A Large Scale Survey of Motivation in Software Development and Analysis of its Validity - Survey Validity

Idan Amit idan.amit@mail.huji.ac.il
Dror G. Feitelson feit@cs.huji.ac.il
Department of Computer Science
The Hebrew University, Jerusalem 91904, Israel

September 29, 2024

1 Analysis of Validity and Reliability

In the previous analysis we analyzed the data as if it is completely reliable. However, the reliability might be limited in many ways. Since the data is given, what we do in this section is to evaluate its reliability from various aspects.

When using the answers of participants, one should check which population they represent. We compare our survey demographics to the demographics of the Stack Overflow survey, answered by around 80 thousand developers worldwide. Distributions are not similar yet close. We used two channels in order to reach participants: direct emails to developers contributing at GitHub and social media. We build supervised learning trying to differ them, whose performance was no better than the positive rate, indicating no big obvious difference. These are detailed in the supplementary materials.

We compare our results to other surveys, the investigate their agreement (Section 1.1). We grouped questions into motivators based on their content, regardless of the answers to them. In Section 1.3 we examine the coherence of the motivators and compare them to grouping based on the answers. The follow-up survey allows us to evaluate the stability of answers, comparing a person's answer in two different dates (Section 1.4). It also provides an additional dataset on which we can check the degree in which our results reproduce. Last but not least, we investigate reliability in the answers themselves, from typos to mistakes and biases (Section 1.2).

1.1 Surveys' Hit Rate Comparison

In Table 1 we compare our survey to prior work on motivation. Most prior work provides hit rates of motivators, forcing us to use the same metric for the

comparison. Note that hit rate indicates how common motivators are and not their relation to general motivation. Moreover, some works provide only the ranking of motivators. In case the hit rate itself was provided, we present it in parentheses. Gap is the difference between the minimal and maximal rank of a motivator in the different surveys, serving as a diversity metric. Also note that in some cases different names or even just overlapping concepts were used in the prior work, detailed in the replication package.

Note that the surveys are from different years, from 1978 to 2022. Herzberg investigated the general population, Fitz and Couger investigated IT personnel, Gerosa investigate open-source developers, and we investigated developers in general, many of which were open-source developers. The instruments were also different in most cases. However, if the surveys generalize to describe a universal behavior, they should agree.

Table 1: Comparison of Motivators Hit Rates and their Ranks in Different Surveys

Motivator	**Avg. Rank**	**Fitz [8]**	**Couger [6]**	**Herzberg [11]**	**Gerosa [9]**	**Our**	**Follow-up**	**Gap**
Enjoyment	2	2	1	3 (22)	3 (18.9)	1 (74)	2 (71)	2
Challenge	2.60	1	2	1 (40)		5 (62)	4 (64)	4
Ownership	3.80	6	6	4 (20)		2 (73)	1 (76)	5
Learning	4	5	7	5 (5)	1 (22.6)	3 (72)	3 (69)	6
Importance	5					4 (63)	6 (59)	2
Self-use	5				2 (21)	6 (53)	7 (53)	5
Recognition	5.17	3	5	2 (30)	5 (10.6)	8 (48)	8 (46)	6
Ideology	5.67				6 (8.8)	6 (53)	5 (61)	1
Payment	6.50	8	4	5 (5)	4 (16.6)	9 (45)	9 (37)	5
Community	8	7	8	7 (4)	6 (8.8)	10 (41)	10 (32)	4
Hostility	11					11 (7)	11 (6)	

We computed the Pearson correlation between the average ranks in all surveys and the ranks in each individual survey. The result was above 0.8 for Fitz and for our survey, about 0.7 for Couger and Herzberg, and 0.64 for Gerosa. Gerosa is closer to us in time and population so a higher correlation is expected. Our ranking and our followup ranks had a correlation of 0.96, Herzberg and Fitz had a correlation of 0.87 and the rest were much lower. For example, Couger that reproduced the work of Fitz 10 years later, had a correlation of 0.67 with it.

When looking at the hit rates themselves, even just browsing the table shows that in general the distributions are quite different. We get a 0.97 correlation

between our survey and our follow-up. Gerosa has 0.65 correlation with ours, 0.51 with the follow-up. Herzberg has 0.24 correlation with ours, 0.38 with the follow-up, and -0.21 with Gerosa.

Enjoyment is ranked high in all surveys, with a small gap between the maximal and minimal rank, *indicating a stable high result*. Community and payment are rather low, and not ranked high in any survey. Ownership, Recognition and Learning have large gaps, hence results regrading them in one survey do not generalize well to the others. It might be possible achieve higher agreement by applying transformations considering different times (e.g., importance of payment is reduced) or different populations (e.g., payment is less important in open source). However, the disagreement is not surprising since Couger [6] already indicated the people in different position report different motivators, and so do people from different countries Herzbereg [11].

1.2 Face Validity of Answers

To check the validity of the answers in our survey, we looked for mistakes, insincere answers, and biases.

The answer to the gender question was a free text field, in which the participant could write any answer. Only 4 (0.8%) of the answers had a typo (e.g., 'mail', 'boi'). Three of the answers were variants of 'Attack Helicopter', a term "used to disparage transgender people"[1]. Hence, these answers were probably not sincere. In the country question, 1.2% of the answers had a typo.

1.3% of the developers said they had 15 years of experience with GitHub, established in 2008, which was impossible when the survey ended in 2021. A single answer (0.2%) of age of 100 years is probably insincere. Note that these error rates are much better than the 8.5% who seemed to have given a wrong answer to a single simple question in [10], and the 10% failure to identify negatively worded (reverse-coded) items discussed in [15].

The job satisfaction questions were taken from a survey of 9,900 Australian clinical medical workers published in 2011 [12]. Amusingly, software developers were on average less satisfied in all questions. More importantly, questions about payment are irrelevant to volunteers and questions about community are irrelevant to people working alone. We explicitly asked to skip these questions if they are irrelevant. However, 57.7% of the people that answered that they are not paid, answered the payment satisfaction question. Therefore, it seems they answered regarding their salary from a different job, not related to the discussed project. Some participants made comments in the open question that support this.

Some developers answered that they work on a public GitHub project. For these projects, we checked the number of developers who committed code. Of five developers who were found to work on single person projects, one answered most of the community-related questions, which are irrelevant to such projects.

There were questions in which the change in the follow-up survey is known

[1] https://en.wikipedia.org/wiki/I_Sexually_Identify_as_an_Attack_Helicopter

in advance: age and experience should grow linearly with time. The follow-up question was about a year after the first one. In order to avoid rounding mistakes (e.g., a 20.5 years old participant might answer either 20 or 21), we consider answers as "unreasonable" only if the follow-up answer was more than a year lower, or at least three years higher. 26% of the answers about experience exhibited such unreasonable differences. Two participants lost 5 years of experience each, somehow compensated by a participant that gained 11 years of experience in about a year. 16% of the answers regarding GitHub experience were unreasonable. However, for age, which has a higher presence in daily life, there were no unreasonable differences.

It seems that the biggest reliability problem comes from human failings [15], bias due to ego defenses[5], or the Dunning–Kruger effect (that people with lower capabilities tend to have higher self-esteem) [13]. Only 5.6% of the participants gave a low answer to 'My code is of high quality', going up to 20.8% when including neutral answers (6 on the 11-point scale). The Pearson correlation with years of experience, a common method to estimate skill [7], was a very low 0.06. Moreover, first degree holders gave answers averaging 9.05, higher than all others. People trained in computer science gave answers averaging 9.3, lower than the 10.5 average in math, yet a bit higher than arts (9.0), science (8.7), technology (8.6), and business (7.5).

Using the participants' GitHub profile, we can compare their actual activity to their self-perception. People that answered that they write detailed commit messages (at least 9 - 'somewhat agree'), had average commit message length of 89 characters, placing them in the 61 percentile of GitHub developers, not very far from the median. Participants saying that they write high quality code have corrective commit probability (CCP) [2] of 0.36 (investing more than a third of their work in bug fixing), worse than 81% of the GitHub developers.

It seems that there is also a bias leading to higher answers about the participant than about the community. The average answer for questions about themselves is 9.1, 24% more than the average answer to questions about the project. A somewhat smaller difference of 4.5% to 17.1% was found when the questions were essentially paired (e.g, 'My code is of high quality' and 'The quality of the code in this repository is better than others').

We cannot accurately aggregate the probability of mistakes. The probability of identifying an insincere answer is low, around 0.2%. Mistakes typically occur in few percent of the answers, yet in specific questions (the job satisfaction in our case) might be around 50%. Biases are the largest threat to validity, as demonstrated by the 79% participants that consider their code to be of high quality. Only 5.6% of the participants gave a low answer to 'My code is of high quality', and 20.8% when including neutral. The 5.6% that think that their code is of low quality seem to be either more modest or more realistic.

1.3 Internal Coherence of Motivators

The motivators were represented in the survey by one or more questions each. The use of multiple questions (e.g. for 'community') allows us to treat them

as labeling functions of the same concept and evaluate their agreement [16, 4]. The agreement, measured by the average Pearson correlation of the related questions, reflects the internal coherence of these motivators. Low coherence might be due to our subjective grouping of questions or due to human nature.

As a reference of the level of correlation that we can expect, we focus on closely related question pairs. For example, 'I am skilled in software development' has a correlation of just 0.62 with 'My code is of high quality'. 'I regularly reach a high level of productivity' and 'I am a relatively productive programmer' have correlation of just 0.57. Table 3 shows that the correlation of the same person answers to the motivation question in the original and follow-up survey is 0.52. Note that a correlation of 0.5 is even higher than the correlation between LOC count and step functions on it [1]. Therefore, coherence of about 0.6 is high.

Table 2: Motivator Coherence

Motivator	**Coherence**	**Follow-up Coherence**
All Questions	0.11	0.07
Community	0.36	0.15
Enjoyment	0.32	0.25
Hostility	0.49	0.60
Ownership	0.58	0.57
Recognition	0.24	0.17

Table 2 presents the coherence of the motivators. 'Coherence' is defined as the average Pearson correlation between all pairs of questions related to the same motivator. The motivators 'Challenge', 'Ideology', 'Importance', 'Learning', 'Payment', and 'Self-use' do not appear in Table 2 since they are based on a single question each, hence our method is not applicable to them. 'Follow-up Coherence' is the same metric as 'Coherence' computed on the follow-up survey. Note that this provides additional support, yet the support is not totally independent since the participants in the follow-up survey also participated in the original one.

The 'All Questions' row represents all the questions together (basically related to motivation) and has rather low coherence. The following motivators all have much higher coherence, indicating that they indeed reflect a meaningful grouping of questions related to specific concepts. The coherence of 'Hostility' and 'Ownership' is relatively high in both surveys, and close to the highest coherence we can expect. The coherence of 'Community', 'Enjoyment', and 'Recognition' is between 0.15 to 0.36 in both surveys.

The creation, selection and grouping of questions to motivators was done by the authors. In case of questions from prior work, we had indications of the intended motivators. Disagreements were discussed and resolved. Though the taxonomy is justifiable and fits our needs, we are aware of the justification of other ones, some considered by us. As a benchmark for our taxonomy we compare it to one *created using the answers, based on actual relations between*

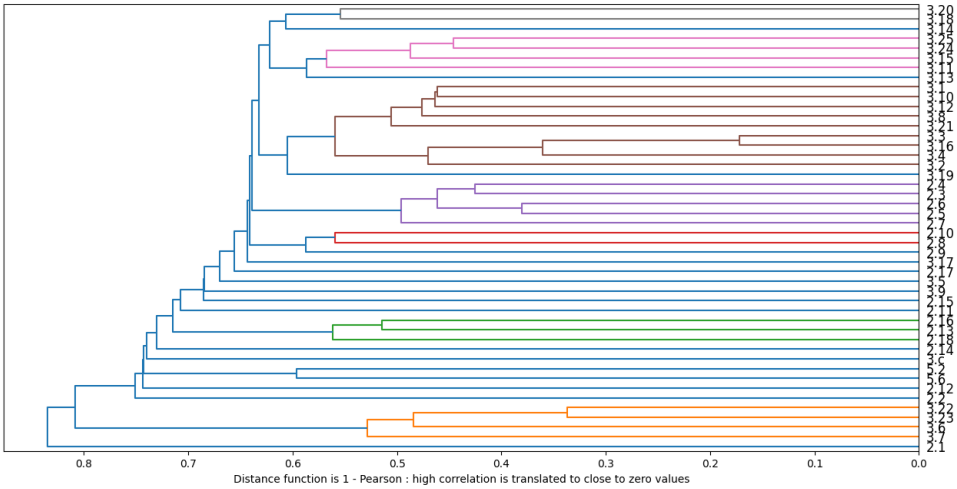

Figure 1: Dendrogram of questions based on their Pearson correlation. See questions text in [3]

them. We build this taxonomy using automatic clustering based on their answer correlations. This will provide additional evidence on whether our grouping was indeed meaningful.

The dendrogram in Figure 1 represents a hierarchical clustering [14] of questions based on the correlations between them. The questions 'I have significant influence on the repository' (3.3) and 'I am a core member of the repository' (3.16) are the most correlated questions with Pearson of 0.83 (transformed to $1 - 0.83 = 0.17$ to represent distance in the figure). As we allow weaker correlations, more questions are clustered together, and when we allow Pearson correlation of only 0.3 most questions are already grouped into one big cluster, which is uninteresting. Some of the clusters match our content-based motivators: The orange cluster matches hostility and the lower brown sub-cluster matches ownership. The purple cluster shows a group that we did not consider: productivity and possible productivity improving elements such as good colleagues, physical conditions, and opportunities to use your abilities. Other clusters may overlap our definitions, but also mix in unrelated questions. For example, the pink cluster contains 3 questions about recognition and one about importance, which we feel is not really related. Since our manually built factors are more coherent with respect to content, we use them and not the hierarchical clusters.

1.4 Answers Stability Between Original and Follow-up Surveys

The follow-up survey, conducted one year after the original survey, allowed us to compare the answers of the same person over time. Table 3 shows stability of questions by motivator.

To compare the answers in the two surveys we first compute the Pearson correlation between them. We also compute the differences between them, both the average absolute average difference (column 'Avg. Abs. Diff') and the average relative difference (difference divided by the question value, column 'Avg. Rel. Diff'). 'Pred(25)' [17] is the probability that the follow-up answer is in the range of 25% of the initial answer.

Note that the distributions of answers are far from uniform, and some answers are much more popular than others. As a result, there is a high probability for getting the same answer even when the answers are independent. 'Pred(25) Lift' computes the lift, i.e. the extra probability above the expected Pred(25) from two independent answers from the answers distribution.

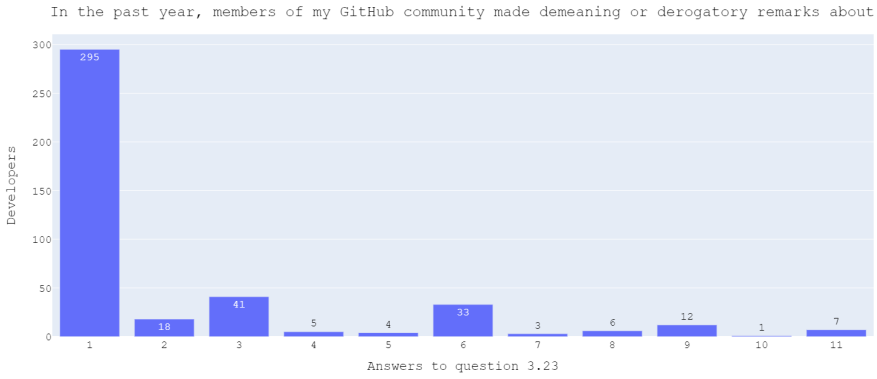

Figure 2: Answers distribution of a hostility question.

Pearson correlation, Pred(25), Pred(25) lift, and relative difference indicate stability for almost all motivators. Hostility has a near zero lift and not a large positive one, indicating less stability than expected. Note that the hostility distribution (e.g. Figure 2) has a strong mode in the lowest value, making the independent distribution benchmark very high. Note also that the lift is close to zero hence more likely to be influenced by noise.

Payment is a binary feature hence its stability should be analyzed with different metrics. The initial and follow-up payment agree in 85% of the cases. 70% of those that were paid in the initial survey were also paid a year later. Only a single person out of the 27 that were not paid in the initial survey got payment in the follow-up.

The Pearson correlations are between 0.42 to 0.68. Though, over time the project, the people, and their motivations change [9], which might result in

Table 3: Similarity of Motivation Type Answers of Same Person in Two Dates

Motivator	Pearson	Avg. Abs. Diff	Avg. Rel. Diff	Pred(25)	Pred(25) Lift
Learning	0.68	0.91	0.04	0.81	0.22
Ownership	0.66	1.02	-0.01	0.83	0.42
Hostility	0.63	1.10	0.38	0.38	-0.02
Enjoyment	0.60	1.06	0.00	0.84	0.29
Ideology	0.57	1.61	0.02	0.74	0.99
Importance	0.54	1.28	0.02	0.74	0.38
Motivation	0.52	1.83	-0.03	0.60	0.30
Challenge	0.51	1.46	0.08	0.70	0.23
Community	0.48	1.46	0.04	0.44	0.04
Recognition	0.45	1.70	0.19	0.54	0.36
Self-use	0.43	2.35	0.04	0.51	0.20

different answers.

References

[1] Amit, I., Ezra, N. B., and Feitelson, D. G. Follow your nose – which code smells are worth chasing?

[2] Amit, I., and Feitelson, D. G. Corrective commit probability: A measure of the effort invested in bug fixing. *Software Quality Journal 29*, 4 (Aug 2021), 817–861.

[3] Amit, I., and Feitelson, D. G. Replication package https://github.com/evidencebp/motivation-survey, Dec 2023.

[4] Amit, I., Firstenberg, E., and Meshi, Y. Framework for semi-supervised learning when no labeled data is given. U.S. patent #US11468358B2, 2017.

[5] Bassett-Jones, N., and C. Lloyd, G. Does herzberg's motivation theory have staying power? *Journal of Management Development 24* (Dec 2005), 929–943.

[6] Couger, J. D. Motivators vs. demotivators in the is environment. *Journal of Systems Management 39*, 6 (1988), 36.

[7] Feigenspan, J., Kästner, C., Liebig, J., Apel, S., and Hanenberg, S. Measuring programming experience. In *20th IEEE International Conference on Program Comprehension* (2012), pp. 73–82.

[8] Fitz-Enz, J. Who is the dp professional. *Datamation 24*, 9 (1978), 125–128.

[9] Gerosa, M., Wiese, I., Trinkenreich, B., Link, G., Robles, G., Treude, C., Steinmacher, I., and Sarma, A. The shifting sands of motivation: Revisiting what drives contributors in open source. In *Proceedings of the 43rd International Conference on Software Engineering* (2021), pp. 1046–1058.

[10] Herbold, S., Trautsch, A., Ledel, B., Aghamohammadi, A., Ghaleb, T. A., Chahal, K. K., Bossenmaier, T., Nagaria, B., Makedonski, P., Ahmadabadi, M. N., et al. A fine-grained data set and analysis of tangling in bug fixing commits. *Empirical Software Engineering 27*, 6 (2022), 1–49.

[11] Herzberg, F. One more time: How do you motivate employees? *New York: The Leader Manager* (1986), 433–448.

[12] Hills, D., Joyce, C., and Humphreys, J. Validation of a job satisfaction scale in the australian clinical medical workforce. *Evaluation & the Health Professions 35* (Mar 2011), 47–76.

[13] Kruger, J., and Dunning, D. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology 77* (1999), 1121–1134.

[14] Murtagh, F., and Contreras, P. Algorithms for hierarchical clustering: an overview. *WIREs Data Mining and Knowledge Discovery 2*, 1 (2012), 86–97.

[15] Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., and Podsakoff, N. P. Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology 88*, 5 (2003), 879.

[16] Ratner, A. J., De Sa, C. M., Wu, S., Selsam, D., and Ré, C. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 3567–3575.

[17] Wen, J., Li, S., Lin, Z., Hu, Y., and Huang, C. Systematic literature review of machine learning based software development effort estimation models. *Information and Software Technology 54*, 1 (2012), 41–59.