How Informative are Job Posting Skill Measures? Evidence from Selection Decisions

December 2024

Abstract

Information embedded within job postings is increasingly central to modern labor markets, serving not only potential employees but also powering matching algorithms, screening tools, and broader workforce analytics. This growing reliance necessitates a thorough validation of skill and job requirement measures derived from job posting content. We provide the first empirical validation of this informativeness by predicting selection for internal jobs within a major firm using a novel measure of skill distance constructed solely from job postings. Our analysis reveals that job posting-derived skill distances strongly predict selection outcomes: the probability of selection is 84% higher when the sought job is in the lowest quintile of skill distance compared to a position in the highest quintile. Furthermore, in 70% of cases where multiple candidates applied, the selected candidate was the one with the shortest skill distance. Notably, these posting-derived skill measures consistently outperform traditional employee characteristics in explaining selection decisions. Building on this validation, we demonstrate how internal application intensity is significantly correlated with the average skill distance to emerging opportunities, yielding novel insights directly relevant to human capital management and scholarly inquiry on employee mobility in contemporary labor markets. Beyond validating the predictive power of job posting content for selection, our analysis unveils rich possibilities for future scholarly inquiry and the development of advanced analytics leveraging the readily available and informative skill content within job postings.

Acknowledgment: We wish to disclose that while AI writing tools were used as assistive technologies for manuscript preparation and editing, all intellectual contributions, including research design, analysis, findings, and conclusions are solely the work of the authors.

1. Introduction

Major corporations like Google, Walmart, and IBM, along with other public and private agencies, have committed to skills-based hiring policies, focusing on capabilities rather than traditional credentials such as degrees or work experience (Fuller et al., 2022; McKinsey Global Institute, 2020; World Economic Forum, 2020). This shift comes at a time when employees increasingly pursue less fixed career paths both within and across firms, and when the skills required to perform modern work evolve rapidly. These forces compel firms to articulate skill requirements and job responsibilities with greater precision in their postings. This information is not only consumed by potential applicants—it is embedded in the algorithms used by platforms like Indeed and LinkedIn to curate and match opportunities to candidates, and ingested by sophisticated screening tools that shortlist applicant resumes based on fit against these requirements. Job postings and their content are also systematically harvested by specialized firms like Revelio and Lightcast, who package this data into analytical products that provide insights into workforce skill requirements across industries[1]. These developments underscore the growing importance of job postings as a key source of valuable information at the analytical, strategic, and policy levels.

Despite this growing centrality of job posting content and significant investment in its analysis, no rigorous validation exists of whether posting-derived skill measures predict actual selection and mobility outcomes. While firms are investing in more detailed skill descriptions, and sophisticated tools are being built to extract and analyze this content, consumers of posting-based analytics and tools operate without clear evidence of informativeness. Legitimate concerns exist - postings might be aspirational rather than realistic, could be template-driven or overly generic, and might list “nice to have" skills rather than true requirements. Understanding how informative these postings are for measuring skill requirements is crucial for the growing ecosystem of tools and analytics being built on this content. The ability to validate posting-derived measures against selection decisions would inform both current applications and future innovations.

We partner with the IT division of a major financial services firm to study the informativeness of job postings by linking it to actual workers and their mobility process. This setting provides an ideal laboratory - the firm’s hiring process generates detailed position requirements through job postings, which multiple internal candidates consider and apply to. The firm’s internal job board offers comprehensive records of both successful and unsuccessful applications over multiple years. Crucially, we can link the applicants

[1]The workforce analytics provided by firms like Revelio and Lightcast have not only supported businesses but also spurred a growing body of academic research that study trends in skill demands and job creation (Goldfarb, Taska, and Teodoridis, 2020; Azar et al., 2020), and other critical labor market phenomena (Hershbein and Kahn, 2018; Forsythe et al., 2020; Braxton and Taska, 2023).

to the postings of their current job and observe both successful and unsuccessful applications, providing the opportunity to assess the informative content in job postings to predict selection. By structuring a prediction problem, we can examine whether the skill distances derived from posting content meaningfully predict selection outcomes, thereby contributing a first of its kind estimate of the nature and level of informativeness in job postings; an information source of increasing application inside and outside the firm.

Our analysis reveals substantial predictive power in posting-derived skill measures. The probability of selection decreases by 84% when comparing applications in the lowest versus highest quintile of skill distance, indicating posting content captures meaningful differences in worker-job fit. This predictive power is particularly evident when examining vacancies with multiple applicants - in 70% of these cases, the candidate with the shortest skill distance was selected. The predictive relationship holds both across vacancies and within applicants (across their different applications). Notably, traditional employee characteristics like age, tenure, and work experience contribute minimally beyond the skill distance measure - suggesting the rich information in job postings dominates standard observable characteristics in explaining selection outcomes.

The first step in our machine learning pipeline converts the key information in a job posting to a job skill-embedding by mapping it in the language space of a pre-trained language model. While these embeddings capture rich semantic relationships in general language, our context of technical skills and responsibilities likely lies in a much lower dimensional space. Moreover, with relatively few selection decisions compared to the embedding dimensions, we need dimension reduction to enable effective learning of the distance metric. The distance metric learning, akin to fine-tuning the general language model's notion of similarity for our specific context, reveals meaningful skill relationships. This learned representation's validity is demonstrated through its strong predictive power on held-out selection decisions, and its application reveals novel patterns in how employees navigate opportunities based on skill-fit.

We apply the skill distance measure to investigate internal job application patterns, revealing another dimension of job posting informativeness. Those whose skills closely match available vacancies adopt a relatively passive approach despite higher chances of success, while employees facing larger skill distances actively pursue positions requiring skill development. By structuring standard posting and mobility data, we show how posting-derived analytics can generate actionable insights about whether reducing search frictions or investing in skill development should be prioritized. Such capabilities suggest a promising direction for HR analytics, a field that has seen limited advancement despite significant interest (Tambe, Cappelli, and Yakubovich, 2019) - the combination of skill information in job postings and mobility patterns enables rich workforce analytics. Our analysis reveals novel patterns about internal mobility, providing evidence about job

search behavior in an environment characterized by employed workers navigating evolving skill requirements - a setting distinct from traditional unemployment-focused search models. This ability to measure skill requirements from job postings and track mobility decisions opens new possibilities for personnel economics research to systematically investigate career navigation and skill development choices.

Our paper illuminates the informative content in job postings by linking their skill and responsibility descriptions to selection outcomes and application decisions. Through selection prediction, we provide first validation that posting-derived skill measures capture meaningful differences in worker-job fit - a crucial validation given the growing use of posting content in platforms, analytics, and research. Through our inquiry into application patterns, we illuminate how this posting information relates to mobility behavior. This analysis demonstrates significant potential for firms to gain insights about internal mobility and its relationship to skill-fit by leveraging posting information, while opening new research possibilities in understanding how employees navigate careers in environments characterized by evolving skill requirements.

To summarize, this paper illuminates how findings validate a fundamental data source driving major market decisions while demonstrating rich analytical possibilities. By showing posting content contains meaningful skill information that predicts outcomes, we inform the growing ecosystem of screening tools and analytics built on this content. The results challenge pessimism about HR analytics by showing how posting information can guide specific interventions - from reducing search frictions for well-matched workers to targeting skill development where needed. This posting-based approach outperforms traditional collaborative filtering methods by leveraging the rich content in job descriptions, opening new avenues for research in labor economics and personnel practices.

1.1. Literature Review

Our work advances a rich technical literature on job recommendation, worker-job matching, and skill representation learning—supporting platforms that facilitate job search, candidate-opening matching, and career trajectory modeling[2]. While these advances demonstrate sophisticated ways to process job posting content, their focus on platform functionality leaves open the fundamental question of whether derived skill measures predict actual selection outcomes. Our work provides the first systematic validation of these measures through observed selection decisions.

Our research contributes directly to the Information Systems literature studying al-

[2]The technical literature has included work on recommender systems (Shaha and Mourad, 2012; Siting et al., 2012) and representation learning (Heap et al., 2014; Zhu et al., 2018; Liu et al., 2019; Bian et al., 2020) to support platforms facilitating job search (Heap et al., 2014; Giabelli et al., 2021), matching jobs to candidates (Zhu et al., 2018; Qin et al., 2020) and modeling career paths and skill recommendations (Maurya and Telang, 2017; Kokkodis and Ipeirotis, 2021).

gorithms that match workers to jobs and recommend skill development (Kokkodis and Ipeirotis, 2021, 2023), while addressing fundamental questions about assessing the value of data for algorithmic applications (Lei, Chen, and Sen, 2024). Particularly relevant is Raghavan et al. (2020)'s documentation of limited transparency in automated screening systems. By demonstrating that posting-derived skill measures strongly predict selection, we provide both validation of these measures and a framework for evaluating hiring algorithms in practice.

Our work enriches the study of internal labor markets, where different approaches to measuring employee-position fit have emerged. De Vos et al. (2024) develops recommender systems using collaborative filtering based on observed transitions and employee features, while Cowgill et al. (2024) examines team assignment mechanisms using survey-based skill scores validated by management. We demonstrate how the rich skill information contained in job postings can generate validated measures of fit that complement these approaches, requiring neither extensive mobility histories nor expert validation.

We provide crucial validation for a growing literature in management and economics that leverages job posting text to study labor market dynamics. Researchers have extracted signals from job descriptions to examine wage premia (Bana, 2021), analyze technology's impact on skill demands (George et al., 2024), and forecast effects of generative AI (Eloundou et al., 2024; Acemoglu et al., 2024). Our demonstration that posting-derived skill measures strongly predict selection outcomes provides important validation for this literature's growing use of posting text to measure skill requirements.

Our analysis advances research on career moves and incentives in personnel economics, where studies have documented patterns in internal moves (Bidwell and Keller, 2024) and efficiency gains from internal hiring (Bidwell, 2011). This literature spans from early work on career incentives (Baker, Gibbs, and Holmstrom, 1994a,b) to recent studies of internal labor markets (Tambe, Ye, and Cappelli, 2020; Huitfeldt et al., 2023). By validating posting-derived skill measures and showing how they predict both selection and application patterns, we provide new tools for studying modern career navigation in environments where skill requirements evolve rapidly.

We proceed as follows: In Section 2, we describe the research setting, the firm, data and the internal mobility process. In Section 3, we outline the conceptual ideas of the prediction problem. Section 4 presents the machine learning pipeline, the algorithms and details of data processing and implementation. In Section 5, we evaluate the effectiveness of the skill distance metric in predicting selection outcomes in detail. In Section 6 we apply the distance measure to new positions to study application intensity to new positions vary with skill fit and concluding in Section 7.

2. Research Setting

2.1. Context

This section outlines our research setting within the Information Technology division of a prominent U.S.-based financial services institution. The division, comprising multiple specialized units, provides technical support ranging from trading systems and wealth management platforms to enterprise-wide initiatives like system modernization and compliance reporting. Several distinctive features make this setting ideal for validating posting informativeness: precise skill specifications required for technical roles, diverse units generating varied requirements, and structured mobility processes yielding both successful and unsuccessful applications.

Within the division, specialized technical roles form the core workforce—application programmers developing trading platforms, data analysts supporting risk management, and quality assurance specialists maintaining enterprise systems. Each position demands specific technical competencies, from programming languages and analytical tools to testing frameworks and development methodologies. Such specialized requirements necessitate detailed articulation of skills in job postings, generating rich content for examining whether posting-derived measures predict selection outcomes.

A centralized job board facilitates internal mobility, where positions are posted with comprehensive skill and responsibility specifications (see Box 2.2 for a representative example). The organization maintains minimal restrictions on internal applications, allowing employees to pursue roles aligned with their interests without prior managerial approval. Following application submission, recruiting units conduct initial screening and interviews to identify suitable internal candidates. External candidates are considered only after a specified timeframe if no appropriate internal candidate emerges.

The organization's structured internal mobility process generates substantial variation in both applications and outcomes across positions. The standardized posting and evaluation procedures, combined with comprehensive data capture of both successful and unsuccessful applications, provide the essential elements for validating posting informativeness through actual selection decisions. This combination of detailed technical postings, standardized processes, and comprehensive outcome data makes the setting particularly suitable for our research objectives.

2.2. Data

Our analysis draws on comprehensive data from the firm's Human Resource Information System (HRIS), which captures all job requisitions, applications, and selection outcomes. Each record includes complete job posting content, application details, and final decisions. A crucial feature for our research design lies in our ability to link internal applications

TABLE 1. Job Postings and Internal Mobility (2018-2021)

Year	Total Positions	Internal Applications (%)	Internal Fills (%)
2018	1,326	40.3	8.3
2019	1,370	50.1	17.3
2020	1,606	75.7	42.5
2021	2,240	62.5	33.9
Total	6,542	58.6	27.3

Notes: Data from HRIS covering 2018-2021, supplemented by 1,638 pre-2018 position postings. Internal Applications (%) represents positions receiving at least one internal application. Internal Fills (%) represents positions filled by internal candidates.

TABLE 2. Structure of HRIS Data

Date	Requisition ID	Job Description	Applicant ID	Employee ID	Internal	Selection
2021-03	REQ2021-456	Hadoop Developer	APP-7K89	EMP-123	Yes	Selected
2021-03	REQ2021-456	Hadoop Developer	APP-7K90	EMP-124	Yes	Not Selected
2021-03	REQ2021-456	Hadoop Developer	APP-7K91	–	No	Not Selected
2021-04	REQ2021-457	Data Analyst	APP-7K92	–	No	Selected
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2021-05	REQ2021-459	Sr. Developer	APP-7L95	EMP-127	Yes	Selected
2021-05	REQ2021-459	Sr. Developer	APP-7L96	–	No	Not Selected
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2021-06	REQ2021-460	ML Engineer	APP-7M01	EMP-130	Yes	Not Selected

Notes: Sample entries from HRIS. Employee ID is populated for internal applicants or when external candidates are selected. Job Description field contains complete posting content (see Box 2.2) for representative example of full posting content. Internal flag indicates current employee status.

to the detailed posting of the applicant's current position at application time, enabling direct comparison of skill requirements between current and sought positions.

Table 1 presents our dataset covering individual contributor positions posted between 2018-2021. The observation window captures 6,542 positions, with 3,836 (58.6%) receiving internal applications and 1,789 (27.3%) filled internally. This data is supplemented by 1,638 pre-2018 position postings, enabling broader coverage when analyzing internal applicants' current positions. The data shows a notable increase in internal mobility over the observation period, with the proportion of positions receiving internal applications rising from 40.3% in 2018 to 62.5% in 2021.

The structure of our HRIS data appears in Table 2, which illustrates the comprehensive information available for each application. The system records application timing, position details, applicant information, and selection outcomes. Crucially, for internal applicants, employee identifiers enable linking to their current positions and corresponding job postings. This linkage capability is essential for analyzing how differences between

TABLE 3. Sample Job Requisitions and Internal Applications

Job Requisition No.	Current Position	Position Applied To	Date	Selected
REQ2021-456	Data Engineer	Hadoop Developer	2021-03	Yes
REQ2021-456	ML Engineer	Hadoop Developer	2021-03	No
⋮	⋮	⋮	⋮	⋮
REQ2021-458	Data Analyst	ML Engineer	2021-04	Yes
REQ2021-459	Software Engineer	Sr. Developer	2021-05	Yes
⋮	⋮	⋮	⋮	⋮

Notes: Derived from HRIS data by linking internal applicants' Employee IDs to their current positions. Both current and applied-to positions contain detailed requirements as shown in Box 2.2.

current and sought position requirements relate to selection outcomes.
Notes: This is a sample vacancy posting with information about the firm, the sub-division, location etc. redacted. The posting demonstrates the rich content available for analyzing skill requirements and job responsibilities.

Box 2.2 demonstrates the typical depth of position postings in our dataset. As shown, each posting contains detailed specifications spanning multiple dimensions: role overview, process context, specific responsibilities, and both mandatory and desired skills. This rich content enables precise measurement of skill requirements and job characteristics, providing the foundation for validating posting informativeness.

Table 3 illustrates how we leverage the HRIS data structure to track internal mobility patterns. By linking internal applicants to their current positions through employee identifiers, we can observe both the positions they apply to and detailed requirements of their current roles. Among positions receiving internal applications, we identified 1,370 cases where we could observe the detailed posting of the applicant's current position at application time, forming the core sample for our analysis of posting informativeness.

Sample Job Description

Overview

We are one of the world's leading financial institutions ... and risk management products and services.

Process Overview

Build and evolve a consistent Authorized Data Source within Consumer & Small Business Bank (CSBB) ... both the strategic and tactical analytics needs of the Consumer Bank.

Job Description

Hadoop developer for multiple initiatives. Develop Big Data Strategy and Roadmap for the Enterprise. Experience in Capacity Planning, Cluster Designing and Deployment. Benchmark systems, analyze system bottlenecks, and propose solutions to eliminate them. Develop highly scalable and extensible Big Data platform, which enables collection, storage, modeling, and analysis of massive data sets from numerous channels. Continuously evaluate new technologies, innovate and deliver solution for business-critical applications.

Responsibilities

Assists the team with the design of the architect layer to ensure re-usable metrics and attributes within the reporting layer. Responsible for creating and maintaining necessary documentation (MDR) to ensure audit readiness where necessary. Prototype improvement ideas. Work effectively with the global team. Expected to play technical leadership as an individual contributor. Articulate challenges, propose and drive solutions ...

Mandatory Skills

Extensive knowledge of Hadoop stack and storage technologies HDFS, MapReduce, Yarn, HIVE, sqoop, Impala, spark, flume, kafka and oozie. Extensive Knowledge on Bigdata Enterprise architecture (Cloudera preferred). Experience in No SQL Technologies (Cassandra, Hbase).

Desired Skills

Experience in Real time streaming (Kafka). Experience with Big Data Analytics & Business Intelligence and Industry standard tools integrated with Hadoop ecosystem. (R, Python). Visual Analytics Tools knowledge (Tableau). Data Integration, Data Security on Hadoop ecosystem. (Kerberos). Awareness or experience with Data Lake with Cloudera ecosystem.

The combination of detailed technical postings, comprehensive observation of application outcomes, and our ability to link current and sought positions makes this setting particularly valuable for validating posting informativeness. The rich content in job post-

ings reflects precise skill requirements for technical positions. Moreover, access to both successful and unsuccessful applications enables robust testing of posting content's predictive power. Our unique ability to observe detailed postings for applicants' current positions at the time of application allows direct measurement of skill distances, providing compelling evidence of whether these measures meaningfully predict selection outcomes.

3. Objective and Approach

Our central aim is to understand the extent to which job postings convey meaningful information about the actual requirements of a position, specifically regarding skill fit. We investigate this by analyzing patterns in internal hiring decisions, under the premise that these decisions implicitly reveal the skills and demands prioritized by the organization.

Job postings utilize varied language to describe complex technical skills. Similar underlying requirements might be expressed differently across postings, while seemingly disparate descriptions could indicate similar skill needs. Internal mobility offers a valuable context for studying posting informativeness, as we observe both the textual content of position descriptions and the outcomes of employee transitions between roles. If postings effectively communicate skill requirements, we expect differences in their content to correlate with these transitions.

To formalize our investigation, consider an internal applicant being evaluated for a new role. Let j_c denote the text of the applicant's current job posting and j_v the text of the vacancy posting. The hiring manager considers various factors, some unobserved by us, represented by the feature vector z_i. We can model the selection decision as dependent on a distance measure $\delta(j_v, (j_c, z_i))$ between the vacancy and the applicant's profile, with selection occurring if this distance is below a threshold τ:

$$S(j_v, (j_c, z_i)) = \begin{cases} 1, & \text{if } \delta(j_v, (j_c, z_i)) < \tau \\ 0, & \text{if } \delta(j_v, (j_c, z_i)) \geq \tau \end{cases}$$

Given that we do not observe the full set of features z_i considered in the actual selection process, our ability to perfectly predict individual hiring decisions based solely on job posting content is limited. We use the Area Under the Receiver Operating Characteristic Curve (AUC) to assess the extent to which the distance between job postings, derived from their text, aligns with observed hiring outcomes, recognizing that this metric provides a measure of how informative the postings are in accounting for skill fit, even with the presence of unobserved factors.

Our approach is guided by two key principles in representing and comparing job postings. First, we leverage the sophisticated language understanding capabilities of pre-

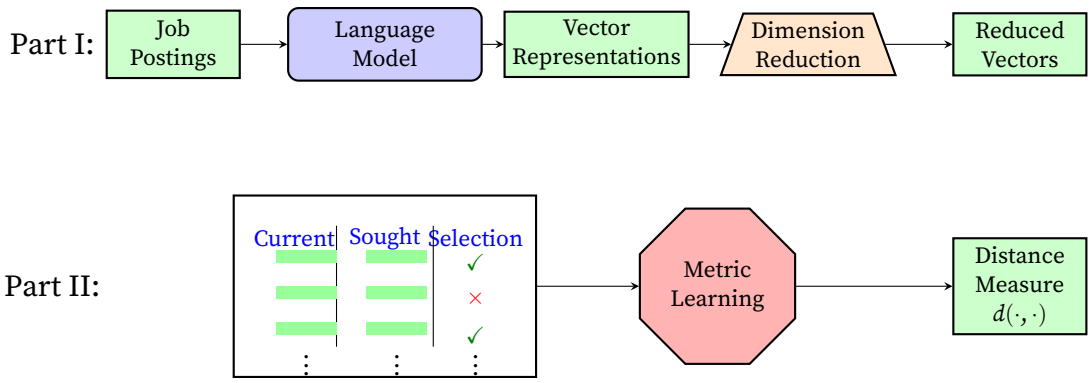


FIGURE 1. Overview of our measurement approach. Part I illustrates the transformation of job posting text into vector representations: job postings are processed through a language model to create embeddings, which are then reduced to a lower-dimensional space. Part II shows how we learn a distance metric from the data: pairs of position vectors (current and sought positions) together with their selection outcomes are used to learn a distance measure that captures meaningful differences between positions.

trained transformer models to generate numerical representations (embeddings) of the posting text. The ability of these models to capture nuanced semantic relationships within large text corpora makes them well-suited for representing the specialized vocabulary of IT job descriptions, and their pre-trained nature offers significant practical advantages. Second, we utilize our historical selection decisions in a supervised manner to learn a distance metric between these embeddings. This process of metric learning allows us to fine-tune the distance measure, weighting differences in the embeddings according to their relevance in predicting observed hiring outcomes. This supervision allows us to extract signals related to the organization's implicit valuation of skills, as revealed through its hiring choices.

A critical aspect of our methodology is addressing the challenges associated with the high dimensionality of embeddings generated by large language models. One key motivation for dimensionality reduction is that learning a meaningful distance metric, which involves assigning appropriate weights to different dimensions, becomes increasingly difficult in high-dimensional spaces, particularly with a limited number of observed selection decisions. Furthermore, the embeddings of job postings, even for distinct roles within a single organization, are anticipated to exhibit a tendency to cluster within the broader embedding space. Therefore, we employ dimensionality reduction techniques to map these high-dimensional representations to a lower-dimensional space, aiming to distill the most salient features relevant to distinguishing between job requirements and facilitating effective metric learning.

The following section details our implementation pipeline. We begin by generating high-dimensional embeddings of job postings using pre-trained transformer models.

4. Machine Learning Pipeline

In this section, we explain the machine learning pipeline that converts vacancy posting text into continuous numerical vectors. We leverage the structured nature of job descriptions to extract high-signal information efficiently. Job descriptions, despite their length, follow a consistent format. We focus on the easily detectable Mandatory and Desired Skills sections, which provide concentrated information about skill and competency requirements. Using the Spacy natural language processing library, we parse these sections to extract key technical skills such as programming languages (e.g., Java, Asp.Net), web frameworks (e.g., Django), and testing tools (e.g., Selenium).

While knowledge of technologies is important, the Responsibilities section conveys crucial functional aspects of the role. These can vary notably even between positions requiring similar technology skills, such as developers, testers, and support engineers. The section typically lists tasks to be performed, which might include 'develop big data strategy,' 'conduct code reviews,' 'engage with stakeholders' or 'application development.' We extract such task descriptions from job responsibilities by employing natural language processing techniques. This includes part-of-speech tagging to identify action verbs and dependency parsing to locate their direct objects, which are then combined to form concise verb-object phrases representing key tasks. By leveraging the structured format and our knowledge of the postings, we effectively eliminate low-signal words in the description. In the following subsections, we describe our machine learning pipeline which adheres to two key ideas: leveraging capabilities of a pre-trained language model and utilizing past selection decisions.

4.1. Pre-trained Embeddings

Here we describe the approach to mapping the extracted text from job postings to the embedding space of a pre-trained language model. This section focuses on generating robust embeddings for informativeness assessment, rather than achieving state-of-the-art performance on general semantic similarity tasks. Prior to embedding, standard pre-processing steps such as lowercasing, stop word removal, and punctuation normalization were applied to refine the text.

To create vector representations of the two parts we employ Sentence-BERT (SBERT), a modification of the pre-trained BERT network (Devlin et al., 2019). SBERT is designed to derive semantically meaningful sentence embeddings, making it more efficient for comparing longer text strings than BERT embeddings, which are typically word vectors, although the BERT word-level embeddings vary with the context, unlike their predecessors (Reimers and Gurevych, 2019). SBERT converts text into dense vector representations, effectively capturing the semantic meanings and relationships between words. This model

is particularly suitable for our needs because it is pre-trained and fine-tuned for similarity tasks, ensuring that it can generate fixed-size embeddings that are both informative and consistent. The embeddings are generated using the SentenceTransformers library (model: paraphrase-MiniLM-L6-v2). By leveraging SBERT, we ensure that our embeddings capture the nuanced language used in job postings, which is critical for accurately representing the skills and tasks.

For every job posting j_i, we have the set of words j_i^s (skills) and j_i^t (tasks) and obtain SBERT embeddings e_i^s (384×1) and e_i^t (384×1) respectively. This separation allows for more granular embeddings and a focused semantic representation of each aspect. To evaluate the benefit of this separation, we also conducted a baseline experiment where the 'Mandatory Skills,' 'Desired Skills,' and 'Responsibilities' sections were concatenated before embedding (referred to as the 'no-split' approach). These embeddings encapsulate the semantic content of the job postings, allowing us to compare and analyze them effectively. The dual-embedding approach provides a comprehensive view of each position, highlighting both the required competencies and the associated responsibilities. This detailed representation forms the foundation for developing a robust distance metric that can accurately reflect the similarity between job postings, ultimately aiding in the assessment of internal mobility and selection probabilities.

4.2. Dimension Reduction: A Two-Stage Approach with PCA and Fuzzy C-Means

Dimension reduction is a critical step in our machine learning pipeline, transforming the initial high-dimensional embeddings into a more manageable and informative representation. To achieve this, we employ a deliberate two-stage process: Principal Component Analysis (PCA) followed by Fuzzy C-Means (FCM) clustering. This sequential approach combines PCA's strength in capturing linear relationships with FCM's ability to model complex, overlapping characteristics of job postings.

Our initial step employs Principal Component Analysis (PCA), a well-established technique for reducing the dimensionality of high-dimensional data while preserving the most significant variance. The initial Sentence-BERT embeddings, for both skills (e_i^s) and tasks (e_i^t), reside in a 384-dimensional space. Working directly with this high dimensionality poses challenges for learning a robust and tailored distance metric. PCA addresses this by identifying principal components—orthogonal linear combinations of the original features that capture maximum variance in the data. By projecting the data onto these top principal components, we effectively reduce dimensionality while retaining the most salient linear information and mitigating the impact of noise. We apply PCA separately to the skill and task embeddings, recognizing that these represent distinct aspects of a job posting with potentially different underlying structures. Retaining the components that explain 90% of the variance reduces the skill embeddings to 49 dimensions (p_i^s) and the

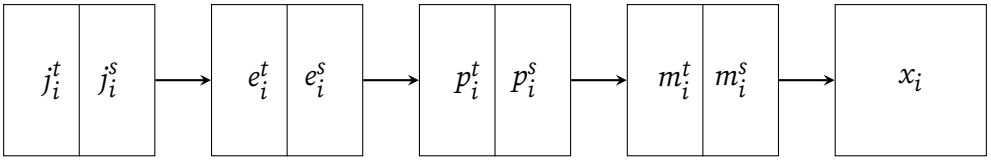
task embeddings to 124 dimensions (p_i^t). This difference in dimensionality suggests that skill descriptions exhibit more structured patterns, likely due to the specific vocabulary used for technical skills compared to the broader language of job responsibilities. While PCA effectively captures linear relationships, the potential presence of non-linear patterns motivates our subsequent step using Fuzzy C-Means clustering.

Building upon PCA's dimensionality reduction, we employ Fuzzy C-Means (FCM) clustering to create a representation capturing non-linear relationships and allowing nuanced characterization of job postings. Unlike traditional clustering methods where each posting belongs to a single cluster, FCM enables membership in multiple clusters with varying degrees of association. For example, a posting for a "Senior Java Developer" position might show 70% membership in a "Java" cluster, 15% in "Cloud Technologies," and 10% in "Backend," while having minimal membership in clusters representing "Networking" or "QA/QC." Similarly, a "Network Administrator" posting might demonstrate strong membership in "Network Infrastructure," moderate membership in "System Administration," and lower membership in "Security Protocols," with negligible association to software development clusters. These membership probabilities, ranging from 0 to 1, provide a nuanced representation of each position's skill and task composition. FCM identifies cluster centers in the PCA-reduced data that represent prototypical skill and task profiles, then assigns membership probabilities based on proximity to these centers.

In this dimension reduction step, FCM transforms the PCA-reduced embeddings into vectors of membership probabilities. We apply FCM separately to the PCA-reduced skill embeddings (p_i^s) and task embeddings (p_i^t). For skills, FCM identifies k_s clusters, representing each job posting with a membership probability vector (m_i^s) of dimension k_s. Each element represents the posting's alignment with a specific skill cluster—higher values indicating stronger alignment with that cluster's skill profile. Similarly, for tasks, FCM identifies k_t clusters, producing a membership probability vector (m_i^t) of dimension k_t. Concatenating these vectors creates our final reduced-dimensional representation (x_i) of dimension $(k_s + k_t)$. While the precise interpretation of individual clusters merits further investigation, this representation effectively captures each posting's relationship to prevalent skill and task combinations in our data. This provides a foundation for learning a distance metric tailored to our organizational context, enabling meaningful comparisons between positions based on their skill and task profiles.

4.3. Distance Metric Learning

Having established a numerical representation of job descriptions through our two-stage dimension reduction, we now focus on calibrating a distance metric to effectively capture meaningful differences between positions. Our approach combines the rich representations derived from pre-trained language models with historical selection decisions as a



Notation	Description	Dimensions
$j_i^{t/s}$	Split to task and skill components of j_i	text
$e_i^{t/s}$	SBERT($j_i^{t/s}$)	384×1
$p_i^{t/s}$	PCA($e_i^{t/s}$)	124×1 (tasks), 49×1 (skills)
$m_i^{t/s}$	FuzzyC($p_i^{t/s}$)	$k_t \times 1$, $k_s \times 1$
x_i	$m_i^t \oplus m_i^s$	$k_t + k_s \times 1$

FIGURE 2. This is a pictorial representation of the vectorization process that captures the information in a job posting to a numerical vector. The table provides a summary of the notation and dimensions used at each step of the process.

supervisory signal. These past decisions serve as ground truth, helping us identify which aspects of job posting content are truly informative for predicting successful hires. Importantly, we do not inject selection outcome information into the job postings themselves; rather, historical data guides our metric to emphasize the inherently predictive differences in posting content.

Learning a custom metric offers significant advantages over generic distance measures in assessing job posting informativeness. Traditional metrics like Euclidean distance treat all dimensions equally, failing to capture the varying importance of different skills and responsibilities in actual hiring decisions. Through metric learning, we calibrate the weights of our representation vector (x_i) based on how informatively different aspects of posting content differentiated between successful and unsuccessful applications. Consider a software development position requiring both core technical skills (e.g., expertise in Mainframe development) and general workplace skills (e.g., presentation software proficiency). Our metric learns to emphasize substantial differences in core technical requirements, which historically strongly predicted selection outcomes, while de-emphasizing variations in general skills that had minimal impact on hiring decisions.

We employ the family of Mahalanobis distances, parameterized by a positive semi-definite matrix (M), where dimensions are weighted according to their informativeness in distinguishing successful applications. In the diagonal case, where (M) contains only diagonal entries, the distance calculation represents an independent scaling of each dimension—stretching those that historically differentiated outcomes and compressing less informative ones. When (M) is non-diagonal, it can capture more complex relationships through both scaling and rotation of the feature space, revealing subtle but informative patterns in the posting content. Formally, we define our distance function as:

$$d(x_v, x_c; M) = \sqrt{(x_v - x_c)^T M (x_v - x_c)} \tag{1}$$

where (x_v) represents the vector of the sought vacancy, (x_c) represents the vector of the current job, and (M) is our calibrated Mahalanobis matrix.

To calibrate M, we utilize a dataset (A) of tuples $((x_v, x_c, S_{v,c}))$, comprising vector representations of vacancy and current job descriptions with their corresponding selection outcomes. We employ the Mahalanobis Metric for Clustering (MMC) algorithm (Xing et al., 2002), which solves:

$$\text{Maximize:} \quad \sum_{(v,c):S_{v,c}=0} d(x_v, x_c; M)$$

$$\text{Subject to:} \quad \sum_{(v,c):S_{v,c}=1} d(x_v, x_c; M)^2 \leq 1$$
$$M \succeq 0$$

This formulation calibrates the metric to distinguish between successful and unsuccessful applications by maximizing distances between rejected pairs while maintaining bounded distances for selected pairs. Implementation uses Python 3.11 and scikit-learn, leveraging 1,060 observations (75% of our selection decisions) for training. The dimensionality of (M), determined by our FCM clustering parameters, influences the granularity of our representation of informative skill and task profiles, which we optimize through parameter tuning as discussed next.

4.4. Parameter Tuning

To implement the Mahalanobis Metric for Clustering (MMC) algorithm effectively, we need to determine the optimal number of clusters for the Fuzzy C-Means (FCM) step: (k_t) for task embeddings and (k_s) for skill embeddings. These parameters directly determine how finely our representation can distinguish between different types of job postings. The choice of cluster numbers influences our ability to capture meaningful distinctions between positions while avoiding overly granular distinctions that might not generalize well. Selecting appropriate values is crucial for creating a representation that effectively captures the underlying patterns in our hiring data and leads to the most accurate distance metric for predicting selection probabilities.

Utilizing our dataset of 1,370 internal applications and their outcomes (excluding cases where internal candidates were rejected in favor of other internal applicants), we employ a rigorous 5-fold cross-validation strategy on the application dataset (A). For each fold, we test combinations of (k_s) and (k_t), where $k_s, k_t \in \{2, 3, \ldots, 20\}$, subject to the constraint $k_s + k_t < 25$. This constraint serves a practical purpose: with limited training data, allowing the dimensionality of the Mahalanobis matrix (M) to grow too large risks overfitting and poor generalization. For each combination, we train the MMC algorithm on the reduced-dimensional representations derived from the SBERT embeddings and FCM, evaluating performance on the held-out fold using the Area Under the Receiver Operating Characteristic Curve (AUC).

Our cross-validation analysis, visualized in Figure 3, reveals that optimal performance is achieved with $k_s = 17$ skill clusters and $k_t = 3$ task clusters, yielding an average AUC of 0.62. This substantial imbalance between the number of skill and task clusters—17 versus 3—aligns with our intuition about job postings: while technical skills tend to be numerous and specific (e.g., programming languages, tools, frameworks), job tasks often fall into broader categories (e.g., development, maintenance, management). The data-

driven optimization of these parameters ensures our representation captures meaningful patterns while maintaining generalizability.

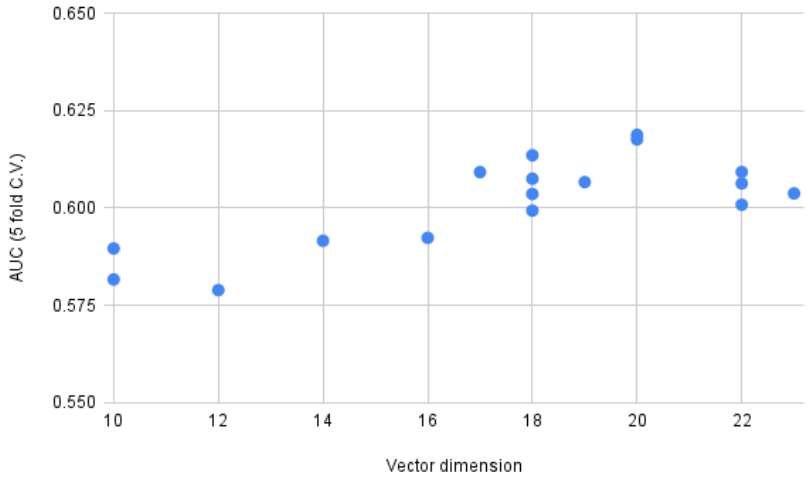


FIGURE 3. Cross-validated AUC scores across different total dimensionalities (k), showing optimal performance at $k = 20$

4.5. Evaluating the Value of Distance Metric Learning

To assess the effectiveness of our learned distance metric, we conduct a comparative analysis against a strong baseline approach. This evaluation serves two purposes: first, to validate that our metric learning process captures meaningful patterns in selection decisions, and second, to quantify the value added by learning an organization-specific distance metric compared to using general semantic similarity measures.

Baseline Implementation: Our baseline leverages the semantic capabilities of Sentence-BERT, using cosine similarity between job posting embeddings as a measure of job similarity. We implement this baseline using the SentenceTransformers library (model: paraphrase-MiniLM-L6-v2), which generates 384-dimensional embeddings. For direct comparison with standard approaches in natural language processing, we process each job posting as a single text, concatenating the 'Mandatory Skills,' 'Desired Skills,' and 'Responsibilities' sections. The cosine similarity between two postings is computed as the normalized dot product of their embeddings, a standard approach for measuring semantic similarity between texts.

Experimental Design: We evaluate both approaches—the baseline cosine similarity and our learned Mahalanobis distance—using the Area Under the Receiver Operating Characteristic Curve (AUC). The AUC in our context represents the probability that our metric will correctly rank a randomly chosen successful application above a randomly

chosen unsuccessful one. This evaluation metric is particularly appropriate as it assesses the relative ordering of job similarities rather than absolute threshold values. We conduct this evaluation using our held-out test set of selection decisions to ensure unbiased assessment of generalization performance.

Results and Analysis: Our evaluation reveals three key findings:

1. The baseline SBERT cosine similarity achieves an AUC of 0.562, significantly above chance (0.5). This indicates that even general semantic similarity between job postings contains some signal about potential selection outcomes.

2. Our learned Mahalanobis distance, using the optimal cluster parameters ($k_s = 17, k_t = 3$), achieves an AUC of 0.62. This substantial improvement over the baseline demonstrates the value of learning an organization-specific metric.

3. The magnitude of improvement (0.058 in AUC) represents approximately a 10% increase in predictive accuracy. In practical terms, this means our learned metric is about 10% more likely to correctly identify suitable internal candidates compared to using general semantic similarity.

These results validate our core premise that job posting text contains valuable information about position requirements that can be extracted through careful methodology. The superior performance of our learned metric suggests that it successfully captures organization-specific patterns in hiring decisions that go beyond general semantic relationships. Our metric has learned to emphasize aspects of job descriptions that are truly predictive of successful hires within our organizational context, leading to more accurate assessments of job fit.

The improvement in predictive accuracy is particularly meaningful given our use of a strong baseline. While the SBERT cosine similarity represents a sophisticated approach to semantic textual similarity, it treats all aspects of similarity equally. Our learned metric, in contrast, has adapted to recognize which differences between job postings are most informative for predicting selection outcomes. This targeted emphasis on relevant features explains the substantial gain in predictive performance.

5. Evaluation: Does Skill Distance Predict Selection?

This section evaluates whether our job postings based skill-distance measure meaningfully captures skill requirements by testing how well our derived distance metric predicts selection of internal candidates for new positions. Our primary interest lies in understanding the extent to which posting content predicts selection, and in assessing the relative contribution of the skill distance measure compared to a few other observed features. While an Area Under the Curve (AUC) of 0.62 might appear modest in a traditional classification setting, the key contribution of this section is demonstrating the surprisingly rich and informative signal captured by a distance metric derived solely from job posting content.

Even with this level of predictive power, we reveal that job postings contain substantial information about the skills relevant to selection decisions.

Our test set comprises 308 cases where internal applicants applied to vacancies (approximately 25% of our total cases). We examine whether skill distances computed purely from posting content capture predictable variation in selection outcomes. It's important to note that the information contained in an applicant's job description (j_c) may not fully capture the skills and abilities of that position, let alone those of the candidate. This potential measurement loss informs our approach.

5.1. Skill Distance and Selection Probability

To assess the relationship between skill distance and selection probability, we employ a logistic regression to quantify the association:

$$\text{logit}(P(Sc, v = 1)) = \beta_0 + \beta_1 \times Q_k(dj_c, j_v) \tag{2}$$

Here, $Q_k(d_{j_c,j_v})$ represents the quintile of the skill distance between the applicant's current job (j_c) and the vacancy (j_v). We discretize the distance into a categorical variable, as the relationship between selection probability and distance is unlikely to be simply linear. This discretization offers stability and robustness to extreme values while retaining easy interpretation. Table 4 shows that the coefficient for skill distance is negative and significant ($\beta_1 = -0.3268,\ p < 0.01$), indicating that applicants with lower skill distance to a vacancy have a higher likelihood of being selected.

TABLE 4. Logistic Regression Results for Skill Distance and Selection Probability

Variable	**Coefficient**	**Std. Error**
Intercept	1.1171***	0.265
Quintile	-0.3268***	0.086
Observations	308	
Pseudo R-squared	0.03551	

*** p<0.01

The regression suggests that, on average, an applicant-vacancy pair in the lowest skill distance quintile has an 84% higher probability of selection compared to a pair in the highest quintile. This substantial difference in selection probability across skill distance quintiles strongly suggests that our posting-derived metric is effectively capturing informative variations in the skill alignment between applicants and vacancies.

While skill distance accounts for a small component of the variation in selection decisions, its ability to capture differences in selection probability is unambiguous. We

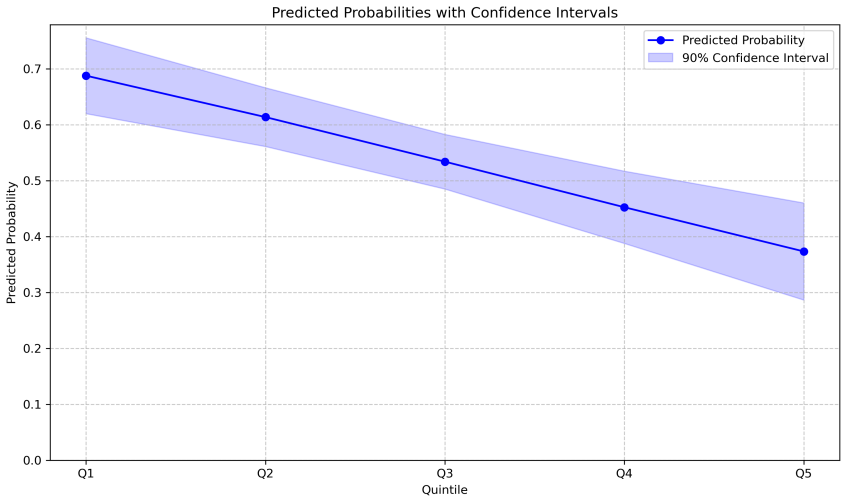


FIGURE 4. Selection Probability by Skill Distance Quintile

will now exploit the structure in our test sample to examine the association.

5.2. Skill Distance: Within-Vacancy and Within-Applicant Variation

When multiple internal candidates apply to the same vacancy, we have an opportunity to verify if selection among different applicants for the same position is predicted by their respective skill distances. Correlation between skill distance and selection measured at each vacancy level would convey a more direct link of the measured distance. Similarly, we have cases where an applicant has applied to different vacancies. This allows us to focus on applicant-level selection data to examine whether the applicant has a higher selection probability for vacancies to which they have a low skill distance. Focusing on a single applicant allows us to control for unobserved factors that influence their selection chances across vacancies, such as broadly relevant skills they have acquired or performance in their current job. While we have already established a notable negative association between skill distance and selection, these analyses would further demonstrate that the measured skill distance captures significant selection variation at both the applicant and vacancy levels, thereby reinforcing the measure's informativeness.

5.2.1. Within-Vacancy Selection

To assess the predictive power of the skill distance metric, we evaluated the proportion of vacancies where the applicant with the lowest skill distance was selected. Our analysis encompassed 261 observations across 110 vacancy groups from our test set, each group comprising a vacancy with multiple internal applicants and one selection. The analysis revealed that the candidate with the lowest skill distance was chosen in 69.92% of cases,

suggesting the metric's reliability in predicting selection outcomes. We further analyzed the role of skill distance in the selection probability among applicants competing for the same vacancy using a conditional logit model. The model incorporated vacancy-specific fixed effects (α_v) to concentrate on within-vacancy variation. The results indicated that being in a higher skill distance quintile decreased the probability of selection by 35% ($\beta = -0.4293$, $p < 0.01$). The initial exploratory analysis demonstrates the skill distance metric's capacity to predict selection outcomes in a majority of cases. Building on this, the more formal conditional logit model enables a rigorous examination of the role of skill distance in differentiating applicants within the same vacancy. Together, these findings highlight the relevance and importance of the skill distance measure in the selection process.

$$\text{logit}(P(S_{v,c} = 1)) = \alpha_v + \beta Q_k(d_{j_c, j_v}) \tag{3}$$

TABLE 5. Conditional Logit Model Results for Within-Vacancy Analysis

Variable	**Coefficient**	**Std. Error**
Quintile	-0.4293***	0.122
Observations	261	
Groups	110	

*** p<0.01

These results underscore the relevance and importance of the skill distance measure in the selection process. Crucially, the ability of skill distance to differentiate between candidates applying for the same position highlights the fine-grained informativeness embedded within job postings regarding the specific requirements of that role.

5.2.2. Within Applicant Selection

Now, we analyze the ability of our skill distance measure to account for selection decisions within an applicant's multiple vacancies. We again employ a conditional logit model, this time incorporating applicant-specific fixed effects (α_c).

$$\text{logit}(P(S_{c,v} = 1)) = \alpha_c + \beta Q_k(d_{j_c, j_v}) \tag{4}$$

The test dataset now comprises 111 observations from 33 applicant groups, each applying to multiple vacancies. The results reveal that applicants in higher skill distance quintiles experience a 50% reduction in selection probability ($\beta = -0.695$, $p = 0.017$). This

demonstrates that the measured skill distance effectively accounts for selection decisions within applicants.

TABLE 6. Conditional Logit Model Results for Across-Applicant Analysis

Variable	Coefficient	Std. Error
Quintile	-0.695**	0.290
Observations	111	
Groups	33	

** p=0.017

These patterns demonstrate that the skill distance from an applicant's current job provides valuable information for predicting selection to new internal vacancies. While selection typically depends on an applicant's skill-set and preparedness for the new position's demands, our findings reveal that a distance measure constructed from job postings and past selection decisions correlates strongly with observed selection patterns. This correlation opens an analytical window into the firm, linking skills to internal mobility. Before exploring these links we examine how important is the skill distance measure relative to other observed features in predicting selection to vacancies.

5.3. Other Observed Features

While our paper primarily focuses on leveraging information from job postings and selection decisions, we recognize the importance of evaluating other predictors of internal candidate selection. To enhance our understanding of selection patterns, we examine additional features available in the firm's information system. We compare the performance of our skill distance metric against other observed applicant characteristics, including tenure at the organization, total work experience, job category, vacancy (job sought) category, and proficiency rating. Table 7 presents summary statistics for these features.

TABLE 7. Summary Statistics – Other Observed Features

Notation	Description	Unit	25th Percentile	75th Percentile
t_c	Tenure at c	days	661	1346
T_c	Total experience	days	1927	3009
l_c	Job category	4 levels	2	3
l_v	Vacancy category	4 levels	2	3
p_c	Proficiency rating	4 levels	2	2

The data shows notable differences among applicants. Tenure ranges from about 2 years to 3.5 years, while total experience spans from roughly 5 years to 8 years at the 25th

and 75th percentiles. Job and vacancy categories both range from level 2 to 3, indicating movement across organizational tiers. Proficiency ratings, however, remain consistent at level 2 for at the 25th and 75th percentiles. These observed features, as shown in Table 7, display clear variation across the applicant pool. Their potential to explain selection patterns merits further exploration, complementing our primary skill distance metric.

To assess the relative importance of the skill distance measure we employ Random Forest classifiers. The well known ensemble learning algorithm captures complex interactions between variables. We train two Random Forest models on our training dataset: one incorporates all features including skill distance, while the other excludes skill distance. This method isolates and quantifies the predictive power of the skill distance metric relative to other observed features. We evaluate these models on our test set and will again employ the AUC metric. Table 8 presents the AUC scores for both models on the test set.

TABLE 8. Model Comparison: Random Forest Classifiers with and without Skill Distance

Model	**AUC Score**
Full Model (with Skill Distance)	0.600
Model without Skill Distance	0.500

The full model, which includes skill distance, achieves an AUC of 0.600, while the model without skill distance yields an AUC of 0.500 - equivalent to a classifier with no predictive ability. This dramatic difference in performance powerfully illustrates the substantial informative content captured by the skill distance measure. Despite the clear variation in other observed features such as tenure, total experience, job categories, and proficiency ratings, skill distance alone accounts for nearly all predictable variation in selection outcomes.

This finding is particularly noteworthy given that we derived our measure from job descriptions, which are generally less sensitive than other HR data. While firms often possess additional potentially predictive features, these may be subject to privacy concerns or legal restrictions. Our approach demonstrates that valuable insights can be derived from less sensitive data sources, advancing the field of people analytics while maintaining ethical data practices. Moreover, the granularity of our skill distance measure appears to capture nuanced differences between roles that categorical variables like performance ratings may miss, especially when applicants cluster in just one or two groups.

6. Decoding Internal Mobility Patterns

Having developed a measure of skill distance between positions based on job postings and validating it, we explore how this can be applied to understand mobility patterns within the firm. The measure quantifies the distance between skills and tasks enumerated in job

postings and assesses selection chances of candidates for new vacancies. In this section, we apply the skill distance metric to study how internal candidates' application behaviors vary based on their prospects at the time of application. By capturing predictable variation in selection, this measure enables a more nuanced examination of internal transitions and application patterns. The firm can now pose more sophisticated queries about internal mobility. For instance, it can now assess for any employee the proportion of new vacancies where their skill distance suggests a high probability of selection based on historical patterns. Central to our empirical investigation is the average skill distance measure, which assesses an employee's prospects for movement within the firm. From an internal applicant's perspective, if many vacancies share skills similar to their current job, their prospects are good; if most vacancies are distant, their prospects are weaker. We first develop this average skill similarity measure, then proceed to a detailed empirical analysis of application patterns and internal mobility. This analysis opens up new possibilities for understanding and supporting workforce development.

6.1. Average Skill Distance

To formalize the concept of average skill distance introduced above, we define the following metric:

$$\bar{d}_{j_c,t} = \frac{1}{|J_t|} \sum_{j_v \in J_t} d(j_c, j_v) \tag{5}$$

where $\bar{d}_{j_c,t}$ represents the average distance to job postings at time t, J_t is the set of job postings in time window t, and $d(j_c, j_v)$ is the distance between the current job j_c and a vacancy j_v. In our empirical analysis, we compute this measure for each internal applicant, considering all vacancies in a 30-day window prior to their observed application. This approach allows us to quantify an employee's prospects at the time of application, as discussed in the preceding paragraph. The distribution of $\bar{d}_{j_c,t}$, illustrated in Figure 5, reveals notable heterogeneity in applicants' average skill distances, providing a foundation for our subsequent analysis of application patterns and internal mobility.

6.2. Empirical Analysis

Our focus is to develop an understanding of how skill match to vacancies, which we can now measure, affects application patterns. Applying is a key step in the internal mobility process and the stage of the search process that can be observed. We will look at two aspects of the application process: its directness and intensity. Directness corresponds to whether the applicant is applying to positions where they have a higher probability of securing a job,

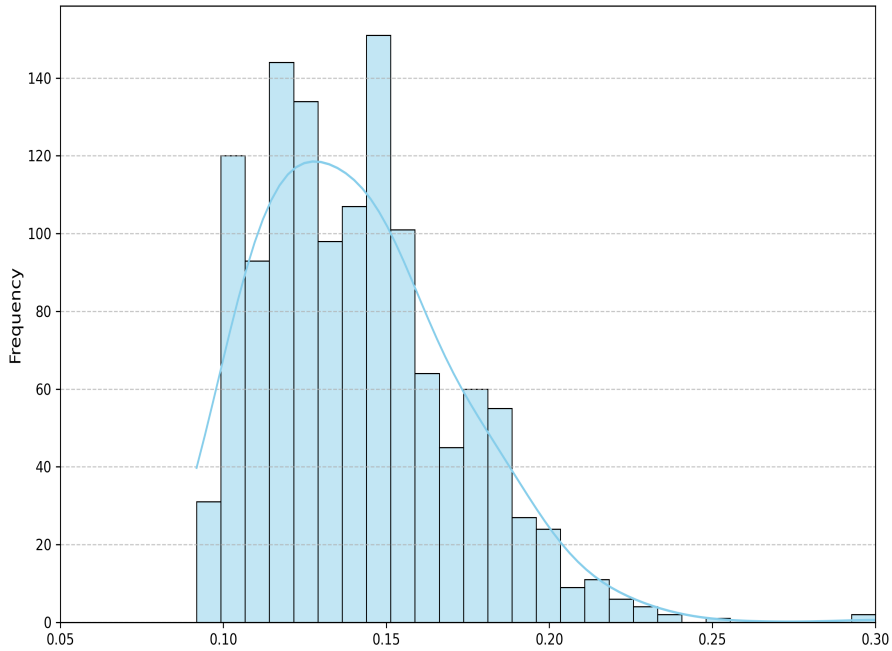


FIGURE 5. Distribution of Average Skill Distance ($\bar{d}_{j_c,t}$)

measured using the skill distance. Intensity captures how many positions the applicant applies to. Both directedness and intensity improve the probability of moving to a new position within the firm. From the firm's perspective, understanding how application behavior predictably varies enables them to tweak processes to support employee mobility within the organization. Our empirical investigation focuses on three key patterns in application behavior as they relate to the average skill distance measure, and we present our findings for each of these patterns in the following sub-sections.

6.2.1. Average Skill Distance and Application Intensity

The relationship between an applicant's average skill distance to available vacancies and the positions they choose to apply for provides crucial insights into internal mobility patterns. We investigate whether a higher average skill distance is associated with applying to closer or farther positions in terms of skill requirements. This analysis helps us understand how employees navigate their career paths within the firm when faced with varying degrees of skill alignment with available opportunities.

Applicants with a high average skill distance to vacancies face a potentially smaller set of opportunities that closely align with their existing skillset. This situation presents a dilemma: do these employees constrain their applications to fewer, more closely aligned positions, or do they make more exploratory applications to positions that require signifi-

cant skill development? Conversely, applicants whose skillsets are closer to the available vacancies might be less inclined to seek positions that require learning new technologies or skills.

To empirically examine how skill remoteness affects application behavior, we estimate the following model:

$$d_{j_v,j_c} = \beta_0 + \beta_1 \bar{d}_{j_c,t} + \epsilon \tag{6}$$

d_{j_v,j_c} is the standardized skill distance between the current job j_c and the job applied for j_v, and $\bar{d}_{j_c,t}$ is the standardized average skill distance to vacancies.

TABLE 9. Average Skill Distance and Application Intensity

Variable	**Coefficient**	**Std. Error**
Constant	2.945e-16	0.022
$\bar{d}_{j_c,t}$ (std)	0.6093***	0.025
R-squared	0.371	

*** p<0.01, ** p<0.05, * p<0.1

The results presented in Table 9 reveal a strong positive association between $\bar{d}_{j_c,t}$ and d_{j_v,j_c}. Specifically, we find that a one standard deviation increase in average skill distance to vacancies is associated with a 0.6093 standard deviation increase in the distance to the position applied for. This substantial and statistically significant effect suggests that applicants with higher skill remoteness are more likely to apply to positions that are farther away in terms of skillset. The application patterns here confirm that when faced with a higher average distance to available vacancies, applicants make more exploratory choices. Rather than constraining their applications to a narrower set of closely aligned positions, employees appear to broaden their search, potentially seeking opportunities that require attaining new skills and capitalize on the firm's knowledge of the applicant's performance in the current role which cannot be easily conveyed when applying to a position outside the firm. We already know that applying to a more distant position reduce selection chances, but will examine this more closely.

6.2.2. Average Skill Distance and Selection Probability

We've already seen that skill distance is inversely linked to selection probability, and that applicants facing higher average skill distance to the set of vacancies tend to apply to more distant positions. Connecting these dots, we expect that applicants dealing with higher average skill distances to vacancies are less likely to get selected. We examine this here:

$$\text{logit}(P(S_{v,c} = 1)) = \beta_0 + \beta_1 \times \mathbb{I}[\text{above}]_{j_v, j_c} \tag{7}$$

$S_{v,c} = 1$ is a binary variable indicating whether an applicant with current job j_c is selected for the job vacancy they applied to j_v. The variable $\mathbb{I}[\text{above}]_{j_v, j_c}$ is an indicator function that equals 1 if the skill distance between j_c and j_s is above the median, and 0 otherwise. Table 10 presents the results of this logistic regression:

TABLE 10. Logistic Regression Results: Impact of Skill Distance on Selection Probability

Variable	**Coefficient**	**Std. Error**
Constant	-0.0061	0.078
Above Median	-0.4083***	0.112
Log-Likelihood	-895.62	

*** p<0.01, ** p<0.05, * p<0.1

The results align with our expectations. Applicants with above-median skill distance to the vacancies have a 33.5% lower likelihood of being selected into their position. With this we clearly see that the avg. skill distance to the vacancies affect the applicant's ability to secure an internal position with the same effort. The natural follow up and closely aligned question is of the intensity of the application process which we take up next.

6.2.3. Average Skill Distance and Application Intensity

We will now examine whether the application intensity—the number of vacancies an applicant applies to—varies with their average skill distance to the vacancy set. We already know that applicants faced with a high $d_{c,t}$ apply to more distant positions and are less likely to be selected. Our goal here is to inquire whether they also apply to more positions. For this analysis, we slightly modify our approach to constructing application windows. We define each window as a sequence of applications where the gap between any two consecutive applications does not exceed 30 days. This allows for windows that may span longer than 30 days, provided no two consecutive applications within the window are more than 30 days apart. We estimate the following regression:

$$A_{i,t} = \beta_0 + \beta_1 \mathbb{I}[\text{above}]_{i,t} + \epsilon \tag{8}$$

Here, $A_{i,t}$ is the number of applications person i submits in time window t, and $\mathbb{I}[\text{above}]_{i,t}$ indicates if their skill remoteness is above the median.

Table 11 shows that applicants with above-median skill remoteness apply to about 1.5

TABLE 11. Average Skill Distance and Application Intensity

Variable	Coefficient	Std. Error
Constant	1.8210***	0.137
Above Median	0.5334***	0.205
Observations	637	
R-squared	0.011	

*** p<0.01, ** p<0.05, * p<0.1

times more positions on average. Faced with diminished prospects for their skillset within firm opportunities, they adopt a more exploratory approach when seeking a new position. Putting all our findings together, we get a clearer picture of how skill remoteness to the new vacancies shape internal job applications within the firm. Applicants with higher skill distance to the vacancies tend to apply to jobs that are more distant in terms of skills, face lower chances of being selected, and submit more applications on average. In other words our skill distance measure has allowed to bring out distinct difference in the application patterns to new vacancies. With this understanding, the natural follow up question is how we can fine-tune internal mobility processes in the firm? which we take up next.

6.3. From Mobility Patterns to Practice

In a technology-focused firm, understanding and supporting internal mobility requires a clear view of how employees' current skills align with emerging requirements. Our analysis of job posting content reveals striking patterns in how employees navigate opportunities based on their skill-fit - patterns that would be difficult to discern without validated measures of skill distance. Employees whose skills closely match emerging opportunities often adopt a passive approach despite higher chances of success, while those facing larger skill distances pursue positions more actively but with lower success rates.

These empirically-documented patterns point to a clear quantitative framework for workforce development. When employees' skills align well with vacancies but application rates are low, the primary barrier appears to be search friction rather than skill gaps. As Minni (2023) note, managerial initiative in surfacing opportunities can be crucial in large firms. Conversely, when employees face high average skill distances to vacancies and apply actively but unsuccessfully, the evidence suggests skill development needs rather than search frictions limit mobility. This ability to distinguish between friction and skill gap challenges using posting content enables firms to develop targeted interventions.

While our analysis illuminates these distinct patterns, moving from measurement to practice requires careful consideration. For employees with low average skill distance who do apply, the data supports proactive opportunity identification. However, for those

not observed applying, questions remain about how best to surface relevant opportunities. Similarly, the application patterns of employees facing larger skill distances provide valuable signals about career directions they see as feasible or desirable - opening new possibilities for studying how employees navigate skill requirements in modern careers. This suggests rich possibilities for building systematic, evidence-based approaches to workforce development by leveraging the skill information in job postings. Rather than relying on generic mobility programs, firms can use posting content to quantify where reducing search frictions versus supporting skill development would be most valuable, while gaining deeper insight into how employees approach career development in environments of evolving skill demands.

7. Conclusion and Discussion

Our major contribution lies in providing a rigorous empirical test and validation of the informativeness of job posting content, particularly for technical roles in the modern digital economy. Analyzing internal mobility within a major corporation's technology division, we demonstrate that skill descriptions in job postings possess significant predictive power for selection decisions. Crucially, our methodology leverages information from both the applicant's current role and the sought-after position, revealing that the calculated skill distance strongly discriminates between candidates' fit to new opportunities. Indeed, the probability of selection is 84% higher when comparing applications with the closest versus furthest skill distance. This finding, derived from observing both successful and unsuccessful applications, provides robust validation of job postings as a meaningful signal of actual skill requirements, at least within the context of technology-related roles in large organizations.

Furthermore, our work illuminates how this validated informativeness of job postings unlocks key insights into internal mobility and human capital analytics within the firm. By establishing that posting content reliably reflects skill demands, we demonstrate its value as a primary data source for understanding employee transitions and career pathways. The measured skill distances between current and potential positions serve as a powerful lens for identifying areas where search frictions impede well-matched candidates and where genuine skill gaps exist relative to emerging opportunities. The observed heterogeneity in application intensity, directly linked to skill alignment with current vacancies, further underscores how employees actively navigate the internal market, balancing immediate fit with aspirations for skill development.

The demonstrated informativeness of job requirements validates and strengthens the broader trend in systematic collection and analysis of job postings data, both within and beyond organizational boundaries. This trend, fueled by specialized firms and academic researchers alike, seeks to understand the dynamics of the labor market and provide

strategic intelligence to firms. Our findings underscore that firms' internal repositories of job postings, alongside increasingly available external data on skill requirements across their competitive landscape, represent a valuable, and now empirically validated, resource. This access enables a more nuanced understanding of evolving skill demands, informing strategic decisions related to talent acquisition, training, and workforce planning.

Our results suggest firms can move beyond conventional HR metrics to develop validated measures of worker-job fit, leveraging readily available posting data to guide specific interventions and foster more effective internal mobility. This capability to link rich posting content to tangible selection outcomes firmly positions job postings as a powerful, and often underutilized, source of intelligence for navigating the complexities of human capital management in modern, skill-driven labor markets.

References

Acemoglu, Daron, David Autor, Jonathon Hazell, and Pascual Restrepo. 2024. "Artificial Intelligence and Jobs: Evidence from Online Vacancies." *Journal of Labor Economics* 40 (S1).

Azar, José, Ioana Marinescu, Marshall Steinbaum, and Bledi Taska. 2020. "Concentration in US labor markets: Evidence from online vacancy data." *Labour Economics* 66: 101886.

Baker, George, Marian Gibbs, and Bengt Holmstrom. 1994a. "The internal economics of the firm: Evidence from personnel data." *The Quarterly Journal of Economics* 109 (4): 881–919.

Baker, George, Michael Gibbs, and Bengt Holmstrom. 1994b. "The Wage Policy of a Firm." *The Quarterly Journal of Economics* 109 (4): 921–955.

Bana, Sara. 2021. "Using Language Model to Understand Wage Premia." Working Paper.

Bian, Shuqing, Xu Chen, Wayne Xin Zhao, Kun Zhou, Yupeng Hou, Yang Song, Tao Zhang, and Ji-Rong Wen. 2020. "Learning to match jobs with resumes from sparse interaction data using multi-view co-teaching network." In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*,: 65–74.

Bidwell, Matthew. 2011. "Paying more to get less: The effects of external hiring versus internal mobility." *Administrative Science Quarterly* 56 (3): 369–407.

Bidwell, Matthew, and J. R. Keller. 2024. "Stepping Sideways to Step Up: Lateral Mobility and Career Advancement Inside Organizations." *Management Science* 0 (0).

Braxton, J Carter, and Bledi Taska. 2023. "Technological change and the consequences of job loss." *American Economic Review* 113 (2): 279–316.

Cowgill, Bo, Jonathan M.V. Davis, B. Pablo Montagnes, and Patryk Perkowski. 2024. "Stable Matching on the Job? Theory and Evidence on Internal Talent Markets." *Management Science* 0 (0).

De Vos, Simon, Johannes De Smedt, Marijke Verbruggen, and Wouter Verbeke. 2024. "Data-driven internal mobility: Similarity regularization gets the job done." *Knowledge-Based Systems* 295: 111824.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*,: 4171–4186.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2024. "GPTs are GPTs: Labor market impact potential of LLMs." *Science* 384 (6702): 1306–1308.

Forsythe, Eliza, Lisa B Kahn, Fabian Lange, and David Wiczer. 2020. "Labor demand in the time of COVID-19: Evidence from vacancy postings and UI claims." *Journal of public economics* 189: 104238.

Fuller, Joseph, Manjari Raman, Eva Sage-Gavin, and Kristen Hines. 2022. "Skills-Based Hiring Is on the Rise." *Harvard Business Review*. https://hbr.org/2022/02/skills-based-hiring-is-on-the-rise. Accessed on 15 November 2024.

George, Nikhil, Christophe Combemale, Ramayya Krishnan, and Rahul Telang. 2024. "Linking Skills to Mobility: Insights for Workforce Dvelopment.". https://ssrn.com/abstract=4857198. Working Paper.

Giabelli, Anna, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Andrea Seveso. 2021. "Skills2job: A recommender system that encodes job offer embeddings on graph databases." *Applied Soft Computing* 101: 107049.

Goldfarb, Avi, Bledi Taska, and Florenta Teodoridis. 2020. "Artificial intelligence in health care? Evidence from online job postings." In *AEA Papers and Proceedings*, vol. 110: 400–404, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Heap, Bradford, Alfred Krzywicki, Wayne Wobcke, Mike Bain, and Paul Compton. 2014. "Combining career progression and profile matching in a job recommender system." In *PRICAI 2014: Trends in Artificial Intelligence: 13th Pacific Rim International Conference on Artificial Intelligence*, Proceedings 13: 396–408. Gold Coast, QLD, Australia: Springer.

Hershbein, Brad, and Lisa B Kahn. 2018. "Do recessions accelerate routine-biased technological change? Evidence from vacancy postings." *American Economic Review* 108 (7): 1737–1772.

Huitfeldt, Ingrid, Andreas R Kostøl, Jan Nimczik, and Andrea Weber. 2023. "Internal labor markets: A worker flow approach." *Journal of Econometrics* 233 (2): 661–688.

Kokkodis, Marios, and Panagiotis G Ipeirotis. 2021. "Demand-aware career path recommendations: A reinforcement learning approach." *Management science* 67 (7): 4362–4383.

Kokkodis, Marios, and Panagiotis G. Ipeirotis. 2023. "The Good, the Bad, and the Unhirable: Recommending Job Applicants in Online Labor Markets." *Management Science*. Published online in Articles in Advance.

Lei, Xiaoxia, Yixing Chen, and Ananya Sen. 2024. "The Value of External Data Capabilities in the Search Market: Evidence from a Field Experiment." *Available at SSRN*. https://ssrn.com/abstract=4452804.

Liu, Mengshu, Jingya Wang, Kareem Abdelfatah, and Mohammed Korayem. 2019. "Tripartite vector representations for better job recommendation." *arXiv preprint arXiv:1907.12379*.

Maurya, Abhinav, and Rahul Telang. 2017. "Bayesian multi-view models for member-job matching and personalized skill recommendations." In *2017 IEEE International Conference on Big Data (Big Data)*,: 1193–1202, IEEE.

McKinsey Global Institute. 2020. "The future of work in Europe.". https://www.mckinsey.com/featured-insights/future-of-work/the-future-of-work-in-europe. Accessed on 15 November 2024.

Minni, Virginia. 2023. "Making the Invisible Hand Visible: Managers and the Allocation of Workers to Jobs." Working Paper.

Qin, Chuan, Hengshu Zhu, Tong Xu, Chen Zhu, Chao Ma, Enhong Chen, and Hui Xiong. 2020. "An enhanced neural network approach to person-job fit in talent recruitment." *ACM Transactions on Information Systems (TOIS)* 38 (2): 1–33.

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. "Mitigating bias in algorithmic hiring: Evaluating claims and practices." In *Proceedings of the 2020 conference on fairness, accountability, and transparency*,: 469–481.

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*,: Association for Computational Linguistics. http://arxiv.org/abs/1908.10084.

Shaha, T Al-Otaibi, and Ykhlef Mourad. 2012. "A survey of job recommender systems." *International Journal of Physical Sciences* 7 (29): 5127–5142.

Siting, Zheng, Hong Wenxing, Zhang Ning, and Yang Fan. 2012. "Job recommender systems: A survey." In *2012 7th International Conference on Computer Science & Education (ICCSE)*,: 920–924.

Tambe, Prasanna, Peter Cappelli, and Valery Yakubovich. 2019. "Artificial Intelligence in Human Resources Management: Challenges and a Path Forward." *California Management Review* 61 (4): 15–42.

Tambe, Prasanna, Xuan Ye, and Peter Cappelli. 2020. "Paying to program? Engineering brand and high-tech wages." *Management Science* 66 (7): 3010–3028.

World Economic Forum. 2020. "The Future of Jobs Report 2020.". https://www.weforum.org/reports/the-future-of-jobs-report-2020. Accessed on 15 November 2024.

Xing, Eric P., Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. 2002. "Distance Metric Learning, with Application to Clustering with Side-information." In *Advances in Neural Information Processing Systems*, vol. 15.

Zhu, Chen, Hengshu Zhu, Hui Xiong, Chao Ma, Fang Xie, Pengliang Ding, and Pan Li. 2018. "Person-job fit: Adapting the right talent for the right job with joint representation learning." *ACM Transactions on Management Information Systems (TMIS)* 9 (3): 1–17.