Erasmus School of Economics

International Bachelor Economics and Business Economics (IBEB)

Bachelor Thesis

The value of maintaining diversity when selecting experts

Name: Yari Baars

Student ID: 492810

Supervisor: A. C. Peker

Second Assessor: D. R. Gonzalez Jimenez

Date of final version: 22-07-2021

Abstract

Ever since Francis Galton (1907) discovered that the aggregated estimate of the crowd can outperform individual estimates, the 'wisdom of crowds'-concept has been replicated many times. Recently, research has explored the selection of experts within the crowd. In this paper, I intend to investigate the role diversity plays within the 'wisdom of crowds'-concept. I propose a method, which preserves diversity while selecting experts. This is done by selecting diverse samples from a subsample with a baseline level of expertise based on one demographic characteristic at a time.

In the paper, the estimation ability of 162 participants is analysed. First, I test the benchmark performance of participants selected solely on past performance. The results show that this group does not outperform the crowd. Furthermore, I research the predictive power of identity diversity. Selecting identity diverse samples, based on a single demographic characteristic at a time, does not lead to significantly lower correlations for the diverse samples when compared to homogenous groups. Lastly, the selection of a diverse sample with a baseline level of expertise, does not outperform the expert group. The findings are in line with De Oliveira and Nisbett (2018) and suggest a more complex definition of diversity in future research.

Table of Contents

1. Introduction

In 1907, Francis Galton discovered something rather peculiar during his visit to the Stock and Poultry Exhibition in Plymouth. Visitors to the exhibition could participate in a competition to guess the weight of a dead ox. They could buy a ticket for half a shilling on which they would write their guess. The participants with the most accurate guesses received prizes. Because they paid an entry fee, the participants were likely trying to guess correctly. Nevertheless, most of them were no experts on the topic. After the competition had finished, Galton decided to collect all the tickets. When he aggregated the guesses on all these tickets, he found that the median guess of the 787 participants was only 9 pounds, or 0.8 percent, higher than the actual weight. This finding formed the basis for a new field of research.

The accuracy of the aggregated guesses of a group of people has been replicated many times (Armstrong, 2001; Clemen, 1989; Surowiecki, 2004). In 2004, Surowiecki named the phenomenon 'The Wisdom of Crowds'. Since then, researchers have focused on different subtopics such as the selection of experts within a crowd using a variety of methods (Budescu & Chen, 2015; Mannes et al., 2014); Mellers et al., 2015). The intuition is that some people are better at estimating than the rest of the crowd. By focusing on these 'experts', we can get an even more accurate aggregated estimate. This can either be done by assigning different weights to all the participants (Budescu & Chen, 2015) or by excluding the rest of the crowd entirely (Mannes et al., 2014).

Selecting expertise has the potential to improve the accuracy of aggregated estimates. However, by focusing solely on participants who signal competence, either by past performance or another method, the participants in the selected group naturally become very similar (Hong & Page, 2004) (Broomwell & Budescu, 2009). This loss of diversity is a problem as Hong and Page (2004) show that more diverse crowds tend to outperform homogenous crowds.

All in all, selecting expertise has potential to improve accuracy of estimates, but the loss of diversity is a concern. To analyse the role of diversity when selecting expertise, my research question is the following:

"*How does focusing on diversity when selecting expertise in a crowd affect accuracy of the aggregated estimates made by the selected experts?*"

This paper builds on the existing literature by focusing on the accuracy-diversity trade-off in a more practical way compared to Davis-Stober et al. (2014). Their paper builds a mathematical model to determine the trade-off between inter-judge correlation and accuracy. This paper focuses on a participant's observable characteristics instead. Furthermore, the role of diversity in the accuracy of

crowds is still unclear. Hong and Page (2004) find that a higher level of diversity improves the performance of a crowd by comparing randomly selected groups and groups based on past performance. Meanwhile, De Oliveira and Nisbett (2018) conclude that groups that are diverse along a demographic characteristic do not significantly outperform homogenous groups. This paper finds similar results to De Oliveira & Nisbett (2018). The results show no significant improvement in performance by maintaining demographic diversity while selecting experts. The findings of this paper could contribute to the current literature by supporting existing findings. Furthermore, they could help to increase the understanding of the value of diversity in the selection of experts.

The implications from the paper are not only of academic importance. Diversity in organizations, either based on gender or ethnicity, has become an increasingly important point of discussion in society. However, the benefits and costs of implementing the idea of diverse teams are still unclear. On the one hand, members in diverse groups tend to have different perspectives and opinions, improving the quality of the final decision or solution (Hong & Page, 2004; Phillips & Loyd, 2006). On the other hand, a diverse team, either based on demographics or past performance, has the potential to lead to worse outcomes due to the inability or unwillingness of team members to communicate effectively (Mellers et al., 2015; Polzer et al., 2002; Watson et al., 1993). The results of this paper contribute to this discussion by researching the way in which diversity based on demographic characteristics translates into changes in performance.

To answer the research question, I will use the following structure. In Section 2, I will define several hypotheses based on existing literature to make aspects of the research question statistically testable. For testing the hypotheses, I rely on responses collected in an online survey. Section 3 explains the way this data was collected. I will also discuss similarities to De Oliveira & Nisbett (2018) here. Section 4 aims to explain the way these responses are used to statistically test the hypotheses. The methodology is put into practice in section 5 where I will test the hypotheses by comparing the performance and correlations of groups selected based on past performance and/or demographic characteristics. The conclusion is meant as a way to summarize the findings and discuss their implications. I will also mention limitations of my research design and potential directions for future research.

2. Theoretical Framework

In this section, I will elaborate on findings in existing literature. These findings will be used to define the hypotheses in support of the research question. In doing so, I explain my expectations and their implications. Furthermore, I will explain the domain-specific concepts used in the paper.

2.1 The performance of experts

Previous literature has used several methods of defining experts. Generally speaking, these methods can be categorized into two different groups. On the one hand, there are methods purely focusing on the experts. On the other hand, there are methods applying different weights to participants based on their level of expertise. Sections 2.1.1 and 2.2.2 provide an overview of these methods.

2.1.1 Experts selected on past performance

The idea behind selecting experts within a crowd is simple. By focusing on the best estimators within a crowd and combining their estimates, one tries to improve the aggregated result. The most straightforward way to accomplish this is to preserve the estimates of the experts and to ignore the rest. Mannes et al. (2014) implemented this idea by selecting the top judges based on several recent judgements. They show that selecting the average of the top five judges gives robust improvements over the whole crowd's estimate. Another interesting observation is that only selecting the best judge underperformed both the combination of five judges and that of the entire crowd.

Mellers et al. (2015) confirm this finding in a team setting. In their research study, participants are part of a forecasting competition. Here they were asked to make probabilistic estimates of future events. Defying expectations, participants who performed very well in the first year, continued to do so in the following year. The performance of these superforecasters, or experts, was further improved by putting them together in groups with other superforecasters. Again, by purely focusing on the predictions of the experts, the accuracy was improved compared to the entire crowd's estimate.

Budescu and Chen (2015) take a slightly different approach in their way of aggregating the estimates. Whereas Mannes et al. (2014) excluded the estimates of the rest of the crowd entirely, Budescu and Chen (2015) put different weights on all the estimates. The weight for each estimate was decided by looking at the positive contribution a given estimate made to improving the aggregated estimate. This meant that more estimates were considered, although some were still excluded with a weight of 0. The paper shows that the weighted selection based on positive contribution

outperformed both the unweighted aggregate estimates of top judges or the entire crowd. This suggests that considering relative quality of judgements is better than absolute quality.

2.1.2 The loss of diversity

Although the focus on experts is shown to improve aggregate estimates, it usually leads to a loss of diversity within the selected group (Hong & Page, 2004). This is a problem as these experts tend to rely on the same information (Budescu & Chen, 2015).

Broomwell and Budescu (2009) examine a similar phenomenon by focusing on the inter-judge correlation between experts. As expected, there is a strong correlation between the estimates of experts. The researchers determine two main sources of the correlation, namely the environment and the characteristics of the judges. Because of a similar environment, experts tend to have access to the same cues. This leads to them using the same information in making a decision or prediction. The correlation is further increased by the characteristics of the judges. Experts tend to select and weigh the importance of the available cues in the same manner.

Diversity is one of the most important aspects in the wisdom of crowds. By giving uncorrelated answers, participants cancel out each other's mistakes (Budescu & Chen, 2015). Therefore, by focusing a bit more on the diversity of a selected group, it might be possible to improve the performance of the selected experts further. However, defining diversity is not that easy as I will discuss in the next section.

2.2 The performance of diverse groups

2.2.1 Diversity based on correlation

The first way of identifying diversity is done by looking at the correlation between different judges as was implemented by Davis-Stober et al. (2014). By looking at the correlation of the judges within the sample, one can determine how similar the estimates are. Broomwell and Budescu (2009) show that this tends to happen when experts are selected. Davis-Stober et al. (2014) show that there is a balance between the quality of individual estimates and their correlation. On the one hand, one wants to select the most accurate estimate. On the other hand, it is important that the errors in the estimates are able to cancel out. This means that one is sometimes better off selecting a slightly less accurate participant if this leads to less correlated estimates.

2.2.2 Diversity based on functional diversity

A second way of defining diversity is by using either identity and functional diversity. In their paper, Hong and Page (2004) discuss the distinction between the two as follows. They define identity diversity as the combination of differences in people's demographic characteristics, cultural identities and ethnicity, and training and expertise. Functional diversity is defined as differences in how people go about solving problems.

There are many factors which influence the level of functional diversity. For example, The information and tools people have can affect the way people solve a problem. Furthermore, personality traits can influence the way you use this information. Some people are very precise and base their solution on facts, while others are more creative in their way of thinking. One could also consider the difference in solutions given by people with either a optimistic or pessimistic world view. Although there are many more factors, let us consider an extreme example of functional diversity.

Suppose two people are asked to predict which football team will win the Champions League during the next year. One of these people is an experienced banker, the other one is a housewife. The banker has been taught to base such a prediction on facts. As an experienced banker, he is used to looking for reliable sources before making his prediction. He might determine the net value of every football team and look at past years to make an accurate prediction. The housewife has had little formal education and prefers to rely on her gut instinct. Besides, newspapers and television are the only source of information she has. She might rely on things she has heard within her social circle to make a prediction. Although neither method guarantees success, it is clear that these individuals solve the problem in vastly different ways.

The definitions of identity diversity and functional diversity will be used in the rest of the paper. Hong and Page (2004) use functional diversity to determine the role of diversity in crowd performance. They do so by comparing the performance of a group of the best-performing problem-solvers with that of a randomly selected group. The paper concludes that the randomly selected group outperformed the experts. It is important to note that all the participants were qualified in solving these problems. Because of the baseline problem-solving ability, differences in performance within the crowd were likely smaller. Furthermore, participants interacted with the rest of the selected group to come to a final solution.

Jain et al. (2011) also used functional diversity in their paper. While collecting predictions, the researchers also asked questions related to personality. Using the Big Five personality test, developed by Digman (1990), they determined pairs of diverse personalities. As discussed, a difference in personality likely leads to a difference in the way people solve problems as well. The results show that combining the predictions of diverse pairs is indeed more accurate than combining those of

homogeneous pairs. This means that the effectiveness of considering functional diversity is robust while determining it in different ways.

2.2.3 Diversity based on identity diversity

Although identity diversity is likely to be correlated with functional identity, this is not guaranteed. Using identity diversity has two main benefits when compared to functional diversity. First of all, it is easier to observe. Whereas the determination of functional diversity can only be identified using several questions (Jain et al., 2011) or a specific setting (Hong & Page, 2004), identity diversity can easily be determined by a few simple questions or information already available. But the evidence of its effectiveness is not as strong.

De Oliveira and Nisbett (2018) compare the performance of diverse groups with that of homogenous groups using identity diversity as their measure. They run this comparison for every variable separately (e.g. gender and race) randomly selecting four participants from two different categories. After comparing the average estimate with randomly selected homogenous groups, meaning eight people from the same group, they conclude that there is no significant difference. Furthermore, they find that the variance in estimates within groups is bigger than the variance across groups. This indicates that identity does not necessarily lead to significantly different estimates. These findings will be important when explaining methodology used in this paper (See section 4.3).

2.3 Identity diversity as a predictor of functional diversity

To use identity diversity as a measure of diversity when researching the wisdom of crowds, one needs to assume that identity diversity is a predictor of functionality. This is the case, because the wisdom of crowds functions optimally when there is no group bias and errors are cancelled out (Davis-Stober et al., 2014). To accomplish this, it is important that participants think about a problem independently and/or consider different information. This is the case for groups that are functionally diverse. However, groups that are identically diverse (e.g. different ethnicity) might still consider problems in the same way, ultimately leading to a functional homogenous group.

Again, De Oliveira and Nisbett (2018) believe it is unlikely that this is the case given the findings of their paper. However, there is evidence to suggest the opposite. Phillips and Loyd (2006) compare the performance of identically diverse groups with that of homogeneous groups. The authors find that identity diversity, which the authors refer to as surface-level diversity, impacts the decision of groups in two different positive ways. Firstly, identically diverse groups tend to voice their opinion more which leads to more opinions being considered. Furthermore, task engagement among the surface-level

diverse groups was higher. Both factors likely contributed to a better task performance. It is important to note that the groups used in the study were functionally quite similar. The findings of the paper show that identically diverse groups tend to have different perspectives, even when functional diversity is limited. Furthermore, the findings suggest that selecting groups based on identity diversity might be more effective in practice. All in all, the value of identity diversity in the context of the wisdom of crowds is worth exploring further.

2.4 Combining diversity and expertise

It is clear that both diversity and the selection of experts have the potential to improve the entire crowd's estimate. It is however quite unclear how the two interact. Hong and Page (2004) conclude that maintaining functional diversity is important for a group's task performance. As the participants are all of an acceptable skill-level, this is an indication that the interaction between expertise and diversity affects performance. However, this setting did not provide the opportunity to see what the isolated effects of diversity and expertise were. Therefore, I believe this paper can make a significant contribution to the literature by studying the isolated and interaction effect of these two factors.

2.5 The hypotheses

Based on the findings in the current literature, I have formulated several hypotheses. I will now explain their basis in the literature and the implications this would have.

H1: The median estimate of a group selected based on past performance significantly outperforms the median estimate of the crowd.

This hypothesis is based on the literature in section 2.1. Not rejecting this hypothesis would suggest that a selection of experts indeed outperforms the entire crowd. This would lay the foundation for the idea that it might be possible to further improve performance by including diversity. The decision to use the median as a measure of performance will be explained in section 4.4.1.

H2.A: Groups selected on the basis of gender diversity are more functionally diverse than homogenous groups.

H2.B: Groups selected on the basis of age diversity are more functionally diverse than homogenous groups.

H2.C: Groups selected on the basis of nationality diversity are more functionally diverse than homogenous groups.

H2.D: Groups selected on the basis of educational diversity are more functionally diverse than homogenous groups.

Hypotheses H2.A through H2.D based on the findings in section 2.3. The findings suggest predictive power of identity diversity for functional diversity, although the strength of the evidence is debatable. Not being able to reject these hypotheses would further support the use of identity diversity in research. Furthermore, it helps to support the idea that the combination of identity diversity and expertise has potential to improve aggregate estimates over isolated methods. The hypothesis is split into sub hypotheses focusing on individual characteristics. This allows me to investigate the role of individual demographic characteristics in more detail. The operationalization will be elaborated on in section 4.3.

H3.A: A group selected on expertise while maintaining identity diversity based on gender outperforms a group of experts.

H3.B: A group selected on expertise while maintaining identity diversity based on age outperforms a group of experts.

H3.C: A group selected on expertise while maintaining identity diversity based on education outperforms a group of experts.

H3.D: A group selected on expertise while maintaining identity diversity based on nationality outperforms a group of experts.

Hypotheses H3.A through H3.D illustrate the expectation that combining the selection identity diversity and expertise will lead to more accurate estimates than selection of experts alone. Both the performance of the balanced and the expert group is determined using the median estimation. Not being able to reject these hypotheses would support the findings of Hong and Page (2004). More

importantly, it would have numerous real life implications on the way teams are selected. Again, the hypothesis is split into sub hypotheses to explore individual aspects of identity diversity.

3. Data

3.1 General Description

The data used for this paper was collected using an online survey with a total of 162 complete responses. The survey consisted of 15 estimation questions equally divided into five different themes: food, money, geography, world records and random measurements. After the participants made estimates for all 15 questions, they were asked questions related to their demographics. Partial responses were deleted after 72 hours of inactivity. The full survey can be found in Appendix E.

3.2 Links to De Oliveira & Nisbett (2018)

The setup of the survey is based on studies conducted by De Oliveira and Nisbett (2018). In their paper, they study the effects of identity diversity on estimation accuracy using data from seven studies. Each study focused on a different topic and recorded estimates and demographics on that specific topic.

I decided to collect my own data for two reasons. Firstly, only two out of the seven studies recorded a sufficient number of different demographics to use in answering my research question. As every study used in their paper focused on a relatively narrow subject, only using two studies would lead to very similar question topics. This threatens the external validity of any significant effects I might find while analysing my results.

Secondly, collecting new data provides the opportunity to either support or question the findings in their paper. As mentioned, the role of diversity on performance is uncertain. De Oliveira and Nisbett (2018) find no significant role, while findings by Hong and Page (2004) suggest that diversity increases performance. Using new data could help increase the understanding of the role of diversity on performance.

In the online survey used to collect my own data, I made two basic changes to the original survey design by De Oliveira and Nisbett (2018). Firstly, I used only one type of estimation question. Participants were asked to write down a number that was as close as possible to the real measurement. Because of this, I am able to assign an equal weight to all answers given in the survey. This makes a participant's performance easier to rank.

Furthermore, I made a slight change to the demographics. In their study, De Oliveira and Nisbett (2018) record the political orientation of participants. As the responses are collected in North-America, the Republican/Democratic divide makes it straightforward to inquire on this topic.

However, in Europe the diversity of political orientations makes it harder to record and likely adds little value to the analysis. Thus, I have decided to remove this question.

3.3 The use of two versions

In designing the online survey, I ran a pilot to receive general feedback and determine the appropriate amount of estimation questions. Based on the feedback, I decided to make two versions of the survey. The two versions both contain 15 questions divided into the same five themes. Appendix E shows how the survey is split into these two versions. The participants were randomly shown one of the two versions. I will now explain the reasoning behind my decision.

3.3.1 Number of data points

An important concern for all types of data collection is the number of data points collected. Having an insufficient number of data points likely leads to insignificant results, when an effect could have been observed with additional data. By using two versions, the number of questions is effectively doubled. However, this approach does lead to a reduction of answers recorded per version and thus per question. In an ideal scenario, I would have asked all participants to answer all 30 questions. This was not feasible as the pilot showed that asking more than 15 questions led to a decrease in concentration among participants.

3.3.2 Concentration and Attrition

A reduction in concentration likely decreases the accuracy of participants. As a result, it becomes harder to find significant differences between groups. The reduction in concentration also risks an increase in attrition. If participants are less engaged with the survey, they are more likely to quit without finishing the survey. To further limit concerns regarding attrition rates, the 15 questions are split into five themes. This serves to make the survey shorter as participants are continuously making progress. The exact attrition rates were not measurable as many respondents refreshed the survey when encountering errors. This resulted in partially completed responses, making the attrition rate unknown.

4. Methodology

In this section, I will discuss the operationalization of the key concepts and hypotheses of this paper. In doing so, I will refer to the design of the survey described in the previous section. Furthermore, I will discuss robustness checks to improve the validity of my results.

4.1 Expertise

In this paper I will focus on the selection of expertise based on past performance. I will determine the participant's performance by first ranking all the answers given by all the members of the crowd for a given question according to the absolute error. After ranking the performance for every single question, I take the average of all the ranks received by a participant to determine their overall performance. I am able to average the ranks as all the questions have an equal weight. Because the ranks represent relative performance in comparison to the rest of the crowd, I do not have to worry about differences in absolute errors across questions. If two participants have the same absolute error for a particular question, they will both receive the average of the two neighbouring ranks as their score. The ranks will be used to determine the experts, for example selecting the top five by determining the five participants with the lowest average ranks. In section 4.4.1, more information is given on the way the performance of the experts is aggregated and compared with other groups. The survey is split into two versions, which prevents comparison of performance across the two different versions. However, both versions can still be used simultaneously to compare performance between groups.

4.2 Identity Diversity

In the survey, I measure identity diversity based on four different characteristics: age, gender, nationality and educational attainment. When analysing the role of identity diversity on estimation accuracy, I will focus on one characteristic at a time. This approach is based on the methodology of De Oliveira and Nisbett (2018). Focusing on one characteristic at a time makes it easier to define the level of diversity within a subsample. An optimally diverse sample is equally distributed along the different categories of a characteristic. For example, an optimally diverse sample of six participants contains two participants for all three categories. In order to use this approach, I have divided age into four categories (<25, 25-50, 50-75, >75). Nationality will be split up into 'Dutch', 'Mixed Nationality' and 'Non-Dutch'. A participant is considered to have a mixed nationality when he or she has two

nationalities of which one is the Dutch nationality. Education is divided into three categories: low, middle and high. The survey in Appendix E shows an overview of the options for education attainment. Option 1 and 2 belong to the 'low' category. Option 3 and 4 are form the 'middle' category and option 5 and 6 are considered to be a high level of educational attainment. This transformation is made to ensure enough observations per category and make it easier to create optimally diverse subsamples. Gender is divided into three categories, where the option 'Prefer not to say' is left out of the analysis. In section 4.4.2 I will discuss how diverse samples are selected and what method is used in case an optimally diverse sample is impossible.

4.3 Functional Diversity

The concept of functional diversity is based on the paper by Hong and Page (2004). In their paper they create functional diversity by taking a random sample of the entire crowd instead of only selecting the best performers. Because I need to measure the level of diversity, the paper does not offer a feasible method that I can copy. Instead I will follow Davis-Stober et al. (2014) who calculate the inter-judge correlation. As the correlation indicates the extent to which participants approach a question in different ways, I will use this to mimic the concept of functional diversity.

4.4 Hypotheses

Having defined the key concepts of this paper more precisely, I will now focus on the operationalization of the hypotheses.

4.4.1 The first hypothesis

In my approach I will follow Mannes et al. (2014). In the paper, the authors find that a group of five experts leads to the most accurate aggregate estimates. Selecting more/less judges gradually decreases the accuracy. This result is robust across various domains. The selection of five experts on the basis of past performance will form the basis for testing my first hypothesis. I will not test different sizes of the expert group to find the optimal group size in this specific setting. The performance of the experts is not the focus of the paper, it merely serves as a benchmark for the performance of the balanced group considered in the third hypothesis.

The first hypothesis will not be rejected if the error of the aggregate estimate of the five preselected experts is significantly lower than that of the crowd. The error of the aggregated estimate is calculated in several steps.

Firstly, experts are selected per question based on the average rank obtained on the other 14 questions. Thus, experts are selected based on 14 out of 15 questions, using the remaining question as the out-of-sample question in order to test the hypothesis. This idea is based on a method called jack knifing used by Budescu and Chen (2015). By using this method, I make full use of the available data. By creating the maximum number of predictions, I aim to increase the probability of finding significant results.

Secondly, the estimates of the selected experts are aggregated. As shown in section 5.1.1, the estimation data for most questions contain outliers. Given the limited number of responses and the extremity of the outliers, the aggregated estimate will be highly influenced by the outliers when using the average. This would not only lead to large errors, but also to a misrepresentation of the crowd's wisdom. The estimate error of almost all participants would be lower than that of the aggregate estimate. Therefore, I have chosen to use the median as the aggregate measure. The median is more robust to outliers and will give a better representation of the crowd and preselected experts. Averaging the rank of the participants within a group, instead of using the median, is not an option. The 'Wisdom of Crowd'-concept (Surowiecki, 2004) is based on the idea that errors cancel each other out across participants. When averaging ranks, this is not necessarily the case as the resulting average rank does not change when participants' individual errors are biased in the same way or not. By using the median, the basic idea of the 'Wisdom of Crowds'-concept is maintained.

Thirdly, the errors of the aggregate estimates of the experts are compared to the crowd's estimate. The crowd's estimate is calculated by taking the median of all participants who answered the question. For both the expert group and the crowd this results in one median value per question. To prevent the influence of the naturally different sizes of absolute errors per question, all errors are expressed as relative errors. The relative error is obtained by dividing the absolute error by the correct answer of the corresponding question. A paired t-test allows me to test the difference between the relative errors of the experts and crowds. The hypothesis is not rejected if the relative errors of the experts are lower than those of the crowd. As a robustness check, I will also apply the Wilcoxon signed-rank test. If this test suggests the same result, this will further increase the robustness of the result.

4.4.2 The second hypothesis

To test the second hypothesis, I will be focusing on the inter-judge correlation (Davis-Stober et al., 2014). The hypothesis cannot be rejected if the inter-judge correlation of subsamples based on

identity diversity is lower than that of identity homogenous subsamples. To select participants based on identity diversity, I will focus on one characteristic at a time. I will then select an optimally diverse group, a group of participants which is equally distributed along the different categories. Comparing their inter-judge correlation with that of the homogenous subsamples will allow me to reject or not reject the second hypothesis. For every identity characteristic I will randomly select 15 diverse and 15 homogenous subsamples per version. As there are two versions, this results in 30 diverse and 30 homogenous subsamples per characteristic. The samples are selected with replacement. This means that a participant can be part of multiple sample. For every sample the inter-judge correlation of the participants is calculated based on the 15 estimation questions, resulting in a correlation matrix. To get a single correlation measure, I average all correlations in the correlation matrix. Thus, for every characteristic, I end up with 30 correlation measures for both the diverse and homogenous groups. I will now explain the decisions I have made regarding the sample selection.

4.4.2.1 Selection of diverse samples

For the diverse subsamples, I will randomly select samples of six participants with equal representation across the two or three categories. A sample size of six was chosen as this is divisible by both two and three, the number of categories with sufficient observations (See section 5.1.2). When it is impossible to create optimally diverse subsamples, I plan to use two different methods to make samples as diverse as possible. Both methods will maintain the sample size of six. This is important while calculating inter-judge correlations as using smaller sample sizes risks obtaining unrealistically high or low average correlations. This is caused by a lower number of correlations in the correlation matrix, making the average correlation less reliable. However, keeping the sample size consistent does result in a lower level of diversity. This can be illustrated with a simple example.

Suppose we select a diverse sample based on education. The three categories are ‘low’, ‘middle’ and ‘high’. For now we assume there was only one participant in the ‘low’ category (this is not true in the actual data). When we try to randomly select two observations from every category, we end up with a sample size of five. Of this sample, 40% belongs to the category high, 40% to ‘middle’ and 20% to ‘low’. We can keep the sample size constant by adding in a participant from the ‘middle’ category. However, now the percentages of the categories ‘high’, ‘middle’ and ‘low’ shift to 33.33%, 50% and 16.67% respectively. This division leads to a decrease in identity diversity, but is necessary to keep the sample size constant. Now I will discuss the methods I will use to select the samples.

The first method consists of selecting three participants of all three categories (instead of two) and randomly dropping one or two observations to end up with a sample size of six. This means that I will sometimes add an observation from one category with sufficient observations and other times

add an additional observation for both categories with sufficient observation while dropping the observation from the category with insufficient observations. I chose to add a chance of dropping the category entirely to limit the dependence on a single observation while calculating correlations. When all categories contain a sufficient number of observations, no observations are added or dropped.

The second method serves as a robustness check for the first method. As I explained, I sometimes drop a category entirely using the first method to limit the dependence on a single observation. For the second method, I drop the category with insufficient observations for every subsample. This allows me to test the effect of dropping entire categories when selecting subsamples. It is important to keep any differences between the two methods in mind when interpreting the results. If dropping the entire category leads to higher correlations, this should lead to higher correlations for the first method as well for the samples where categories are dropped.

4.4.2.2 Selection of homogenous samples

Homogenous subsamples consist of six participants belonging to the same category within a characteristic. The total of 15 subsamples are equally divided between the categories. This means that, for a characteristic with three categories, I will randomly select five homogenous subsample within every subsample. When it is not possible to select six participants within a given category, I will only select subsamples within the other categories.

4.4.2.3 Comparing correlations

Because the values of the correlations can vary between -1 and 1, the distribution is not suited to apply a t-test to directly. Therefore, I will transform the correlation to Z-values first using Fisher Z Transformation (Meng et al., 1992). Then, I obtain the mean and standard deviation of the combined sample of the two groups being compared. Lastly, I use these two measures to transform the correlations into Z-values. This allows me to apply an unpaired t-test. Firstly, I will test the difference between the diverse samples obtained using the first method and the homogenous samples. Secondly, I will compare the same homogenous samples with the correlations obtained using the second method. I plan to test the differences for every characteristic individually and for all characteristics combined. The robustness of the results will again be tested using the Wilcoxon signed-rank test.

4.4.3 The third hypothesis

I intend to strike a balance between expertise and diversity in two steps. Firstly, I will select the top 25% best performing participants in the same way explained in section 4.4.1. The boundary condition

allows for more participants to be selected. This is important as the sample needs to be big enough to select a diverse subsample.

Again, I will focus on one characteristic of identity diversity at a time. As the performance will be compared to the top five experts, there is no need to select homogenous subsamples. The diverse subsamples will be selected differently compared to section 4.4.2.1. When comparing correlations, it is important to maintain equal sample sizes as differences in sample sizes could influence inter-judge correlation. However, this is not the case when focusing on performance by calculating the median. Compared to the correlation measure, the median is affected less by a decrease in the sample size. This is the case as the correlation measure is an average of multiple correlation. As discussed, averages tend to be more susceptible to outliers than the median. In the example in section 4.4.2.1, keeping the sample size constant leads to a decrease in diversity. Therefore, I have decided to allow sample size to differ for the balanced samples. Thus, some balanced sample will contain only four or five observations instead of six. Although the median measures becomes less reliable as a result, it does not outweigh the loss in diversity.

After selecting a subsample from the top 25% experts, the relative errors of the resulting medians are estimated in the same way as explained in section 4.4.1. The relative errors are then compared to the relative error of the top five experts using two separate unpaired t-tests. Again, Wilcoxon signed-tests are used as robustness checks. Not being able to reject the hypothesis suggests that balancing expertise and diversity could improve aggregate estimates further.

5. Results

In the following sections, I will discuss the results obtained through the analysis plan laid out in section 4. First, in section 5.1, I will focus on the descriptive statistics of both the estimations and demographic characteristics of the participants. The descriptive statistics serve to support specific choices made for the analysis. In sections 5.2 through 5.4 I focus on the t-tests and Wilcoxon rank-sum tests used to test the hypotheses.

5.1 Descriptive statistics

5.1.1 Estimation questions

Table A-1, which can be found in Appendix A, shows the mean, median, standard deviation, minimum value and maximum value of the 30 estimation questions included in the survey. The correct answer provided at the end serves as a reference point for the other values.

The maximum and minimum value for the estimations in both versions show that the answers contain extreme outliers. For example, the maximum value of question 3 is bigger than the correct answer by a factor of 100,000. Although these extreme outliers are rare, they have a significant influence on the mean value. As most participants are more accurate than the average estimation, I will not use the average.

Instead I will use the median as the aggregation method. The median is more robust to these outliers and gives a better representation of the wisdom of the crowd. Although the average is a common aggregation method in the existing literature (De Oliveira & Nisbett, 2018; Mannes et al, 2015), using the median is not a new approach (Galton, 1907).

The number of observations and the division of participants along the categories suggests that the randomization across the two versions worked properly. This means that the level of expertise is likely similar in both groups, allowing me to combine the errors of experts in both versions to test the difference with the general crowd. As mentioned, it is important to transform the medians into relative errors as both the medians and correct answers of the 30 questions have vastly different magnitudes.

5.1.2 Demographic characteristics

For the selection of diverse subsamples, it is important that the demographic characteristics are distributed relatively equally across the different categories. The distributions in table 2 indicate that not all categories are well-represented.

Table 1

Descriptive Statistics of the Demographic Characteristics (Both Versions Combined)

Variable	Categories	Frequency	Percentage (%)
Gender	Female	79	48.77
	Male	81	50.00
	Non-binary/ third gender	1	.62
	Prefer not to say	1	.62
Age	Under 25	98	60.49
	25 - 50	32	19.76
	50-75	30	18.52
	Over 75	2	1.23
Nationality	Dutch	131	80.86
	Mixed-Nationality	5	3.09
	Non-Dutch	26	16.05
Education	Low	13	8.07
	Middle	106	65.84
	High	42	26.09

For diverse samples based on gender, all optimally diverse samples will contain the non-binary participant as this category only contains a single observation. The 'Prefer not to say' category will be left out during the analysis as it contains no information about gender. The dependence on the single non-binary observation could lead to problems regarding the validity of the results. This will be discussed in further detail in section 6.4.1.

All but one of the age categories are well represented, with the 'Over 75' category only having two observations. I have decided to exclude this category for two reasons. Firstly, the two observations are spread across the two different versions leaving only one observation per version (see table A-2 and A-3). Furthermore, using only three categories makes it possible to create an optimally diverse sample with two observations per category.

Both 'Nationality' and 'Education' have one category that is underrepresented. This should not lead to problems for the second hypothesis. However, when selecting diverse samples from a smaller sample only including experts, there will likely be insufficient observations to create optimally diverse samples. I will use the methods discussed in section 4.4.2.1 to obtain diverse samples when needed.

5.2 Performance of experts

Table 3.1 and 3.2 show the results regarding the performance of preselected experts compared to the general crowd. The paired t-test indicates that there are no significant differences between the relative errors of the two groups (*P*= .174). The mean of 0.233 shows the average difference between the two groups. A value of 0.233 indicates that the error of the median was on average bigger for the crowd by a margin of 23.3% of the correct answer. However, given the standard deviation of 0.916, this difference is not significant. This means that I reject the hypothesis '*The median estimate of a group selected based on past performance significantly outperforms the median estimate of the crowd.'* based on the results of the t-test.

Table 3.1

Two Sample T-test Comparing the Performance of Experts and the Crowd

	Paired Differences							
				95% Confidence Interval of the Difference				
Relative errors	Mean	Std. error mean	Std. Deviation	Lower	Upper	t	df	Sig. 2-tailed
crowd - expert	.233	.167	.916	-.109	.575	1.394	29	.174

Note. **p* < .10. ***p* < .05. ****p* < .01

This result is supported by the Wilcoxon Signed-ranks Test in table 3.2. Again, the test indicates that there are no significant differences between the two groups with a p-value of .167. The positive and negative ranks show the frequency of the crowd having either a higher or lower relative error compared to the experts. Ties indicates the frequency with which the relative errors are equal. There are 18 instances in which the crowd obtained a higher relative error than the expert group and 11 instances in which the relative error was lower. Although this suggests a better performance of the expert group, the difference is not significant. The finding of this second test adds to the validity of the t-test. Although the first hypothesis is rejected, the expert group will still serve as a benchmark when testing the performance of the balanced group.

Table 3.2

Wilcoxon Signed-ranks Test Comparing the Performance of Experts and the Crowd

Ranks

		N	Mean Rank	Sum of Ranks
crowd - expert	Positive Ranks	18[a]	16.67	300.00
	Negative Ranks	11[b]	14.90	164.00
	Ties	1[c]		
	Total	30		

a. crowd > expert
b. crowd < expert
c. crowd = expert

Test Statistic[b]

	crowd - expert
Z	1.399[a]
Assymp. Sig (2-tailed)	.167

a. Based on positive ranks
b. Wilcoxon Signed Ranks Test

Note. *p < .10. **p < .05. ***p < .01

5.3 Predictive value of identity diversity

5.3.1 The different tests

The second hypothesis focuses on the difference in correlation between diverse and homogenous samples. Both the second and third hypotheses are split into sub hypotheses in line with the four different demographic characteristics. The differences are tested separately for every characteristic to provide a more complete overview. In addition to the separate test per category, I also included tests on the differences when combining all characteristics. For the second hypothesis, the characteristics are combined by using all the diverse samples, based on the different characteristics, and comparing the correlations with all the homogenous samples.

The results of the combined characteristics are shown in table 4.1 and 4.2. I will also discuss the results of the tests on the separate characteristics. These tests can be found in Appendix B and C. As mentioned in section 4.4.2.1, the diverse samples are selected in two different ways. This results in a two t-test and two Wilcoxon Signed-ranks Tests per characteristic. For the characteristics 'Nationality' and 'Education' only the results of the first method are shown. As there was no need to readjust any of the diverse samples, both methods result in the same selection process.

5.3.2 The results

The unpaired t-test of the combined characteristics in table 4.1 indicates that, using the first method, the correlations of the diverse samples are significantly lower than those of homogenous samples at a 10% significance level (P =.075). The mean value of -0.23 shows that the average z-value of the correlations of the diverse samples was 0.23 lower than the average of the homogenous samples. All t-tests used to test the second hypothesis are unpaired as the samples are not linked to a counterpart

directly. Furthermore, I did not assume equality of variances while testing the differences as the samples were not collected from a similar population.

Table 4.1

Two Sample T-test on the Difference in Correlations Between the Combined Characteristics and the Homogenous Samples (First Method)

	Unpaired Differences						
			95% Confidence Interval of the Difference				
Z- values correlations	Mean	Std. error mean	Lower	Upper	t	df	Sig. 2-tailed
overall - homogenous	-.230	.129	-.483	.023	-1.790	237.01	.075*

Note. *p < .10. **p < .05. ***p < .01

The Wilcoxon Signed-ranks Test indicates significantly lower correlations for the diverse samples at a 5% significance level (P =0.044) using the first selection method. The combination of both tests add robustness to the finding that diverse samples have significantly lower correlations.

Table 4.2

Wilcoxon Signed-ranks Test on the Difference in Correlations Between the Combined Characteristics and the Homogenous Samples (First Method)

Ranks

		N	Mean Rank	Sum of Ranks
overall - homogenous	Positive Ranks	49[a]	58.43	2863.00
	Negative Ranks	71[b]	61.93	4397.00
	Ties	0[c]		
	Total	120		

a. overall > homogenous
b. overall < homogenous
c. overall = homogenous

Test Statistic[b]

	overall - homogenous
Z	-2.009[a]
Assymp. Sig (2-tailed)	.044**

a. Based on positive ranks
b. Wilcoxon Signed Ranks Test

Note. *p < .10. **p < .05. ***p < .01

However, it is also important to consider the results of the second selection method and the tests on the separate characteristics as well. When sampling the diverse samples using the second method, both the t-test and Wilcoxon Signed-ranks Test of the combined characteristics show insignificant differences between the correlations with a p- value of .387 and .331 respectively (See table C-5 and C-6). On average, the correlations using the second method were higher than those using the first

method. This indicates that dropping categories entirely results in higher correlation measures. The implications of this finding are further discussed in section 6.4.1.

The results for the specific characteristics are mixed (table B-1 through B-8). The characteristics 'Gender', 'Nationality' and 'Education' all show insignificant differences using the first method. However, the correlations for diverse samples selected based on age have significantly lower correlations compared to the homogenous groups. This result holds for both the t-test and the Wilcoxon Signed-ranks Test and is significant at a 1% level (See table B-3 and B-4). It is important to note that the results for the individual characteristics, using the second method, were all insignificant. This supports the difference between the two methods found for the combined characteristics.

The tests suggest potentially significant differences for both the overall diverse samples and diverse samples specifically focused on age. I therefore do not reject the second hypothesis '*Groups selected on the basis of age diversity is more functionally diverse than homogenous groups.'* using the first method. The results are somewhat less reliable because of the findings using the second method. The difference between the two methods shows that the result is reliant on a specific methodology. Furthermore, including a small number of participants in many samples could lead to overdependence on these observations in calculating correlations. This is discussed in further detail in section 6.4.1. The other sub hypotheses are rejected given the insignificant differences in correlation with the homogenous samples.

5.4 Performance of a balanced sample

Table 5.1 through 5.4 show the results regarding the relative performance of balanced samples compared to the expert group. The tables shown below concern the tests with all categories combined. The categories are combined by averaging the median value obtained from the characteristics-specific samples. The tests for the separate characteristics are provided in Appendix D.

The paired t-test in table 5.1 shows that there is no significant difference between the balanced group and the expert group with a p-value of .787. Similar to the results in section 5.2, the mean value of the paired t-test indicates the average difference in the relative error between all balanced samples and the expert group. In comparison to the expert group, the average relative error is higher by 3.5% of the correct answer.

Table 5.1

Two Sample T-test on the Performance of the Expert and Balanced Group (Combined Characteristics)

	Paired Differences					t	df	Sig. 2-tailed
				95% Confidence Interval of the Difference				
Relative errors	Mean	Std. error mean	Std. Deviation	Lower	Upper			
overall - expert	.035	.128	.700	-.226	.296	0.272	29	.787

Note. *p < .10. **p < .05. ***p < .01

The Wilcoxon Signed-ranks Test in table 5.2 indicates a significant difference between the balanced group and the expert group with a p-value of .038. The test in table 5.2 indicates that the relative errors of the balanced group are significantly higher than those of the expert group. This suggests that the performance of the balanced group is worse than those of the experts. This is the opposite of the third hypothesis discussed in section 2.5

Table 5.2

Wilcoxon Signed-ranks Test Comparing the Performance of the Crowd and the Balanced Group Based on the Combined Characteristics

Ranks

		N	Mean Rank	Sum of Ranks
overall - expert	Positive Ranks	20[a]	16.65	333.00
	Negative Ranks	10[b]	13.20	132.00
	Ties	0[c]		
	Total	30		

a. overall > expert
b. overall < expert
c. overall = expert

Test Statistic[b]

	overall - expert
Z	2.067[a]
Assymp. Sig (2-tailed)	.038**

a. Based on positive ranks
b. Wilcoxon Signed Ranks Test

Note. *p < .10. **p < .05. ***p < .01

Turning our attention to the specific characteristics, we see that the comparisons of relative errors for samples based on gender are insignificant (See table D-1 and D-2).

The categories 'Age', 'Nationality' and 'Education' show a similar pattern to the tests on the combined characteristics. For these three categories, the t-tests indicate insignificant differences in relative errors. However, the Wilcoxon Summed-ranks Tests suggest that the relative errors of the balanced groups are significantly higher than those of the expert group (see table D-4, D-6 and D-8). This means that the balanced groups performed significantly worse than the experts.

The tests on both the combined characteristics and the separate characteristics show mixed results. Still, it is clear that none of the evidence supports the third hypothesis '*A group selected on expertise while maintaining identity diversity based on ... outperforms a group of experts'* for any of the characteristics. Therefore, the hypotheses are rejected.

6. Conclusion

6.1 Main results

The results obtained in the previous section allow me to either reject or not reject the three main hypotheses stated in section 2.5. The first hypothesis on the performance of experts compared to the crowd is rejected. Although the results suggest potentially better performance of the expert group due to the mean difference and the number of positive ranks, the results are not significant.

The results are mixed for the second hypothesis. For the 'Age' characteristic and the combined characteristics, the first method indicates significantly lower correlations for the diverse samples. For the 'Age' characteristic both tests indicate a significance level of 1%. However, results for all characteristics using the second method are all insignificant. The null hypothesis for the second hypothesis is the equality of correlations across the identity diverse and homogenous groups. When focusing on the first selection method, we cannot reject the null hypothesis. However, the results of the second method suggest that the results might be reliant on a specific methodology.

The results in section 5.4 lead to a rejection of the third hypothesis. When comparing the performance of the balanced samples with the experts, the results either indicate insignificant differences or significantly better performance of the expert group. This means that we reject the null hypothesis that the performance of the balance group is equal to that of the expert group.

6.2 Relation to the existing literature

6.2.1 The first hypothesis

The results for the first hypothesis are not in line with the findings by Mannes et al. (2015), who find that the top five experts perform significantly better than the entire crowd. There are several potential explanations for this difference.

Firstly, the difference could be explained by the difference in methodology. The estimates in Mannes et al. (2015) are averaged, both for the experts and the general crowd. As explained in section 5.1, I have decided to use the median to aggregate the estimate. Using the average would have resulted in bigger errors and a misrepresentation of the groups. As the expert group is less likely to have extreme outliers, using the average would almost certainly have resulted in a significantly better performance of the experts.

Secondly, the number of data points or diversity in the data points might be insufficient. As the difference using the median is likely smaller, it requires more data to determine a significant difference in performance. Having more data points would naturally reduce the variance and thus lead

to a smaller confidence interval. Furthermore, having a more diverse sample leads to bigger differences in expertise. This makes it easier to measure a difference in performance between the two groups.

A third explanation is that the finding by Mannes et al. (2015), that five experts is the optimal number, does not hold in this setting. As it is not the main focus of this paper, I have decided not to determine the optimal number of experts. However, it is possible that a different number of experts would have resulted in a significant difference.

6.2.2 The second hypothesis

The results for the second hypothesis are in line with the finding by De Oliveira and Nisbett (2018), who report only a very marginal effect of social diversity on group performance. However, there are important differences between the two studies.

My findings suggest a potentially strong correlation between age diversity and inter-judge correlation, while finding no significant evidence for the other characteristics. The strong finding for 'Age' is likely mostly responsible for the significant findings for the tests on the combined characteristics. It would therefore be wrong to conclude that my results indicate that general identity diversity leads to lower inter-judge correlations. To the contrary, De Oliveira and Nisbett (2018) are able to make more generalized statements as the different characteristics show relatively similar patterns.

The difference in results could be caused by a limited number of data points used for my analysis. As discussed in section 5.1.2, some categories contain very few observations. For example, the category 'non-binary/third gender' only contains one observation. Including the same observations in multiple samples can lead to highly variable outcomes based on the specific properties of that observation. This could, at least partly, explain the variance in results between the 'Age' characteristic and other characteristics. De Oliveira and Nisbett (2018) did not experience this problem due to a sample size of at least 200 participants for a given characteristic.

Another difference between the studies is the methodology. De Oliveira and Nisbett (2018) use a method called bracketing. The researchers calculate the number of participants below and above the correct answer. If there is no group bias, these percentages across multiple questions should be around 50%. If the general crowd is often on the same side of the correct answer, this indicates a group bias. By measuring this group bias, the researchers determine the functional diversity. As I used the inter-judge correlation for determining functional diversity, the difference in methodology could also have resulted in the slightly different findings.

6.2.3 The third hypothesis

For the third hypothesis, there are no studies to directly compare my findings to. However, on the general topic of the performance of a diverse group, two main studies have been mentioned throughout the paper. Hong and Page (2004) find that, when selecting a diverse group from capable individuals, performance is increased. On the other hand, De Oliveira and Nisbett (2018) find only very weak evidence to support this idea.

Although other existing literature suggest a positive effect of identity diversity on group performance (Phillips & Loyd, 2006), the studies do not link identity diversity to functional diversity directly. Both De Oliveira and Nisbett (2018) and myself used only one identity characteristic at a time. Combining different identity characteristics into a single measure of diversity, could help to bridge the gap between identity diversity and functionality. In turn, this could result in significant differences between the identity diverse and identity homogenous groups.

Furthermore, both studies mentioned above took place in a setting in which participants could communicate. This could increase the effect of diversity on performance. Allowing the randomly selected groups based on identity diversity to communicate could potentially lead to significantly better performance for diverse samples as well.

6.3 Implications of the findings

Due to the insignificant results, the findings of this paper alone do not lead to strong implications. However, combining the results with the existing literature enables me to confirm existing findings and their implications.

Firstly, the findings confirm those of De Oliveira and Nisbett (2018). Selecting samples based on a single characteristic does, on its own, not result in better performance of aggregated estimates. In the current discussion around diversity, it is often the case that the focus is limited to a single identity characteristic. For example, there is often a focus on gender diversity in the boardroom and on racial diversity in the police force. Broadening the pursuit of diversity across multiple dimensions could lead to more functional diversity and therefore the associated increase in performance.

The research by Hong and Page (2004) and Phillips and Loyd (2006) suggests that the diverse groups need to interact to make full use of the existing diversity. However, as mentioned in the introduction, an ability or unwillingness among diverse group members to communicate could lead to decreases in performance (Mellers et al., 2015; Polzer et al., 2002; Watson et al., 1993). Given the insignificant findings, it is still unclear how organizations should balance these two factors.

6.4 Limitations and suggestions for future research

6.4.1 Limitations

Some limitations that I will discuss have already been mentioned briefly throughout the paper. This section serves to summarize the limitations. It will also allow me to discuss the suggestions for further research more easily.

The first limitation concerns the limited number of observations. This affects the findings in two different ways. A relatively low number of observations leads to relatively high variance when testing differences. A higher variances, and thus a higher standard deviation, results in a wider confidence interval. Potentially significant results could remain undetected due to a wide confidence interval.

Furthermore, a limited number of observations leads to a low number of observations within certain categories. As discussed in section 6.2.2, this potentially causes tests to be overdependent on a small number of observations present in (almost) all samples. Using the second method of selection subsamples leads to higher correlation measures. This suggests that the correlations would have been higher without these limited number of observations. Therefore, the overdependence on these observations is a serious concern.

Secondly, the collected sample of participants is not optimally diverse. Because of the way the survey was distributed, my own social circles are necessarily overrepresented among the participants. These participants belong to similar categories with regards to characteristics such as age and educational attainment. Therefore, these categories are also overrepresented. A more diverse sample could have resulted in more significant results.

Lastly, the methodology contains some limitations which can be overcome in future research. As discussed in 6.2.3, I measured identity diversity using one characteristic at a time. To mimic functional diversity, it is likely more effective to create one diversity measure by combining the different characteristics. Besides the measurement of identity diversity, the methodology could be improved by including extra robustness checks. The top five experts did not significantly outperform the crowd in my paper. Determining the optimal number of experts for this setting would have provided a better benchmark for the balanced samples to be compared to. Although the number of experts was not determined arbitrarily, there is no existing literature on the specific setting used in my paper. Furthermore, using ‘bracketing’ in addition to the inter-judge correlation measure would have allowed a better comparison of my results with those of De Oliveira and Nisbett (2018).

6.4.2 Suggestions for future research

For my suggestions for future research, I mainly focus on the methodology. In a larger scale research study, the first two limitations regarding the sample size discussed in the previous section will likely not be a problem. For example, the selection of optimally diverse samples, when there is a lack of observations, will not be a problem for future research. The suggestions are aimed at giving identity diversity the best chance of enhancing a group's performance given the proper collection of the data needed.

The main suggestion concerns the measurement of identity diversity. A combination of different characteristics has not yet been used in the existing literature. This is understandable as there is no straightforward way of creating such a model. However, the model, or different versions of that model, would help tremendously in determining the value of identity diversity in group performance.

Another suggestion stems from the discussion in section 6.2.3. Future research could test two settings, one setting with and one without communication among participants. This is especially interesting in the context of identity diversity as communication could improve the value of identity diversity in samples.

Lastly, I suggest implementing the robustness checks discussed in the previous section. The implementation has two main goals. First, it adds validity to the obtained results. Furthermore, it allows for better comparison to the existing literature. This is useful as it allows for a stronger result when a similar pattern is present in related research. Furthermore, existing findings can be questioned more easily.

Bibliography

Armstrong, J. S. (2001). Combining Forecasts. *International Series in Operations Research & Management Science,* 417-439.

Broomwell, S. B., & Budescu, D. V. (2009). Why are experts correlated? Decomposing correlations between judges. *Psychometrika*, 74, 531–553.

Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science,* 61(2), 267-280.

Clemen, R. T. (1989). *Combining forecasts: A review and annotated bibliography. International journal of forecasting*, 5(4), 559-583.

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomwell, S. B. (2014). When is a crowd wise? *Decision*, 1(2), 79.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual review of psychology,* 41(1), 417-440.

De Oliveira, S., & Nisbett, R. E. (2018). Demographically diverse crowds are typically not much wiser than homogeneous crowds*. Proceedings of the National Academy of Sciences*, 115(9), 2066-2071.

Galton, F. (1907). Vox populi. *Nature*, 75, 450 – 451.

Hong, L., & Page, S. E. (2004). Groups of diverse problem-solvers can outperform groups of high ability problem-solvers. *Proceedings of the National Academy of Sciences*, USA, 101, 16385–16389.

Phillips KW, Loyd DL (2006) When surface and deep-level diversity collide: The effects on dissenting group members*. Organ Behav Hum Decis Process*, 99, 143–160.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of personality and social psychology*, 107(2), 276.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., ... & Tetlock, P. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, 10(3), 267-281.

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological bulletin*, 111(1), 172.

Phillips KW, Loyd DL (2006) When surface and deep-level diversity collide: The effects on dissenting group members. *Organ Behav Hum Decis Process*, 99, 143–160.

Polzer, J. T., Milton, L. P., & Swarm Jr, W. B. (2002). Capitalizing on diversity: Interpersonal congruence in small work groups. *Administrative Science Quarterly*, 47(2), 296-324.

Surowiecki, J. (2004). *The wisdom of crowds.* New York, NY: Doubleday.

Watson, W. E., Kumar, K., & Michaelsen, L. K. (1993). Cultural diversity's impact on interaction process and performance: Comparing homogeneous and diverse task groups*. Academy of management journal*, 36(3), 590-602.

Appendix

Appendix A: Descriptive Statistics

Table A-1

Descriptive Statistics of the Estimation Questions (Both Version Combined)

Question	N	Mean	Median	SD	Min	Max	Correct Answer
1	80	6773	58	39307	0	250000	32
2	80	125143	100	1118018	6	1.00e+07	99.7
3	80	6.57e+09	1550017	3.62e+10	10	3.00e+11	291490
4	80	438442	180000	1107199	50	9200000	154900
5	80	741	47.5	2781	3	20000	18.9
6	80	318	48	1565	1	10000	105
7	80	408	363	270.82	50	1500	443
8	80	2656	1200	7857	200	70000	1331
9	80	2.19e+09	40000	1.95e+10	0	1.75e+11	40075
10	80	1600	175	7867	8	50007	237
11	80	7569	5000	9207	120	60000	5768
12	80	327	35	1356	1	10000	48
13	80	386	200	767	8	6000	495
14	80	1836	985	2865	150	16987	3500
15	80	136	50	315	0	2353	13.59
16	82	2022	100	11137	0	100000	3025
17	82	14.9	10	14.7	1	82	10.19
18	82	64.2	37	76.7	.20	500	39
19	82	1090050	1000	8828288	5	8.00e+07	2755
20	82	1898809	22	1.08e+07	.50	8.70e+07	106.3
21	82	128720	75000	266899	.40	1800000	81190
22	82	5124	1950	11794	20	100000	6650
23	82	126897	17500	310216	187	2000000	37000
24	82	947	52.5	5088	0	45000	500
25	82	40.0	35	18.7	15	100	54
26	82	27598	45	221388	2	2000000	227
27	82	48.7	42.5	22.7	2	90	88
28	82	123	50	398	4	3500	41.42
29	82	497	2589	905	30	6000	1191
30	82	24.2	18	23.1	1	125	43

Table A-2

Descriptive Statistics of the Demographic Characteristics (Version 1)

Variable	Categories	Frequency	Percentage (%)
Gender	Female	38	47.50
	Male	41	51.25
	Non-binary/ third gender	1	1.25
	Prefer not to say	0	0
Age	Under 25	51	63.75
	25 - 50	17	21.25
	50-75	11	13.75
	Over 75	1	1.25
Nationality	Dutch	66	82.50
	Mixed-Nationality	3	3.75
	Non-Dutch	11	13.75
Education	Low	5	6.33
	Middle	56	70.89
	High	18	22.78

Table A-3

Descriptive Statistics of the Demographic Characteristics (Version 2)

Variable	Categories	Frequency	Percentage (%)
Gender	Female	41	50.00
	Male	40	48.78
	Non-binary/ third gender	0	0.00
	Prefer not to say	1	1.22
Age	Under 25	47	57.32
	25 - 50	15	18.29
	50-75	19	23.17
	Over 75	1	1.22
Nationality	Dutch	65	79.27
	Mixed-Nationality	2	2.44
	Non-Dutch	15	18.29
Education	Low	8	9.76
	Middle	50	60.98
	High	24	29.27

Appendix B: Tests on difference in correlations using the first method

Table B-1

Two Sample T-test on the Difference in Correlations Between Gender Diverse and Homogenous Samples (First Method)

Z- values correlations	Unpaired Differences				t	df	Sig. 2-tailed
	Mean	Std. error mean	95% Confidence Interval of the Difference				
			Lower	Upper			
gender - homogenous	-.285	.258	-.802	.231	-1.108	53.69	.273

Note. **p* < .10. ***p* < .05. ****p* < .01

Table B-2

Wilcoxon Signed-ranks Test on the Difference in Correlations Between the Gender Diverse and Homogenous Samples (First Method)

Ranks

		N	Mean Rank	Sum of Ranks
gender-homogenous	Positive Ranks	13[a]	14.15	184.00
	Negative Ranks	17[b]	16.53	281.00
	Ties	0[c]		
	Total	30		

a. gender > homogenous
b. gender < homogenous
c. gender = homogenous

Test Statistic[b]

	gender - homogenous
Z	-0.998[a]
Assymp. Sig (2-tailed)	.329

a. Based on positive ranks
b. Wilcoxon Signed Ranks Test

Note. **p* < .10. ***p* < .05. ****p* < .01

Table B-3

Two Sample T-test on the Difference in Correlations Between Age Diverse and Homogenous Samples (First Method)

Z- values correlations	Unpaired Differences				t	df	Sig. 2-tailed
	Mean	Std. error mean	95% Confidence Interval of the Difference				
			Lower	Upper			
age - homogenous	-.980	.226	-1.43	-.527	-4.329	57.74	.000***

Note. **p* < .10. ***p* < .05. ****p* < .01

Table B-4

Wilcoxon Signed-ranks Test on the Difference in Correlations Between the Age Diverse and Homogenous Samples (First Method)

Ranks

		N	Mean Rank	Sum of Ranks
age - homogenous	Positive Ranks	3[a]	8.00	24.00
	Negative Ranks	27[b]	16.33	441.00
	Ties	0[c]		
	Total	30		

a. age > homogenous
b. age < homogenous
c. age = homogenous

Test Statistic[b]

	age - homogenous
Z	-4.288[a]
Assymp. Sig (2-tailed)	.000***

a. Based on positive ranks
b. Wilcoxon Signed Ranks Test

Note. *p < .10. **p < .05. ***p < .01

Table B-5

Two Sample T-test on the Difference in Correlations Between Nationality Diverse and Homogenous Samples (First Method)

	Unpaired Differences						
			95% Confidence Interval of the Difference				
Z- values correlations	Mean	Std. error mean	Lower	Upper	t	df	Sig. 2-tailed
nationality - homogenous	-.091	.260	-.612	.430	-0.348	56.65	.729

Note. *p < .10. **p < .05. ***p < .01

Table B-6

Wilcoxon Signed-ranks Test on the Difference in Correlations Between the Nationality Diverse and Homogenous Samples (First Method)

Ranks

		N	Mean Rank	Sum of Ranks
nationality - homogenous	Positive Ranks	15[a]	15.27	229.00
	Negative Ranks	15[b]	15.73	236.00
	Ties	0[c]		
	Total	30		

a. nationality > homogenous
b. nationality < homogenous
c. nationality = homogenous

Test Statistic[b]

	nationality - homogenous
Z	-0.072[a]
Assymp. Sig (2-tailed)	.952

a. Based on positive ranks
b. Wilcoxon Signed Ranks Test

Note. *p < .10. **p < .05. ***p < .01

Table B-7

Two Sample t-test on the Difference in Correlations Between Educational Diverse and Homogenous Samples (First Method)

Z- values correlations	Unpaired Differences				t	df	Sig. 2-tailed
	Mean	Std. error mean	95% Confidence Interval of the Difference				
			Lower	Upper			
education - homogenous	.270	.261	-.254	.793	1.035	50.81	.305

Note. **p* < .10. ***p* < .05. ****p* < .01

Table B-8

Wilcoxon Signed-ranks Test on the Difference in Correlations Between the Educational Diverse and Homogenous Samples (First Method)

Ranks

		N	Mean Rank	Sum of Ranks
education - homogenous	Positive Ranks	18[a]	16.22	292.00
	Negative Ranks	12[b]	14.42	173.00
	Ties	0[c]		
	Total	30		

a. education > homogenous
b. education < homogenous
c. education = homogenous

Test Statistic[b]

	education - homogenous
Z	1.224[a]
Assymp. Sig (2-tailed)	.229

a. Based on positive ranks
b. Wilcoxon Signed Ranks Test

Note. **p* < .10. ***p* < .05. ****p* < .01

Appendix C: Tests on difference in performance

Table C-1

Two Sample T-test on the Difference in Correlations Between Gender Diverse and Homogenous Samples (Second Method)

Z- values correlations	Unpaired Differences				t	df	Sig. 2-tailed
	Mean	Std. error mean	95% Confidence Interval of the Difference				
			Lower	Upper			
gender - homogenous	.124	.260	-.397	.644	0.476	57.54	.636

Note. **p* < .10. ***p* < .05. ****p* < .01

Table C-2

Wilcoxon Signed-ranks Test on the Difference in Correlations Between the Gender Diverse and Homogenous Samples (Second Method)

Ranks

		N	Mean Rank	Sum of Ranks
gender-homogenous	Positive Ranks	16[a]	15.75	252.00
	Negative Ranks	14[b]	15.21	213.00
	Ties	0[c]		
	Total	30		

Test Statistic[b]

	gender - homogenous
Z	0.401[a]
Assymp. Sig (2-tailed)	.700

a. gender > homogenous
b. gender < homogenous
c. gender = homogenous
d. Based on positive ranks
e. Wilcoxon Signed Ranks Test

Note. **p* < .10. ***p* < .05. ****p* < .01

Table C-3

Two Sample t-test on the Difference in Correlations Between Age Diverse and Homogenous Samples (Second Method)

Z- values correlations	Unpaired Differences				t	df	Sig. 2-tailed
	Mean	Std. error mean	95% Confidence Interval of the Difference				
			Lower	Upper			
age - homogenous	.138	.153	-.169	.434	0.867	233.32	.387

Note. **p* < .10. ***p* < .05. ****p* < .01

Table C-4

Wilcoxon Signed-ranks Test on the Difference in Correlations Between the Age Diverse and Homogenous Samples (Second Method)

Ranks

		N	Mean Rank	Sum of Ranks
age - homogenous	Positive Ranks	16[a]	15.56	249.00
	Negative Ranks	14[b]	15.43	216.00
	Ties	0[c]		
	Total	30		

a. age > homogenous
b. age < homogenous
c. age = homogenous

Test Statistic[b]

	age - homogenous
Z	0.339[a]
Assymp. Sig (2-tailed)	.746

a. Based on positive ranks
b. Wilcoxon Signed Ranks Test

Note. *p < .10. **p < .05. ***p < .01

Table C-5

Two Sample T-test on the Difference in Correlations Between the Combined Characteristics and the Homogenous Samples (Second Method)

	Unpaired Differences						
			95% Confidence Interval of the Difference				
Z- values correlations	Mean	Std. error mean	Lower	Upper	t	df	Sig. 2-tailed
overall - homogenous	.112	.129	-.143	.366	0.867	233.32	.387

Note. *p < .10. **p < .05. ***p < .01

Table C-6

Wilcoxon Signed-ranks Test on the Difference in Correlations Between the Combined Characteristics and the Homogenous Samples (Second Method)

Ranks

		N	Mean Rank	Sum of Ranks
overall - homogenous	Positive Ranks	65[a]	61.58	4003.00
	Negative Ranks	55[b]	59.22	3257.00
	Ties	0[c]		
	Total	120		

a. overall > homogenous
b. overall < homogenous
c. overall = homogenous

Test Statistic[b]

	overall - homogenous
Z	0.977[a]
Assymp. Sig (2-tailed)	.331

a. Based on positive ranks
b. Wilcoxon Signed Ranks Test

Note. *p < .10. **p < .05. ***p < .01

Appendix D: Tests on difference in performance

Table D-1

Two Sample T-test on the Performance of the Experts and Balanced Group (Gender Diversity)

	Paired Differences							
				95% Confidence Interval of the Difference				
Relative errors	Mean	Std. error mean	Std. Deviation	Lower	Upper	t	df	Sig. 2-tailed
gender - expert	-.091	.126	.688	-.348	.166	-0.725	29	.474

Note. *p < .10. **p < .05. ***p < .01

Table D-2

Wilcoxon Signed-ranks Test Comparing the Performance of the Experts and the Balanced Group Based on the Gender Diversity

Ranks

		N	Mean Rank	Sum of Ranks
gender - expert	Positive Ranks	17[a]	17.12	291.00
	Negative Ranks	11[b]	15.55	171.00
	Ties	2[c]		
	Total	30		

a. gender > homogenous
b. gender < homogenous
c. gender = homogenous

Test Statistic[b]

	gender - expert
Z	1.234[a]
Assymp. Sig (2-tailed)	.225

a. Based on positive ranks
b. Wilcoxon Signed Ranks Test

Note. *p < .10. **p < .05. ***p < .01

Table D-3

Two Sample T-test on the Performance of the Experts and Balanced Group (Age Diversity)

	Paired Differences							
				95% Confidence Interval of the Difference				
Relative errors	Mean	Std. error mean	Std. Deviation	Lower	Upper	t	df	Sig. 2-tailed
age - expert	.202	.244	1.338	-.298	.702	0.827	29	.415

Note. *p < .10. **p < .05. ***p < .01

Table D-4

Wilcoxon Signed-ranks Test Comparing the Performance of the Experts and the Balanced Group Based on the Age Diversity

Ranks

		N	Mean Rank	Sum of Ranks
age - expert	Positive Ranks	22[a]	14.95	329.00
	Negative Ranks	8[b]	17.00	136.00
	Ties	0[c]		
	Total	30		

a. age > homogenous
b. age < homogenous
c. age = homogenous

Test Statistic[b]

	age - expert
Z	1.985[a]
Assymp. Sig (2-tailed)	.047**

a. Based on positive ranks
b. Wilcoxon Signed Ranks Test

Note. *p < .10. **p < .05. ***p < .01

Table D-5

Two Sample T-test on the Performance of the Experts and Balanced Group (Nationality Diversity)

	Paired Differences							
				95% Confidence Interval of the Difference				
Relative errors	Mean	Std. error mean	Std. Deviation	Lower	Upper	t	df	Sig. 2-tailed
nationality - expert	.180	.199	1.090	-.226	.587	0.907	29	.372

Note. *p < .10. **p < .05. ***p < .01

Table D-6

Wilcoxon Signed-ranks Test Comparing the Performance of the Experts and the Balanced Group Based on the Nationality Diversity

Ranks

		N	Mean Rank	Sum of Ranks
nationality - expert	Positive Ranks	20[a]	16.25	325.00
	Negative Ranks	9[b]	15.44	139.00
	Ties	1[c]		
	Total	30		

a. nationality > homogenous
b. nationality < homogenous
c. nationality = homogenous

Test Statistic[b]

	nationality - expert
Z	1.913[a]
Assymp. Sig (2-tailed)	.056*

a. Based on positive ranks
b. Wilcoxon Signed Ranks Test

Note. *p < .10. **p < .05. ***p < .01

Table D-7

Two Sample T-test on the Performance of the Experts and Balanced Group (Educational Diversity)

	Paired Differences					t	df	Sig. 2-tailed
				95% Confidence Interval of the Difference				
Relative errors	Mean	Std. error mean	Std. Deviation	Lower	Upper			
education - expert	.027	.113	.620	-.204	.259	0.238	29	.813

Note. *p < .10. **p < .05. ***p < .01

Table D-8

Wilcoxon Signed-ranks Test Comparing the Performance of the Experts and the Balanced Group Based on the Educational Diversity

Ranks

		N	Mean Rank	Sum of Ranks
education - expert	Positive Ranks	18[a]	17.56	316.00
	Negative Ranks	9[b]	15.89	143.00
	Ties	3[c]		
	Total	30		

a. education > homogenous
b. education < homogenous
c. education = homogenous

Test Statistic[b]

	education - expert
Z	1.780[a]
Assymp. Sig (2-tailed)	.077*

a. Based on positive ranks
b. Wilcoxon Signed Ranks Test

Note. *p < .10. **p < .05. ***p < .01

Appendix E: The Survey

Figure E-1

Opening Screen – Language Selection

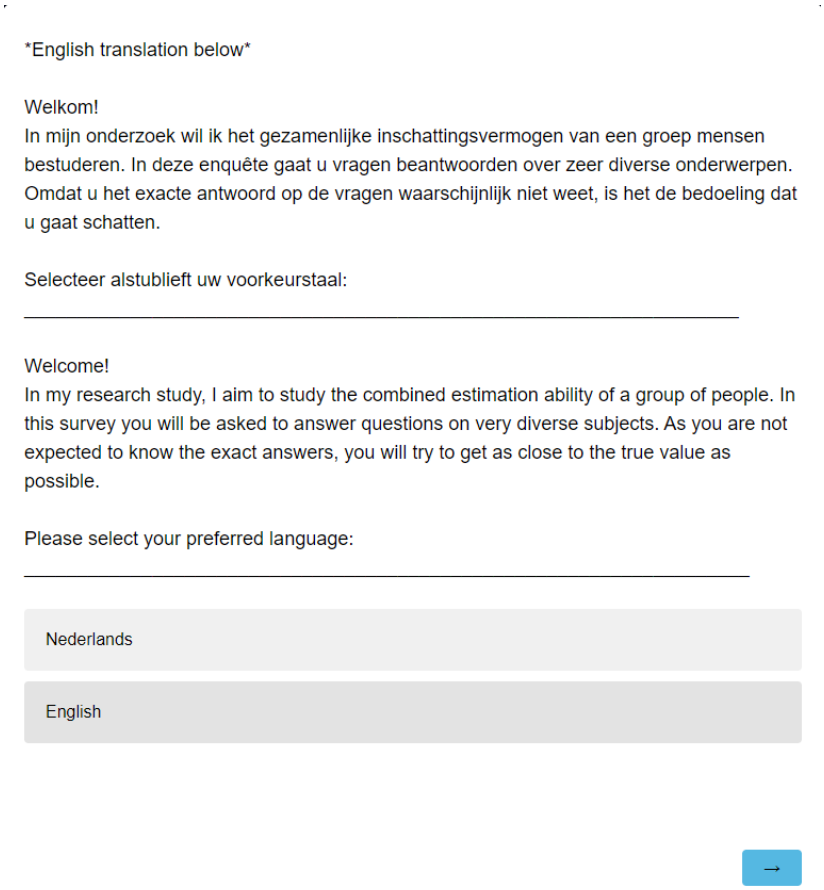


Figure E-2

Consent Page

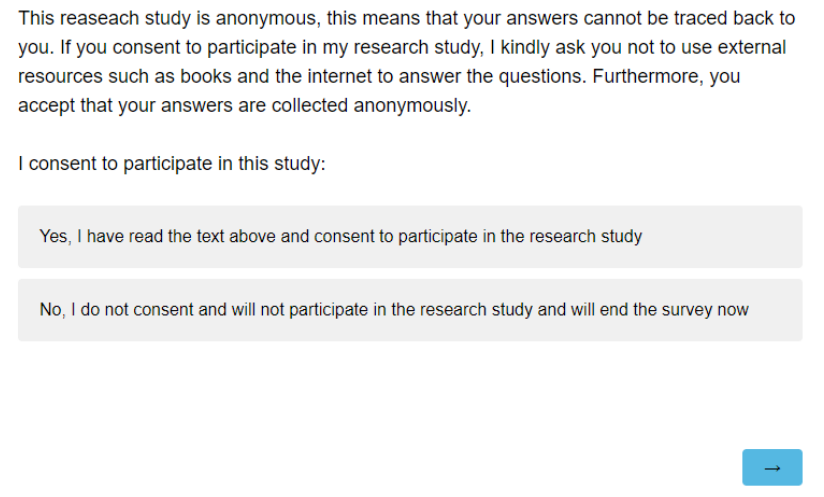
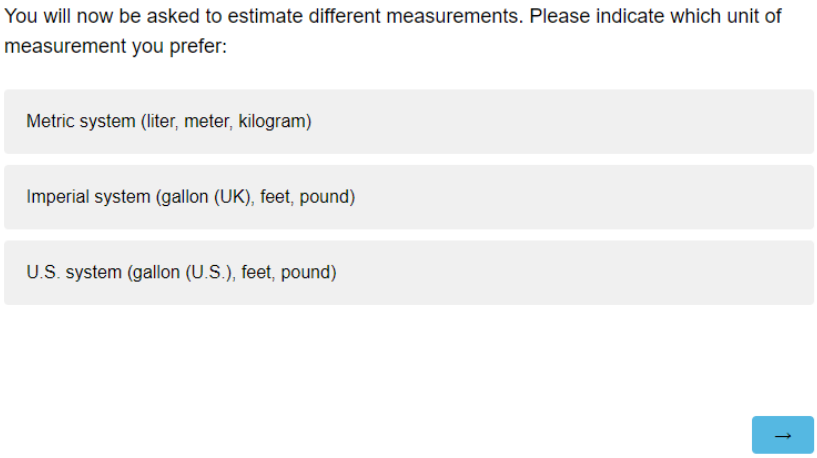


Figure E-3

Choice of Measurement Used in the Estimation Questions



Estimation Questions (Version 1)

Theme 1: Food

Q1

How many calories does 100 gram of strawberries contain?

Q2

How many liters of beer did the average German adult drink in 2019?

Q3

How many M&M's were used for this M&M mosaic?

Theme 2: Money

Q4

How much does an Audi R8 V10 Spyder (2021) convertible cost in US Dollars? (1 Euro = 1,21 US Dollar)

Q5

How tall is a stack of 1000 dollar bills?

Q6

How much money did Cristiano Ronaldo earn between 1 June, 2019 and 1 June, 2020 (after taxes)? (1 Euro = 1,21 US Dollar)

Theme 3: Random Measurements

Q7

How tall is the Empire State Building including the spire and antenna?

Q8

How much does a Ford Focus (2013) weigh in kilogram?

Q9

What is the circumference of the earth in kilometers measured at the equator?

Theme 4: Geography

Q10

How many times does the Netherlands fit into the United States based on surface area?

Q11

What is the distance between New York and Madrid in kilometers?

Q12

How many national parks are there in Canada?

Theme 5: World Records

Q13

What is the longest consecutive time in an abdominal plank position?

Q14

What is the record for the most toothpicks in a beard?

Q15

What is the record for the fastest time to pull a caravan (735 kilogram) 50 meters?

Estimation questions (Version 2)

Theme 1: Food

Q16

How many liters does it take to grow 1 kilogram of olives?

Q17

How much does this carrot weigh?

Q18

How much sugar does a 350 mililiter Coca Cola can contain?

Theme 2: Money

Q19

How many billionaires are there in the world?

Q20

How much money did Roger Federer earn between 1 June, 2019 and 1 june 2020 (after taxes)? (1 Euro = 1,21 US Dollar)

Q21

How much does a Tesla model S long-range (2021) cost in US Dollars? (1 Euro = 1,21 US Dollar)

Theme 3: Random Measurements

Q22

How long is the river Nile in kilometers?

Q23

How many kilometers of bike lane does the Netherlands have?

Q24

How many people are estimated to be killed by hippos every year?

Theme 4: Geography

Q25

How many official countries are there in Africa?

Q26

How many inhabited islands does Greece have?

Q27

What percentage of the world's population is estimated to live in the northern hemisphere? (estimated in 2017)

Theme 5: World Records

Q28

What is the record for the longest bicycle in meters?

Q29

What is the record for the heaviest pumpkin in kilogram?

Q30

What is the record for the oldest goldfish ever recorded?

Questions on demograpic characteristics

Please indicate your gender:

1. Female
2. Male
3. Non-binary/ third gender
4. Prefer not to say

Please indicate your age:

1. Under 25 years old
2. 25-50 years old
3. 50-75 years old
4. Over 75 years old

Please indicate your nationality by selecting the corresponding country:

(drop down menu containing all countries)

Please indicate your second nationality (optional):

(drop down menu containing all countries)

Please indicate your level of education:

1. Some high school or less
2. High school diploma
3. Some college/ university, but no degree
4. Bachelor's degree
5. Master's degree
6. Ph.D. or higher