Forming Teams of Learners Online in a User as Wizard Study with Openness, Conscientiousness, and cognitive Ability

Federica Lucia Vinella
Utrecht University
Utrecht, The Netherlads
f.l.vinella@uu.nl

Sanne Koppelaar
Utrecht University
Utrecht, The Netherlads
a.c.koppelaar@students.uu.nl

Judith Masthoff
Utrecht University
Utrecht, The Netherlads
j.f.m.masthoff@uu.nl

ABSTRACT

Forming teams of learners is a task that presents numerous challenges for educators increasingly relying on automated tools to optimize the process. The problem increases in difficulty in online classroom settings, where educators have little familiarity with the students. In this work, we present a User as Wizard study where 108 online crowd participants formed four teams of three teammates each from a pool of twelve dummy learner profiles. The profiles contained information about the learners' Conscientiousness, Openness, and cognitive ability levels. These attributes were derived from a pre-study with a smaller sample of crowd participants (N=52) rating the relevance of the Big Five personality traits and cognitive ability in team formation for educational purposes. The User as Wizard study shows that most people tend to form within (meaning most attributes of the teammates even out) and between (meaning the teams have similar attributes averages) balanced teams. It also shows that people perceive Conscientiousness and Openness as two of the most relevant personality traits when profiling learners for team formation. We compare these results to the probability of them being random and discuss the findings in the light of human-centered modeling of system designs and automation in education.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing systems and tools**; **User interface toolkits**.

KEYWORDS

team formation, personality, cognitive ability, user as wizard

ACM Reference Format:
Federica Lucia Vinella, Sanne Koppelaar, and Judith Masthoff. 2022. Forming Teams of Learners Online in a User as Wizard Study with Openness, Conscientiousness, and cognitive Ability. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling,Adaptation and Personalization (UMAP '22 Adjunct), July 4–7, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3511047.3537660

UMAP '22 Adjunct, July4–7, 2022, Barcelona, Spain

ACM ISBN 978-1-4503-9232-7/22/07…$15.00
https://doi.org/10.1145/3511047.3537660

1 INTRODUCTION

The team formation problem (TFP) is the problem of allocating multiple individuals in a way that matches a required set of skills to maximize one or several social positive attributes [20]. It optimizes human resource allocation in diverse settings such as work, socialization, and education. Factors such as the task type, the team size, personal attributes, and context all play a part in crafting the collaboration and lend themselves to combinatorial optimization approaches. Personality matters, with traits such as Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, influencing one's attitudes and behavior toward team performance and satisfaction [55]. Similarly, other aspects of individuals, namely emotional intelligence [51] and thinking skills [45], contribute to teamwork in equal measure. In educational settings, where knowledge acquisition is central to the learners' objectives and where team formation is normally for academic courses, having insight about the individual characteristics affecting teamwork is of the essence.

In this research, we approach the TFP in education settings in a human-centred way. We observe users interacting with systems as they assign students to teams online [31]. We assess what future automated systems should consider when recommending teammates and team compositions. Automated strategies for the TFP already offer solutions through computed outputs [40]. Some of the most common forms of computed solutions rest on established partitioning approaches (e.g., regression analysis optimization [38], genetic algorithms[7], k-means [3], etc.). Withal, algorithmic modeling of this kind may not always mirror the thinking process behind human choices when forming teams. Our study investigates what happens when users manually form teams without being informed of an ideal strategy. Our work closely follows the research by Odo et al. [40, 41] on group formation for collaborative learning as it considers the individuals traits such as personality in advancing solution to the TFP in education.

In line with Odo et al. [41], we utilize the User as Wizard (UAW) method to observe users' behavior while in charge of the team formation task. UAW, formalized by Masthoff [33], is a method that places the human at the center of the design process as participants take the role of the system performing the task without scripts or instructions to guide them. Our first research question, (RQ1) **"How do users distribute learners' attributes within and between teams?"**, addresses the need for insights into what people do when placing learners into teams knowing little about them through computer-mediated systems. It forwards research on

human-centered design for team formation as it bases system guidelines on the human interaction with an application while executing a given task– in our case, a team formation task for education. Most research on stable team formation considers making teams as similar as possible to each other across multiple attributes a fundamental objective [23, 24, 54]. Through the balancing paradigm, we assume that our participants will strive to distribute learners' attributes (resources) between and within teams in a way to ensure a fair capital share [2]. Therefore, we define balance as the act of distributing individual capital (e.g., cognitive ability, personality traits) between teams to guarantee that all formations have equal starting assets. For this study, we propose the following two hypotheses:

- *H1.1: "We expect people to distribute attributes equally within teams"*. In this study, we focus on the Big Five traits and cognitive ability of the learners and expect that participants will ensure that the aggregation of the teams' traits (made from the learners in that team) is even throughout. For example, we expect teams to have personality traits and ability levels with similar means calculated as the average of the teammates' attributes.
- *H1.2: "We expect people to distribute attributes equally between teams"*. We also expect that the teams formed by each participant will have similar attribute averages. Thus, we foresee that participants' strategy is to avoid imbalanced teams where some have higher attribute means than others.

We also propose a second research question as follows. **RQ2: Which attributes do users consider most important in the team formation of learners?**. *H2:" We expect cognitive ability to be the most influential trait when forming teams of learners online"*. We assume that people perceive the learner's ability as the most crucial in team formation. Corroborating our hypothesis is the relevance of intelligence and cognitive ability in the pedagogical field [48] as well as the public perception of ability as a crucial attribute of students for academic performance [11]. Additionally, cognitive ability often features in learner characteristics used by adaptive learning environments especially online ones [15, 39] The remainder of the paper is structured as follows. Section 2 is dedicated to the related work on team formation of learnerns and profiling attributes such as personality traits and cognitive ability. Section 3 describes the study design including the Wizard as User methodology. Section 4 grounds the choice of using three profiling attributes to form groups of learners based on a pre-study with online participants. Section 5 explains the team formation tool developed for the UAW study, the calculation of all possible combinations of teams with the given set of learners, the study participants, and the results. Section 6 discusses the findings from both studies and the limitations. Section 7 concludes the paper with final remarks for future work.

2 RELATED WORK

2.1 Team formation of learners

The TFP is a usual concern for educators as they are in charge of classroom activities and need to decide who should be teaming with whom. With insufficient resources such as narrow timelines, classroom size, and academic objectives (e.g., facilitating new collaborations between students, sharing ability levels across teams, etc.), educators face constraints limiting their investment in the TFP. Furthermore, forming teams, in the conventional sense, can be thought of as a pen and paper problem. Flexibility and cost-effectiveness are generally two advantages of solving the TFP of learners manually. However, the growth of online classrooms (e.g., MOOCs – Massive Open Online Courses) and remote education have transformed team formation for learners reducing it into an intractable problem when solved manually. More complications arise from the lack of time and familiarity with the students. Thus, many online tutors resort to either letting online learners form teams by themselves or relying on tools that automate the TFP for them.

One advantage of automated tools for team formation of learners is the computerization of matchmaking. The algorithm in charge of the team composition treats attributes as variables and distributes them according to quantifiable objectives such as an equal spread of academic grades across multiple teams. To date, many tools offer automated solutions to the TFP in education and utilize several criteria to profile learners [22]. The recent systematic literature review by Maqtary et al. [31] (2019) shows a great variety of team formation attributes and techniques within the educational domain to automate the team formation task. One example of such a system is CATME [58] and its "Team-Maker" tool that automatically forms teams based on student responses to a variety of categories such as demographics, performance metrics, and convenience. In the large-scale online education setting, research (e.g, [56, 64]) has experimented with criteria-based team formation algorithms yielding mostly positive results.

2.2 Learners' profiling attributes

For the past 200 years or so, education was mainly mass schooling with little to no adjustment to the individual's characteristics [57]. However, a recent growing trend of personalized education meant that schools and universities are acquiring recommender systems approaches to tailor education [60]. Personalized education is the systematic adaptation of instruction to individual learners [61]. Nowadays, teaching bodies can profile, classify, and assign students to courses and teams with relatively inexpensive methods and computational costs. The essential aspect of personalized education – and subsequently ad hoc team formation with learners – is the capacity to gather information about each individual and classify it in a meaningful way. From collecting information about the student before the course starts to documenting their performance and engagement, modeling profiles can be a static procedure (one-off) or a dynamic process (ongoing). Profiling attributes can be several depending on their relevance for the teaching agents and the different timescales [61]. According to Drachsler and Kirschner [13], there are at least four types of characteristics that differentiate learners, namely personal, academic, social/emotional, and cognitive.

Personal characteristics often relate to demographic information such as age, gender, maturation, language, social-economic status, cultural background, as well as specific needs (e.g., disabilities and impairments to learning). *Academic* characteristics are learning goals, knowledge, educational type, and educational level.

Social/emotional characteristics deal sociability, self-image (including self-efficacy and agency), mood, etc. Lastly, *cognitive characteristics* relate to attention, memory, mental procedures, and cognitive skills. Another critical set of characteristics is personality traits. In the broadest sense, personality traits are the aspects of individual differences that affect the human behaviors in different states [15]. There are many personality models and instruments to classify people based on their differences.

Some of the most used models in the education setting are the Five-Factor Model (FFM) or Big Five Model (OCEAN) [35], the Dominance, Influence, Steadiness, and Conscientiousness model (DISC) [59], the Myers-Briggs Type Indicator (MBTI) [34], the HEXACO model of personality structure personality inventory [4], the Revised NEO Personality Inventory (NEO-PI-R) [10], the Eysenck personality inventory [49], the Minnesota multiphasic personality Inventory [8], the Birkman method [16], and many more. Each model and inventory provides a different perspective about motivations, strengths, and weaknesses. It can shed light on the student's preferred thinking and working styles, communication, learning, managing, and team-working. Understanding personalities means determining the learner's motivation and how they relate to teammates, team roles, and shared workload. In this paper, we focus on the Big Five model and its five personality traits, namely Openness to experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

2.3 The Big Five personality traits in education

The Big Five framework of personality traits (also known as the Five-Factor Model (FFM) and OCEAN model) [35] is a robust model for understanding the relationship between personality and various academic behavior [46]. The first version of the Big Five came from Tupes and Christal [63] in 1961. However, only in the 80s and 90s did it reach an academic audience through the work of Digman [12] and Goldberg [18]. Nowadays, the five personality traits are considered the elementary structure behind all personality traits [42]. The Big Five traits, or dimensions, are: Openness to experience (inventiveness and curiosity; its opposite is consistency and caution), Conscientiousness (organization, efficiency, and responsibility; its opposite is extravagance and carelessness), Extroversion (assertiveness, sociability; its opposite is introversion), Agreeableness (compassion, friendliness, trust in others; its opposite is a criticism), and Neuroticism (tendencies toward sensitivity and anxiety; its opposite is confidence and resilience). In their meta-analysis of the Five-Factor Model of personality and academic performance, Poropat [46] lists theory-grounded arguments justifying the relationship between personality and learner achievement across academic subjects. The first theoretical basis for the Big Five is the lexical hypothesis that behavior and work outcome are related. According to this theory, performance in academic settings is determined by factors relating to capacity, opportunity, and willingness to perform [6].

2.4 Cognitive ability in education

Cognitive ability is the collection of skills needed to complete tasks such as thinking, learning, reading, remembering, speaking, listening, and focusing; it is the capacity to think in the abstract, reason, problem-solve, and comprehend [44]. Over a century of scientific research has shown that general cognitive ability (or *g*) predicts a broad spectrum of critical life outcomes, behaviors, and performances [26]. Considering academic achievement as a type of life outcome is no surprise that the educational setting presents several domain-specific and general cognitive ability tests [29, 50]. Cognitive ability instruments (e.g., the Miller Analogies Test [37]) are often present in educational admissions decisions as they estimate the relationship between cognitive ability and performance. Research has shown that cognitive ability tests handling cognition as a fixed property (entity theorists) rather than malleable (incremental theorists) produce very different results [17]. In the teamwork context for education, Liu et al. [27] proposes a two-stage framework to apply cognitive diagnosis for collaborative learning team formation. One quantifies the student skill proficiency (or cognitive ability); the other optimizes team formation based on dissimilarity-based and gain-based objectives. Experimenting with the framework with student teams produced better results than the baselines. More work using cognitive ability as a modeling feature [1, 9, 30, 65] investigate the TFP as a simulation and do not formally test their approaches through real-world experimental studies. In this work, we propose to compare intellectual ability with personality traits (Big Five) by assessing the ways users assemble teams of learners manually given a set of dummy profiles.

2.5 The User as Wizard method in team formation and education

Most research on team formation in education and academic performance optimization relies on top-down algorithmic methods based on learners' modeling and predefined objectives. However, human-centered approaches to the TFP have been proposing the principle of co-design and user engagement in the system design process. One of the most established approaches of this kind that examines users interacting with computers to facilitate rapid iterative development is the Wizard of Oz method [19]. Conventionally, the technique takes two machines linked together, one for the subject and one for the experimenter (the *Wizard* pretending to be a computer typing replies). One of the first implementations of the Wizard of Oz dates back to a study in 1985 [19]. The study featured the IBM Personal Computer used in several experiments with simulated user interfaces for an easy-to-use home computer banking program. Since then, the Wizard of Oz has been part of numerous other studies on human factors and human-computer interaction design.

In the TFP, the Wizard of Oz is mainly used to evaluate automated processes in cooperative scenarios such as human-robot-interaction [32, 62] and human-autonomy teaming [25, 36]. An alternative to the Wizard of Oz method is the User as Wizard method (UAW) formally introduced by Masthoff [33]. UAW predominately focuses on developing human-centered research to inspire algorithm adaptation. It places participants in the role of the Wizard and leaves them completely free to perform the task without a script to follow [33]. The method consists of two stages. One, called *Exploration stage*, sees participants taking the role of the adaptive system. The other named *Consolidation stage* requires participants to judge the

performance of others. We focus on the exploration stage by presenting participants with a scenario (team formation in education) and fictitious users (dummy profiles of learners). Three steps follow. 1) Giving participants the task the adaptive system is supposed to perform. 2) Finding out participants' reasons for their decisions and actions. 3) Repeating the previous steps for several scenarios (optional). The consolidation phase, which we do not consider in this study, comprises six steps. 1) Presenting participants with both the scenario and the fictional users as well as the users' intentions with the associated task. 2) Showing participants a human or system performance on the given task. 3) Asking participants to judge the performance. 4) Investigating the participants' reasons for their judgment. 5) Repeating steps 2 to 5 for a set of task performances. 6) Repeating steps 1 to 5 for many scenarios (optional).

In the TFP in education, the method is present in a user-led study by Odo et al. [41] investigating automatic group formation to improve the effectiveness of academic (and non-academic) collaboration. The study used a combination of Conscientiousness, Agreeableness, and ability levels to characterize twelve learners. It then asked twenty-four participants to form different-sized groups to ensure that they would work well together. The study shows that users account for personality and ability characteristics as they assemble teams of learners. Conscientiousness is overall weighted more than Agreeableness and Ability in the distribution of traits. Another interesting finding from the study by Odo et al. [41] is that team attributes such as cohesion and balance are taken into account as users form teams of learners.

3 METHODOLOGY

In our study setup, we consider the learners' profiling attributes as the independent variables and the teams' averaged attributes the dependent ones as the result of the team formation task. For our first part of the study, we run a crowdsourcing task to evaluate which of the Big Five personality traits users consider overall the most relevant to education and team formation with learners (Section 4). This was done using surveys with multiple-choice and open questions. In the second half of the study, we ask crowd workers to assemble teams of learners from a pool of fictitious profiles using a drag and drop application and following the Exploration stage of the UAW method (Section 5). From the latter, we gathered information on how users form teams given a set of learners and profiling attributes and their justification according to the answers given at the end of the task.

4 PRE-STUDY: OPENNESS AND CONSCIENTIOUSNESS THE MOST RELEVANT PROFILING ATTRIBUTES

Before running the UAW study, we carried out an exploratory study with a batch of participants (N=52) assessing which of the Big Five traits [53] users would consider most relevant in the education domain [1]. We included cognitive ability as a profiling characteristic known to affect performance [14]. The results from the exploratory study allowed us to narrow down the attribute lists of the learners to a smaller subset (Openness, Conscientiousness, and Ability) and reduced feature congestion [52]. In the survey, participants had to indicate on a Five-point Likert Scale how much they perceived each personality trait as important when forming teams of learners. The order of the attributes was shuffled to prevent presentation bias. Out of the Big Five attributes, Conscientiousness ($mean$=4.05, sd=0.82), and Openness ($mean$=4.09, sd=0.99) were the top two preferred attributes (Table 2). In comparison, Agreeableness scored lower ($mean$=3.84, sd=0.89), followed by Extraversion ($mean$=3.30, sd=0.94), and Neuroticism ($mean$=2.61, sd=1.25). According to these findings, Openness and Conscientiousness are the most important personality traits when profiling learners for team formation. We used these attributes plus cognitive Ability (as it is typically another known attribute in education) to profile learners in the follow-up UAW study.

[1]The reason for looking at a sub-set of profiling attributes rather than using them all concurrently was to avoid excessive feature congestion [52] occurring when too many elements clutter the UI. We run a pre-study with a small batch of crowd workers (N=52) recruited from the crowdsourcing platform Prolific [43]

5 MAIN STUDY: USER AS WIZARD WITH A DRAG-AND-DROP TEAM FORMATION ONLINE TOOL

5.1 Team formation tool

To enable participants to form teams, we developed a web-based application with Javascript and Flask. After registering with a username and password, participants were introduced to the task with a visual example and could read the explanation about each profiling attribute. Next, they formed four teams of three learners with a drag-and-drop card-based user interface. We used the results of the team formation task to address RQ2. Finally, participants answered the following questions.

- *Explain why you teamed the learners the way you did.* This was an open question, in which they could elaborate on what they thought was their strategy when dragging and dropping learners into the four teams of three.
- *Which attribute did you find most important when forming teams?* This was a multiple choice answer (Conscientiousness, Openness, and Ability).

5.1.1 Learners' profiling attributes. We created twelve fictitious learner profiles (Table 3) comprising value-neutral culture names [21] and three profiling attributes (Conscientiousness, Openness, and Ability). Indicating differences between learners were three attributes scores shown as low (red, 1/3 of the progress bar), medium (yellow, 2/3 of the progress bar), and high (green, 3/3 of the progress bar). Participants were informed about the meaning of these scores in the introduction part of the study (Table 1). We distributed the attribute scores according to the following criteria: a) half of each attribute scores (6/12) were medium, three were low (3/12), and the remaining three were high (3/12), b) no learner profile had more than one low and one high attribute score. The learners dummy profiles are listed in Table 3.

From these profiles, for the analysis of the results, we calculated the average for each team attribute (using high=3, medium=2, low=1) and classified the results within bounded ranges namely low ($LowTA \in [1, 1.33]$), medium ($MediumTA \in [1.34, 1.67]$), and high ($HighTA \in [2, 3]$). Since we used colors in the UI of the drag and drop team formation tool to represent the team averages, users

Attribute	**Low**	**High**
Openness	Dislikes changes Does not enjoy new things Resists new ideas Not very imaginative Dislikes abstract or theoretical concepts	Very creative Open to trying new things Focused on tackling new challenges Happy to think about abstract concepts
Conscientiousness	Dislikes structure and schedules Makes mess and doesn't care about things Fails to return things or put them back where they belong Procrastinates important tasks Fails to complete necessary or assigned tasks	Spends time preparing Finishes important tasks right away Pays attention to detail Enjoys having a set schedule
Ability	Low ability to produce ideas Struggles with cognitive problems	Excels at producing ideas Excellent at solving cognitive problems

Table 1: Short descriptions of low and high profiling attributes as shown in the User as Wizard study with crowd participants. The medium range was not included in the table as it was explained to be an equidistant point between the two extremes.

Learners' Attributes	*Mean*	*SD*	*SE*
Openness	4.09	0.99	0.13
Conscientiousness	4.05	0.82	0.11
Extraversion	3.30	0.94	0.13
Agreeableness	3.84	0.89	0.12
Neuroticism	2.61	1.25	0.17

Table 2: Mean, Standard Deviation (SD), and Standard Error (SE) of each Big Five personality traits according to the pre-study participants (*N*=52). Their preference of profiling attributes for team formation of learners indicates that Openness and Conscientiousness are the most favored traits.

Learner's name	**Openness**	**Conscientiousness**	**Ability**
Andy	low	medium	high
Bo	medium	low	medium
Carl	high	medium	low
Darrel	medium	medium	low
Edwin	medium	low	high
Finn	high	low	medium
Grant	medium	high	medium
Hunter	low	medium	medium
Ian	medium	medium	high
Josh	low	high	medium
Karter	high	medium	medium
Liam	medium	high	low

Table 3: Learners' dummy profiles used for the User as Wizard study with their profiling attributes' scores (low, medium. high).

would see the aggregated team attributes in the form of a bar above each team divided into three equally sized sections (see Figure 1). These sections had labels with background colors changing according to the computed averaged team attribute (red=low, yellow=medium, and green=high). For example, with low team openness, medium Conscientiousness, and high ability, the bar would have sections colored in red, yellow, and green accordingly. The visualization would indicate that the team attributes are not entirely 'balanced' without showing more descriptive information such as numeric averages.

Combinatorics and probabilities. To consider the validity of our findings, we first calculated all possible combinations, so that we would know the likelihood of our outcomes being due to randomness. The first team of learners had 220 possible combinations calculated with the formula:

$$C_{n,r} = \frac{n!}{r!(n-r)!}$$

where r is the size of each team (3 in our study), and n is the number of possible people to include in a team at the start (12 in our study). By calculating the averages of the attribute scores of all

Figure 1: Overview of the team formation card-based drang-and-drop UI. It allowed users to form teams by placing learners into four separate containers representing four teams and to adjust their compositions by dragging the learners' cards between them.

these possible teams (called *TA*), we obtained the probability of all three team attributes being of *LowTA*, *MediumTA*, and *HighTA* average at once ($P(TA_{open}) = P(TA_{cons}) = P(TA_{able})$). Then, we compared these probabilities with those of 2/3rd of the attribute means classification being the same, and with none of them being the same. For example, suppose a team contains Andy, Edwin, and Ian. Looking at the dummy profiles in Table 3, we can derive the averaged team attributes as follows. TA_{able}=Average(3,3,3)=3; TA_{open}=Average(1,2,2)=1.67; TA_{cons}=Average(2,1,2)=1.67. Next, we can classify the team averages according to the bounded ranges introduced in Section 5.1.1. In this case, the given combination has a high Team Ability (since its TA_{able}=3), a medium Team Openness (TA_{open}=1.67) and a medium Team Conscientiousness (TA_{cons}=1.67). More exactly, we can state that this team has

$HighTA_{able}$, $MediumTA_{open}$, and $MediumTA_{cons}$ and that it is semi-balanced since 2/3 of its averaged attribute are the same.

Our calculation of combinations and dependent probabilities is presented in Table 4. The results show that the first team always has the same n=220 possible combinations. However, the second team has n=84 possible combinations for each possible first team. The third has n=20 possible combinations for each second and first possible team and lastly, the fourth has only one possible combination dependent on the third, second, and first team combinations. In the table, we show the calculation of possible combinations given one first team [2]. The results are summarized as follows.

LowTA. As shown in the *LowTA* column of Table 4, we see that the first team has approximately 26% probability of being unbalanced on *LowTA*. It is of the first teams. The remaining 163 possible first teams (approximately 74% of the total) do not have any *LowTA*. Therefore, it is a higher likelihood to randomly form first teams with no *LowTA* than to form first teams with one *LowTA*. It is not possible to randomly form teams with two or more *LowTA*. Similar results show across the other three teams where the only *LowTA* is in the form of unbalanced team composition (where 1/3rd of the attributes is *LowTA*). There is a 17% probability of randomly forming the second team with one *LowTA* and an 83% probability of having no *LowTA*. The third team has a 15% probability of being randomly made with one *LowTA* (3 out of 20 possible teams) and a higher probability (85%) of not having a single *LowTA*. Finally, the fourth team has no probability of getting *LowTA* in the given probabilities.

MediumTA. Looking at the column corresponding to *MediumTA*, we notice that when forming the first team randomly, there is a higher probability of that being semi-balanced (approximately 49%) than fully-balanced (16.4%) or unbalanced (27.3%) on *MediumTA*. The second team has an even higher probability of being randomly semi-balanced (54.4%), followed by unbalanced (27.4%) and full-balanced (14.3%). For the third and fourth teams, there is no probability of being fully-balanced on *MediumTA* in this case [3]. The third team has 55% probability of being semi-balanced on *MediumTA* while the last team has a 100% probability of being semi-balanced on the same attribute. Unbalanced third teams on *MediumTA* are also possible by random formation with a 35% probability. The same is not possible for the fourth team. There is a small probability that none of the four teams have *MediumTA* (see column *MediumTA* in Table 4).

HighTA. The high attribute average *HighTA* is most likely to show in randomly formed first teams but only as an unbalanced team (approximately 63% of the time). The first team also has a smaller probability of not having *HighTA* (19.1%) and being semi-balanced on *HighTA* (approximately 18%). Unbalanced *HighTA* teams are also more probable for the second teams (64.3%) more so than semi-balanced *HighTA* (20.2%) and fully-balanced on the same attribute (just above 1%). There is approximately a 14% probability that the second team has no *HighTA*. The third teams have the highest probability of being un-balanced on *HighTA* (65%) and lesser of being semi-balanced (30%) and fully-balanced (5%) on the same attribute average. Finally, the fourth team has a certain probability of being an un-balanced *HighTA* team in the given combination and no probability of it being fully or semi-balanced.

5.2 Participants

Similar to the pre-study, we recruited test subjects through Prolific. Out of a preliminary batch (*N*=120), most participants (*N*=108) successfully completed the task [4]. Of the valid subset of participants, almost half were male (*N*=55) and the other half female (*N*=53). Participants were mostly European (*N*=80) followed by African (*N*=14), South American (*N=12*), and Middle Eastern (*N=2*). From their contribution, we yielded 432 teams of learners (four for each of the 108 participants) which we used in our analysis.

5.3 Results

5.3.1 *Forming teams of learners is a balancing act of their within and between attributes.*

Within-Teams. In this section, we address RQ1 **'How do users distribute learners' attributes within and between teams?'**. We analyze the way participants distributed the learners' attributes both within and between teams. For the second hypothesis (*'H1.1: "We expect people to distribute attributes equally within teams'*) we considered the *TA* within-team distribution including *LowTA*, *MediumTA*, and *HighTA*. The majority of the fully-balanced teams consisted of *MediumTA* (*N*=115, or 26% of all teams), meaning that all three team's attributes were of medium average ($TA_{open} = TA_{cons} = TA_{able}$). *HighTA* (*N*=5) consisted of a minority of the fully-balanced teams. No fully-balanced teams were possible with *LowTA*. Counting the teams that had 2/3rd of the same *TA*, we noted that 60% (*N=220*) had two *MediumTA* and 10% (*N*=47) had two *HighTA*. No partially balanced teams with *LowTA* were found.

Between-teams. To test the hypothesis (*H1.2: "We expect people to distribute attributes equally between teams"*), we compared the *TA* differences between the four teams formed by the participant. Grouping the results for the fully-balanced, semi-balanced, and unbalanced teams, we formed an overview of how many participants managed to assemble all fully-balanced teams and how many others distributed the attributes differently (i.e., semi-balanced or unbalanced).

- **Fully balanced teams.** Observing the number of fully-balanced *TA* between teams, we note that there was a similar number of fully-balanced first, second, and third teams ($N_{T_{1st}} = 35$, $N_{T_{2nd}} = 31$, $N_{T_{3rd}} = 30$). The fourth team tended to be less fully-balanced ($N_{T_{4th}} = 24$). This may indicate that the number of fully-balanced teams slightly decreased with the order of the teams. Together (*N*=120) the fully-balanced teams made up 27% of all teams.
- **Semi-balanced teams:** Considering the number of semi-balanced *TA* between teams, we observe that the first teams were slightly less semi-balanced ($N_{T_{1st}} = 57$) than the second

[2]For the entire simulation of all dependent probabilities we recommend dedicating a separate study.
[3]However, the probabilities herein presented depend on only one combination of the first team. We discuss this limitation later in the paper.
[4]Test subject compensation complied with the Prolific recommended minimum wage (6.18/hour GBP) [47]. On average, the participants spent 10 minutes on the task and thus received approximately 1 GBP pp.

	LowTA				*MediumTA*				*HighTA*			
	Full-B	**Semi-B**	**Un-B**	**None**	**Full-B**	**Semi-B**	**Un-B**	**None**	**Full-B**	**Semi-B**	**Un-B**	**None**
1st (N=220)	n=0	n=0	n=57 (25.9%)	n=163 (74.1%)	n=36 (16.4%)	n=108 (49.1%)	n=60 (27.3%)	n=16 (0.073%)	n=1 (0.5%)	n=39 (17.7%)	n=138 (62.7%)	n=42 (19.1%)
2nd (N=84)	n=0	n=0	n=14 (16.7%)	n=70 (83.3%)	n=12 (14.3%)	n=44 (52.4%)	n=23 (27.4%)	n=5 (0.06%)	n=1 (1.2%)	n=17 (20.2%)	n=54 (64.3%)	n=12 (14.3%)
3rd (N=20)	n=0	n=0	n=3 (15%)	n=17 (85%)	n=0	n=11 (55%)	n=7 (35%)	n=2 (10%)	n=1 (5%)	n=6 (30%)	n=13 (65%)	n=0
4th (N=1)	n=0	n=0	n=0	n=1 (100%)	n=0	n=1 (100%)	n=0	n=0	n=0	n=0	n=1 (100%)	n=0

Table 4: Dependent probabilities of *LowTA*, *MediumTA*, and *HighTA* occurring for at least one attribute in the four teams, and how often those teams had all attributes with the same average classification (indicated as Full-B), 2/3rd of the attributes with the same average classification (Semi-B), or no attributes with the same average (Un-B). None indicates the number of teams in which the classification did not occur at all.

and third teams ($N_{T_{2nd}} = 69$, $N_{T_{3rd}} = 65$). The fourth team was the most semi-balanced ($N_{T_{4th}} = 76$). In total, *N*=267 teams were semi-balanced which made up almost 62% of all teams.

- **Unbalanced teams:** Similar to the semi-balanced teams, the numbers of unbalanced combinations of *TA* between the first, second, third, and fourth teams did not greatly differ ($N_{T_{1st}} = 16$, $N_{T_{2nd}} = 8$, $N_{T_{3rd}} = 13$ and $N_{T_{4th}} = 8$) and do not seem to follow a linear trend. The total number unbalanced teams was *N=45* and this made up 10% of all teams.

Finally, we analyzed the open responses to the survey question *'Explain why you teamed the learners the way you did.'*. We stripped the answers from stopwords and counted the frequency of terms. The most frequent words related to the subject of attributes distribution were: *'balance' (N=50), 'medium' (N=10), 'equal' (N=10)* and *'average' (N=9)*. Some examples of sentences were: *'I grouped the learners by similar attributes levels'*, *'I focused on balancing the attributes and grouping learners with attributes that would enhance the group and, make sure that they can perform well.'*, *'I tried to group everyone so it results in medium stats across the board.'*, and *'everything is evenly distributed and no group has a clear advantage'*.

5.3.2 Conscientiousness matters more than Ability when forming teams of learners. By analyzing the responses to the survey question *"Which attribute did you find most important when you created teams?"* – where participants indicated the most relevant attribute – we addressed RQ2 (**"Which attributes do users consider most important in the team formation of learners?"**) and evaluated H2 (*"We expect cognitive ability to be the most important trait when forming teams of learners online"*). The descriptive results show that Conscientiousness was the most preferred trait (*N*=45) followed by cognitive Ability (*N*=38) and Openness (*N*=25). These results demonstrate that Conscientiousness first, and cognitive Ability after, are considered the most relevant by people forming teams of learners online.

6 DISCUSSION AND LIMITATIONS

We investigated team formation in education from the human-centered User as Wizard approach. With the pre-study and the main study findings, we could gather insight into what people do when assembling teams of learners having only a limited amount of knowledge about the individual's characteristics. Even though the probability of forming semi-balanced teams was high (as seen from the calculation of the dependent probabilities), participants frequently expressed their intent to balance teams when asked to explain their approach. Comparing the likelihood of forming balanced teams randomly with the percentage of balanced teams according to the participants, we can conclude that people strive to balance teams even when they do not know the learners or the exact task requirements (e.g., optimal strategy). These results conform with the perception that imbalances in teams' attributes negatively correlate with performance [5]. Comparing our results with those from Odo et al. [41], we note that balance is one of the most common strategies in both studies where balance is considered a between-teams attribute and where all teams have approximately equal conditions/opportunities. With the goal of within-balanced team attributes, the participants had to redistribute learners to achieve similar results across the formations. This balancing act is known in previous research to count in situations where personality and complimentary matter [28].

We discuss some limitations that have affected the study design and the results. a) *One too many profiling attributes.* Despite reducing the number of attributes displayed on the learners' profiles, it is likely that participants still struggled to decide on which attribute to base their team formation on; b) *Not enough students and educators in the sample.* For this study, we asked crowd workers to participate as our sample. Although a small number (*N*=47) were said to be students, there were no requirements to be involved in the educational domain as either a learner or an educator. The monetary objective of partaking in a crowdsourcing study may have stirred and biased participants away from the purpose of the study. c) *Limited degrees of freedom when forming teams.* We used twelve dummy profiles. This number is a limiting factor to team composition as there are only a (relatively) small amount of possible combinations. The same applies to the scale of the attributes scores that strongly limited the variability of the compositions.d0 *The calculation of probabilities depended on one first team.* The study mainly focused on the UAW method to observe users interacting with systems while executing a team formation problem. Hence, we limited the calculation of random probabilities to one case where the first team was known and the other three dependent team formation probabilities were derived. Future work is needed to extend the calculation of all probabilities where all teams are known.

7 CONCLUSION AND FUTURE WORK

Forming teams of learners can be a daunting task for educators resorting to relying on automated tools to do the job. In this work, we investigate the popular approach when forming teams of learners manually to extract a generic method. Our results indicate that

people tend to prefer balancing teams by attributes by distributing them equally within and between teams of learners. We also noted that people explicitly find certain profiling traits more relevant than others. cognitive Ability, Conscientiousness, and Openness were the most favored. Some of our findings (Conscientiousness is relevant to profiling learners and balancing is important in forming teams of students) are supported by similar research on team formation in education. However, future work should experiment with more than a small batch of learners and four teams and use more degrees to represent the attributes levels of the learners' profiles.

REFERENCES

[1] Rakesh Agrawal, Behzad Golshan, and Evimaria Terzi. 2014. Grouping students in educational settings. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1017–1026.

[2] Wayne M Alves and Peter H Rossi. 1978. Who should get what? Fairness judgments of the distribution of earnings. *American journal of Sociology* 84, 3 (1978), 541–564.

[3] Sofiane Amara, Joaquim Macedo, Fatima Bendella, and Alexandre Santos. 2016. Group formation in mobile computer supported collaborative learning contexts: A systematic literature review. *Journal of Educational Technology & Society* 19, 2 (2016), 258–273.

[4] Michael C Ashton and Kibeom Lee. 2007. Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and social psychology review* 11, 2 (2007), 150–166.

[5] Paul Blowers. 2003. Using student skill self-assessments to get balanced groups for group projects. *College Teaching* 51, 3 (2003), 106–110.

[6] Melvin Blumberg and Charles D Pringle. 1982. The missing opportunity in organizational research: Some implications for a theory of work performance. *Academy of management Review* 7, 4 (1982), 560–569.

[7] Steffen Brauer and Thomas C Schmidt. 2012. Group formation in elearning-enabled online social networks. In *2012 15th international conference on interactive collaborative learning (ICL)*. IEEE, 1–8.

[8] James N Butcher. 2010. Minnesota multiphasic personality inventory. *The Corsini Encyclopedia of Psychology* (2010), 1–3.

[9] Christos E Christodoulopoulos and K Papanikolaou. 2007. Investigation of group formation using low complexity algorithms. In *Proc. of PING Workshop*. Citeseer, 57–60.

[10] Paul T Costa Jr and Robert R McCrae. 2008. *The Revised Neo Personality Inventory (neo-pi-r)*. Sage Publications, Inc.

[11] Jean-Claude Croizet and Theresa Claire. 1998. Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin* 24, 6 (1998), 588–594.

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

[13] Hendrik Drachsler and Paul A Kirschner. 2011. Learner characteristics.

[14] Jan J Elshout and Marcel VJ Veenman. 1992. Relation between intellectual ability and working method as predictors of learning. *The Journal of Educational Research* 85, 3 (1992), 134–143.

[15] Somayeh Fatahi, Hadi Moradi, and Leila Kashani-Vahid. 2016. A survey of personality and learning styles models applied in virtual environments with emphasis on e-learning environments. *Artificial Intelligence Review* 46, 3 (2016), 413–429.

[16] Sharon Birkman Fink and Stephanie Capparell. 2013. *The Birkman method: Your personality at work*. John Wiley & Sons.

[17] Adrian Furnham, Tomas Chamorro-Premuzic, and Fiona McDougall. 2003. Personality, cognitive ability, and beliefs about intelligence as predictors of academic performance. *Learning and individual Differences* 14, 1 (2003), 47–64.

[18] Lewis R Goldberg. 1990. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology* 59, 6 (1990), 1216.

[19] Paul Green and Lisa Wei-Haas. 1985. The rapid development of user interfaces: Experience with the Wizard of Oz method. In *Proceedings of the Human Factors Society Annual Meeting*, Vol. 29. SAGE Publications Sage CA: Los Angeles, CA, 470–474.

[20] Jimmy H Gutiérrez, César A Astudillo, Pablo Ballesteros-Pérez, Daniel Mora-Melià, and Alfredo Candia-Véjar. 2016. The multiple team formation problem using sociometry. *Computers & Operations Research* 75 (2016), 150–162.

[21] Matthew W Hahn and R Alexander Bentley. 2003. Drift as a mechanism for cultural change: an example from baby names. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270, suppl_1 (2003), S120–S123.

[22] Farnaz Jahanbakhsh, Wai-Tat Fu, Karrie Karahalios, Darko Marinov, and Brian Bailey. 2017. You want me to work with who? Stakeholder perceptions of automated team formation in project-based courses. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 3201–3212.

[23] Amir Karimi and Randall D Manteufel. 2020. Performance Balanced Team Formation for Group Study and Design Projects. In *2020 ASEE Virtual Annual Conference Content Access*.

[24] Ryutaro Kawaguchi, Masashi Hayano, and Toshiharu Sugawara. 2015. Balanced Team Formation for Tasks with Deadlines. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Vol. 2. IEEE, 234–241.

[25] Ali Hosseini Khayat. 2009. *Distributed Wizard of Oz Usability Testing for Agile Teams*. UNIVERSITY OF CALGARY.

[26] Nathan R Kuncel, Sarah A Hezlett, and Deniz S Ones. 2004. Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of personality and social psychology* 86, 1 (2004), 148.

[27] Yuping Liu, Qi Liu, Runze Wu, Enhong Chen, Yu Su, Zhigang Chen, and Guoping Hu. 2016. Collaborative learning team formation: a cognitive modeling perspective. In *International Conference on Database Systems for Advanced Applications*. Springer, 383–400.

[28] Ioanna Lykourentzou, Angeliki Antoniou, Yannick Naudet, and Steven P Dow. 2016. Personality matters: Balancing for personality types leads to better outcomes for crowd teams. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 260–273.

[29] Nils Machts, Johanna Kaiser, Fabian TC Schmidt, and Jens Moeller. 2016. Accuracy of teachers' judgments of students' cognitive abilities: A meta-analysis. *Educational Research Review* 19 (2016), 85–103.

[30] Barati Jozan M Mahdi and Taghiyareh Fattaneh. 2013. A semi-pareto optimal set based algorithm for grouping of students. In *4th International Conference on e-Learning and e-Teaching (ICELET 2013)*. IEEE, 10–13.

[31] Naseebah Maqtary, Abdulqader Mohsen, and Kamal Bechkoum. 2019. Group formation techniques in computer-supported collaborative learning: A systematic literature review. *Technology, Knowledge and Learning* 24, 2 (2019), 169–190.

[32] Matthew Marge, Claire Bonial, Brendan Byrne, Taylor Cassidy, A William Evans, Susan G Hill, and Clare Voss. 2017. Applying the Wizard-of-Oz technique to multimodal human-robot dialogue. *arXiv preprint arXiv:1703.03714* (2017).

[33] Judith Masthoff. 2006. The user as wizard: A method for early involvement in the design and evaluation of adaptive systems. In *Fifth workshop on user-centred design and evaluation of adaptive systems*, Vol. 1. Citeseer, 460–469.

[34] Robert R McCrae and Paul T Costa Jr. 1989. Reinterpreting the Myers-Briggs type indicator from the perspective of the five-factor model of personality. *Journal of personality* 57, 1 (1989), 17–40.

[35] Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality* 60, 2 (1992), 175–215.

[36] Nathan McNeese, Mustafa Demir, Erin Chiou, Nancy Cooke, and Giovanni Yanikian. 2019. Understanding the role of trust in human-autonomy teaming. In *Proceedings of the 52nd Hawaii international conference on system sciences*.

[37] Don Meagher. 2006. Introduction to the Miller Analogies Test.

[38] Amir Mujkanovic, David Lowe, Keith Willey, and Christian Guetl. 2012. Unsupervised learning algorithm for adaptive group formation: Collaborative learning support in remotely accessible laboratories. In *International Conference on Information Society (i-Society 2012)*. IEEE, 50–57.

[39] Nur Baiti Afini Normadhi, Liyana Shuib, Hairul Nizam Md Nasir, Andrew Bimba, Norisma Idris, and Vimala Balakrishnan. 2019. Identification of personal traits in adaptive learning environment: Systematic literature review. *Computers & Education* 130 (2019), 168–190.

[40] Chinasa Odo, Judith Masthoff, and Nigel Beacham. 2019. Group formation for collaborative learning. In *International Conference on Artificial Intelligence in Education*. Springer, 206–212.

[41] Chinasa Odo, Judith Masthoff, and Nigel A Beacham. 2019. Adapting Online Group Formation to Learners' Conscientiousness, Agreeableness and Ability.. In *SLLL@ AIED*. 1–7.

[42] Brian P O'Connor. 2002. A quantitative review of the comprehensiveness of the five-factor model in relation to popular personality inventories. *Assessment* 9, 2 (2002), 188–203.

[43] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (2017), 153–163.

[44] Robert Plomin. 1999. Genetics and general cognitive ability. *Nature* 402, 6761 (1999), C25–C29.

[45] NF Plotnikova and EN Strukov. 2019. Integration of teamwork and critical thinking skills in the process of teaching students. *Cypriot journal of educational sciences* 14, 1 (2019), 1–10.

[46] Arthur E Poropat. 2009. A meta-analysis of the five-factor model of personality and academic performance. *Psychological bulletin* 135, 2 (2009), 322.

[47] Prolific. 2022. Prolific: Pricing. https://www.prolific.co/pricing.

[48] Malcolm James Ree and James A Earles. 1992. Intelligence is the best predictor of job performance. *Current directions in psychological science* 1, 3 (1992), 86–89.

[49] Thomas Rocklin and William Revelle. 1981. The measurement of extroversion: A comparison of the Eysenck Personality Inventory and the Eysenck Personality Questionnaire. *British journal of social psychology* 20, 4 (1981), 279–284.
[50] Treena Eileen Rohde and Lee Anne Thompson. 2007. Predicting academic achievement with cognitive ability. *Intelligence* 35, 1 (2007), 83–92.
[51] Juan Pablo Román-Calderón, Sara Aguilar-Barrientos, Juan Esteban Escalante, Jaime Barbosa, and Alejandro Arias Salazar. 2021. The effect of student work group emotional intelligence on individual task performance in teams. *Journal of Experiential Education* 44, 2 (2021), 121–136.
[52] Ruth Rosenholtz, Yuanzhen Li, Jonathan Mansfield, and Zhenlan Jin. 2005. Feature congestion: a measure of display clutter. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 761–770.
[53] Sebastiaan Rothmann and Elize P Coetzer. 2003. The big five personality dimensions and job performance. *SA Journal of industrial psychology* 29, 1 (2003), 68–74.
[54] Paul A Rubin and Lihui Bai. 2015. Forming competitively balanced teams. *IIE Transactions* 47, 6 (2015), 620–633.
[55] Eduardo Salas, Dana E Sims, and C Shawn Burke. 2005. Is there a "big five" in teamwork? *Small group research* 36, 5 (2005), 555–599.
[56] Luisa Sanz-Martínez, Alejandra Martínez-Monés, Miguel L Bote-Lorenzo, Juan A Munoz-Cristóbal, and Yannis Dimitriadis. 2017. Automatic group formation in a MOOC based on students' activity criteria. In *European Conference on Technology Enhanced Learning*. Springer, 179–193.
[57] Evan Schofer, Francisco O Ramirez, and John W Meyer. 2021. The societal consequences of higher education. *Sociology of Education* 94, 1 (2021), 1–19.
[58] Chelsey S Simmons. 2015. Using CATME team-maker to form student groups in a large introductory course. In *American Society for Engineering Education Southeast Section Conference*.
[59] Jeffrey Sugerman. 2009. Using the DiSC® model to improve communication effectiveness. *Industrial and Commercial Training* (2009).
[60] Chuan Sun, Hui Li, Xiuhua Li, Junhao Wen, Qingyu Xiong, and Wei Zhou. 2020. Convergence of recommender systems and edge computing: A comprehensive survey. *IEEE Access* 8 (2020), 47118–47132.
[61] Leonard Tetzlaff, Florian Schmiedek, and Garvin Brod. 2021. Developing personalized education: A dynamic framework. *Educational Psychology Review* 33, 3 (2021), 863–882.
[62] Thomas Trainer, John R Taylor, and Christopher J Stanton. 2020. Choosing the Best Robot for the Job: Affinity Bias in Human-Robot Interaction. In *International Conference on Social Robotics*. Springer, 490–501.
[63] Ernest C Tupes and Raymond E Christal. 1992. Recurrent personality factors based on trait ratings. *Journal of personality* 60, 2 (1992), 225–251.
[64] Miaomiao Wen, Keith Maki, Steven Dow, James D Herbsleb, and Carolyn Rose. 2017. Supporting virtual team formation through community-wide deliberation. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–19.
[65] Virginia Yannibelli and Analía Amandi. 2012. A deterministic crowding evolutionary algorithm to form learning teams in a collaborative learning context. *Expert systems with applications* 39, 10 (2012), 8584–8592.