The designing of Performance Management Systems in OLPs

Author: Nataly Minassian
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

ABSTRACT

This paper investigates how Performance Management Systems (PMSs) are developed today in Online Labor Platforms (OLPs), as Artificial Intelligence (AI) is taking over the Performance Management Process to assess Gig Workers' Performance. This is interpreted through the PMS developed in five steps: 1) Identifying Performance Dimensions, 2) Developing Performance Measures, 3) Evaluating Employee Performance, 4) Providing Feedback, and 5) Developing Action Plans To Improve Workers' Performance by using Lepak and Gowan (2010) theory to come up with conclusions accordingly. This paper concludes with PMSs development being very complex to ensure an accurate and clear system by finding out many factors and inconsistencies that lead to such complexity. In a system where a customer is "Queen/ King", the whole system relies on each customer's satisfactory factors, making it difficult to identify Performance Dimensions and further develop the PMSs "appropriately". This difficulty contributes to the system's complication in ensuring validity, reliability, specificity, and being free of any bias throughout the whole process. Furthermore, this paper exposes and elaborates on how OLPs' Performance Management is different from theory and the possible reasons.

Graduation Committee members:
1st Supervisor: DR. J.A. HÜLLMANN
2nd Supervisor: DR. J.G. MEIJERINK

Keywords
Human Resource Management (HRM), Online Labor Platforms (OLPs), Algorithms, Gig Workers (GWs), Performance Management Systems, Rating Systems

1. INTRODUCTION

The adoption of AI algorithms to Human Resource Management (HRM) in OLPs is intrinsically new, bringing to the surface several questions and highlighted imperfections regarding the algorithm design (Park et al., 2022) in the context of Performance Management of Gig Workers (GWs). OLPs refer to organizations that rely on information technology, such as Artificial Intelligence (AI) and software algorithms, to matchmake between requesters (i.e., organizations or consumers) and on-demand short-term labor (i.e., gig workers) via an online marketplace. OLPs are present within many various industries like transportation (Uber), food delivery (Deliveroo) or freelancing (Upwork) (Meijerink & Arets, 2021; Kuhn & Maleki, 2017; Meijerink & Keegan, 2019). The AI algorithm here is enabled by Machine Learning, a category of algorithms, statistical models, and procedures, which are trained with data to make predictions, determine actions, and improve automatically through experience, data collection, and training (Ray, 2019; Nitzberg & Zysman, 2022). With the rise of OLPs, different ways of working have been developed. One of these is the gig economy, where short-term or temporary workers, also known as Gig Workers (GWs), are required to complete individual tasks, assignments, or jobs (Kumar et al., 2022). This free-market system has seen significant growth over the past few years (Frenken et al., 2020). While GWs are perceived as independent, autonomous workers, technology has a significant influence over their work. This is especially true through HRM activities, such as recruitment, appraisal, task allocation, compensation, and job design (Immonen, 2023; Kuhn and Maleki, 2017; Connelly et al., 2021; Keegan and Meijerink, 2021; Meijerink et al., 2021b; Waldkirch et al., 2021).

OLPs, such as Uber, Lyft, Uber Eats, and Deliveroo, have been studied to investigate the effects of algorithmic control and management (Chaouali et al., 2022). This has mostly been done by conducting interviews and getting an insight into gig workers' perspectives and opinions regarding algorithmic management (Mohlmann et al., 2021), with many of these viewpoints concerning performance management. There are many different processes, such as daily work practices, task allocations, or the way work is delivered, which are all fully based on algorithmic management (Mohlmann et al., 2021). OLPs being driven by Information Technology (IT) comes with its advantages, both from an organizational perspective and a GW perspective. GWs benefit from the flexibility of work, different adjustable time schedules to choose from, and overall independence (Chaouali et al., 2022). From an organizational perspective, IT offers the opportunity to maximize profit, be more innovative, and rely on the algorithm to keep workers under control (Mohlmann et al., 2021).

On the contrary, despite the positive aspects, many GWs complain that they are facing struggles regarding algorithmic control when it comes to how their performance is managed (Mohlmann et al., 2021). GWs have expressed that they feel that they have been evaluated falsely because the assessment is many times based on reasons that are beyond their control (Mohlmann et al., 2021). For instance, according to interviews with GWs working for Uber Drivers, GWs mentioned unfair ratings by their customers for reasons such as weather conditions, traffic, or surge pricing (Mohlmann et al., 2021). Furthermore, they have expressed their uneasiness regarding the ability to meet and exceed their requester's expectations to prevent getting a bad rating (Mohlmann et al., 2021). A similar struggle can be found within Quora as well, where there have been more complaints regarding Uber's rating system. Here, Uber drivers talk about how their requesters would give them a falsified rating or how the rating system is stressful for them because of its precariousness (Quora, n.d.). This false evaluation leads to the GWs having bad ratings and, therefore, a "bad reputation" to their requesters. This can result in getting worse match-making to customers by the algorithm (Mohlmann et al., 2021). Evaluations like this raise the GWs' work uncertainty as it can influence their income and overall job opportunities, making the effectiveness of the PMSs' design questionable.

The aforementioned concerns create controversy regarding the designing process of PMSs in OLPs. For example, the GWs being evaluated for reasons that are beyond their control and, nevertheless, negatively affect their overall rating score is conflicting with the GWs, and raises questions regards the validity embedded in the system (Kelly et al., 2016). Therefore, it is necessary to conduct an inquiry when it comes to the process of the PMSs' design to prevent such circumstances. A PMS can be considered effective when performance it's accurately developed through the Performance Management Process, ensuring validity, reliability, and specificity without biases occurring (Lepak & Gowan, 2010). The performance management process consists of activities such as identifying performance dimensions, developing and implementing performance measures, performance evaluation, providing feedback, and creating action plans for the development of performance (Lepak & Gowan, 2010).

Therefore, it is important to study how PMSs are designed in the OLPs context to prevent such errors and false evaluations and what actions are taken in circumstances where measurement errors or any kind of inconsistency are detected within the systems. There is a lack of literature on how the PMSs' processes are fully developed in OLPs, specifically performance dimensions, measures, worker performance evaluation, feedback, and further actions for development. It is also unclear how these aspects are prioritized in the designing of PMSs.

The indicated concerns lead to the following research question this paper will be focusing on:

How are performance management systems (PMSs) designed in online labor platforms (OLPs)?

This study contributes to theory by addressing the gap in the literature surrounding the designing process of PMSs in OLPs, by building upon existing literature and collecting new data from online forums. The existing literature is mostly uncovering information concerning the influences or outcomes the GWs face because of algorithmic control from Performance Management, such as control, changes in their task allocations (Mohlmann et al., 2021), the HR structure model (Keegan & Meijerink, 2022), or the use and definition of HRM activities in OLPs on a more general level (Keegan & Meijerink, 2021). There is limited literature on how decisions are made regarding the design of the PMSs, especially when it comes to how the complaints and concerns about PMSs are addressed (Mohlmann et al., 2021). Hence, this study intends to look further into the decision-making made to develop PMSs and how it is developed through each step, but also investigate the extent these complaints are addressed adequately.

Thus, it contributes by providing new insights into OLPs by addressing the aforementioned concerns for the PMSs to draw assumptions, as it is the most relevant data, but also getting clearer and more accurate information on how the process and the steps are developed in the PMSs (Lepak & Gowan, 2010). The contributions offered to practice are helping to have a more mindful design of PMSs in OLPs, where the PMSs, such as reputation systems and rating systems, are trustful, reliable, valid, and specific, followed by unprejudiced evaluations and developing plans (Garg & Johari, 2021). Since the adoption of AI algorithms to HRM, and specifically PMSs, is new, the paper will bring more awareness to AI designers and people involved with this matter on how decision makings have consequences and help stakeholders take actions for improvements and more strategic decisions regarding the PMSs, and their overall performance management approach. Lastly, OLPs will be able to add more value to their platform by treating employees better through their PMSs. Besides ensuring a PMS design works to prevent errors and such circumstances, a platform will be able to improve its whole performance as well, such as increasing the achievement of results, profits, a better-aligned strategy, and innovativeness (Kourtit & De Waal, 2008).

As for the structure of this paper, the first existing literature and the theoretical concepts will be briefly explained after the introduction of the topic and problem statement. Secondly, the methodology used is discussed, and the data collection and data analysis are explained. Afterward, the findings will be elaborated, followed by discussions and conclusions, to sum up, and present the key takeaways.

2. LITERATURE REVIEW

2.1 OLPs usage of PMSs

The OLPs' PMSs, or otherwise "reputation systems," consist of reputation scores and ratings for the GWs, constructed by customers' or requesters' feedback (Tan et al., 2021). Reputation and rating scores are considered part of the algorithmic management mechanisms set to supervise and monitor the quality of the work delivered. Furthermore, they are used as a trust mechanism to assess and facilitate further interactions (Tan et al., 2021). The goal of OLPs is to generate income by making reliable and consistent match-making with the use of algorithms, and the PMSs are assisting in improving match-making, making it more balanced and reliable (Keegan & Meijerink, 2022). Thus, this concludes how important the use of PMSs in the OLPs is, as they are necessary to operate but also supervise their workers' quality work.

Many GWs are bound to receive feedback, either directly or indirectly, through rankings and/or ratings that are embedded in multiple OLPs. Here, metrics are produced to evaluate and monitor the GWs' performance (Gandini, 2019). Keegan and Meijerink (2022) provided examples of OLPs and how their performance appraisal systems work. Accordingly, many performance appraisal systems are based on behaviors, procedures, and standards the GW has to comply with. The performance appraisal would then be based on inputs such as online ratings, measured worker behavior, and subjective feedback from customers (Keegan & Meijerink, 2022). For example, food delivery platforms use customer feedback to control their workers' performance.

It is argued that many times PMSs leads to unequal results (Tan et al., 2021) and trust problems (Diekmann et al., 2014). For example, a worker could be declined by a customer because they are displayed as a worker who delivers "bad quality" in the system, no matter if that is the case or not (Diekmann et al., 2014). Furthermore, in some cases, such as DoorDash, workers who have an average rating below a certain number (e.g., below 4.2/5) are removed because of a company policy. They often do not know who rated them lowly or the reason for this rating which leads to suspicion. Additionally, some GWs have expressed suspicion that the rating is often biased and that issues such as racism play a role in receiving low ratings (Hanrahan et al., 2017).

2.2 Performance Management Process

The PMSs of the OLPs will be analyzed using Lepak and Gowan's (2010) theory on the Performance Management process). In this paper, the structure of the OLPs' PMSs will be interpreted using the authors' performance management theory, which outlines a five-step process for developing and implementing an effective performance management system. These steps include 1) Identifying Performance Dimensions, 2) Developing Performance Measures, 3) Evaluating Employee Performance, 4) Providing Feedback, and 5) Developing Action Plans to Improve Employee Performance. It is important to note that the theory used by Lepak and Gowan (2010) is adapted to a more traditional organizational context, with the exclusion of such advanced technology.

Within the five steps, some errors might occur, such as biases and concerns when it comes to validity, reliability, and specificity, which are also explained by Lepak and Gowan (2010) and will be elaborated on briefly. These concerns are important to address and key to ensuring effective PMSs. With the ability of OLPs to maximize the use of AI adoption in HRM activities, especially PMSs practices, the OLPs will be able to add more value to their platform. This can happen by treating their workers better, as it can support the platform in achieving better results, profit, a better strategic alignment, more innovation., and achieve overall better performance (Kourtit & De Waal, 2008).

Furthermore, it is important to address that all the steps can influence one another in the process. For example, to ensure better measurements and better feedback, there need to first be clear performance dimensions. If the performance dimensions are clear, then also more precise measures can be developed, followed by more explicit feedback, minimizing the risk of inaccuracy, invalidity, unreliability, or other potential errors (Lepak & Gowan, 2010).

Step 1: Identifying Performance Dimensions

The first step focuses on identifying the tasks and activities for which the workers are responsible. These tasks and activities define the performance dimensions that the worker should be evaluated on. Therefore, the goal is to define the specific tasks and responsibilities the workers are evaluated for, or not evaluated for, the motivation for setting these tasks and responsibilities, and how that is achieved. It is critical to firstly have a clear division of the tasks and responsibilities that are assigned to the workers to get evaluated so that later on in the following step, the measure developments are more comprehensive and explicit. Then that would also assist in preventing "unfair" rating systems, for example, where workers are evaluated for irrelevant reasons (Mohlmann et al., 2021).

Step 2: Developing Performance Measures

The second step focuses on developing a performance measuring system. To evaluate performance effectively, performance dimensions first need to be identified. In addition, developing specific measures for each dimension and measuring the performance level in each dimension is essential. The performance measures must be developed in a way that is valid, reliable, specific, and has clear standards. These aspects are crucial in the process of this step and are key to preventing any of the following errors from occurring. Within the step, validity, reliability, and specificity propose important concepts. Therefore the three concepts are explained in more detail.

Validity in measurement refers to "the extent to which a measurement is representative of true scientific value, taking "true" to mean an exact representation of what happened, free from all possible sources of error or bias" (Kelly et al., 2016, p.2).

Reliability in measurement is defined as "the extent to which a tool gives measurements that are consistent, stable, and repeatable" (Kelly et al., 2016, p.2). For instance, gig workers may receive a poor evaluation due to factors beyond their control, which negatively affects their reputation among customers, even though they may not be bad workers (Mohlmann et al., 2021). Therefore, the system needs to have a structure that ensures consistent and accurate evaluations across different customers who rate the same worker over time. If every customer has the freedom to choose a rating based on their standards, beliefs, perceptions, and expectations, then the OLPs systems need to have a structure that can support reliability and validity (Mohlmann et al., 2021).

A measurement that includes factors that are not relevant to the intended construct is a contaminated performance measure, which produces irrelevant indicators. A measurement that fails to measure important aspects of individual performance is a deficient performance measure. Step 1 is crucial for avoiding deficient or contaminated performance measurements. By identifying and clarifying the performance dimensions, the risk of producing deficient or contaminated performance measures can be reduced.

Specificity means that the expected level of performance (performance standards) is clear to both the worker and the evaluator. The functionality of performance evaluations relies on the clarity of the performance standards. Specificity can help achieve a balance between reliability and validity in the evaluation of workers. For example, it can enhance the evaluation process by providing more detailed and discrete performance standards, reducing the risk of measuring irrelevant or inconsistent factors. While validity reflects the performance standards, specificity focuses on the clarity of these performance standards (i.e., responsibilities and tasks). Higher levels of specificity can also guide different employees on how to perform different aspects of their job.

Step 3: Evaluating Employee Performance

This step involves evaluating the workers effectively. There are various approaches to achieve this, such as the ranking approach or the absolute approach, which compares employees with 'absolute' standards based on multiple performance dimensions. Moreover, there are behavior-based approaches as well since some argue that focusing solely on a worker's attributes could be somewhat misleading (Lepak & Gowan, 2010).

The workers' performance evaluation phase is prone to performance measurement errors, which can undermine the effectiveness of the evaluation. For example, biases and discriminations based on race, gender, or other factors can occur in this step. Bias and discrimination against certain special groups are observed in OLPs, especially in OLPs that have gotten more popular (Monachou & Ashlagi, 2019). Bias in the OLPs context can, for example, be how people's ratings are influenced by their gender, the evaluator's gender, and the displayed scores/ ratings of the workers (Jahanbakhsh et al., 2020).

These can lead to a halo effect or a horn error, where a single positive or negative trait of a person influences the whole evaluation, resulting in a higher or lower rating than the worker deserves. It is important to note that bias does not always result in discrimination.

Halo effect/ error: The halo effect refers to a bias occurring when a positive characteristic of a person affects the evaluation of other attributes of the person.

Horn error: Horn error refers to the error occurring when one negative aspect of a worker harms the whole evaluation, resulting in a lower evaluation than what the worker merits.

It should be noted that the halo or horn effect can vary depending on each evaluator. Different evaluators may have different perceptions of what constitutes a positive or negative trait, and they may also define a trait differently. For example, one customer may consider race as a factor, while another may consider beauty. This can affect the validity and consistency of the measurements as well. A study found that gig workers with low performance were more vulnerable to gender bias than those with high performance (Jahanbakhsh et al., 2020). This illustrates the importance of ensuring a system without any bias or discrimination in the evaluation process of gig workers. It is also important to communicate any preferred skills or characteristics to the workers.

Step 4: Providing Feedback

In step four, the focus is on how employees are reached out to and how feedback is provided to the workers. It is important to ensure that the feedback is communicated effectively with the workers in a nice professional way but also promptly. Furthermore, it is important to ensure that the workers are aware of their mistakes and get to learn from them, but also understand the impact that their work has (Kittur et al., 2013). Ideally, a system would offer peer-to-peer and expert feedback but also encourage the workers to make self-assessments and ensure or enhance the willingness of the requesters to provide feedback to the workers (Fieseler et al., 2017). Providing feedback that is not "private" and states the reasonings for an evaluation is crucial to the workers. The workers wish to know what they did do wrong or good and how they can improve. In many cases, gig workers are barely getting sufficient feedback, not knowing what they did incorrectly, but also, in many cases, gig workers are receiving feedback that is anonymously received (Mohlmann et al., 2021).

Step 5: Developing Action Plans To Improve Workers' Performance

Finally, in the fifth step, it is crucial to address how further actions are taken to develop the workers. It is crucial to address how workers with bad evaluations are dealt with. Do workers with bad evaluations get the opportunity to develop and improve themselves, or are they completely prevented from continuing their work? In many circumstances, gig workers would immediately get banned from working by the platform without an explanation or a given reason (Mohlmann et al., 2021). According to Lepak and Gowan (2010), after managers provide feedback, a good manager would take it one step further, getting involved to help the worker understand their mistakes. The workers need to understand the causes of poor performance to be able to address them and improve them. The manager will get involved and help to identify the performance problems, help understand the cause, review the performance standards and measures to ensure their development is accurately and correctly developed and help achieve improved performance.

3. METHODOLOGY

3.1 Data Collection

This research employs a qualitative research method. A literature review was conducted to provide a background and a basis for this research by collecting existing information on OLPs and their PMSs to gain a deeper understanding of how OLPs function and address the issues they face. Moreover, the theory by Lepak and Gowan is used as the theoretical framework for the analysis of the process and conclusions, here the process of the PMSs in OLPs is examined based on the five steps of the Performance

Management process (Lepak & Gowan, 2010). The goal is to gain a clearer vision regarding the designing process and decision-making embedded in the systems for the Performance Management of the gig workers through collecting data from GWs' experiences and perceptions of the PMS. Accordingly, assumptions will be created for the Performance Management Process and address upcoming concerns. The data collection is done qualitatively, using secondary data from online forums. The data collection focuses on Upwork and Uber, as these platforms are most relevant for this research since many of the problem statements identified in the literature review relate to the performance management of gig workers on these platforms, especially Uber.

An online forum is defined as a site on the internet where users discuss various topics. Online forums in this research include websites such as Reddit, Quora, Upwork community, UberPeople.net, YouTube, and any other online forums, social media, community blogs, and/or FAQS that relevant data can be collected from for answering the RQ. On these forums, platform workers share their experiences, questions, complaints, or challenges related to their work and receive responses from other workers who offer suggestions or share similar experiences or from moderators who offer assistance, creating online communities. Therefore, these online forums will be used to observe group dynamics, interactions, and personal experiences and perceptions, without any of the researcher's involvement. Posts will be read through, interactions will be explored, and the relevant data concerning performance management issues will be collected. Performance management issues relate to tasks, responsibilities, the evaluation process, feedback, and development planning or involvements (Im & Chee, 2012). A table is presented in the Appendix summarizing all the data collection sources (Table 1).

Additionally, though with many constraints, there is the opportunity to complement the secondary data by having the opportunity to conduct a semi-structured in-depth interview with company X and collect primary data. The interview consisted of seventeen questions and lasted around an hour and a half. The data was then stored in an audio recording. Due to consent and confidential information, the company and interviewee remain anonymous. The questions are presented in detail in the appendix (Table 2). The interview was structured based on Lepak and Gowan's theory, starting from the first step of developing performance dimensions and then followed by the next steps of the process to gain insights for each step in the design of the PMSs. While an interview has been conducted to gather insights, this research mainly relies on the secondary data collected from online forums, focusing on workers' posts and comments.

3.2 Data Analysis

After the data collection was completed, the data was then analyzed. For both the online forums and the interview, the data analysis was conducted using the computer program software ATLAS.ti. For the interview, all the data was transcribed in ATLAS.ti. Since an audio recording was used for the data storage, the data was then converted from audio to text format and transcribed into ATLAS.ti. As for the data collected by online forums, all of it was processed in the software program directly. The approach to the data analysis was a deductive data process, "deductive reasoning commences with generalizations and seeks to see if these generalizations apply to specific instances" (Hyde, 2000). The theoretical framework was used to define the "standard" performance management process, and by conducting the data analysis and observations, the gap between the standard performance management process and the one in OLPs' context is addressed following the theory, and conclusions are drawn accordingly (Platform, 2021). Subsequently, when all the data was converted and transferred to the software program, codes were applied to identify the quotations to each of the steps of Lepak and Gowan's Performance Management process. When all the data were identified and coded, new patterns were further recognized to create subcategories. Lastly, conclusions were drawn, based on comparing the findings to the theory, about the designing process of the PMS in OLPs and about how the gig workers' performance management is handled in the OLPs.

4. FINDINGS

This section is structured after the five steps identified by Lepak and Gowan (2010). Collected data has been analyzed concerning each step, and relevant findings are presented accordingly.

4.1 Performance Dimensions

Among the studied platforms, the performance dimensions can greatly vary per customer and platform worker. In some cases, such as Uber, it seems that the performance dimensions are not clearly defined. Even though an Uber driver's main task is to transport a rider (customer) from point A to B, there are more factors to consider. Uber drivers are expected to adhere to certain standards. The Uber website provides tips for drivers to maintain their 5-star rating, including keeping a clean and scent-free car, choosing the preferred route to the destination (even if it differs from the navigation), being polite and engaging in respectful conversation while providing excellent service to riders by helping with doors and luggage. Additionally, drivers must follow local traffic laws and speed limits (5-Star Trips | Driver App | Uber, n.d.). In addition to these standard tips provided by the official website, experienced Uber drivers suggest that there are many other indicators of responsible driving. For example, "The Rideshare Hub," a YouTube channel that provides tips and information for new drivers, mentions common mistakes that can affect driver ratings, such as not greeting passengers with a smile, talking too much, playing music too loudly, picking up unaccompanied minors (which is illegal and puts the driver's job at risk), and not dressing professionally (The Rideshare Hub, 2019). Despite the standards provided by the official website and the tips by experienced Uber drivers, many drivers also argue that the platform is one-sided and prioritizes customers over drivers. Consequently, the experience of each rider and their evaluation of their driver depends on their perceptions, standards, beliefs, and many other external or internal factors.

"Only in a perfect world without bias and many other human imperfections. There are no guidelines no rate the driver, and many riders will rate it according to their own perceptions, that will vary wildly, and even will give a rate based on the traffic conditions, among many unrelated circumstances. The same goes for low rating as for high rating. Driver will appeal for a 5 stars, and even will try to buy it with amenities for the rider." (P1)

In other cases, such as Upwork, a platform where matchmaking is happening between freelancers and clients or employers, freelancers are responsible for building a professional and attractive profile that clearly indicates their skills, capabilities, and proposals. Furthermore, an employer is also responsible for providing clear skills and requirements acquired by the freelancers. Therefore, matchmaking can be more efficient based on alignment between requirements and acquirements and good communication between the two parties for further arrangements. A Quora user explains how freelancers should approach clients or employers, providing tips to increase the possibility of getting a job, which could be relevant to a freelancer's responsibilities:

"Dont copy paste the proposal - You should write every proposal differently.

Ask questions- Make your proposal engaging. Involve the client.

Dont write essay- dont make your proposal to lengthy. […] Add samples- If you dont have past samples of your work try to make some and attach them along with your proposal." (P2)

Moreover, another important factor aside from the skills to offer, a different Quora user highlights the importance of a freelancer bringing up the problem-solving that they can offer.

"Talk about the problems you will solve and not what skills you have. Benefits not features." (P3)

Another user in Upwork Community shared their experience of quitting a job because the client changed the requirements frequently. This resulted in a score dip for the freelancer. This case illustrates how changing requirements can mislead freelancers, especially if they have already seen a specific agreement beforehand on how to carry out a specific task.

"The Upwork score of 67%. The score of 2 was because I ended the contract when the client kept changing the requirements" (P4)

Lastly, company X is a match-making platform that uses a talent pool. Unlike the other platforms, company X does not directly use or include Performance Management systems in its platform. It leaves this issue to the match-makers, such as recruiters, organizations, and job candidates. However, company X still requires the users to provide their skills and experiences for the match-making process. This process can be adapted to the performance indicators of a job candidate based on the theory. The interviewee explained that the platform allows for a diverse and flexible range of skills across different industries. To use the platform, the users have to upload their CVs and create their profiles, where they add their skills, education, and experience, but also any achieved certificates to support the skills and experiences. The platform then processes this information into a talent pool, from which recruiters or organizations can select the most suitable candidates.

"It's on the user that you can like how you update your LinkedIn? Okay, these are the work experiences you have. Similarly, you have to add by yourself, okay, these are the skills these are the places you have worked, and this is a certain type of education, which you have acquired, and put it in there. That will help you." (Interviewee)

4.2 Performance Measures

Concerning Performance Measures, the data analysis revealed that most of the applied codes were related to the performance measures of the platforms, especially Validity. Many complaints and concerns were about how the OLPs handle these measures. For example, Uber has embedded measurements that are not directly linked to a driver's overall rating or evaluation. These include cancellation rates and acceptance rates, which affect the drivers differently. When a trip request pops up, the driver can accept or decline it. Declining affects only the acceptance rate while accepting and then canceling affects both rates. Moreover, the cancellation rate can determine whether a driver can continue working or not, whereas the acceptance rate does not have such consequences.

"Excessive cancelation rate because of cherry picking rides. It is OK to ignore requests. That affects acceptance rate. Passenger never sees you ignored her request. It is not OK to cancel rides after accepting. That affects cancelation rate and also acceptance rate. It results in terrible customer experience. Passenger cannot rate the driver, but the company will fire the driver for excessive cancelation." (P5)

Uber measures gig workers comprehensively, depending solely on customer experience and satisfaction based on their own perceptions and standards, through the 5-star rating scale. The 5-star scale can be based on the drivers' cleanliness, driving, navigation, pickup, car quality and smell, conversation (Uber Editor, 2022), and any other factor that can have an impact on the customer experience and satisfaction accordingly.

Concerning Upwork, the algorithm uses the Job Success Score (JSS) measure. A moderator replied to a freelancer in the forums regarding concerns about the way JSS is applied in practice:

"A freelancer's Job Success score is a measure of their work and reputation on Upwork based on client feedback and indicators of client satisfaction. It includes their public and private feedback, long-term relationships, rehires, and contracts that do not result in work delivered." (P6)

Though moderators explain how the JSS works on a "high level," many argue that there are many other indicators affecting JSS, but it is not being disclosed. The disclosed calculation of the JSS score on a general level is (successful contract outcomes – negative contract outcomes/ total outcome) (Job Success Score, 2023.).

When it comes to company X, according to the interviewee, there are not any direct measurements, but the interviewee has explained that a ranking of a candidate can depend on their certificates, which also clarifies the level of skill they have, level of experience, and education.

"particular recruiter is looking for a particular skill, a particular job, we know that for this job role, these skills are very important. So we will rank a person who have better skills, and who are more experienced in those skills than the one which addresses it also changes how much experience the recruiter is looking for" (Interviewee)

4.2.1 Validity

When it comes to Validity, through the data collection, many comments of workers were concerned regarding the validity of the measurements. Most importantly, many workers would complain about getting a score that shouldn't be the case.

"so many pastors will rate you a 1 star just because they want to get a free ride" (P7)

Some other cases reported that they had been accused of alcohol consumption, though it was not true. In this case, the driver had to immediately take a test to prove their sobriety, and luckily enough, Uber allowed the driver to reactivate:

"I one time was accused by rider with Uber that I was either drunk or under the influence of drugs. I believe the rider said this to get a free ride. As I neither do drugs or drink." (P8)

In many other cases, drivers are fully, permanently deactivated, regardless of the reasonings being truthful or not:

"Hi So I have completed eight jobs on Upwork so far and have gotten all 5-star ratings from my clients except for two who were not responsive, so I had to close the job after a month of not working. Please solve my issue. It's not my fault if those two clients were not responsive. My profile is now affected due to such a low job success score." (P9)

Many other freelancers have expressed their concerns regards their score calculation. Despite Upwork's disclosure of the JSS calculation, many freelancers question its transparency and accuracy as they struggle to understand the reasons for their score decrease.

"I am in the exact same boat! My JSS inexplicably dropped to 89% and I cannot fathom a reason for that. I had some fairly good-quality jobs in the interim that were received well by clients and paid for quite promptly." (P10)

The company X interviewee has indicated that they have never faced any issues regarding validity. The worker has been questioned on how they ensure the stated skills and work experiences mentioned by the job seekers are true and valid. The interviewee responded:

"This is, you're actually good at the totally on the user side that if they changes the label for themselves, it will, but at the end of the day, it will affect them, because when there will be interaction between recruiters and talent, and if they don't pass that particular interview or something like that, that probably not be good for that they might come back and change it in the system." (Interviewee)

Therefore, any false information would affect the people within the match-making; the company cannot act in any way regarding that concern. Furthermore, the interviewee was asked how the candidates' skills were ensured to be true and further explained that to ensure that the mentioned skills, education, or experiences, the submission of certificates achieved that support the individual's report are used to the candidate's advantage.

4.2.2 Reliability

Reliability refers to the consistency of different evaluations. For example, Uber riders may have different standards and perceptions of the same driver, or Upwork clients may have different requirements and expectations for the same freelancer.

Uber refers on their website that "where applicable, there is a minimum average rating in each city. This is because there may be cultural differences in the way people in different cities rate each other." (Legal | Uber, 2023). However, it is not explained how this is applied and what factors are considered to conclude the specific minimum average in each city.

Upwork uses two types of feedback: Public and Private. Public feedback includes star ratings and comments that are visible to the freelancer. Private feedback is not visible to the freelancer, and it affects the agency's or talent's job success and any confidential information that the client does not want to share publicly. However, some freelancers question the consistency of the two feedbacks, wondering if the client gives different feedback in private than in public. Others suspect that Upwork uses "private feedback" to hide potential errors or secrets.

"They have bugs in their calculation score and private feedback looks like a blanket term for anything they can't explain or don't want to." (P11)

4.2.3 Specificity

When it comes to specificity, there are many complaints regarding the system not being specific or transparent with measurements and what factors are considered for these measures. A YouTuber, "Melbourne Uber driver," explains in a video his concerns regarding Uber's rating system and its elaborateness towards the drivers (Uber driver ratings which rider gave you a 1 or 2 star - YouTube).

"System gives you the time and how much you got paid. How are you supposed to know who gave you that one star, who gave you that two star, why did you get that one star why did you get that 2 star, you actually don't know […] The system should be more elaborative to us, about who gave us a 1star so we sort of know how to improve […] is it navigation, is it clean, is it cleanliness of the car, is it driver behaviour, conversation" (P12)

In some cases in Upwork, freelancers are asking for a disclosure regards a specific JSS, but the moderators in response are avoiding answering directly to solve such issues. For example, some of the responses to such concerns were:

"I'm sorry to hear about your score drop, but we're afraid that we won't be able to comment on an individual score or how specific contracts affect it. Factors such as private and public feedback, the dollar weight of the projects you've completed, and the long-term relationships with your clients can have an impact on your overall score. You may want to check this article to learn more." (P13)

4.3 Performance Evaluation

When it comes to Performance Evaluation findings, data was found concerning biases. In Uber, the rider defines the rating score of an Uber driver. In case of bad ratings, many complain that the platform is one-sided, with most of the cases neglecting the "story side" of the Driver, with the rider being in favor.

"When we looked at it, Uber's platform seems to focus on one user — the person who wants a ride — somewhat at the expense of the drivers." (P14)

A previous Uber Driver who has been deactivated has also complained about how Uber puts customers first before its drivers:

"Very sorry to hear that! Sure Uber does have some bad characters driving for them, but they always believe the passenger before the driver. I once had a complaint that my car stunk, it was the catalytic converter of a car ahead of us, you know the rotten egg smell. Last week I got permanently deactivated because they said my selfie didn't look like me. And just like that I'm no longer driving for them." (P15)

Some other issues that are arising when it comes to Performance Evaluation, In Uber, many drivers experience bias and discrimination since a score will be "suspiciously" low and leave the drivers wondering about the reasoning.

"I got super angry when I was deactivated from uber for community guidelines. I was driving for uber for more than 5 years and I was involved in a discussion with the passenger just because he doesn't likes Hispanic people. He recorded our conversation and I really was mad with him for his way to treat me. I reach out to Brave notch on the internet and they got it reactivated immediately" (P16)

It could also be argued that this rider exhibited a horn effect by judging the driver negatively based on their Hispanic ethnicity, affecting the driver's whole evaluation.

4.4 Feedback

The feedback process appears to be insufficient in many cases. Some workers receive anonymous or unexplained evaluations, while others get deactivated without any feedback at all. These workers express their frustration about the low score ratings and the lack of reasons for their deactivation:

"They just deactivated me for fraudulent activity without even telling me what I did wrong and. I called Uber support, and they said it's permanent" (P17)

The system does not provide adequate feedback on their rides, which upsets Uber drivers. They receive low ratings without knowing who gave them to them or why; therefore, they cannot improve their performance or prevent further low ratings. For instance, the YouTuber "Melbourne Uber driver" expressed their disappointment about a low rating and wondered what the reason was: "Is it navigation, cleanliness of the car, driver behavior, conversation?". (Melbourne Uber driver, 2021).

In Upwork there, no particular remarkable issues concerning feedback have been found. However, there are still some small issues arising, causing suspicions. As mentioned earlier, Upwork uses two types of feedback: Public and Private. Public feedback includes star ratings and comments that are visible to the freelancer. Private feedback is not visible to the freelancer, and it affects the agency's or talent's job success and any confidential information that the client does not want to share publicly. This creates doubt about the alignment of the two feedbacks. Furthermore, freelancers, in many cases, wonder why their JSS score drops suddenly since they do not receive any sort of feedback explaining the reasoning for this occurrence. These changes, without explanation lead many freelancers to think that there are some mistakes in the system regarding the way it measures their JSS.

"I've been having the same issue with Upwork, in the past month my score JSS went from 94% to 85%. I honestly do not trust the way it's calculated now, I've been top rated consistently for 3-years. And makes no sense because in the last month I had no negative contracts. My last client left great public feedback and I know privately too, other client left good public feedback and based on the fact that he intends to hire me later, I assume good private feedback as well. I think Upwork has some issue they won't admit to having." (P18)

Furthermore, moderators would reply to the freelancer to address the issue but avoid giving further details of the reason for the score dip and instead mention some standard reasons that can lead to a score drop.

4.5 Developing Action Plans to Improve Workers' Performance

Uber suspends drivers who get low ratings, regardless of whether the drivers are truly bad or not. Many drivers complain that Uber does not care about them. The platform seems to exploit the high demand for drivers by getting rid of the "bad" ones and keeping the good ones to ensure customer satisfaction.

"Uber could care less about their drivers because if they deactivate you there is always someone else to take your place." (P19)

In addition, many drivers could not reach out to Uber Support when their accounts were suspended without a known reason. Only a few drivers managed to get their accounts reactivated. Some of the suspended drivers had years of experience and thousands of rides, but they still faced deactivation due to unknown causes.

"I've tried to contact investigations at least 6 times now and each time I've tried they refuse to transfer me over. In fact, one of the support techs transferred me over to my own phone number! Honestly it's been one of the most ridiculous run arounds I've ever been given, especially considering I never broke any of the terms and conditions." (P20)

The evaluation ratings also affect how the algorithm treats the workers. If the ratings are low, the workers will have a "worse" experience on the platform. For example, the algorithm will influence their future work opportunities and match-makings, which will also affect their compensation.

"Rating is very crucial for driver to get rides, better the rating, they have chances to get more rides […] When I book a cab, say there are 5 cab near my pick up location, so the highest rating driver among the 5 cabs near you gets notified first, it goes to the next only if the 1st doesn't accept the request or rejects it. This is the way Uber promotes good drives and get rids of bad drivers." (P21)

Low JSS scores also affect freelancers' chances of getting job proposals on Upwork. This upsets freelancers who expect a good score but get a drop without knowing the reason.

"That should not lead to a drop in my JSS Score! On the contrary, I was fairly confident that my JSS Score would see an uptick this Sunday In the "real world", these things don't matter, since one is not merely a "score", but given the unrelenting competition, innumerable proposals, and understandably short attention spans of clients here, the JSS Score serves as a quick benchmark for them to filter applicants. Dropping below the 90% mark certainly leads to a significant diminution in job prospects. I have already written to Upwork Support on this matter and would appreciate clarity on it." (P22)

This step suggests that the OLPs do not directly involve themselves in developing their workers' performance. Instead, they easily dispose of the "bad workers" and replace them with others. Moreover, the OLPs often do not provide elaborate feedback or give the workers a chance to continue working after some "bad" evaluations.

5. DISCUSSION

5.1 The process applied in OLPs

First, regarding the identification of the performance dimensions, it can be assumed, based on the findings, that OLPs are customer-based, meaning that the customer or client has priority and defines the performance dimensions. This happens "directly" or "indirectly." When the performance dimensions are defined "directly," the requirements for a specific task or job are clear and visible to the worker. For example, in Upwork, a client posts the job or tasks, the skills required, and the way they want the work done. Therefore, a freelancer knows what skills they need and what responsibilities they have. When the performance dimensions are defined "indirectly," there are some expectations or behaviors that a worker should follow or avoid, but they are not officially set or fixed. For example, Uber's official website suggests some standards for a comfortable and safe ride, but these standards are not enforced or specified by the system. They can greatly vary depending on each customer's experiences, perceptions, beliefs, and standards. In contrast, in a more traditional organizational setting, the performance dimensions and the role of a job would be more stable and clear, accordingly to the job position and the higher people within the company itself (Lepak & Gowan, 2010).

According to Lepak and Gowan (2010), managers are responsible for identifying the workers' responsibilities and defining the performance dimensions. It is crucial to break down the performance into different dimensions to be able to later identify areas of the individual's performance that need improvement. However, in OLPs, each customer or client has different priorities and preferences, having the responsibility of identifying the workers' tasks and responsibilities themselves, making the process more unstable and uncertain. This contrasts with traditional settings, where the performance dimensions are more stable and certain, and the evaluator is more consistent. This difference may result from rapid technology changes (Snell & Morris, 2021). and the relatively new adoption of AI in OLPs (Park et al., 2022). The gig economy was introduced mainly alongside technological advancements. The way the OLPs are operating seems to be very different compared to the traditional organizations, bringing different ways of defining the performance dimensions but also bringing different ways through the whole performance management process.

Second, the development of performance measures follows. The OLPs seem to limit the disclosure of details in the calculations and factors involved in the overall assessment of the workers. These constraints make it difficult to assume how the measurements are exactly developed and what factors they include or prioritize in the calculations collecting data by online forums, which leads to the level of transparency of the system. However, it seems that there are many circumstances where there are invalid measures, but with the constraints of disclosure, and the possibility of the evaluators to freely choose what to include for the measurements, it can be used to the OLP's advantage by not placing the organization in a risky position, for example avoiding being sued or other legal procedures.

This affects the validity, reliability, and specificity of the performance measures. Specificity is especially important, as it can influence the validity and reliability as well. Considering that the performance dimensions are differently "defined" in each match-making, which involves different specializations, tasks, and responsibilities, depending on the customer experience, which is in turn based on their perceptions and standards, the OLPs face a challenge in designing a measurement system that can cover all kinds of match-makings outcomes. Moreover, the lack of clear and fixed performance dimensions in some cases hinders the process of ensuring valid and reliable measures. This creates complexity for the whole measurement process by the OLPs toward the gig workers. Nevertheless, it should still be ensured that a measurement system is reliable, valid, and specific for different types of (gig) workers. In cases where a worker has complete proof of an "unfair" assessment, the OLPs should have no space of choice but to get involved to remove or "fix" a bad measurement.

Third, regarding the performance evaluation, it seems there is only individual comparisons, with the most common being by rating, a rank, or a forced distribution (i.e., the JSS score). Therefore the system has embedded measurements that will conclude the customer's evaluation process, though they are not disclosed, except for Upwork, disclosing its JSS measurement but in broad terms. This is overall creating a lot of uncertainty and inconsistency regarding the performance evaluation of the PMSs, as the system lacks providing reasoning behind the performance evaluation process, also referring to an issue of transparency and highlighting its importance (Alloa & Thoma, 2018).

Moreover, it is observed that there is bias embedded in the system. The OLPs will choose customers/clients over the workers by neglecting the worker's point of view in case of a bad evaluation. Therefore, in cases where the customer will evaluate a worker poorly, the OLPs will not confirm that in any way but rather choose to believe that evaluation. Furthermore, if a customer/client gives a bad evaluation because of them being biased or discriminating, it is difficult to detect it or know. It is detected that there are two types of biases in OLPs: a) a bias that is embedded in the system itself, and b) a bias that comes from the client/ customer evaluation. Nevertheless, the system needs to be free of bias or any kind of discrimination itself, but it is also important for the system to have the capability to detect and prevent bias from customers' / clients' input. This emphasizes the value and importance of making the performance dimensions clear. If the performance dimensions are clear, it would be easier to detect factors that are irrelevant to the performance evaluation and identify and avert biases or discrimination (Roselli et al., 2019).

Fourth, regarding feedback, it is crucial to mention that in OLPs seem to not involve feedback systems as much. Many workers are complaining about not getting elaborative feedback and therefore not knowing why they got a bad evaluation neither by whom, since there are only cases in which it is anonymous. This makes it difficult for the workers to know their wrongs and how to improve. This leads to the assumption that the OLPs structure their feedback systems in ways that it doesn't push the evaluators to be more elaborate. This comes to a high-level contrast compared to the theory, as the theory suggests a "timely manner" and "professional and positive" feedback that involves appraisal of the workers but also detailed feedback of the wrongs for the worker to be able to understand and improve, which is followed by the final step, the development/ action plans of the workers (Lepak & Gowan, 2010, p.273).

Finally, concerning the development / action plans of the workers after the evaluation, it is observed that it is often the first option to suspend a worker, in case of a bad evaluation, rather than allowing them to do better. OLPs do not get involved in the development of the workers, except for some cases where a worker is persistent and has proof, leaving no space for the OLPs to get involved. On the other hand, according to the theory, the organization should get involved to improve a worker's performance and communicate new strategies for further actions. According to Lepak and Gowan (2010), the employer actually gets involved and is responsible for reviewing the performance standards with the employee, ensuring and re-checking the accuracy of performance measures, and evaluating potential role concerns. After the reflection, there is a planned development for ways to improve the performance of the employee.

Overall, it seems that the PMSs does not provide a system that enhances elaborateness and clarity towards the workers. It is important to note that using Lepak and Gowan's (2010) theory, it is presumed that the Performance Process steps are dependent on each other. It is observed that through steps, the next step is building up upon the previous one. Therefore, it seems that the main issue within the OLPs' PMSs, is the inability of the system not having fixed, clear performance dimensions in the first place, making it complex to further build valid, reliable, specific measures, make an efficient evaluation, and accordingly provide proper feedback, and come up with development/ action plans, creating inconsistency and uncertainty.

It could be argued that these are consequences of the adaption of AI to HRM activities. The gig economy is relatively new, and due to the rapid changes in the environment because of technology (Snell & Morris, 2021). Furthermore, if there are no specific regulations/laws pushing OLPs to be structured in specific ways when it comes to gig workers' performance management, it wouldn't "advantage" the OLPs to restructure their designing process to improve in specific aspects. For example, for the OLPs to take action to change their PMSs will need an investment to do so, which would lead to money loss. Furthermore, if there are indeed important underlying reasons behind the constraints on transparency, that would increase the risk of the platform and might lose customers or even not be able to operate anymore.

The OLPs should focus more on the development of their PMS and be more elaborate and clear towards the workers. On the other side, being elaborative and disclosed can bring the OLPs to a disadvantage. But they could restructure the PMS in a way that they push the evaluators to be more clear with their measurement and elaborative over their overall assessment of the worker. For example, when providing feedback, the system could provide more options regarding the performance dimensions that are being evaluated or structure the system in a way that the evaluator has no option but to provide sufficient, elaborative feedback.

5.2 Contributions to theory and practice

Some studies have highlighted the consequences of the Performance Management of OLPs on the GWs, through algorithmic management, how it impacts their daily life, and how they perceive it (Mohlmann et al., 2021). Different studies are studying general HRM practices of OLPs, adapting it from theory (Keegan & Meijerink, 2022). This study takes one step further and digs deeper into HRM practices adapted to OLPs, specifically PMSs, and defines their Performance Management Process in a better way by breaking down every step of the process. Furthermore, it contributes to theory by providing a clearer understanding of how PMSs are assessed through each step of the Performance Management Process and also of the drawbacks and the gaps that divide how theory suggests the application of Performance Management and how it is actually in practice in the context of OLPs. This study uses Lepak and Gowan's (2010) theory of Performance Management Process as a benchmark to identify how PMSs are different compared to Lepak and Gowan’s (2010) standards and give insights into how the PMS systems are designed in the OLPs, accordingly. Though this study could mostly make assumptions based on Online Forums, some important patterns have been identified, bringing to the surface more questions regarding the PMSs’ designing process. For instance, the findings and discussion reveal that Performance Management in OLPs is very different compared to the theory. Lastly, this study contributes to bringing more awareness as regards overall Performance Management.

5.3 Limitations and Future Research

This study has some limitations. First, the literature review on OLPs and their adoption of HRM activities are limited, as this is an intrinsically new and fast-changing field (Snell & Morris, 2021). This study did not collect primary data by contributors of OLPs’ algorithms regarding their PMSs and instead relied mainly on secondary data from online forums and limited primary data. Though this data collection method has been challenging to adapt to theory and make assumptions, future research collecting primary data can validate the findings of this study. Moreover, because of the secondary data collection, the findings are based on assumptions that are based mainly on the GWs perspective, though future research studies can support this by adding more perspective to ensure a full picture and understanding of the PMSs. Finally, the assumptions and conclusions are more limited and specific to the studied OLPs. Therefore, they should not be generalized to other OLPs.

It is suggested for future research to collect primary data to obtain more accurate, reliable, and insightful data on the PMS of the OLPs. In addition, this research study shows that Performance Management is very different in the OLPs context than in the traditional organizational context. Future research should investigate the optimal design of performance management systems (PMSs) in online learning platforms (OLPs). This could be achieved by incorporating multiple perspectives into the PMSs and examining each stage of development individually to determine the most effective approach within the context of OLPs. Furthermore, there are many constraints and inconsistencies in the disclosure of the PMSs, the factors included in the measurements, their development, and the feedback. Future research should investigate what causes OLPs to be so limited in revealing some processes or in providing more elaborate feedback to gig workers on their evaluations. Future research should also explore the risks and benefits of disclosing or concealing the PMSs’ processes for the OLPs. Ultimately, future studies could focus on developing an adaptable framework for the OLPs, and how their PMSs should be designed.

6. CONCLUSION

This study offers valuable insights into the structure of OLPs’ PMSs, using Lepak and Gowan’s (2010) theory as a benchmark to draw assumptions. The research concludes and provides a deeper understanding of the underlying issues that lead to gig workers’ complaints about OLPs’ algorithmic performance management (Mohlmann et al., 2021). By analyzing the data in greater detail, the study defines which performance management processes are of concern, identifies new patterns, and explicitly addresses the sources and nature of these concerns. The main findings in this study have been the performance dimensions of GWs being mostly defined by the customers/ clients themselves, but also the drawbacks that come alongside this possibility, such as the lack of definition and clarity, creating inconsistencies throughout all the following steps. Furthermore, another important finding is how complex the development of the Performance Management Process is in the context of OLPs, due to the uncertainty and instability, making it difficult to build a valid, reliable, specific system accordingly. This study addresses the concerns of the GWs regarding their performance management and contributes a different perspective by making assumptions and connections, adapting it to the actual process of the Performance Management Process.

7. REFERENCES

*5-Star Trips | Driver App | Uber. (n.d.). Uber. https://www.uber.com/nl/en/drive/basics/5-star-pro-tips/?uclick_id=eca7de70-57a4-4ee9-a79f-760deb0aabb5

*Alloa, E., & Thoma, D. (2018). Transparency, Society and Subjectivity. In Springer eBooks. https://doi.org/10.1007/978-3-319-77161-8

*Bamel, U. K., Kumar, S., Lim, W. M., Bamel, N., & Meyer, N. (2022). Managing the dark side of digitalization in the future of work: A fuzzy TISM approach. Journal of Innovation & Knowledge, 7(4), 100275.

https://doi.org/10.1016/j.jik.2022.100275

*Chen, M. K., Chevalier, J. A., Rossi, P. H., & Oehlsen, E. (2019). The Value of Flexible Work: Evidence from Uber Drivers. Journal of Political Economy, 127(6), 2735–2794. https://doi.org/10.1086/702171

*Connelly, C. E., Fieseler, C., Černe, M., Giessner, S. R., & Wong, S. I. (2021). Working in the digitized economy: HRM theory & practice. Human Resource Management Review, 31(1), 100762. https://doi.org/10.1016/j.hrmr.2020.100762

*Diekmann, A., Jann, B., Przepiorka, W., & Wehrli, S. (2014). Reputation Formation and the Evolution of Cooperation in Anonymous Online Markets. American Sociological Review, 79(1), 65–85. https://doi.org/10.1177/0003122413512316

*DoorDash Dasher Support. (n.d.). https://help.doordash.com/dashers/s/article/Dasher-Ratings-Explained?language=en_US

*Ferrer, X., Van Nuenen, T., Such, J. M., Coté, M., & Criado, N. (2021). Bias and Discrimination in AI: A Cross-Disciplinary Perspective. IEEE Technology and Society Magazine, 40(2), 72–80. https://doi.org/10.1109/mts.2021.3056293

*Fieseler, C., Bucher, E., & Hoffmann, C. P. (2017). Unfairness by Design? The Perceived Fairness of Digital Labor on Crowdworking Platforms. Journal of Business Ethics, 156(4), 987–1005. https://doi.org/10.1007/s10551-017-3607-2

*Frenken, K., Vaskelainen, T., Fuenfschilling, L., & Piscicelli, L. (2020). An Institutional Logics Perspective on the Gig Economy. Emerald Publishing Limited eBooks, 83–105. https://doi.org/10.1108/s0733-558x20200000066005

*Gandini, A. (2019). Labour process theory and the gig economy. Human Relations, 72(6), 1039–1056. https://doi.org/10.1177/0018726718790002

*Garg, N., & Johari, R. (2021). Designing Informative Rating Systems: Evidence from an Online Labor Market. Manufacturing & Service Operations Management, 23(3), 589–605. https://doi.org/10.1287/msom.2020.0921

*Gaskell, A. (2018, May 8). How Uber Drivers Feel About Being Managed By Machines. Forbes. https://www.forbes.com/sites/adigaskell/2018/05/08/how-do-uber-drivers-feel-about-being-managed-by-machines/?sh=651e89ec6754

*Hanrahan, B. V., F, N., MA, & Yuan, C. W. (2017). The Roots of Bias on Uber. In European Conference on Computer Supported Cooperative Work (pp. 1–17). https://doi.org/10.18420/ecscw2017-27

*How can I make a good freelancer profile in Upwork or Freelancer or Elance? (n.d.). Quora. https://www.quora.com/How-can-I-make-a-good-freelancer-profile-in-Upwork-or-Freelancer-or-Elance/answer/Zara-Razaq?ch=10&oid=20992456&share=adfee150&srid=uanbNj&target_type=answer

*How Does Uber do Performance Management? (n.d.). https://www.performyard.com/articles/how-does-uber-do-performance-management

*Hyde, K. F. (2000). Recognising deductive processes in qualitative research. Qualitative Market Research: An International Journal, 3(2), 82–90. https://doi.org/10.1108/13522750010322089

*Im, E., & Chee, W. (2012). Practical Guidelines for Qualitative Research Using Online Forums. Cin-computers Informatics Nursing, 1. https://doi.org/10.1097/nxn.0b013e318266cade

*Immonen, J. (2023). HRM models of online labor platforms: Strategies of market and corporate logics. Frontiers in Sociology, 7. https://doi.org/10.3389/fsoc.2022.980301

*Is the Uber rating system fair? (n.d.). Quora. https://www.quora.com/Is-the-Uber-rating-system-fair

*Is the Uber rating system fair? (2015). Quora. https://www.quora.com/Is-the-Uber-rating-system-fair

*Jahanbakhsh, F., Cranshaw, J., Counts, S. E., Lasecki, W. S., & Inkpen, K. (2020a). An Experimental Study of Bias in Platform Worker Ratings: The Role of Performance Quality and Gender. https://doi.org/10.1145/3313831.3376860

*Keegan, A., & Meijerink, J. G. (2021). Online labour platforms, human resource management and platform ecosystem tensions: an institutional perspective. In Edward Elgar Publishing eBooks. https://doi.org/10.4337/9781839100284.00016

*Keegan, A., & Meijerink, J. G. (2022). Dynamism and realignment in the HR architecture: Online labor platform ecosystems and the key role of contractors. Human Resource Management, 62(1), 15–29. https://doi.org/10.1002/hrm.22120

*Kelly, P., Fitzsimons, C., & Jepson, R. (2016). Should we reframe how we think about physical activity and sedentary behaviour measurement? Validity and reliability reconsidered. International Journal of Behavioral Nutrition and Physical Activity, 13(1). https://doi.org/10.1186/s12966-016-0351-4

*Kittur, A., Nickerson, J. V., Bernstein, M. S., Gerber, E. M., Shaw, A., Zimmerman, J., Lease, M., & Horton, J. R. (2013). The future of crowd work. https://doi.org/10.1145/2441776.2441923

*Kourtit, K., & De Waal, A. (2008). Strategic Performance Management in Practice: Advantages, Disadvantages and Reasons for Use. Social Science Research Network. https://doi.org/10.2139/ssrn.1192264

*Kuhn, K. M., & Maleki, A. (2017). Micro-entrepreneurs, Dependent Contractors, and Instaserfs: Understanding Online Labor Platform Workforces. Academy of Management Perspectives, 31(3), 183–200. https://doi.org/10.5465/amp.2015.0111

Legal | Uber. (n.d.). https://www.uber.com/legal/en/document/?name=general-community-guidelines&country=netherlands&lang=en

*Lepak, D. P., & Gowan, M. A. (2010). Human Resources Management : Managing Employees for Competitive Advantage.

*Meijerink, J. G., & Arets, M. (2021). Online labor platforms versus temp agencies: what are the differences? Strategic Hr Review, 20(4), 119–124. https://doi.org/10.1108/shr-12-2020-0098

*Meijerink, J. G., & Keegan, A. (2019). Conceptualizing human resource management in the gig economy. Journal of Managerial Psychology, 34(4), 214–232. https://doi.org/10.1108/jmp-07-2018-0277

*Meijerink, J. G., Keegan, A., & Bondarouk, T. (2021). Having their cake and eating it too? Online labor platforms and human resource management as a case of institutional complexity. International Journal of Human Resource Management, 32(19), 4016–4052. https://doi.org/10.1080/09585192.2020.1867616

*Melbourne Uber driver. (2021, February 4). Uber driver ratings which rider gave you a 1 or 2 star [Video]. YouTube. https://www.youtube.com/watch?v=k1tkdApqOPQ

*Mohlmann, M., Zalmanson, L., Henfridsson, O., & Gregory, R. W. (2021). Algorithmic Management of Work on Online Labor Platforms: When Matching Meets Control. Management Information Systems Quarterly, 45(4), 1999–2022. https://doi.org/10.25300/misq/2021/15333

*Monachou, F., & Ashlagi, I. (2019). Discrimination in Online Markets: Effects of Social Bias on Learning from Reviews and Policy Design. In Neural Information Processing Systems (Vol. 32, pp. 2145–2155). https://papers.nips.cc/paper/2019/file/e00406144c1e7e35240afed70f34166a-Paper.pdf

*Mousa, M., Chaouali, W., & Mahmood, M. (2022). The Inclusion of Gig Employees and their Career Satisfaction: Do Individual and Collaborative Job Crafting Play a Role? Public Organization Review. https://doi.org/10.1007/s11115-021-00596-4

*Nitzberg, M. J., & Zysman, J. (2022). Algorithms, data, and platforms: the diverse challenges of governing AI. Journal of European Public Policy, 29(11), 1753–1778. https://doi.org/10.1080/13501763.2022.2096668

*Park, H., Ahn, D., Hosanagar, K., & Lee, J. (2022). Designing Fair AI in Human Resource Management: Understanding Tensions Surrounding Algorithmic Evaluation and Envisioning Stakeholder-Centered Solutions. In CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3491102.3517672

*Platform, C. K. (2021, December 7). How to use the Theoretical Domains Framework - KnowledgeNudge - Medium. Medium. https://medium.com/knowledgenudge/how-to-use-the-theoretical-domains-framework-e26b81d64f0c

*Ray, S. D. (2019). A Quick Review of Machine Learning Algorithms. https://doi.org/10.1109/comitcon.2019.8862451

*Roselli, D., Matthews, J., & Talagala, N. (2019). Managing Bias in AI. https://doi.org/10.1145/3308560.3317590

*Sainato, M. (2023, March 10). 'It was traumatic': Uber, Lyft drivers decry low pay and unfair deactivations. The Guardian. https://www.theguardian.com/business/2023/mar/10/uber-lyft-driver-suspension-deactivation-pay

*Snell, S. A., & Morris, S. S. (2021). Time for Realignment: The HR Ecosystem. Academy of Management Perspectives, 35(2), 219–236. https://doi.org/10.5465/amp.2018.0069

Sustainability in the Gig Economy. (2022). https://doi.org/10.1007/978-981-16-8406-7

*Tan, Z., Aggarwal, N., Cowls, J., Morley, J., Taddeo, M., & Floridi, L. (2021). The ethical debate about the gig economy: A review and critical analysis. Technology in Society, 65, 101594. https://doi.org/10.1016/j.techsoc.2021.101594

*The Rideshare Hub. (2019, May 12). 25 MISTAKES UBER DRIVERS MAKE EVERYDAY! [Video]. YouTube. https://www.youtube.com/watch?v=fo-Wa24diRw

*The Rideshare Hub. (2022, June 12). Why Uber is Deacivating so many drivers and How To Get Reactivated [Video]. YouTube. https://www.youtube.com/watch?v=ocsttmQDySo

*Uber Editor. (2022, January 11). Ratings, rider feedback, and what it means. | Uber Blog. Uber Blog. https://www.uber.com/blog/rider-feedback/

*Upwork Community. (n.d.). https://community.upwork.com/

What Is Uber and How Do You Use it? (n.d.). Uber. https://www.uber.com/us/en/ride/how-it-works/

*Wright, C., Wailes, N., Bamber, G. J., & Lansbury, R. D. (2017). Beyond National Systems, Towards a 'Gig Economy'? A Research Agenda for International and Comparative Employment Relations. Employee Responsibilities and Rights Journal, 29(4), 247–257. https://doi.org/10.1007/s10672-017-9308-2

8. APPENDIX

Table 1: Overview of data collection sources

Website name	Person identification	Link to source
Quora	P1	Is the Uber rating system fair?
	P2	How do I get work from Upwork?
	P3	How can I make a good freelancer profile in Upwork or Freelancer or Elance?
	P5	Why would a driver with good ratings get dropped by Uber?
	P8	Why would a driver with good ratings get dropped by Uber?
	P21	How does the uber rating affect the drivers?
Youtube Videos	P12	Uber driver ratings which rider gave you a 1 or 2 star
Youtube Comments	P7	Why Uber is Deactivating so many drivers and How To Get Reactivated
	P15	Why Uber is Deactivating so many drivers and How To Get Reactivated
	P16	Why Uber is Deactivating so many drivers and How To Get Reactivated
	P17	Why Uber is Deactivating so many drivers and How To Get Reactivated
	P19	Why Uber is Deactivating so many drivers and How To Get Reactivated
	P20	Why Uber is Deactivating so many drivers and How To Get Reactivated
Upwork Community	P4	Re: My JSS went from 100% to 69%
	P6	Help Center Support
	P9	Job Success dropped from 100% to 82%
	P10	Re: JSS dropped and has not increased since
	P11	Re: JSS dropped and has not increased since

	P13	Moderator responding to a concern
	P18	Re: JSS dropped and has not increased since
	P22	Re: JSS dropped and has not increased since
Forbers.com	P14	How Uber Drivers Feel About Being Managed By Machines

Table 2: Interview Questions

Step	Concepts	Question
1. Identifying Performance Dimension	Global performance measure/ or performance dimension	• Q1: Which specific tasks and responsibilities that workers are responsible for does your platform/organization evaluate? Why? How? • Q2: Which specific tasks and responsibilities that workers are responsible for does your platform/organization not evaluate? Why? How?
2. Developing Performance Measures	Validity & Reliability	• Q3: What metrics / measures are used to measure the performance on these tasks / responsibilities? Why these?
		• Q4: How do you make sure these measures are consistent and accurate over time and across different customers that evaluate an individual worker? • Q5: How do you make sure to measure what it is intended to measure?
	Contaminated & deficient perf. measure	• Q6: How do you make sure to develop measures that take into account only important aspects of an individual and not irrelevant information?
	Specificity	• Q7: How do you make sure these measures are specific and clear rather than being very broad to the worker and evaluator?
3 . Evaluating Employee Performance	Halo effect/ Horn effect	• Q8: How do you prevent the bias that occurs when a positive/ negative characteristic of a person affects the evaluation of the person's performance? • Q9: How can specific negative or positive characteristics overtake the performance evaluation of a worker?
	Bias (more general)	• Q10: How do you prevent bias in the system when evaluating a worker's performance? • Q11: Have you ever dealt with complains about an "unfair" appraisal system? if yes → how did you deal/ address the complaint, elaboration on the case
4. Providing Feedback		• Q12: How is feedback communicated with the workers? • Q13: How do you make sure feedback is communicated in such a way that workers also learn from it? (timely, professional positive)

5. Action plans to improve worker's performance		• Q14: What actions are taken in order to improve worker performance? • Q15: How would a worker with a bad evaluation would be dealt with? → Is there a chance given for further improvement or development?
6. Optional	general	• Q16: What are the main challenges faced in the designing of appraisal systems? • Q17: Are there any other challenges that you face in the performance management?