A Cognitive Model of Human Bias in Matching

Rakefet Ackerman[1], Avigdor Gal[1], Tomer Sagi[2], and Roee Shraga[1](✉)

[1] Technion – Israel Institute of Technology, Haifa, Israel
ackerman@ie.technion.ac.il, avigal@technion.ac.il,
shraga89@campus.technion.ac.il
[2] University of Haifa, Haifa, Israel
tsagi@is.haifa.ac.il

Abstract. The *schema matching* problem is at the basis of integrating structured and semi-structured data. Being investigated in the fields of databases, AI, semantic Web and data mining for many years, the core challenge still remains the ability to create quality matchers, automatic tools for identifying correspondences among data concepts (*e.g.*, database attributes). In this work, we investigate human matchers behavior using a new concept termed *match consistency* and introduce a novel use of cognitive models to explain human matcher performance. Using empirical evidence, we further show that human matching suffers from predictable biases when matching schemata, which prevent them from providing consistent matching.

Keywords: Schema matching · Data integration · Human-in-the-loop

1 Introduction

Schema matching is at the basis of integrating structured and semi-structured data. The schema matching task revolves around providing correspondences between concepts describing the meaning of data in various heterogeneous, distributed data sources, such as SQL and XML schemata, entity-relationship diagrams, ontology descriptions, interface definitions, and forms format [28].

Schema matching research originated in the database community [28] and has been a focus for other disciplines as well, from artificial intelligence [10,20], to semantic web [17] to data mining [18]. Schema matching research has been going on for more than 30 years now, focusing on designing high quality matchers, automatic tools for identifying correspondences among database attributes. Initial heuristic attempts (*e.g.*, COMA [11]) were followed by theoretical grounding (*e.g.*, see [5,16]).

Recently, the information explosion (a.k.a Big Data) has provided many novel sources for data and with them the need for efficient and effective integration. Crowd-sourcing has allowed pay-as-you-go frameworks for data integration (*e.g.*, [21,35]), to make flexible use of human input in the matching process.

A. C. Nayak and A. Sharma (Eds.): PRICAI 2019, LNAI 11670, pp. 632–646, 2019.
https://doi.org/10.1007/978-3-030-29908-8_50

A basic tenet of the matching process, present from its inception, is that an algorithmic matcher provides a set of definite (true or false) correspondences to be then validated by a human expert. Human validation of algorithmic results assumes the superiority of human matchers over algorithms, which may be naïve, partially because different human matchers may have different opinions and may differ in the way they match schemata [31]. The emergence of crowd-based solutions has not changed this assumption, but merely extended the validation phase to include additional individuals.

A popular contemporary trend involves developing human-level AI. We believe it is equally important to understand human's strengths and predictable biases when determining the appropriate sharing of responsibility with the machine. Hence, in this work we focus on analyzing human's performance in matching. The central new concept in this work is *match consistency*, which we use, aided by cognitive principles, to show that human behavior in matching vary along consistency dimensions, namely temporal, consensuality, and control (as defined in this work). Given a set of human matchers, we assess their abilities, much like traditional models do for algorithms. Additionally, however, human matchers have biases that we can detected and accounted for when making use of human matching.

We present theoretical analyses, using cognitive models, of human matchers strengths and biases (Sect. 3) as well as empirical results on match consistency (Sect. 4) to support our framework. Additionally, Sect. 2 presents background on matching and metacognition. We review of related work in Sect. 5 and conclude in Sect. 6.

2 Background

We next present a formal matching model (Sect. 2.1) and models for human involvement in matching (Sect. 2.2).

2.1 Schema Matching Model

Let S, S' be two schemata with attributes $\{a_1, a_2, \ldots, a_n\}$ and $\{b_1, b_2, \ldots, b_m\}$, respectively. A matching process matches S and S' by aligning their attributes using *matchers* that utilize matching cues such as attribute names, instance data, and schema structure (see surveys *e.g.*, [6] and books *e.g.*, [16]). A matcher's output is conceptualized as a similarity matrix $M(S, S')$ (M for short), having entry $m_{i,j}$ (typically a real number in $[0, 1]$) represent a degree of similarity between $a_i \in S$ and $b_j \in S'$. A *match*, denoted σ, between S and S' is a subset of M's entries.

Matching is a stepped process of applying algorithms, rules, and constraints. Matchers can be separated into *first-line matchers – 1LMs*, which are applied directly to the problem, returning a similarity matrix, and *second-line matchers – 2LMs*, which are applied to the outcome of matchers, receiving similarity matrices and returning a similarity matrix.

Table 1. A similarity matrix example

$S_1 \longrightarrow$ $\downarrow S_2$	1 cardNum	2 city	3 arrivalDay	4 checkIn Time
1 clientNum	**0.84**	0.32	0.32	0.30
2 city	0.29	**1.00**	0.33	0.30
3 checkInDay	0.34	0.33	**0.35**	**0.64**

Example 1 (Matchers). To illustrate the variety of available matchers, consider three 1LMs. Term [16] compares attribute names to identify syntactically similar attributes (*e.g.*, using edit distance and soundex). WordNet uses abbreviation expansion and tokenization methods to generate a set of related words for matching attribute names. Token Path [27] integrates node-wise similarity with structural information by comparing the syntactic similarity of full paths from root to a node.

Example 2 (Similarity Matrices). Table 1 provides an example of an outcome of a matching process between fragments of two reservation systems' schemata, one (S_1) with four attributes and the other (S_2) with three attributes, conceptualized in a similarity matrix. S_1 consists a CardNum attribute with long data-type and a city attribute, which contains some example instances (city names). Attributes may be independent of other attributes or composable, creating compound attributes. *E.g.*, ArrivalDay and CheckInTime attributes can be composed to a compound arrival day/time attribute. S_2 has clientNum, city, and checkInDay attributes.

2.2 Human Involvement in Matching Models

Human schema matching is a complex decision making process, which involves a series of interrelated tasks. Each attribute in one schema is examined to decide whether and which attributes from the other schema correspond. Humans either validate an algorithmic result or locate a candidate attribute unassisted. Human matchers may choose to rely upon superficial information such as string similarity of attribute names (*e.g.*, qty is similar to quantity) or explore additional information such as data-types, instances, and position within the schema hierarchy. The decision whether to explore additional information relies upon self-monitoring of confidence.

Most of the works in schema matching over the years assume that an algorithmic matching system provides a set of definite (true or false) correspondences to be then validated by a human expert who can provide the ultimate matching. Human validation is typically prohibitively large. High matching costs and limited expert availability spawned research into crowd sourcing usage by breaking the matching task into small-sized tasks, suitable for unskilled workers with minimal compensation (pay-as-you-go approach) [23]. McCann *et al.* proposed methods to validate algorithmic matchers, ranging from direct match validation (*e.g., does ccost match pcost?*) to constraint validation (*e.g., does bDate* $<$ 2007 *always hold?*) [26]. Zhang *et al.* provided tools for validation task selection [35], Bozovic and Vassalos used feedback to tune matcher weights [8], and Hung *et al.* suggested methods to select conflicting matches in a network of schemata [22]. Sagi and Gal proposed the *Expert Sourcing* model [29], which we follow in this work, in which knowledgeable humans examine substantial parts of the matching task or the entire task (*e.g.*, for small-medium scaled schemata).

3 Match Consistency

Historically, humans (relative) strong matching abilities put them as final decision makers, disregarding biases that affect their ability to provide accurate matches. To capture the impact of human biases on matching, we present a formal notion of a consistent matcher, and use cognitive models to explore human matcher variability.

3.1 Consistent Match Definition

Matchers are typically measured using a global matching evaluation measure, *e.g.*, precision and recall. Such a measure evaluates the similarity matrix a matcher generated against some reference matrix. Given a similarity matrix M of $n \times m$ entries and an evaluation measure E, we define matcher consistency with respect to a consistency dimension $D = \{d_1, d_2, \ldots, d_k\}$ using a dimension function $F_D : \{m_{i,j} \mid m_{i,j} \in M\} \longrightarrow D$ that maps each entry in M into a value in D. In Sect. 3.3 we present a classification of dimensions and give four examples.

Given a similarity matrix M of $n \times m$ entries, a dimension $D = \{d_1, d_2, \ldots, d_k\}$ induces a partition $M_1, M_2, \ldots, M_k$ over M such that $M_l = \{m_{i,j} \in M \mid F_D(m_{i,j}) = d_l\}$. We apply the evaluation measure E over each partition $E(M_1), E(M_2), \ldots, E(M_k)$ and define match consistency using coefficient of variation as follows:

Definition 1 (Match Consistency). *Let M be an $n \times m$ similarity matrix and $D = \{d_1, \ldots, d_k\}$ a consistency dimension. Let $E \in [0, 1]$ be a random variable, with an expected value of $\mu(E)$ and a standard deviation of $sd(E)$, representing an evaluation of partitions $M_1, \ldots, M_k$ over M. MC is a* match consistency *measure of M wrt D, computed as follows:*

$$MC(M, D, E) = 1 - \frac{sd(E)}{\mu(E)} \tag{1}$$

A higher MC value should correlate with increasingly consistent match performance across the partitions induced by D. Coefficient of variation, which was chosen as a best practice measure of data consistency (see [33]), achieves this correlation through its standard deviation component. A higher standard deviation increases the coefficient of variation and reduces the value of MC. This is in line with our understanding of consistency, where a lower standard deviation means a more consistent match. The second parameter is the average performance of the different partitions. Here, we take into account not only consistency but also our desire to achieve a good match, which entails an overall high evaluation measure.

3.2 Self Monitoring of Performance

Cognitive psychology has been examining factors impacting humans when performing knowledge intensive tasks [4]. The metacognitive approach, traditionally applied for learning and answering knowledge questions [7], highlights the role of subjective confidence in regulating efforts while performing challenging tasks.

Metacognition research was recently applied to reasoning and decision making tasks [2]. It suggests that online monitoring of subjective confidence regulates the cognitive effort invested in each task (*e.g.*, identifying a correspondence). The *Discrepancy Reduction Theoretical Framework (DRTF)* explains learning effort investment by suggesting that people set a target knowledge level as a stopping criterion. They continue to invest time and effort, while subjectively monitoring their confidence level, until meeting the stopping criterion [7].

Metacognitive models use three basic components of effort regulation measures, which we use for the matching task:

(1) **Subjective confidence:** Human matchers report matching confidence as their performance monitoring.
(2) **Invested time:** Elapsed time from selection of a term to the final matching decision is used as an objective measure that presumably reflects the metacognitive control decision to either continue or terminate a task, based on the ongoing monitoring of the chance of success.
(3) **Objective performance evaluation:** We use the well accepted precision and recall to evaluate performance.

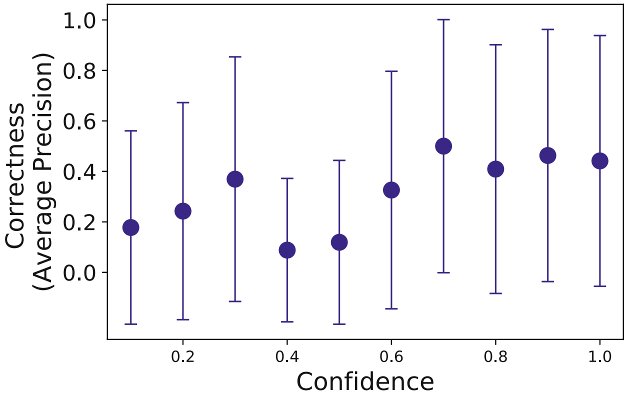


Fig. 1. Correctness by confidence, partitioned into buckets of 0.1

By way of motivation, we provide an illustration (Fig. 1) of the relationship between human confidence in matching and correctness (in terms of precision) based on our experiments (see Sect. 4). It is clear that human subjective confidence cannot serve as a good predictor to matching correctness. In this work we show how human biases affect confidence levels via consistency dimensions.

3.3 Consistency Dimensions

Consistency dimensions can be classified as continuous or discrete and may be performed using individual or collective matchers, as illustrated in Fig. 2. We introduce four consistency dimensions, namely (local and global) temporal, consensuality, and control, as examples to the full set of dimension possibilities.

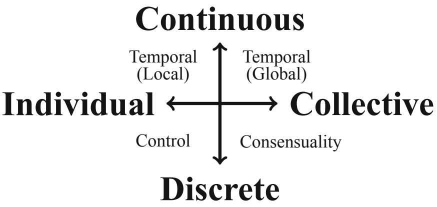


Fig. 2. Consistency dimensions

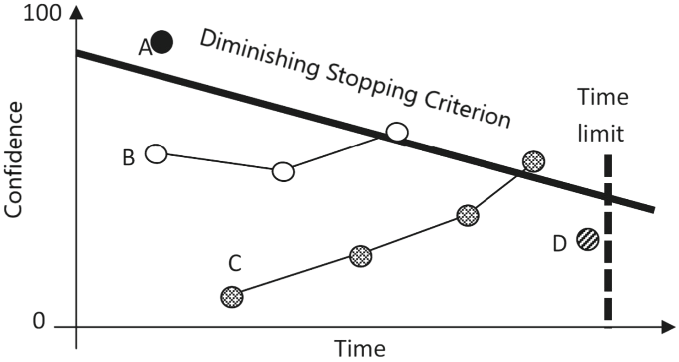


Fig. 3. DCM with hypothetical confidence ratings for four items and a self-imposed time limit (adapted from [1]).

(Local and Global) Temporal Dimension: This dimension is continuous (can be discretized into buckets to fit Definition 1) with both individual and collective variations.

The motivation to analyze the temporal dimension for human biases is rooted in the *Diminishing Criterion Model (DCM)* [1], a DRTF-based model (see Sect. 3.2) that models a common bias in human confidence judgment. DCM stipulates that the stopping criterion of a DRTF model is relaxed over time.

Thus, a human matcher is more willing to accept a low confidence level after investing some time and effort on finding a correspondence.

Figure 3 illustrates hypothetical confidence ratings while performing a schema matching task. Each dot in the figure represents a possible solution to a matching decision (*e.g.*, attributes a_i and b_j correspond), and its associated confidence, which changes over time. The search for a solution starts at time $t = 0$ and the first dot for each of the four examples represents the first solution a human matcher reaches.

As time passes, human matchers continuously evaluate their confidence. In case A, the matcher has a sufficiently high confidence after a short investigation, thus decides to accept it right away. In case B, a candidate correspondence is found quickly but fails to meet the sufficient confidence level. As time passes, together with more comprehensive examination, the confidence level (for the same or a different solution) becomes satisfactory (although the confidence value itself does not change much) and thus it is accepted. In Case C, no immediate candidate stands out, and even when found, its confidence is too low to pass the confidence threshold. Therefore, a slow exploration is performed until the confidence level is sufficiently high. In Case D, an unsatisfactory correspondence is found after a long search process, which fails to meet the stopping criterion before the individual deadline passes. Thus, the human matcher decides to reject the correspondence. When fitting a model based on the temporal dimension we can address human matchers individually, fitting a model for each human matcher separately (Local temporal), or collectively by fitting a general DCM model based on a group of human matchers (Global temporal).

Consensuality Dimension: This dimension models agreement among matchers. Metacognitive studies suggest that the frequency in which a particular answer is given by a group of people predicts confidence strongly [24].

The consensuality principal serves as a strong motivation to use crowd sourcing for matching, and was indeed proposed, *e.g.*, [35]. Although consensuality does not ensure accuracy [25], in this study we examine whether the number of people who chose a particular match can be used as a predictor of its chance to be correct. This can also support using majority voting based solutions as indication of correctness [3].

Consensuality requires multiple opinions to measure matchers agreement and a repetition of choices. We therefore classify this dimension as collective and discrete.

Control Dimension: Control analyzes the consistency of human matchers when assisted by a result of an algorithmic solution. This dimension is discrete (binary, in fact). In this work we consider control as an individual dimension, although it can be easily extended, using a general model for assisted/non-assisted matchers, to be collective.

Metacognitive control decisions are the regulatory decisions people take, given a self-assessment of their chance for success [7]. In the context of this study, the use of algorithmic output for helping the matcher in her task is taken as a control decision.

Variability in this dimension may be attributed to the predicted tendency of humans who do not use system suggestions to be more engaged in the task and recruit more mental effort than those who use suggestions as a way to ease their cognitive load (see [32]). Shraga *et al.* showed that human matchers who rely on algorithmic support are likely to follow the algorithm suggested [31].

4 Empirical Evaluation

The experiments analyze match consistency along consistency dimensions. Results indicate variability along consistency dimensions with varying trends of correctness.

4.1 Experiment Setup

Dataset and evaluation measures are presented next.

Human Matching Dataset: The dataset contains match results of 106 human matchers, all Industrial Engineering undergraduates who studied logistics and database management courses. Participants were briefed in schema matching prior to the task. Four pilot participants completed the task prior to the study to ensure its coherence and instruction legibility. Participants were trained on a pair of small schemata (9–12 attributes) from the *Thalia* dataset[1] prior to the main task.

The main schema matching task was chosen from the *Purchase Order* dataset [11]. The schemata used are medium size, with 142 and 46 attributes, and with high information content (labels, data types, and instance examples). Correspondences are of differing difficulty levels, with both easy matches and complex relationships, which may yield low precision and recall, even when using the strongest of matchers. Potentially, a maximum number of 6,532 correspondences are possible **per human matcher**, by (impossibly) evaluating each and every pair of attributes. In reality, each matcher chose to evaluate 51 correspondences on average, creating a dataset of ~5,600 human matcher's correspondences (1,229 distinct correspondences). A reference match for evaluation was compiled by domain experts over the years in which this dataset has been used for testing.

A side-by-side view of the two schemata and a dynamic match table were provided. The system records the time it takes for a matcher to determine on a correspondence. Match confidence was inserted by participants directly into the match table as a value between 0 and 1, displayed as a percentage.

Participants were randomly assigned to one of four conditions, differing by the algorithmic support provided. **No suggestions** (0), where participants perform the task with no algorithmic assistance; **limited suggestions** (1a), where

[1] www.cise.ufl.edu/research/dbintegrate/thalia/howto.html.

participants are allowed a limited (8 clicks) use of a lifesaver button. A counter of the suggestions used vs. remaining is presented; **unlimited suggestions** (1b), where participants are allowed an unlimited use of a lifesaver button; and **validate algorithmic result** (2), where algorithmic suggestions are pre-entered, letting participants validate, override, or complete them. The latter represents the classic "humans as validators" approach. The algorithmic matcher we used to create suggestions was Term (see Example 1) with typical performance (F1 ≈ 0.5) for automatic schema matchers on difficult instances.

To analyze the control dimension, we further separated participants into two groups. The first contains those participants who did not have suggestions (condition 0) or did not use the suggestion (from conditions 1a and 1b). The second contains those who actively requested suggestions from conditions 1a and 1b, in addition to the participants from condition 2.

Duplicate ratings for the same correspondence were removed, taking the latest. Out of the 106 participants, 6 were discarded due to technical faults, leaving 100 valid results. Elapsed time outliers (over 2 standard deviations from the mean of each participant) were removed due to the sensitivity of our measures to outliers, which may occur due to methodical pauses, unrelated to the matching task.

We created a group of the top 10% performing human matchers, considered as performance idealization of humans as validators, to show that even they suffer from biases and therefore are non-distinguishable from others a-priori.

Evaluation Measures: Let M^e be a reference matrix, such that $m_{i,j} = 1$ whenever the correspondence (a_i, b_j) is part of the reference match and $m_{i,j} = 0$ otherwise. The precision (P) and recall (R) evaluation measures are defined as follows:

$$P(\sigma) = \frac{\mid \sigma \cap M^{e+} \mid}{\mid \sigma \mid}, R(\sigma) = \frac{\mid \sigma \cap M^{e+} \mid}{\mid M^{e+} \mid} \tag{2}$$

where M^{e+} represent non-zero entries of M^e and recalling that σ is a subset of M's entries. The F1 measure, $F(\sigma)$, is calculated as the harmonic mean of $P(\sigma)$ and $R(\sigma)$.

Given a consistency dimension $D = \{d_1, d_2, \ldots, d_k\}$, precision and recall can be defined similarly per value d_i, by replacing σ with $\sigma \cap M_i$ (see Sect. 3.1). To compute match consistency we use Eq. 1 by estimating μ as the sampled average and sd as the sampled standard deviation over the evaluation measure of choice (*e.g.*, precision).

To analyze human matcher confidence we use metacognitive measures of calibration and resolution, based upon *performance monitoring.*

$$Calibration(\sigma) = \overline{\sigma} - P(\sigma), Resolution(\sigma) = \gamma(\sigma, M^{e+}) \tag{3}$$

Table 2. Resolution, (P)recision, (R)ecall, and (F)1 of matchers.

Matcher	Resolution	Sig. (p-value)	P	R	F
Term	0.63	0.045	0.35	0.80	**0.48**
TokenPath	0.72	0.140	0.25	0.86	0.33
WordNet	0.94	0.035	0.31	**0.87**	0.44
Human matchers	0.16 (SD=0.46)	0.001	**0.63**	0.36	0.45
Top-10 human matchers	0.58 (SD=0.57)	0.104	0.91	0.60	0.71

where $\overline{\sigma}$ is a user average confidence and $\gamma(\cdot,\cdot)$ is GK-Gamma correlation [19].

Positive calibration is interpreted as *overconfidence* and negative calibration as *under-confidence*. *Resolution* measures the extent to which confidence discriminates between correct and incorrect correspondences. GK-Gamma ranges in $[-1, 1]$ where scores of 1 and -1 indicate perfect resolution and 0 indicates no resolution. Negative resolution scores are interpreted as identifying good results as bad and vice-versa.

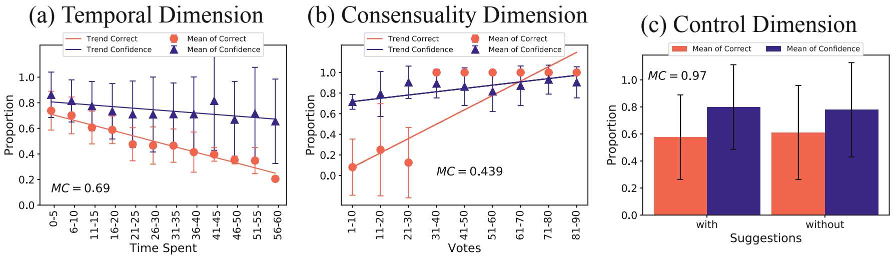
4.2 Results

We present a confidence analysis and empirical evaluation along consistency dimensions as evidence for human matching biases. Experiments show that human matchers are, in general, overconfident with low ability to distinguish correct from incorrect correspondences. In terms of consistency, results demonstrate significance variability along all dimensions with varying trends of correctness.

Confidence Analysis: We begin with a metacognitive evaluation, examining calibration and resolution (Sect. 4.1) in a schema matching setting. Average calibration (over participants) for match decisions was .26. 45 participants had over .3 calibration and 8 had negative calibration. Overall, the calibration levels demonstrate a right skewed distribution, interpreted as overconfidence, which was reported in the literature as a well-established human tendency [13].

Resolution results are given in Table 2. To compute matches, a 0.5 threshold was applied over the results. Human matchers, as a group, have significant, but low positive resolution (.16) with high variance. Only 25 (31%) had significant resolution (.68 average resolution within the group). Of those, 23 had significant positive resolution (average positive resolution was .79) and 2 had significant negative resolution (average negative resolution was $-.57$).

Note that performing the same calculation with algorithmic matchers yields much better resolution (see Table 2). Comparing human and algorithmic matching, overall the former has better precision, while the latter has better recall. However, even matching algorithms with comparatively fair F1 scores such as TokenPath, demonstrate high resolution. This serves as empirical evidence that the traditional view of “humans as validators”, may not be suitable for matching.


(a) Temporal Dimension

(b) Consensuality Dimension

(c) Control Dimension

Fig. 4. Confidence (Blue) and correctness (Red) by dimension (Color figure online)

(Local and Global) Temporal Dimension: We validate that the DCM (Sect. 3.3) reflects human matchers behavior by showing the predicted association between elapsed time and reported confidence. We show evidence of temporal bias and our ability to use elapsed time as a predictor of human matching performance. Support for this model would manifest itself via negative correlation between elapsed time and confidence per participant (local) and all participants (global).

Experimental results support the DCM model both locally and globally. A collective negative mean slope of $-.23$ suggests that on average, confidence decreases with time, which supports global temporal dimension. Zooming in on individual confidence reports, mean slope varied in $[-.274, -.213]$ with 40% individual matchers having significant (negative) correlations. A single-sample t-test was used to reject the null hypothesis of the slopes being random-noise. Also, a one-way ANOVA test was used to reject the null hypothesis of all participants sharing the same confidence mean ($F_{1,80} = 23.6$, p-value $< 10^{-5}$) emphasizing the need for a local model. To support the DCM self-imposed time limit, we followed [34] and found a significant curvilinear relation between time and confidence, reflecting the combination of **two** stopping criteria, which are unique to the DCM [1].

With correlated confidence and elapsed time, we now validate the use of DCM in matching, by examining the accuracy of human matchers as predicted by elapsed time. We tested the correlation between elapsed time and participant chance of providing correct matches. We partitioned the elapsed time into buckets, each of 5 s ($0–5, 6–10$, *etc.*) and examined an aggregative temporal behavior of human matchers. With $MC = 0.42$, the temporal dimension exhibits high variability among the dimension buckets. Correctness was computed as the precision within the bracket's time frame (Sect. 4.1). Mean slope was found to be $-.54$ (statistically significant with $p_{val} < 10^{-5}$). This serves as evidence that time spent is predictive of matching (in)correctness.

Figure 4a compares correctness with confidence, by showing precision (red) and mean of confidence across all human matchers (blue), partitioned according to elapsed time. For each measure we also include a linear trend-line and error bars for each time bucket. As discussed before, as time passes, less decisions made by humans are correct and there is a decline in human confidence. We also note that confidence consistently receives higher values than correctness proportion, which reflects the overall overconfidence, as reported above. The difference between the two becomes more prominent as time elapses, which is a classic finding in metacognitive literature, called hard-easy effect [9]. Note that error bars show variance in the way confidence is determined.

Offline examination of the top 10 human matchers reveals a slight (statistically insignificant) improvement in consistency. Accuracy is not available in real-life scenarios and therefore cannot be used to identify the best human matchers a-priori.

Consensuality Dimension: Next, we validate that the agreement level among matchers is correlated with self-reported confidence, and show evidence to its impact on human matching performance (Sect. 3.3).

We partitioned the number of votes for correspondence into increasing agreement levels (0–5, 6–10, *etc.*). For each level we computed the average confidence of correspondences and proportion of correctly matched correspondences out of all correspondences that were determined within the level. Figure 4b (similarly to Fig. 4a) presents confidence (blue) and correctness (red), partitioned by agreement level (number of votes). For each measure we also include a linear trend-line and error bars.

Overconfidence is demonstrated in lower agreement levels, while for higher levels the human matchers underestimate correctness. The error bars illustrate a significant variance in lower agreement levels and becomes negligible at higher levels, possibly as a result of correspondences that are easier to detect in levels where consensus is higher.

We also tested the correlation between level of agreement among participants and participant chance of providing correct matches. Although recent studies suggest that consensuality does not ensure accuracy [25], mean slopes for accuracy was found to be .13 (statistically significant with a p-value $< 10^{-5}$), showing that consensus among matchers is predictive of matching correctness. Consistency was measured at $MC = 0.36$, which is indicative of high variability. Here we see an improvement among the top 10 human matchers probably because they agree more among themselves on correct matches (evaluated only 95 correspondences compared to 1,229 overall).

Control Dimension: Finally, we show the impact of availability of algorithmic correspondences on human matching performance. To evaluate the performance of human matchers we compare the self-reported confidence and objective performance of participants by the control condition.

Figure 4c presents a comparison between participants who used (actively or passively) suggestions (left side) and those who did not (right side). A statistically significant (Pearson) correlation (p-value $< 10^{-5}$) was found between a binary variable indicating the use of a suggestion given a correspondence and a binary variable indicating whether this correspondence is a part of the reference match. Clearly, the human matcher is overconfident, regardless of the algorithmic assistance. Yet, results show better performance of participants who did not use the system's suggestions versus those who did. This can be explained by the fact that human matchers with machine support are more likely to behave as suggested [31], because of shallower processing than without this opportunity (see [32]). It is worth noting that with $MC = 0.85$, the control dimension demonstrates a more consistent pattern than the other two dimensions. Offline examination shows that the top 10 human matchers exhibit better consistency but show larger difference in confidence levels. The matchers assisted among the top 10 are much more confident (.81 compared to non assisted confidence of .69) but also live up to the expectations, achieving high accuracy levels (.86).

5 Related Work

Section 2.2 outlined the main effort in human involvement in matching. We now focus on demonstrating the contribution of this work on the background of state-of-the-art.

Using humans to answer schema matching validation questions was first proposed in [26]. This work was later extended [21,35] by using crowd sourcing to reduce uncertainty. Sarasua *et al.* suggested mechanical turks to validate matching by providing context information [30]. A recent work [35] also acknowledged the fact that the crowd is not always correct, associating probabilities to answers based on the question hardness (hard-easy effect as addressed in this paper) and worker's trustworthiness, which are estimated empirically, based on [15]. We take the observation that humans are not perfect a step further, analyzing cognitive biases that make human evaluation error prone.

Schema matching and ontology alignment [14] are closely related research areas, both aiming at finding matches between concepts. The two vary in their matching objects (schemata *vs.* ontologies), matching refinement (equivalence *vs.* richer semantics such as inclusion), and the underlying mathematical tools (*e.g.*, similarity matrix analysis *vs.* logic). To date, little work was devoted to the role of human matchers in either research areas. Nevertheless, a recent work in ontology alignment have acknowledged the fact that humans (users) can make mistakes [12]. Although it addresses cognitive oriented issues, *e.g.,* cognitive load, their aim is to avoid them. Further, they propose to collect confidence as a future work, which we collected and showed it may be unreliable (overconfidence). Our research insights can be readily applied to ontology alignment.

6 Conclusions and Future Work

This work introduces match consistency as a measure of human matching variability along potential bias dimensions. We view match consistency as a powerful tool to analyze human matching behavior. In future work we intend to identify additional dimensions, beyond the dimensions identified in this work, namely temporal, consensuality, and control. Our empirical evaluations serve as proof-of-concept that validate the important roles of humans as participants in the matching process, and less so as validators. Therefore, future work will involve collaboration models, supporting both human and algorithmic matchers, jointly performing schema matching considering humans biases.

References

1. Ackerman, R.: The diminishing criterion model for metacognitive regulation of time investment. J. Exp. Psychol.: Gen. **143**, 1349 (2014)
2. Ackerman, R., Thompson, V.: Meta-reasoning: monitoring and control of thinking and reasoning. TiCS **21**, 607–617 (2017)
3. Raykar, V.C., et al.: Supervised learning from multiple experts: whom to trust when everyone lies a bit. In: ICML (2009)
4. Barsalou, L.W.: Cognitive Psychology: An Overview for Cognitive Scientists. Psychology Press, New York (2014)
5. Bellahsene, Z., Bonifati, A., Rahm, E. (eds.): Schema Matching and Mapping. Springer, Berlin (2011). https://doi.org/10.1007/978-3-642-16518-4
6. Bernstein, P.A., Madhavan, J., Rahm, E.: Generic schema matching, ten years later. PVLDB **4**, 695–701 (2011)
7. Bjork, R.A., Dunlosky, J., Kornell, N.: Self-regulated learning: beliefs, techniques, and illusions. Ann. Rev. Psychol. **64**, 417–444 (2013)
8. Bozovic, N., Vassalos, V.: Two phase user driven schema matching. In: ADBIS (2015)
9. Brewer, N., Wells, G.L.: The confidence-accuracy relationship in eyewitness identification: effects of lineup instructions, foil similarity, and target-absent base rates. J. Exp. Psychol.: Appl. **12**, 11 (2006)
10. De Una, D., Rümmele, N., Gange, G., Schachte, P., Stuckey, P.J.: Machine learning and constraint programming for relational-to-ontology schema mapping. In: IJCAI (2018)
11. Do, H.H., Rahm, E.: COMA: a system for flexible combination of schema matching approaches. In: VLDB (2002)
12. Dragisic, Z., Ivanova, V., Lambrix, P., Faria, D., Jiménez-Ruiz, E., Pesquita, C.: User validation in ontology alignment. In: Groth, P., Simperl, E., Gray, A., Sabou, M., Krötzsch, M., Lecue, F., Flöck, F., Gil, Y. (eds.) ISWC 2016. LNCS, vol. 9981, pp. 200–217. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46523-4_13
13. Dunning, D., Heath, C., Suls, J.M.: Flawed self-assessment implications for health, education, and the workplace. Psychol. Sci. Public Interest **5**, 69–106 (2004)
14. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, New York (2007). https://doi.org/10.1007/978-3-540-49612-0
15. Franklin, M.J., Kossmann, D., Kraska, T., Ramesh, S., Xin, R.: CrowdDB: answering queries with crowdsourcing. In: SIGMOD (2011)
16. Gal, A.: Uncertain Schema Matching. Morgan & Claypool Publishers, San Rafael (2011)
17. Gal, A., Roitman, H., Sagi, T.: From diversity-based prediction to better ontology & schema matching. In: WWW (2016)
18. Gal, A., Roitman, H., Shraga, R.: Heterogeneous data integration by learning to rerank schema matches. In: ICDM (2018)

19. Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications. J. Am. Stat. Assoc. **49**, 732–764 (1954)
20. Halevy, A.Y., Madhavan, J.: Corpus-based knowledge representation. In: IJCAI (2003)
21. Hung, N.Q.V., Nguyen, T.T., Miklós, Z., Aberer, K., Gal, A., Weidlich, M.: Pay-as-you-go reconciliation in schema matching networks. In: ICDE (2014)
22. Hung, N.Q.V., Tam, N.T., Miklós, Z., Aberer, K.: On leveraging crowdsourcing techniques for schema matching networks. In: Meng, W., Feng, L., Bressan, S., Winiwarter, W., Song, W. (eds.) DASFAA 2013. LNCS, vol. 7826, pp. 139–154. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37450-0_10
23. Jeffery, S.R., Franklin, M.J., Halevy, A.Y.: Pay-as-you-go user feedback for dataspace systems. In: SIGMOD (2008)
24. Koriat, A.: Subjective confidence in one's answers: the consensuality principle. J. Exp. Psychol.: Learn. Memory Cognit. **34**, 945–959 (2008)
25. Koriat, A.: When reality is out of focus: can people tell whether their beliefs and judgments are correct or wrong? J. Exp. Psychol.: Gen. **147**, 613 (2018)
26. McCann, R., Shen, W., Doan, A.: Matching schemas in online communities: a web 2.0 approach. In: ICDE (2008)
27. Peukert, E., Eberius, J., Rahm, E.: AMC-a framework for modelling and comparing matching systems as matching processes. In: ICDE (2011)
28. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. VLDBJ **10**, 334–350 (2001)
29. Sagi, T., Gal, A.: In schema matching, even experts are human. towards expert sourcing in schema matching. In: IIWeb (2014)
30. Sarasua, C., Simperl, E., Noy, N.F.: CrowdMap: crowdsourcing ontology alignment with microtasks. In: ISWC (2012)
31. Shraga, R., Gal, A., Roitman, H.: What type of a matcher are you?: coordination of human and algorithmic matchers. In: HILDA@SIGMOD (2018)
32. Sidi, Y., Shpigelman, M., Zalmanov, H., Ackerman, R.: Understanding metacognitive inferiority on screen by exposing cues for depth of processing. Learn. Instr. **51**, 61–73 (2017)
33. Simonsen, J.C.: Coefficient of variation as a measure of subject effort. Arch. PM&R **76**, 516–520 (1995)
34. Undorf, M., Ackerman, R.: The puzzle of study time allocation for the most challenging items. Psychon. Bull. Rev. **24**, 2003–2011 (2017)
35. Zhang, C., Chen, L., Jagadish, H., Zhang, M., Tong, Y.: Reducing uncertainty of schema matching via crowdsourcing with accuracy rates. TKDE (2018). https://www.computer.org/csdl/journal/tk/5555/01/08533346/17D45XreC6p