Many metrics used in supply chain performance evaluation have been designed to measure operational performance, evaluate improved effectiveness, and examine strategic alignment of the whole supply chain management [3]. Individual measures of supply chain performance have usually been classified into four categories: quality [3,22], time [4,22], cost [4,12], flexibility [1,3]. Furthermore, they have also been grouped by quality and quantity, cost and noncost, strategic/operational/tactical focus, and supply chain processes [12,22]. However, since many measurement systems lacked strategy alignment, a balanced approach and systemic thinking [3,8], they had difficulty in systematically identifying the most appropriate metrics. To address this problem, some researchers have used Balanced Scorecard (BSC) and Activity Based Costing (ABC) methods to evaluate supply chain performance [16]. Other researchers have also proposed similar balanced frameworks, such as Performance Measurement Matrix, results-determinants framework, performance pyramid, etc. [19]. From a process perspective, the Supply Chain Operations Reference (SCOR) model has been developed to facilitate construction of a systematic supply chain performance measurement and improvement tool; it has often been recognized as a systematic approach for identifying, evaluating and monitoring supply chain performance [17]. In the SCOR model, a balanced performance measurement system at multiple levels, covering five core supply chain processes (i.e. Plan, Source, Make, Deliver, and Return), was developed [4,13,17]. However, measurement models (i.e. extended BSC and SCOR) for supply chain performance evaluation have their limitations. First, there are too many individual measures being used in the supply chain context. For example, Shepherd and Gunter [22] have summarized 39, 22, 35, 28 and 8 single supply chain performance indicators related to cost, time, quality (or reliability), flexibility and innovativeness respectively. Though these measures offer valuable information for decision-making, selecting and trading off so many measures to obtain effective and crucial improvement strategies is a difficult task for different supply chain participants. Second, these models do not provide definite cause–effect relationships among numerous (and hierarchical) individual KPIs. Although existing models (e.g., BSC) do illustrate the cause-and-effect relationships between different goal-related KPIs [18], they are inadequate for quantitative analysis of the intricate intertwined relationships. Traditional BSC and SCOR models generally assume that KPIs are uncoupled. These approaches could describe business operations well, and serve as a good communication tool, but they are not effective in improving overall performance by accomplishing the critical KPIs. For many managers, who try to allocate resources efficiently and achieve multiple supply chain performance goals, this becomes the bottleneck.

供应链绩效评估中使用的许多指标旨在衡量运营绩效、评估提高的效率以及检查整个供应链管理的战略路线 [3]。供应链绩效的单独衡量标准通常分为四类：质量[3，22]、时间[4，22]、成本[4，12]、灵活性[1，3]。此外，它们还按质量和数量、成本和非成本、战略/业务/战术重点和供应链流程[12，22]进行分组。然而，由于许多测量系统缺乏战略一致性、平衡方法和系统思维[3，8]，它们难以系统地确定最合适的指标。为了解决这个问题，一些研究人员使用平衡记分卡 （BSC） 和基于活动的成本计算 （ABC） 方法来评估供应链绩效 [16]。其他研究人员也提出了类似的平衡框架，如性能测量矩阵、结果决定因素框架、性能金字塔等[19]。从流程角度而言，开发了供应链操作参考 （SCOR） 模型，以促进系统供应链性能测量和改进工具的构建：它经常被公认为识别、评估和监测供应链绩效的系统方法[17]。在 SCOR 模型中，开发了多个级别的平衡性能测量系统，涵盖五个核心供应链流程（即计划、来源、制造、交付和返回）[4，13，17]。但是，供应链绩效评估的测量模型（即扩展 BSC 和 SCOR）有其局限性。首先，在供应链环境中使用了太多的单独措施。例如，Shepherd 和 Gunter [22] 分别总结了与成本、时间、质量（或可靠性）、灵活性和创新性相关的 39、22、35、28 和 8 个单一供应链绩效指标。虽然这些措施为决策提供了宝贵的信息，但选择和权衡这么多措施以获得有效和关键的改进战略，对于不同的供应链参与者来说，是一项艰巨的任务。其次，这些模型没有提供众多（和等级）个体 KP 之间的明确因果关系。虽然现有模型（例如BSC）确实说明了不同目标相关 KPI 之间的因果关系 [18]，但它们不足以定量分析错综复杂的相互交织的关系。传统的 BSC 和 SCOR 模型通常假设 KP 是脱钩的。这些方法可以很好地描述业务运营，并作为一个很好的沟通工具，但它们不能有效地通过完成关键的 KPI 来改善整体绩效。对于许多试图高效配置资源并实现多个供应链绩效目标的管理者来说，这成为瓶颈。

The traditional logistics performance measures include ‘hard’and ‘soft’metrics. The first one treats quantitative measures such as order cycle time, fill rates and costs; the second deals with qualitative measures like manager’s perceptions of customer satisfaction and loyalty (Chow, Heaver, and Henriksson 1994; Fugate, Mentzer, and Stank 2010). The ‘hard’ metrics are easily computable with some simple mathematical expressions while the soft metrics require more sophisticated tools of measurement (e.g. Regression analysis, fuzzy logic, Data Envelopment Analysis, etc.). In this paper, we will refer to the ‘hard’ metrics as direct indicators and the soft ones as indirect indicators.

传统的物流绩效衡量标准包括"硬"和"软"指标。第一种是定量衡量标准，如订单周期时间、填充率和成本：第二种是定性衡量指标，如经理对客户满意度和忠诚度的看法（周、希弗和亨里克森1994年;富盖特、门策和斯坦克 2010）。"硬"指标很容易用一些简单的数学表达式进行计算，而软指标则需要更复杂的测量工具（如回归分析、模糊逻辑、数据包围分析等）。本文将"硬"指标称为直接指标，将软指标称为间接指标。

All direct indicators extracted from papers are classified according to four performance evaluation dimensions, commonly used in industries. Each indicator can be only assigned to a single dimension at time. These are: time (Chan and Qi 2003; Frazelle 2001; Gallmann and Belvedere 2011; Gunasekaran and Kobu 2007; Mentzer and Konrad 1991; Neely, Gregory, and Platts 1995; Spencer 1993), quality (Frazelle 2001; Gallmann and Belvedere 2011; Neely, Gregory, and Platts 1995; Stainer 1997), cost (Beamon 1999; Cai et al. 2009; Chan and Qi 2003; Keebler and Plank 2009; Mentzer and Konrad 1991; Neely, Gregory, and Platts 1995) and productivity (Chan and Qi 2003; Frazelle 2001; Gallmann and Belvedere 2011; Keebler and Plank 2009; Stainer 1997). We note that some works prefer to use flexibility instead of productivity as the fourth dimension (Beamon 1999; Gunasekaran and Kobu 2007; Neely, Gregory, and Platts 1995; Stainer 1997), defining it as the ‘ability to respond to a changing environment’ (Beamon 1999). However, flexibility may be intangible and difficult to measure (Gunasekaran and Kobu 2007) in some cases. We present in Section 4.2 that flexibility is preferably measured indirectly rather than directly. Consequently, in this section, productivity will be used as a dimension for direct warehouse performance indicators.

从论文中提取的所有直接指标均按行业常用的四个绩效评估维度进行分类。每个指示器只能及时分配到单个维度。这些是：时间（陈和齐2003年：弗雷泽2001年;加尔曼和贝尔维迪尔 2011;古纳塞卡兰和科布2007年;门策和康拉德1991年;尼利、格雷戈里和普氏1995年;斯宾塞1993年），质量（弗雷泽2001年;加尔曼和贝尔维迪尔 2011;尼利、格雷戈里和普氏1995年;染色器 1997）， 成本 （比蒙 1999;蔡等人2009年：陈和齐2003年：基布勒和普朗克2009年;门策和康拉德1991年;尼利、格雷戈里和普氏1995年）和生产力（陈和齐2003年：弗雷泽2001年;加尔曼和贝尔维迪尔 2011;基布勒和普朗克2009年;斯坦纳1997年）。我们注意到，有些作品倾向于使用灵活性而不是生产力作为第四维度（Beamon 1999：古纳塞卡兰和科布2007年;尼利、格雷戈里和普氏1995年;1997年，施泰纳将其定义为"应对不断变化的环境的能力"（Beamon 1999）。然而，在某些情况下，灵活性可能是无形的，难以衡量（古纳塞卡兰和科布2007年）。我们在第4.2节中指出，灵活性最好是间接衡量的，而不是直接衡量的。因此，在本节中，生产率将用作直接仓库性能指标的一个维度。

The following procedure is used for the classification. Initially, all the direct indicators found in the selected papers are listed. Once the list is completed, two types of aggregations are made: (i) similar indicators are regrouped; (ii) very specific metrics are included in more generic ones. One example of this second group is the work by Manikas and Terry (2010) mentioning the indicator ‘time of quality control in receiving’. This can be considered as a portion of the ‘receiving operation time’. Therefore, we include this indicator together with the class of indicators called the ‘receiving time’. Finally, the indicators are organised according to what they measure (time, quality, cost or productivity). We note that, for the sake of uniformity throughout this literature review, the classifications presented in this article are based on our interpretation, instead of the original category proposed by the selected papers. For example, Banaszewska et al. (2012) consider the ‘number of consignment processed per warehouse employee’ as a productivity indicator. Indeed, the measure is a productivity indicator. In this review, we propose a sub-category, called the labour productivity and Banaszewska et al. (2012) appears in this category (see Table A2). Another example is the article by Saetta et al. (2012), where the authors measure the customer satisfaction as ‘the percentage of orders on time’ and we classify the article under a broader indicator which is the ‘on time delivery’ (see Table A2). The results from this analysis are given in Table 5, which demonstrates all performance indicators comprised in a dimension. The third column of Table 5 shows the number of publications containing each specific indicator. The discussions about the results of Table 5 are presented in the next sections.

分类使用以下程序。最初，在选定的论文中发现的所有直接指标都列出。列表完成后，将进行两种类型的聚合：（一） 重新组合类似的指标：（二） 非常具体的指标包含在更通用的指标中。第二组的一个例子是马尼卡斯和特里（2010年）的工作，其中提到指标"接收时的质量控制时间"。这可以被视为"接收操作时间"的一部分。因此，我们将此指标与称为"接收时间"的指标类别一起包括在一起。最后，这些指标根据它们衡量的内容（时间、质量、成本或生产率）进行组织。我们注意到，为了在整个文献审查过程中的统一性，本文中提出的分类基于我们的解释，而不是选定论文提出的原始类别。例如，Banaszewska等人（2012年）将"每个仓库员工处理的货物数量"视为生产力指标。事实上，该指标是一个生产力指标。在本次审查中，我们提出了一个子类别，称为劳动生产率，Banaszewska等人（2012年）出现在这一类别中（见表A2）。另一个例子是 Saetta 等人的文章（2012 年），作者将客户满意度衡量为"按时订单的百分比"，我们将文章分类为"按时交货"的更广泛指标（见表 A2）。此分析的结果在表 5 中给出，该表演示了以一个维度组成的所有绩效指标。表 5 的第三列显示了包含每个特定指标的出版物数量。关于表 5 结果的讨论在下一节中提出。

To address these issues, the next sections present indicator definitions in Tables 6–9. The definitions come initially from paper database. In the case where several definitions are found for a given indicator, we show both in the table (e.g. order lead time in Table 6). In the case where no definitions are given, we look for these definitions in Warehouse Education and Research Council (WERC 2008) database (available at www.werc.org) (e.g. picking accuracy in Table 7). Finally, when the definition is neither in the papers nor in WERC database, we defined the indicators based on the best common sense that we could infer from the literature and we refer to the authors that have used these measures in their papers. For example, in Table 7, cargo damage rate is used by Lu and Yang (2010) but the indicator is not explicitly defined by the authors. Thus, for the purpose of this work, we provide an indicator definition according to the described procedure, identified by ∗ symbol in the tables.

为了解决这些问题，下一节在表 6–9 中提出了指标定义。这些定义最初来自纸质数据库。在为给定指标找到几个定义的情况下，我们同时在表中显示（例如表 6 中的订单准备时间）。在没有给出定义的情况下，我们会在仓库教育和研究理事会 （WERC 2008） 数据库中查找这些定义（可在 www.werc.org 提供）（例如，在表 7 中挑选准确性）。最后，当定义既不出现在论文中，也不在 WERC 数据库中时，我们根据我们可以从文献中推断出的最佳常识定义了指标，我们指的是在其论文中使用这些措施的作者。例如，在表 7 中，货物损坏率由陆和杨（2010 年）使用，但作者没有明确定义该指标。因此，为了这项工作的目的，我们根据所述程序提供指标定义，该程序由表中的∗符号识别。

传统的仓库绩效衡量标准包括"硬"和"软"指标。"硬"指标又可称为直接指标（如时间，花费等），这些指标可以通过一些简单的数学表达式进行计算。同时，’‘软‘’指标称为间接指标，经理对客户满意度和忠诚度的看法，则需要更复杂的测量工具（如回归分析、模糊逻辑、数据包围分析等）。

Traditional warehouse performance measurement standards include hard and soft indicators. Hard indicators can also be called direct indicators (such as time, cost, etc.). Some simple mathematical expressions can be adopted to calculate these indicators. Meanwhile, soft indicators are indirect indicators, such as managers’ perceptions of customer satisfaction and loyalty. These indicators require more complex measurement tools (e.g., Regression analysis, fuzzy logic, Data Envelopment Analysis, etc.).

许多研究将所有直接指标按照时间，质量，成本，灵活性四个绩效评估维度进行分类

According to this theory, an organization should be able to perpetuate its activities over time with respect to the environment and society and by generating profit.

Torabizadeh proposed identifying sustainable warehouse management system indicators and offering a new weighting method. The performance of a sustainable warehouse management system is a multidimensional concept based on the triple bottom line approach.

托拉比扎德建议确定可持续的仓库管理系统指标，并提供一种新的加权方法。可持续仓库管理系统的性能是基于三重底线方法的多维概念。

The conceptual mode can be divided three orders latent constructs by the triple bottom line approach and formative constructs as high-order constructs that can be supported by PLS appropriately. This conceptual performance measurement model indicates that the aggregation effects of all three aspects of sustainability result in the incorporation of sustainability in the warehouse management system. By using this type of hierarchical model, lower-order constructs are reflectively measured by their indicators, and high-order constructs are formatively measured by their relevant variables (Becker et al., 2012).

The TBL is a framework that evaluates a process from three distinct points of view of sustainability: social aspects, environmental aspects, and economic aspects. According to this theory, an organization should be able to perpetuate its activities over time with respect to the environment and society and by generating profit (Torabizadeh et al., 2020).

TBL 是一个框架，从三个不同的可持续性角度评估流程：社会方面、环境方面和经济方面。根据这一理论，一个组织应该能够随着时间的推移，通过创造利润（Torabizadeh等人，2020年）来延续其在环境和社会方面的活动。

Torabizadeh

The indicators of this cluster refer to the economic value created by the organization. In particular, they indicate the warehouse's performances that directly influence the costs and the profit of the company. Inside this group, we subcategorized the indicators into 5 separated subclusters:该集群的指标是指组织创造的经济价值。特别是，它们表明仓库的表现直接影响到公司的成本和利润。在这个组中，我们将指标细分为 5 个分离的子组：

经济指标是指直接影响到公司的成本和利润的相关KPIs,我们将其细分为5个子集

Activities, products, and services of an organization that interacts with the environment are called "aspects," which can have a negative or positive impact on the environment. Typically, aspects can include emissions to air, discharges to water, and waste generation, which in turn can generate environmental and health impacts such as global warming, water pollution, or contaminated land. Table 5 reports the KPIs that describe the warehouse system's environmental impact, i.e., the condition under it works, like temperature or humidity

与环境相互作用的组织的活动、产品和服务称为"方面"，它们可能对环境产生负面影响或积极影响。通常，方面可以包括排放到空气、排放到水和废物产生，这反过来又会产生环境和健康影响，如全球变暖、水污染或土地污染。表 5 报告了描述仓库系统环境影响的 KPI，即仓库系统下的工作条件，如温度或湿度