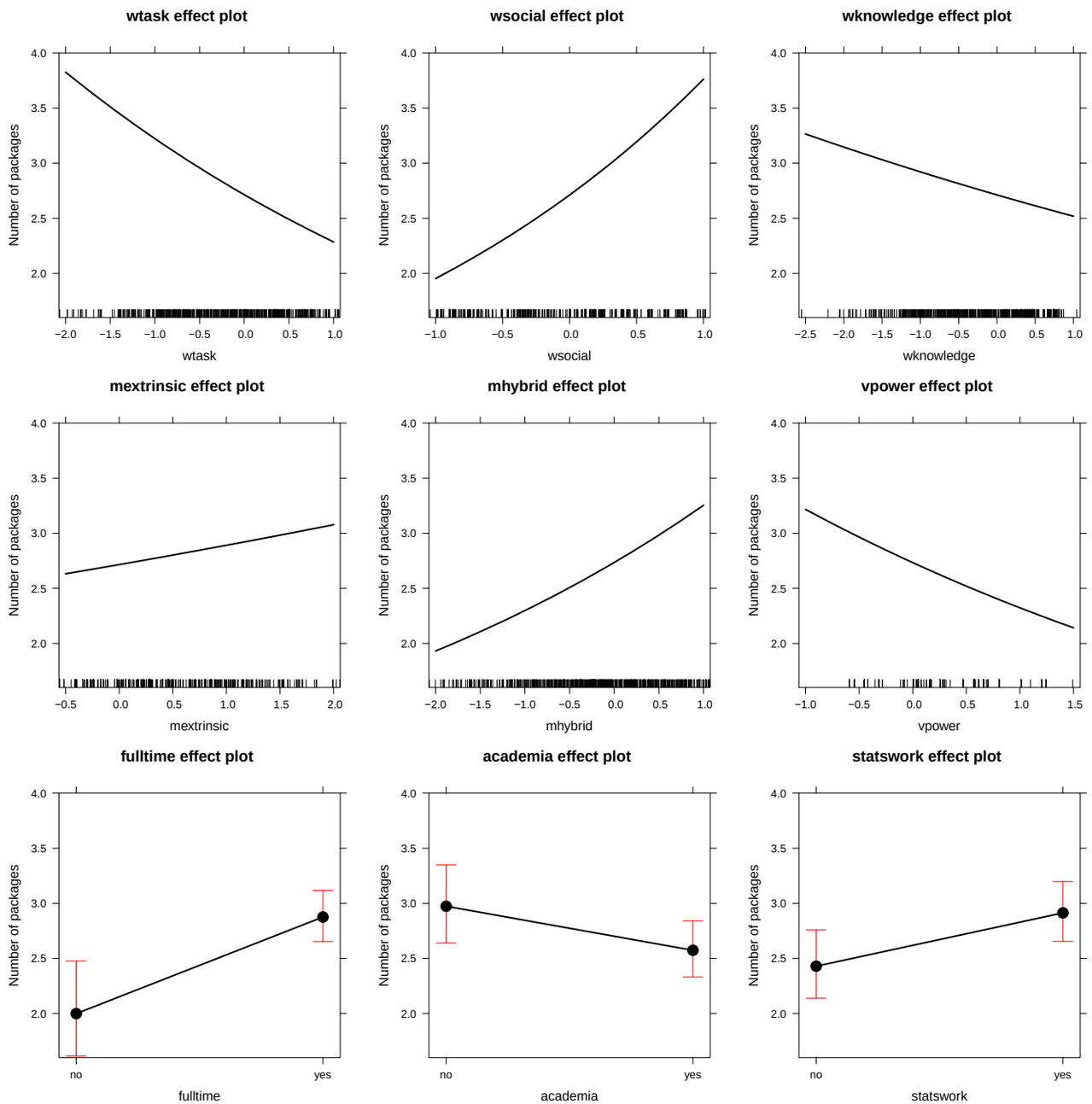

Fig. S4. Effect plots for negative-binomial regression on number of packages (for the variables selected by stepwise regression).

Table S2. Logistic GLM Parameter Estimates for "Participation in Mailing Lists" (standard errors in brackets; significance codes *** for the 0.001 level, ** for the 0.01 level, and * for the 0.05 level).

	Full (ML)	Full (SIMEX)	Step (ML)	Step (SIMEX)
(Intercept)	0.979^{***}	1.095^{***}	0.870^{***}	0.884^{***}
	(0.249)	(0.261)	(0.157)	(0.161)
wtask	-0.255^{*}	-0.469^{**}	-0.237^{*}	-0.425^{**}
	(0.112)	(0.154)	(0.111)	(0.154)
wsocial	0.429^{***}	0.676^{***}	0.421^{***}	0.661^{***}
	(0.114)	(0.160)	(0.113)	(0.154)
wknowledge	0.114	0.152		
	(0.101)	(0.141)		
mextrinsic	-0.361^{**}	-0.623^{***}	-0.344^{**}	-0.543^{***}
	(0.110)	(0.160)	(0.105)	(0.154)
mhybrid	0.443^{***}	0.580^{***}	0.435^{***}	0.559^{***}
	(0.119)	(0.166)	(0.114)	(0.158)
mintrinsic	0.211	0.267	0.216	0.298
	(0.128)	(0.191)	(0.127)	(0.197)
vuniversalism	−0.056	−0.119		
	(0.116)	(0.180)		
vpower	0.046	0.114		
	(0.120)	(0.191)		
vselfdirection	0.043	0.099		
	(0.149)	(0.250)		
phdyes	−0.161	−0.285		
	(0.195)	(0.205)		
statseducyes	−0.289	−0.349		
	(0.208)	(0.215)		
fulltimeyes	0.073	0.025		
	(0.234)	(0.242)		
academiayes	−0.217	−0.196	−0.259	−0.286
	(0.172)	(0.177)	(0.160)	(0.164)
statsworkyes	-0.416^{*}	−0.322	-0.589^{***}	-0.562^{***}
	(0.209)	(0.215)	(0.165)	(0.170)

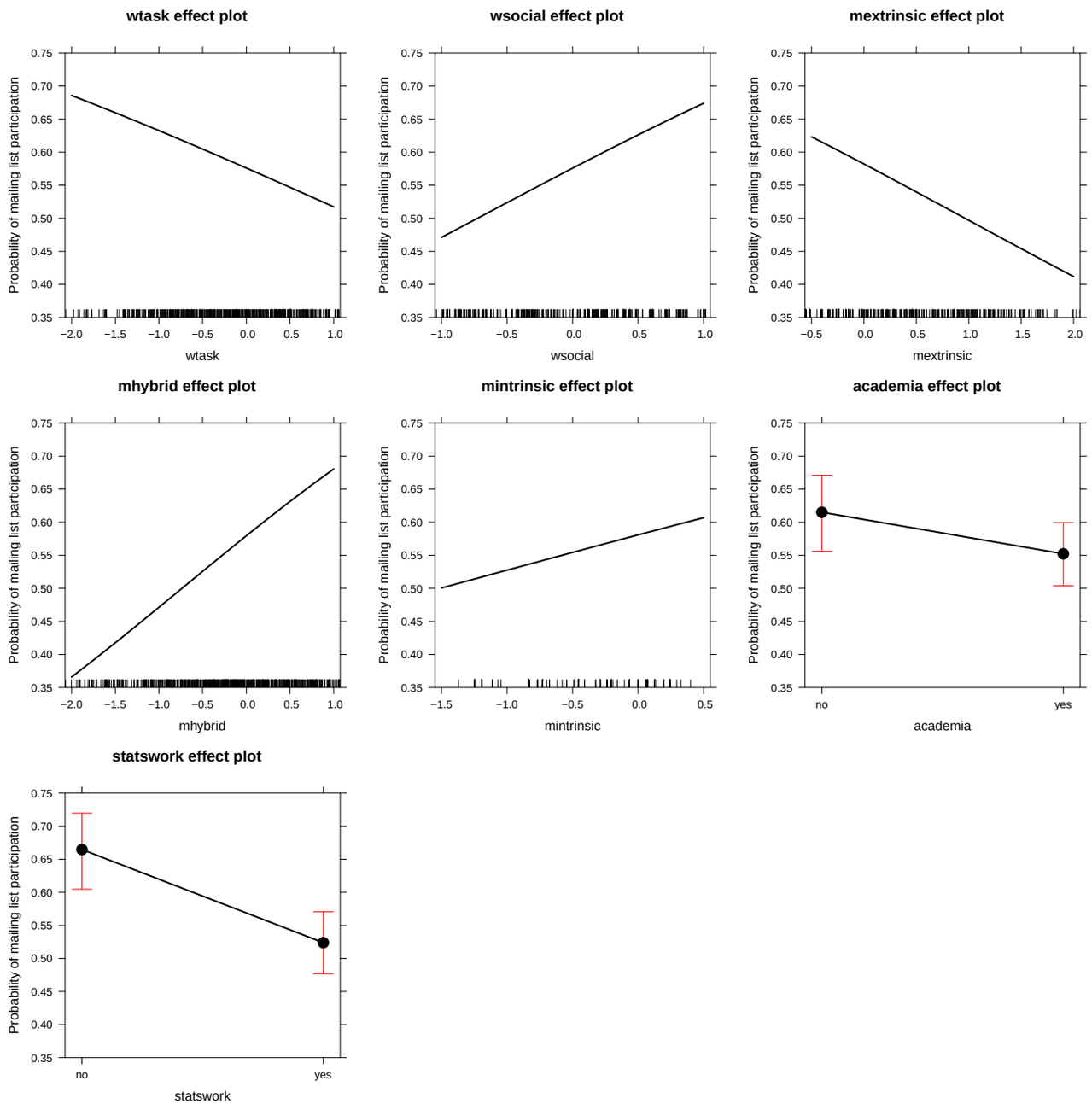

Fig. S5. Effect plots for logistic regression on lists participation (for the variables selected by stepwise regression).

Table S3. Logistic GLM Parameter Estimates for "Participation in Conferences" (standard errors in brackets; significance codes *** for the 0.001 level, ** for the 0.01 level, and * for the 0.05 level).

	Full (ML)	Full (SIMEX)	Step (ML)	Step (SIMEX)
(Intercept)	-1.585^{***}	-1.585^{***}	-1.613^{***}	-1.634^{***}
	(0.283)	(0.289)	(0.267)	(0.272)
wtask	-0.069	-0.129		
	(0.118)	(0.162)		
wsocial	0.458^{***}	0.714^{***}	0.433^{***}	0.650^{***}
	(0.120)	(0.171)	(0.116)	(0.165)
wknowledge	0.026	0.035		
	(0.106)	(0.148)		
mextrinsic	-0.062	-0.109		
	(0.114)	(0.161)		
mhybrid	0.228	0.257	0.276^{**}	0.341^{**}
	(0.124)	(0.176)	(0.098)	(0.122)
mintrinsic	0.119	0.147		
	(0.137)	(0.207)		
vuniversalism	-0.218	-0.421^{*}	-0.238^{*}	-0.432^{*}
	(0.119)	(0.176)	(0.116)	(0.175)
vpower	0.126	0.280		
	(0.126)	(0.189)		
vselfdirection	-0.055	-0.123		
	(0.155)	(0.261)		
phdyes	-0.085	-0.104		
	(0.204)	(0.215)		
statseducyes	0.128	0.082		
	(0.217)	(0.220)		
fulltimeyes	0.714^{**}	0.673^{*}	0.623^{*}	0.603^{*}
	(0.270)	(0.275)	(0.254)	(0.258)
academiayes	-0.165	-0.122		
	(0.180)	(0.185)		
statsworkyes	0.271	0.282	0.313	0.300
	(0.218)	(0.222)	(0.170)	(0.172)

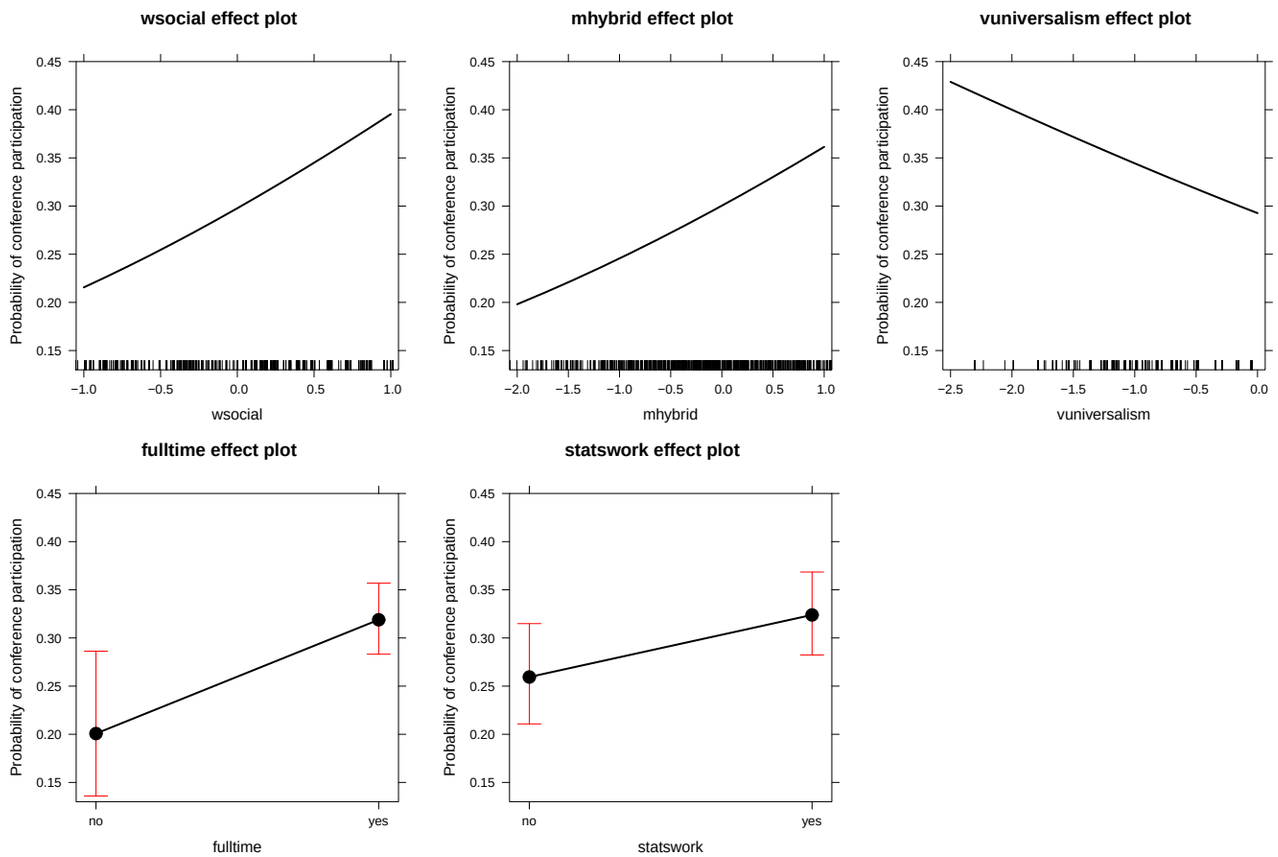

Fig. S6. Effect plots for logistic regression on conference participation (for the variables selected by stepwise regression).

Fig. S1. Sample vs. population proportions for the number of packages. The percentages are based on the conditional relative frequencies (i.e. conditional on authors with more than one packages).

Fig. S2. Distribution of the number of packages the authors are involved in.

Fig. S3. Effect plots for negative-binomial regression on number of packages (for the variables selected by stepwise regression).

Fig. S4. Effect plots for logistic regression on lists participation (for the variables selected by stepwise regression).

Fig. S5. Effect plots for logistic regression on conference participation (for the variables selected by stepwise regression).

Questionnaire. 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.

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 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

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.

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.

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)

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.

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.

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 ()
- RForge ()
- 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, . . .)

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
- . . .