Understanding the Elephant:
The Discourse Approach to Boundary Identification and Corpus Construction for Theory Review Articles

Kai R. Larsen[1], Dirk S. Hovorka[2], Alan R. Dennis[3], Jevin D. West[4]

[1]University of Colorado, USA, kai.larsen@colorado.edu
[2]University of Sydney, Australia, dirk.hovorka@sydney.edu.au
[3]Indiana University, USA, ardennis@indiana.edu
[4]University of Washington, USA, jevinw@uw.edu

Abstract

The goal of a review article is to present the current state of knowledge in a research area. Two important initial steps in writing a review article are boundary identification (identifying a body of potentially relevant past research) and corpus construction (selecting research manuscripts to include in the review). We present a *theory-as-discourse* approach, which (1) creates a theory ecosystem of potentially relevant prior research using a citation-network approach to boundary identification; and (2) identifies manuscripts for consideration using machine learning or random selection. We demonstrate an instantiation of the theory as discourse approach through a proof-of-concept, which we call the automated detection of implicit theory (ADIT) technique. ADIT improves performance over the conventional approach as practiced in past technology acceptance model reviews (i.e., keyword search, sometimes manual citation chaining); it identifies a set of research manuscripts that is more comprehensive and at least as precise. Our analysis shows that the conventional approach failed to identify a *majority* of past research. Like the three blind men examining the elephant, the conventional approach distorts the totality of the phenomenon. ADIT also enables researchers to statistically estimate the number of relevant manuscripts that were excluded from the resulting review article, thus enabling an assessment of the review article's representativeness.

Keywords: Literature Review, Review Article, Research Review, Boundary Identification, Article Identification, Keyword Search, Citation Search, Machine Learning.

1 Introduction

Review articles serve an important knowledge creation function (Templier & Paré 2015; Vessey, Ramesh, & Glass, 2002) by examining, contextualizing and summarizing prior research within a selected research area (Rowe, 2014). As such, review articles present the current state of knowledge about a topic. Review articles are an important element of research training and a research output in their own right (Rivard, 2014; Schultze, 2015; Watson, 2015; Webster & Watson, 2002). Reviews may serve different purposes (Templier and Paré 2015; Rowe 2014; vom Brocke, Simons, Riemer, Niehaves, Plattfaut, & Cleven, 2015), including delineation of topic boundaries (Webster & Watson, 2002), motivation of interest (Ke, Ferrara, Radicchi, & Flammini, 2015), identification of gaps and inconsistencies (Webster & Watson, 2002), and guidance for future research (Schwarz, Mehta, Johnson, & Chin, 2007; vom Brocke et al., 2015).

What constitutes a "good" review article? We argue that a good review article is a form of rational argumentation approaching what Habermas (1990, p. 104) terms "ideal speech": it ensures "that (1) all voices in any way relevant can get a hearing; that (2) the best arguments we have in our present state of knowledge are brought to bear; and that (3) disagreement or agreement on the part of the participants follows only from the force of the better argument and no other force" (see also Chiasson, 2015). If we do not let all voices be heard, we run the risk of the three blind men examining the elephant: The review sees only part of the phenomenon without benefit of a larger context, thereby distorting knowledge of the phenomenon; just as the trunk is not representative of the entire elephant, research published in the top journals is not representative of all research on a phenomenon. That is, research published in top journals differs systematically from research in lower-tier journals, conferences, and unpublished manuscripts (Yong, 2012).

For example, like the blind men, we may focus on the voices of manuscripts at an arbitrary set of "top journals," which often eliminate manuscripts that challenge dominant theory, have unsupported hypotheses or nonsignificant findings and thus did not clear the hurdles of the small set of editors and reviewers at top journals. These research instances are revealing of the phenomena's stability, generalizability, and replicability—they contribute to the overall research discourse. Therefore, creating an inclusive corpus of prior research to be analyzed is critical to the comprehensiveness of the research that is the focus of the review (Boell & Cecez-Kecmanovic, 2014; vom Brocke et al., 2015).

Regardless of the goals or type of review (see Ortiz de Guinea & Paré, 2017), authors strive to not omit articles that meet their inclusion criteria. Instead, they seek to include all relevant manuscripts that fit their criteria because reviews that are unsystematic in including all relevant manuscripts suffer from subjectivity and cannot claim to present a representative understanding of knowledge within the domain of the review (Ortiz de Guinea & Paré, 2017).

Creating a corpus of manuscripts to be analyzed in a review has two major process steps. First, *boundary identification* assesses the size and scope of the *potentially relevant* research manuscripts in the research domain of interest to ensure that all voices are heard by. Second, *corpus construction* utilizes inclusion criteria to determine the relevance of each manuscript for analysis within the boundary to the purpose of the review. [1] The goal of boundary identification is to identify the entire set of potentially

relevant manuscripts, while the goal of corpus construction is to select the manuscripts relevant to the review that will be analyzed. For domains with small boundaries, corpus construction may include all relevant manuscripts; for domains with large boundaries (e.g., thousands of manuscripts) corpus construction may select some subset. If boundary identification or corpus construction are flawed, the resulting corpus may be incomplete or nonrepresentative, and thus the analysis will be distorted, as we may only be reviewing a small part of the metaphorical elephant. Many research domains have thousands of potentially relevant manuscripts and knowledge continues to grow very rapidly (Vom Brocke, Simons, Niehaves, Riemer, Plattfaut, & Cleven, 2009), corpus construction becomes increasingly difficult.

In this article, we describe a way of conducting theoretical reviews in the face of an ever-increasing number of publications. Specifically, we present the automated detection of implicit theory (ADIT), a boundary identification and corpus construction approach, which views theory as ongoing discourse among authors (hence we term it the *discourse approach*). The identification process determines the size and delineation of the corpus and the corpus construction process and is based on machine learning to classify manuscripts as more or less likely to be relevant. We demonstrate ADIT by examining a specific theory domain, TAM.

Our arguments are provided in three sections: In the first section below, we briefly examine the *conventional approach*, which considers theory "an artifact built by humans to achieve some purpose" (Webster & Watson, 2002, p. 4) that is composed of constructs, boundaries, states, and the relationships among them (Baskerville & Pries-Hele, 2010; Gregor, 2006; Weber, 2012).

In the next section, we present ADIT as an alternative approach based on a *discourse view* of theory (Giddens, 2013; Jones & Karsten, 2008). The discourse view suggests that theory development is a historically informed process whereby multiple actors instantiate and revise theory over time in an ongoing discourse of justification, support, extension, and critique. Through a proof-of-concept using TAM, we conclude that the discourse approach performs better than the conventional approach for the TAM research area.

In the final section, we discuss implications for research. The discourse approach bounds potentially relevant publications within an ecosystem (the set of foundational manuscripts, the manuscripts that cite

---

[1] Assessing the quality of manuscripts is done in a later step, when each manuscript is read and analyzed. During this step manuscripts of poor quality are removed from analysis.

them, and every article those manuscripts cite). It ensures that we have a more representative view and that we are more likely to see all the different parts of the elephant, not just the most prominent parts (i.e., top-ranked journal articles). It enables us to use statistical techniques to estimate how much of the elephant we have seen (i.e., what percent of manuscripts relevant to a review article have been included in the review), so we are better able to evaluate the review.

2   The Conventional Approach to Boundary Identification and Corpus Construction

The conventional approach holds that theories are "a particular kind of model,…an abstracted, simplified, concise representation" of things in the world (Weber, 2012, p. 4). This approach has given rise to a theory-boundary identification approach that focuses on the attributes of the theory (e.g., theory name, constructs) and uses them as keywords for searching research databases (e.g., Wu, Zhao, Zhu, Tan, & Zheng, 2011, as well as our own past work—e.g., Dennis, Wixom, & Vanderberg, 2001).

There are two distinct processes: (1) boundary identification (finding all potentially relevant studies in a way that is both comprehensive (i.e., not accidentally omitting relevant studies) and precise (i.e., not including studies that are irrelevant), and 2) corpus construction (selecting a sample of studies from this population for analysis). Many past review articles have not explicitly addressed these processes separately, perhaps because the size of the corpus was small enough to permit complete enumeration. As the number of manuscripts grows, it becomes more difficult to analyze all identified manuscripts (vom Brocke et al. 2015; Webster & Watson, 2002), so the selection process used in corpus construction becomes important. Past review articles have sometimes comingled identification with construction by limiting searches because unconstrained keyword searches can identify more manuscripts than it is practical to read (e.g., on EBSCO, searching using the keyword search string "technology acceptance model" returns more than 15,000 manuscripts; on Google Scholar it is more than 70,000).[2] However, there is a growing consensus that it is necessary to have transparency in separating and documenting the identification and construction processes so that readers have confidence in the

review outcomes (Tate, Furtmueller, Evermann, & Bandara, 2015). It is important to know how much of the potentially relevant past research has been included and excluded (Grant & Booth, 2009).

One common approach that comingles identification and construction is to identify the boundary condition in advance by limiting the search to manuscripts in selected journals. Most ideal review archetypes require search strategies that produce a comprehensive or representative corpus (Templier & Paré, 2015). Therefore, it is not appropriate for boundary identification to be confined to one set of journals (cf. Webster & Watson, 2002). Such an approach violates the principle of ideal speech (Habermas, 1990) by disenfranchising a set of voices from conferences and the "gray literature." Gray literature is the set of books, book chapters, monographs, unpublished manuscripts, and non-peer-reviewed conference presentations that have been evaluated using different review criteria than articles in top journals (Webster & Watson, 2002; terms used in this article are defined in Table 1). Conference papers often present new and novel research and may contain emerging findings. When a journal's ranking is used as a surrogate for quality, individual high-quality manuscripts may be excluded. One reason for rejection in a top-ranked journal is the lack of novel contribution (an ambiguous criterion), not a methodological flaw (Straub, 2009; Yong, 2012). As a result, research published in top-ranked journals favors confirmatory studies, and new theory (Okoli, 2012). Research in conferences, lower-tier journals, and the gray literature may offer critical understanding of the state of knowledge including replications, rigorous but nonsignificant outcomes, and challenges to the dominant viewpoint (Okoli, 2012). Studies with nonsignificant results or those that fail to replicate prior results are unlikely to be published in top journals (Yong, 2012), but are critical to include in a review article; if we omit these articles, we do not see the entire elephant.

Therefore, we argue that the inclusion of papers from conferences, lower ranked journals, and the gray literature is important; it is inappropriate to exclude these manuscripts en masse without a compelling reason. Methodological flaws or nonrepresentative data are reasons for excluding manuscripts, but exclusion should be reasoned and articulated on a case-by-case basis. Otherwise, we are deliberately choosing to disenfranchise what Habermas (1990) would term relevant voices and thereby removing parts of our metaphorical elephant from analyses.

---

[2] Searches conducted on October 14, 2017. EBSCO search conducted on all EBSCO databases available.

Table 1. Definition of Key Terms

Bias: conducting a review with a set of articles that is not representative of the population of articles.

Comprehensiveness—TP/(TP+FN): The number of articles returned by the search strategy or tool that meet the criteria for inclusion divided by the total number of articles that meet the criteria for inclusion. In computer science, this is generally referred to as *recall*.

Corpus construction: The corpus construction process of narrowing the set of all papers within the theory boundary to the set of papers that fit the inclusion criteria.

False negative: An article that meets the criteria for inclusion but is not returned by the search strategy or tool—saves research effort but potentially leads to bias in the review.

False positive: An article that does not meet the inclusion criteria but is returned by the search strategy or tool—requires retrieval and evaluation effort and contributes little to no benefit to the review.

Gray literature is the subset of manuscripts perceived as not having been subjected to as stringent a (peer-)review process. Researchers often disagree about the appropriate cutoff in their definitions of gray literature.

Inclusion criteria: The set of rules determining which sources should be part of the review's analysis. We do not distinguish between inclusion and exclusion criteria because they are often transmutable.

Manuscript: Books, book chapters, journal articles, conference articles, monographs, and unpublished manuscripts. To properly differentiate terms, we refer to "our article" vs. "review articles," and papers that were analyzed by a past review article are referred to as *manuscripts*, unless the context is clearly articles (such as references to "journal articles" in a review article that included only journal articles). This is slightly different than the conventional definition of manuscript (unpublished or handwritten) but we found this differentiation helpful in distinguishing article types.

Precision—TP/(TP+FP): The number of articles returned by the search strategy or tool that meet the criteria for inclusion (i.e., relevant manuscripts) divided by the total number of articles returned by the search strategy or tool.

Relevant manuscripts: The manuscripts that fit the inclusion criteria given the purpose of the review.

Representative body of research: A corpus of research that fully reflects the findings of all research that meets the inclusion criteria, regardless of publication status.

Theory boundary (for boundary identification): The drawing of the boundary between the manuscripts that could *potentially* be relevant given a review goal and those that are definitely not relevant.

Theory ecosystem: A set of foundational manuscripts, the manuscripts that cite them, and every article those manuscripts cite.

True negative: An article that does not meet the criteria for inclusion and is not returned by the search strategy or tool.

True positive: An article that meets the criteria for inclusion and is returned by the search strategy or tool.

Note: TP = true positive, TN = true negative, FN = false negative, FP = false positive

Case-by-case quality assessment is straightforward and can be done during the analysis step that follows corpus construction; after all, each manuscript is read closely during analysis, so including reading for quality is straightforward. Sometimes, the set of manuscripts produced by boundary identification is small, and it is possible to use complete enumeration during corpus construction to select relevant manuscripts for evaluation (e.g., in the case of a new research area). In other cases, the size of the theory ecosystem is so large that complete enumeration is not feasible. If so, we argue that the best approach to corpus construction is the same approach adopted by survey researchers drawing samples from a population. We presume that each response (i.e., manuscript) has been produced in good faith. However, if there is something about a specific response that makes us question its validity, then we can remove it from the sample. The legitimacy of the comprehensive or representative (Templier & Paré, 2015) reviews requires the authors to describe the nature and scope of the

manuscripts that were excluded and provide strong justification for that exclusion, the same way we would require strong justification for the exclusion of data points from a survey or experiment. The burden of proof lies on the decision to exclude, not on the decision to include.

In addition to keyword searches for boundary identification, the conventional approach sometimes includes forward and backward chaining—i.e., searching the reference lists of manuscripts found by keywords (backward) and using databases such as Web of Science to identify papers that cite the found manuscripts (forward) (Webster & Watson, 2002). If an average paper has 50 references and each of those manuscripts has 50 references, this produces a set of about 2,500 manuscripts. Properly evaluating all 2,500 references is a daunting task—even more so when one considers that these 2,500 references represent only the second level of manuscripts on which backward chaining could be performed. The only salient information immediately available to the user of forward and backward chaining is the title, authors, and

journal, which makes it difficult to decide which manuscript to include or exclude without examining the manuscript in more detail.

The goal of corpus construction is to identify all research *relevant* to the goals of the review and to demonstrate that you have done this without introducing bias—here defined as a set of manuscripts that is not representative of the population of manuscripts. Avoiding such bias is challenging: How do you know whether you have identified all relevant research (vom Brocke et al., 2015)? When can you be confident that you have explored the entire elephant? Unfortunately, many review articles do not clearly document their boundary identification and corpus construction processes so we do not know how comprehensive or representative they are (vom Brocke et al., 2009).

To investigate this conclusion, we performed an empirical examination of the conventional approach when applied to the technology acceptance model (TAM). We found a total of 20 TAM review articles in the past research literature, of which 16 were relevant to this paper.[3] Six of these 16 articles had a goal of producing a comprehensive review of all TAM research; the rest focused on one aspect of TAM research. We examined the search strategies and success in identifying/selecting relevant manuscripts of these 16 review articles. The maximum number of manuscripts included by any of these 16 articles was 136 (average 64.7); there was a combined total of 420 unique manuscripts across all 16 articles. We then conducted our own analysis of the total population of manuscripts mentioning TAM and/or its constructs. Our boundary identification process found 5,991 manuscripts. Our corpus construction process identified 1,590 relevant manuscripts (with a 95% confidence interval of 1,378-1,797 manuscripts). The full details of these analyses are presented in Appendix A.

Obviously, 1,590 is much larger than the maximum of 136 manuscripts that these past review articles included—or even than the 420 combined total that all 16 review articles included. Likewise, even the 420 combined total is much larger than the 136 maximum manuscripts included by any one article. We cannot fully assess the quality of a process by examining its outcomes but, nonetheless, we believe that this pattern of outcomes indicates a problem with the conventional approach to boundary identification and corpus construction. It may be that all of these author teams (and all of the editorial teams that reviewed their papers) failed to implement the keyword search appropriately, but we do not believe this is a plausible argument. If every team using an approach misses a large majority of past research, then this is prima facie evidence that the approach itself is inherently flawed. The problem of identifying all relevant

manuscripts will only become worse as the number of publications grows each year (vom Brocke et al., 2015; Webster & Watson, 2002). Therefore, we believe this calls for a new approach.

3   Discourse Approach

The approach to identification and corpus construction that we term the *discourse approach* differs in two fundamental ways from the conventional approach. First, it views research in a domain (e.g., a theory) as a discourse among scholars (Bostrom, Gupta, & Thomas, 2009) rather than a set of characteristics; it is a living process, shaped by many hands, not just a discrete set of things that are a product of the process (Weick 1995). Second, we identify criteria of contribution across the entire set of manuscripts for relevance. Machine learning techniques (Fan, Pathak, & Pathak, 2005; William T. Grant Foundation, 2009) are used to assess the entire set of potentially relevant manuscripts to produce a smaller set of manuscripts that are more likely to contribute to the goals of the review. Where the corpus is too large to be included researchers may use preferred techniques (e.g., random samples) to select manuscripts for inclusion in the review.

As an ongoing discourse, theory (or model, framework, etc.—for simplicity, we just use the term theory) evolves within an ecosystem of research manuscripts. These manuscripts may provide support, replication, and extension, or they may provide critique, failed replications, or contradictory results over time (Bostrom et al., 2009). A theory is contained within and bounded by the corpus of publications that contribute to a theory. Contribution to a theory is determined by whether a manuscript extends, empirically tests, refines, or critiques the theory. Under this theory-as-discourse view, a theory is not fully specified by the original manuscript proposing a theory, nor by the most recent or most cited version, but rather consists of the originating manuscript(s), and all manuscripts that contribute to the theory.

This view promulgates two new ways to think about theory review: (1) the *theory ecosystem* as a citation network containing all manuscripts that potentially contribute to the development and understanding of a theory starting with its proposal; and (2) the *theory-contributing* manuscripts, which will change based on the inclusion criteria for a given project, but will always be bounded by the theory ecosystem. Combining these two components enables a robust approach to boundary identification and corpus construction that improves comprehensiveness and precision.

---

[3] The exclusion criteria are explained in the Appendix for each of the four excluded reviews.
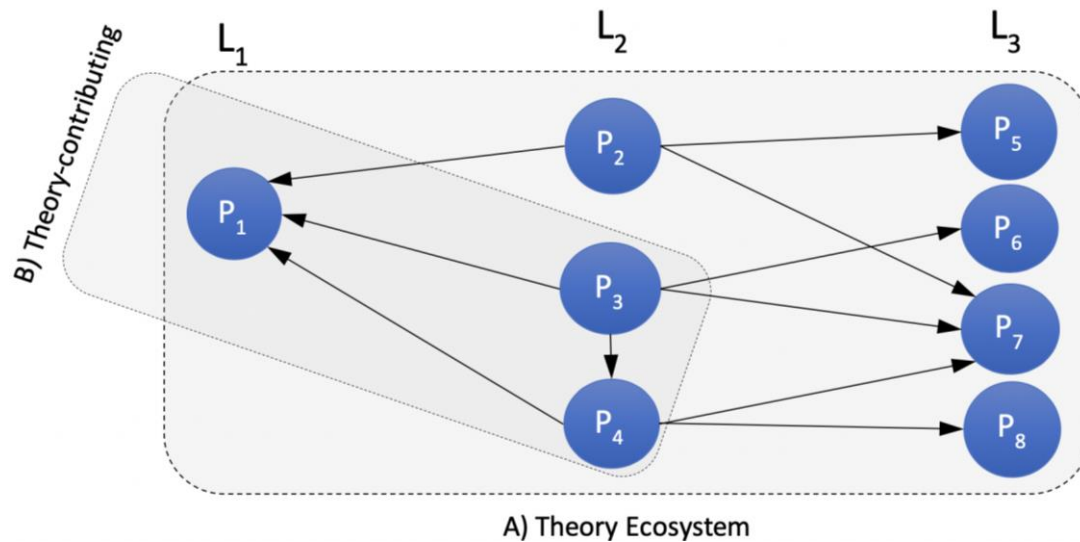
Figure 1. Theory Ecosystem

3.1 Boundary Identification

In academic discourse, manuscripts intending to contribute to the ongoing development of a theory should cite the foundational manuscript(s) that created the theory, show knowledge of the theory discourse by citing key subsequent manuscripts that contribute to the theory, and show knowledge of the current shared conceptualization of the theory, not just its original form (Weber 2012). Over time, the theory discourse produces a set of interconnected manuscripts that form a *theory ecosystem*. Some manuscripts contribute to its development as the theory evolves, others simply invoke or use the theory, while yet others provide methodological or external ideas and were never intended as contributions to the theory. From the perspective of a review article, it is important to identify the set of manuscripts that contribute to a theory and distinguish them from those that cite the theory for other purposes and be able to safely ignore those never intended as contributions to the theory.

In Figure 2, the two gray boxes labeled A and B illustrate two subsets of the overall academic citation network. The network can be divided into three concentric circles of manuscripts (L_1, L_2, and L_3). The first set, (A), is the *theory ecosystem,* which consists of L_1, the theory's foundational manuscripts (for TAM this would be Davis, 1989), L_2 manuscripts that cite the foundational manuscripts, and L_3 manuscripts that influenced the L_2 manuscripts enough to be cited by them. Figure 2 shows an ecosystem that has only one foundational L_1 manuscript (P_1). The second set, (B), comprises the *theory-contributing* manuscripts that contribute to the development of the theory. This set is only found after all L_2 and L_3 manuscripts have been

analyzed, and is composed only of L_1 and L_2 manuscripts. Some L_2 manuscripts cite each other (P3-> P4). Because L_2 manuscripts may be published over a longer period, some L_3 manuscripts may cite L_2 manuscripts (not shown). While L_3 does not contain contributions to the theory in the traditional sense, these manuscripts are included to improve the accuracy of later calculated network metrics. The identified theory boundary between potentially relevant and irrelevant manuscripts is in the discourse process set at the boundary between L_2 and all other manuscripts in existence, regardless of what keywords may exist in those manuscripts outside the boundary.

3.2 Corpus Construction

The second step in the process is to select relevant manuscripts from the population identified in the first step to build a corpus for analysis. The act of citing a foundational manuscript is necessary but not sufficient to indicate that a manuscript contributes to the theory. Not all manuscripts identified by the discourse approach are relevant (and, of course, the same is true for the conventional approach). Our analysis of the TAM benchmark sample found that 26.5% of citing manuscripts made an empirical contribution to the theory (See Appendix A). In a study of UTAUT (Venkatesh, Morris, G. Davis, & F. Davis, 2003), Williams, Rana, Dwivedi, & Lal (2012) found that only 9.6% of the manuscripts citing the foundational UTAUT manuscript made an empirical contribution to the theory. Williams et al. retrieved 52% of theory-citing manuscripts (compared to 98.3% for our sample), covered only journal articles, and used different inclusion criteria.
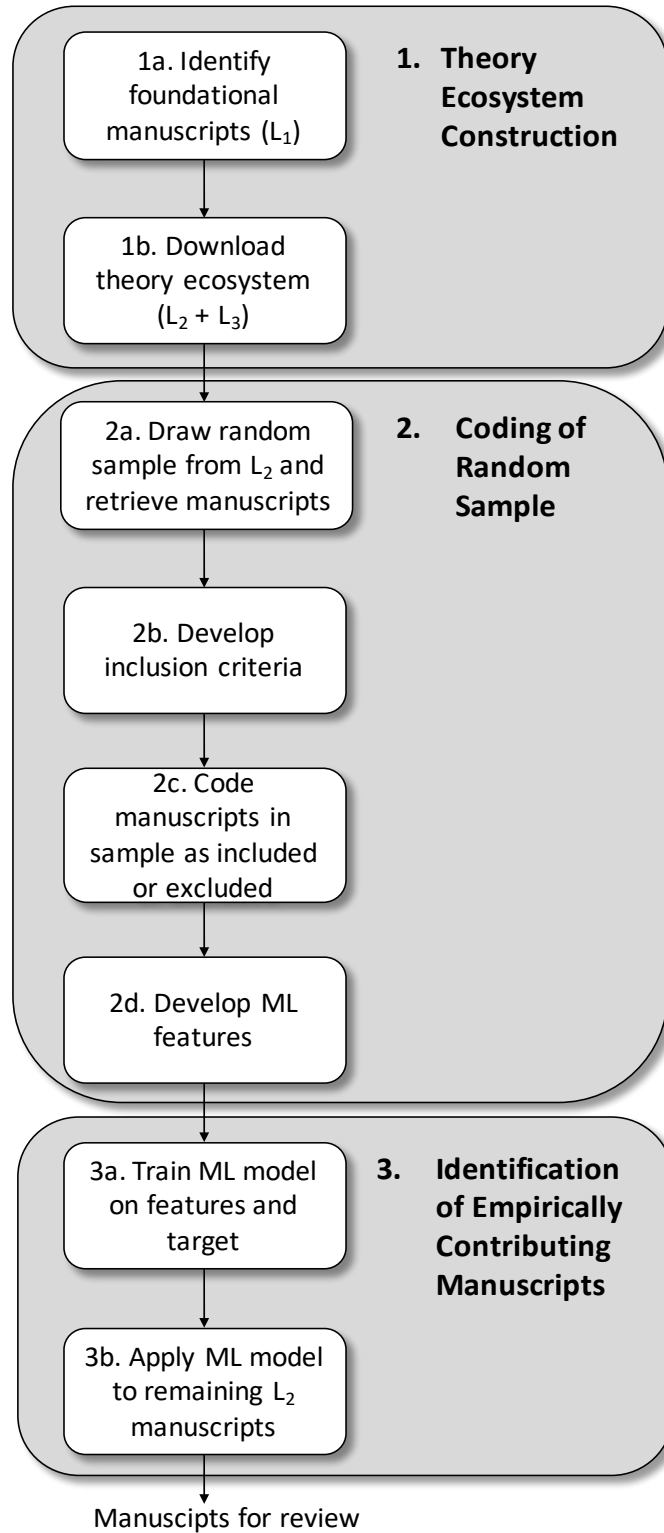
Figure 2. ADIT Review Overview

Therefore, while a theory identification approach that uses the citation of foundational manuscript(s) as an inclusion criterion is comprehensive, it is not likely to be precise; in our case, about 75% of the manuscripts identified did not empirically contribute to the theory. In Williams et al.'s case, it was 90% of the manuscripts. If the size of the population of manuscripts produced in the first step is reasonable, we can simply examine every manuscript to see if it is relevant. The problem comes when the population is so large as to make complete enumeration infeasible. In this case, how do we select the manuscripts to consider for our review? A common approach is to apply simple heuristics based on individual criteria to exclude manuscripts and reduce information overload. For example, manuscripts may be excluded if they do not appear in top journals or conferences, or are "gray" manuscripts (Berger 2003; Han 2003; Hsiao and Yang 2011; King and He 2006; Webster & Watson, 2002). This is not a requirement of the keyword search approach, but a common heuristic that conflates boundary identification with corpus construction.

While these criteria may be well-intentioned, they introduce a systematic bias into the sample (Banks et al., 2015; Kepes, Banks, McDaniel, & Webster, 2012; Kepes & McDaniel, 2015; Webster & Watson, 2002). Journals, especially top-ranked journals, are systematically different from other research publications; they are less diverse in that they are less likely to publish nonsignificant results, results that are not "surprising," results that have many nonsupported hypotheses, and results that challenge the dominant theoretical viewpoint (Dennis & Valacich, 2014; Vessey et al., 2002; Yong, 2012). This is not to say that research in top-ranked journals is "bad." It just is not representative of the larger universe of knowledge (Yong, 2012). Selecting research based on the journal in which it is published is akin to stereotyping and it eliminates a diversity of viewpoints by filtering out manuscripts based on who they are, not what they say. It is better to include these manuscripts and make an individual manuscript-by-manuscript decision in the analysis phase of whether to include them or not. If we exclude them en masse, the resulting pool of research shows only part of the elephant.

What selection criteria should we apply to the theory ecosystem to better enable us to select relevant manuscripts into the corpus for subsequent analysis? We offer two very different approaches that can be used separately or together.

The first approach is simple. It is random sampling or stratified random sampling if researchers prefer a sample with specific proportions of manuscripts by type (e.g., journal article, conference paper, working paper), date (e.g., more recent manuscripts versus older manuscripts), or another characteristic. Random sampling is well known as the least biased approach to selecting a sample from a population (Babbie, 2013; Creswell, 2013). Of course, it is possible that random sampling may still produce a biased sample, but the probability of this is low, and statistics can be used to estimate confidence intervals. If authors are concerned about the inclusion/exclusion of certain parts of the elephant, they can compare the characteristics of the sample to those of the population. For example, one could try Laplace or double exponential sampling if sampling with sharper tails in the sharper peaks in their distribution is desired. As we noted above, stratified random sampling can ensure desired proportional representation based on specific criteria. However, we want to stress that we only suggest stratified sampling if one has strong theoretical reasons to sample from groups rather than the population and is willing to explicitly specify these reasons. We could not formulate a strong enough reason to do this in our case.

If judgment needs to be applied to sort out desirable from undesirable studies, it should be applied in the same way as we apply judgment in survey research—in the analysis phase after the random sample has been drawn to remove outliers or bad data. This way, the judgment is applied in a systematic, thoughtful, manuscript by manuscript basis, rather than using a simplistic heuristic that is imprecise and can exclude useful studies. Applying judgment in the analysis phase also enables researchers to document which studies have been excluded and why, so it is possible to detect and explain any bias (good or bad) that the use of judgment introduces.

The second approach is application of machine learning (Fan et al., 2005; William T. Grant Foundation, 2009; Kattan, Adams, & Parks, 1993), which categorizes manuscripts as likely or not to provide an empirical contribution to the theory based on a set of criteria developed from an analysis of potentially theory-contributing manuscripts (all manuscripts in L_2). This is also heuristic, but is based on empirical analysis of which features relate to the inclusion criteria as well as logic, so it is less likely to be biased. We describe this in the next section.

3.3 The Automated Detection of Implicit Theory (ADIT) Technique

The automated detection of implicit theory (ADIT) technique, which is a design instance suggested by the discourse approach, uses machine learning to select the empirical theory-contributing manuscripts within a theory ecosystem. It has three general steps: (1) construction of a theory ecosystem to provide a comprehensive set of manuscripts for boundary identification; (2) coding of a random sample; and (3) selection of manuscripts for corpus construction. In response to Schryen et al.'s (2017) call for dynamic rather than static literature review tools, ADIT is designed to detect and include theory-relevant

manuscripts even after a review is over. Appendix D provides additional technical details.

1. Theory ecosystem construction (boundary identification): ADIT begins by building a citation network as shown in Figure 2. It starts by (1a) identifying one or more foundational manuscripts for a theory or research area. For example, the foundational articles for TAM are Davis (1989) and Davis, Bagozzi, & Warshaw (1989).[4] It then (1b) searches through a designated research repository to download all manuscripts that cite any of these foundational manuscript(s). Once this second level of the theory ecosystem is downloaded, all references cited by these manuscripts are downloaded to complete the ecosystem, and all citations between manuscripts are established for a full theory ecosystem network. This step requires the user of ADIT to carefully designate a set of publications that then specifies the literature review boundary. Unlike a set of keywords, experts may immediately detect and disagree with the inclusion criteria. In our example, it required us to carefully defend our decision to exclude Davis' (1986) from the foundational list, as well as the decision not to include articles considered foundational for TAM2 and TAM3. The approach also allows later evaluation of the second level of the TAM ecosystem vs. the TAM2 ecosystem as well as the ecosystems of other theories. Finally, we believe that it will eventually allow us to examine the context in which each theory exists (level three) and compare and classify theories through these networks.

We selected Microsoft Academic Search (MAS; academic.research.microsoft.com) as our research repository because of its relative comprehensiveness and its open application programming interface (API). Future versions of ADIT should also work with other research repositories.

2. Coding of a random sample: This step is split into four separate processes. First, (2a) ADIT draws a random sample of manuscripts from L_2 of the ecosystem (See Figure 2). Only this set is sampled because as argued earlier, a manuscript that does not cite an L_1-source was likely not intended as a contribution to the specified theory. A sufficiently large sample is necessary to estimate how many manuscripts contribute to the theory and to lower the bias in estimations in the subsequent machine learning step. Figueroa, Zeng-Treitler, Kandula, & Ngo (2012) tested over 500 models and found sample sizes between 80 to 560 to achieve acceptable error rates, so we selected two separate samples of 300 that could be combined to one sample of 600 should either sample not be sufficient. This also allowed evaluation of

differences between the two samples in the expected case where a sample of 300 was enough. Once the sample set is drawn, the researchers endeavor to retrieve the full-text versions of the manuscripts in the sample.

Next, (2b) inclusion criteria are developed that have the potential to identify relevant manuscripts. These should reflect the goals of the project (Templier and Paré 2015; Rowe 2014). Once the criteria are developed, (2c) two or more coders examine the random sample of full-text manuscripts and code each as contributing (relevant to the review and should be included) or noncontributing (irrelevant to the review and should not be included). As with all literature reviews, the review will rise or fall with the justification for the inclusion criteria, and Appendix B contains the criteria we used to develop a TAM theory corpus for this article. The decision to include or exclude a manuscript in the sample will become the "target" (dependent variable) for our machine learning algorithms. As argued above, the article manuscript is important, but it is managed at a later step by human assessment, not machine assessment.

Then, with a developed sense of the theory and the manuscripts that make an empirical contribution to it, (2d), a set of features that have the potential to distinguish contributing manuscripts from noncontributing manuscripts, is developed for use by the machine learning algorithm. Table D1 contains the features used to test ADIT with TAM. These are features of the manuscripts and their discourse, *not just attributes of the theory*, although in practice the two are deeply intertwined.

Most features fall into one of two categories. The first category is the rhetorical structure of the manuscript itself—in other words, the way that the manuscript employs elements of the theory in its rhetoric. For example, does the manuscript title use the name of the theory? Does the abstract include one or several construct names such as *ease of use* and *usefulness* from the theory? Combined with the year of publication the explicit presence or absence of construct name may be predictive, as the rhetorical structure required to publish may change over time.

The second category comprises attributes that reflect the position of the manuscript within the theory ecosystem, such as its impact, which reflects the citations it has received. One attribute that captures aspects of the network structure around a manuscript is the article-level Eigenfactor for each manuscript in L_2, which evaluates its likelihood of being central in the L_2 network (see Appendix D). In Figure 2, P_4 receives a

---

[4] Arguably, TAM was initiated by Davis' 1986 dissertation (sometimes cited as 1985). Our analysis of the random sample of TAM-citing articles found that less than 1% of

contributing articles were missed when excluding the dissertation.

higher proportion of citations from other L_2 manuscripts ($P_3-> P_4$), suggesting a higher centrality in L_2. We also use an attribute for detecting low-citation manuscripts (e.g., new manuscripts or manuscripts that question the theory) by evaluating the *theory attribution ratio* of a manuscript, that is, the sum of Eigenfactors for cited manuscripts *that exist in L_2* divided by the total number of manuscripts cited. For P_3 this would be

$$\frac{\sum Eigenfactor\ (P_4)}{n}, \tag{1}$$

where n = 3. This way, a manuscript is credited with a higher score because of intimate knowledge of the most influential manuscripts that cite the L_1 manuscripts as well as a focus on this literature over other literatures, even if their own manuscript for any reason is not nor will ever be highly cited. In this example, P_2 will have a lower *theory attribution ratio* than P_3 because it does not cite any manuscripts that could potentially contribute to the theory. The L_1 manuscripts are not included in the algorithm because they are outliers and already known. The *theory attribution ratio* becomes one of several features examined by the machine learning algorithm

3. Identification of empirically contributing manuscripts (corpus construction): ADIT was implemented with a web interface allowing a user to create an account and specify the theory-originating manuscript(s), L_1. A web crawler then downloaded L_2 manuscripts and processed these to specify and download L_3 manuscripts. Once all manuscripts were downloaded, a random sample of L_2 manuscripts was drawn for coding targets to (3a) train the ADIT machine learning algorithm. Once the ML algorithm was finished, the manuscripts found by coders to fit the inclusion criteria and the manuscripts found by the ML to fit the criteria were presented to the user along with statistics of likely retrieval success based on cross-validation. ADIT then continues to monitor MAS regularly and expand the theory ecosystem as necessary and notify the user when a new manuscript likely to fit their inclusion criteria has been published. See Appendix D for more details and a screenshot of the application.

For simplicity, ADIT's machine learning components are implemented through Weka (Bouckaert et al., 2013; Hall et al., 2009). While analysis of the citation network is conducted for levels L_1-L_3, only L_2 manuscripts are assumed to be potential contributors to the theory because they directly cite one or more L_1

manuscript. Our intuition was that it is hard for authors to justify extending a theory without providing a citation for that theory, and our earlier empirical analysis showed that of the 420 potential contributions to TAM listed in the 16 TAM reviews, 418 cited one or both original TAM articles (99.5%).[5] This provides some support for the assertion that if the L_1 manuscripts are carefully selected, only L_2 manuscripts will be theory contributing. Currently, ADIT retains L_3 manuscripts in the database to avoid redownloading these in the future, but once they are used in the network analysis, they are ignored for the purposes of the focal theory review. It is possible that adding another level of citations, L_4, would further improve performance, but this would come at a steep price, as the size of the citation network would grow by an order of magnitude—for TAM, this would mean the download of half a million extra manuscript records.

Finally, ADIT uses the set of attributes and the results of the contributing/noncontributing coding to (3b) apply the machine learning algorithm to categorize the remaining L_2 manuscripts as empirically contributing or noncontributing. The ADIT approach may be used with any number of machine learning algorithms, and our testing indicates that many algorithms provide equivalent results. We used Bayesnet, a versatile approach where nodes represent random variables, often with discrete sets of values. Links in the net represent conditional probabilities for the value of a node given the values of adjacent nodes (Charniak, 1991; Pearl, 2014).

3.4 Assessment of the ADIT Proof-of-Concept

We assessed ADIT using TAM as our theory of interest. We used the benchmark set of manuscripts described previously for our analysis. Specifically, we began with the two sets of 300 manuscripts randomly drawn from L_2 of the TAM ecosystem in Table A1 (Appendix A). As described above, two raters coded these manuscripts according to whether they should be selected as relevant to a research review or not. These codes and the attributes in Table D1 were used to train the machine learning algorithm. The algorithm was then tested using 10-fold validation (See Appendix D).

For both random samples, the machine learning algorithm performed well. Three metrics are commonly used to assess the performance of machine learning algorithms (Jurafsky & Martin, 2008; Larsen & Bong, 2016; Swets, 1988). Comprehensiveness is

---

[5] We note that there is likely selection bias in this evaluation because not citing a foundational manuscript would reduce the probability of a manuscript being included in a review. Nevertheless, we surmise that a manuscript attempting to contribute to a theory that does not cite the theory's

foundational manuscript(s) may also have other structural problems making them less likely to be included in theory reviews.

the number of true positives divided by the number of true positives plus the number of false negatives (see Table 1 for definitions). Precision is operationalized as the number of true positives divided by the number of true positives plus the number of false positives (see Table 1 for definitions). Both comprehensiveness and precision can be thought of as probabilities— comprehensiveness is the probability of finding a manuscript that fits the inclusion criteria and thus should be included in the review article, while precision is the probability that a manuscript identified by the technique is one that fits the inclusion criteria. F_1-score is the harmonic mean of precision and comprehensiveness and represents the overall performance of the technique as a balance between comprehensiveness and precision, punishing underperformance in one measure relative to the other. Area under the curve (AUC) evaluates the success of the algorithm for several cut-off points mapped into a receiver operating characteristics curve; AUC is not applicable to keyword searches. As shown in Table 2, comprehensiveness, precision, and F_1-scores were all above 0.80, indicating that the discourse approach using ADIT is an effective technique (Swets, 1988).

We argued that quality should be assessed on a case-by-case basis (as one does in survey research) so we assessed the quality of all 65 manuscripts identified by ADIT in Random sample #1. This included 5 non-peer-reviewed manuscripts, 15 conference papers, and 45 journal articles. Two authors who have served as a senior editor at a journal in the AIS Senior Scholars' Journal Basket independently assessed whether the methods in each manuscript were of sufficient quality to be included in a review. They agreed on all but one manuscript (98% agreement) and the disagreement was resolved.

If we use a standard of requiring evidence to exclude an article (which is the standard used by survey research that we advocate), 62 manuscripts (95%) of the manuscripts were of sufficient quality. Two manuscripts used a single-item construct, and one had reliabilities below 0.70 and thus would be excluded due to quality concerns. If we use a standard of requiring evidence to include an article, then an additional six manuscripts that failed to report construct reliabilities would be excluded, resulting in 56 manuscripts (86%) being of sufficient quality. All six manuscripts used previously validated items, so we would include these manuscripts if we were doing the analysis, but other researchers might disagree.

Table 2. ADIT and Conventional Approach Assessment Results

Method		Evaluative set	Comprehensiveness	Precision	F_1-score	AUC
Discourse approach Using ADIT		Random #1	**.840**	**.833**	**.835**	.790
		Random #2	**.819**	.815	**.816**	.811
Conventional Approach	Using keywords ("technology acceptance model" *OR* TAM)	Random #1	.500	.727	.593	n/a
		Random #2	.465	.673	.550	n/a
	Using keywords ("technology acceptance model" *OR* TAM) *AND* "usefulness"	Random #1	.219	.700	.333	n/a
		Random #2	.225	**.842**	.356	n/a
	Using keywords ("technology acceptance model" *OR* TAM) *AND* "ease of use"	Random #1	.219	.737	.337	n/a
		Random #2	.239	.810	.370	n/a
Note: Numbers in **bold** indicate the two highest in each column.						

3.5 Assessment of the Conventional Approach

While the scores for ADIT in Table 2 are promising, they are hard to assess in absolute terms; we also need to evaluate how well the conventional approach works using the same metrics. To test the effectiveness of the conventional approach, we conducted our own analysis, as described in Appendix C. We performed keyword searches using

TAM keywords on the databases most commonly used by previous review articles and found millions of manuscripts when we searched the full text of manuscripts (see Table C1 in Appendix C). Terms appearing in the title, abstract, and keywords are intended to convey the central message of the article (Larsen, Monarchi, Hovorka, & Bailey, 2008) and may be more likely to signal that the article contributes to the theory rather than simply including a citation. Many databases enable the user to restrict

the search to the title, abstract, or keywords (but not all—e.g., Google Scholar does not at the time of this writing). Focusing the search in this way reduces the number of manuscripts (see Table C1), but still results in potentially thousands more manuscripts than are relevant.

However, does constraining search to just the title, abstract, and keywords reduce comprehensiveness by unintentionally omitting relevant manuscripts? We used our random samples to evaluate various search strategies and found that using the most common keyword search terms used by the published review articles ("technology acceptance model" *or* TAM) and constraining the search to only the title and abstract would find 50% or less of the relevant TAM contributing manuscripts. When combining this search with another keyword ("technology acceptance model" *or* TAM) *and* "usefulness") less than 23% of the relevant TAM manuscripts are found. Boell and Cecez-Kecmanovic (2014) note that one of the foundation articles for TAM (Davis, 1989), does not include the acronym TAM or the words "technology acceptance model" so it would not be found by keyword search.

Boeker et al. (2013) provided perhaps one of the most striking conclusions on the use of Google Scholar (GS) for systematic literature review, showing an atrocious precision of 0.0013 when evaluating its ability to find a set of manuscripts using the search strategy used in systematic literature searches— 99.87% of all manuscripts found by GS were not relevant to the study. Providing supporting evidence for Boeker et al.'s conclusions, Yousfzai, Foxall, & Pallister (2007), one of the 16 reviews examined for this study, reported a precision of 0.0026 (99.74% of all manuscripts returned were not relevant). If these numbers generalized to a review of TAM manuscripts, finding the 420 manuscripts included in the 16 review articles would require examination of over 400,000 manuscripts. Finding the estimated 1,590 TAM manuscripts available at the end of 2012 would require careful retrieval of 1.5 million manuscripts, suggesting that the main reason the conventional approach is seen by many to be working is that we simply cannot appraise what we do not know. It also provides strong evidence that careful boundary identification through citation analysis can vastly simplify the review process.

Table 2 also shows the performance of the conventional approach using different keyword searches (searching for the listed keywords in the title, abstract or manuscript keywords). Three findings are worth noting. First, there is a striking difference in the comprehensiveness between the two approaches. The discourse approach is noticeably more comprehensive than the conventional approach. The results suggest that the conventional approach

using the widest possible keyword search will miss approximately half of all relevant manuscripts, compared to only about 17% for the discourse approach. As more constrained keyword searches are used to limit the number of manuscripts, comprehensiveness drops even further, so that approximately 75% of relevant manuscripts are missed.

This likely explains our findings that all of the six TAM reviews that claimed comprehensiveness *and* also reported both total manuscripts retrieved by search query and the total included manuscripts, without exception missed the majority of available manuscripts (see Appendix A). For example, when Turner et al. (2010) used the query "technology acceptance model *and* usage" they would have had little ability to understand the implications of this search query—the steep cost to comprehensiveness that came long before they had the opportunity to employ their inclusion criteria.

Second, the approaches differ slightly in precision. While the discourse approach is slightly more precise for the widest possible keyword search, precision does not practically differ between the discourse and conventional approaches for more constrained keyword searches.

Third, the F_1-score is a measure of the overall accuracy that has been commonly used in past information retrieval research. The discourse approach outperforms the conventional approach, regardless of the type of keyword search used. Even though the discourse approach employs vastly more complex search criteria then the conventional approach, increased complexity in the conventional approach is associated with lower comprehensiveness and F_1-scores.

4   Discussion

As the volume of research increases exponentially, how can we as researchers be confident that we are identifying and selecting past research for analysis that is representative of our phenomenon of interest? How can we ensure that we are not like the metaphorical blind men examining the elephant, each of whom is confident in their conclusions, but has missed the entirety of the phenomenon?

We began this research perspective article by adopting the Habermasian (1990) principle of ideal speech (in which all voices relevant to a phenomenon of interest are heard) as an important foundation to identifying and selecting prior research. From our viewpoint, identifying all relevant voices is important; otherwise, we run the risk of omitting important parts of the metaphorical elephant of past research, which can lead to biased conclusions. In situations where the volume

of past research makes it impractical to select all relevant voices, we argue that it is important to listen to a sample that is *representative* of all relevant voices. We acknowledge that there are other viewpoints, but from this starting point, we examined the processes of identifying a corpus of research and selecting a set of relevant manuscripts that should be considered for analysis in a research review article. We examined two distinct approaches: the conventional approach, which uses keyword searches based on elements of the theory, and the discourse approach, which uses citations to build a theory ecosystem and random sampling coupled with machine learning to select relevant manuscripts. Our results show that the discourse approach outperforms the conventional approach; it is more comprehensive and at least as precise. Table 3 provides a summary of the major differences along with an evaluation of the two approaches.

When we have presented our approach to colleagues, the most common objection we hear is to the use of random sampling. There seems to be an inherent belief that authors should deliberately choose manuscripts to be included, rather than leaving the sample to chance, because some elements of the population are more desirable than others and subjective judgment is important (cf. Babbie, 2013). Our response is to ask whether judgment sampling should be used when drawing a sample for a survey. When it comes to survey research, no respectable journal would publish an article that argued that researchers should deliberately use a judgment sample to decide what data to include and what data not to include because such a judgment sample would be inherently biased and not necessarily representative of the phenomenon of interest. Such judgment sampling procedures have led to well-known failures (e.g., the Digest's prediction of Roosevelt's loss in the 1936 election and Gallup's prediction of Truman's loss in the 1948 election). Researchers who use judgment sampling do so for the best of reasons, but unfortunately, the results are biased and it is impossible to know how biased the sample is (Statistics Canada 2013).

There are two overall messages from our results. First, we demonstrated that there are fundamental flaws in the conventional approach when used in large research domains, at least as it was used in the six comprehensive studies that had inclusion criteria equivalent to our own. In these six cases, the conventional approach was neither comprehensive nor precise. Our analysis of these six TAM articles found that they failed to identify 82.5% of relevant prior research on average and require its authors to examine 8.3 irrelevant manuscripts for every relevant manuscript found. For the overall set of 16 reviews, these reviews often focused exclusively on journal articles or even top journal articles, meaning that they were less likely to see the entire metaphorical elephant of past research. As Watson (2015, p. 187)

notes, "the crux of the problem is that we have last century's approach to knowledge management." We believe that this calls into question the validity of the conventional approach for use in large research areas like TAM.

Second, we found that the discourse approach produced a more comprehensive corpus of relevant manuscripts with equivalent or better precision. Because it uses random sampling of manuscripts, the manuscripts in a sample based on the discourse approach are more likely to be representative of the entire theory ecosystem. In contrast, a nonrandom sample, especially one that is deliberately focused on journal articles, will produce a biased sample and lead to poor review articles (Banks, Kepes, & McDaniel, 2015; Kepes et al., 2012; Kepes & McDaniel, 2015; Webster & Watson, 2002). We long ago recognized the fallacy of deliberately selecting survey respondents, as opposed to random sampling followed by a case-by-case quality check. It is time to apply the same approach to literature reviews; otherwise, we end up with samples that accurately identify the most obvious part of the literature (e.g., the metaphorical elephant's legs), while missing the other parts of the elephant. Therefore, we recommend that authors use a discourse approach for large theory ecosystems where complete enumeration is infeasible (See Table 3 for details on appropriate use settings). We also advocate that authors of review manuscripts clearly and transparently articulate the approach to boundary identification and corpus construction and include description of how well the literature reviewed represents the theory ecosystem. Simply knowing how comprehensive a review article is and being able to document this for others represents a major step forward (vom Brocke et al., 2015).

In this article, we have focused on quantitative research, driven primarily by the fact that the research domain we use as an example (TAM) and the review articles about it are primarily quantitative. This is not to disparage reviews undertaken from a qualitative perspective. We believe there are few fundamental differences in the need to identify relevant manuscripts (and avoid accidentally omitting manuscripts), but there may be important differences in the approach to selection. Random selection or machine learning selection means that it is possible that the set of selected manuscripts would omit seminal manuscripts (i.e., early articles that have received many citations over time) or other articles that qualitative researchers may see as critical. When performing what Paré (2015) calls *cumulative reviews* or *aggregative reviews*, omitting the voice of one seminal manuscript in a set of several hundred would have little impact on the validity, especially since the "voice" of that article will have influenced and been repeated in many other manuscripts. There may be articles that researchers deliberately choose to include and thus include them independent of random sampling.

Table 3. The Pros (+) and Cons (-) of Conventional and Discourse Approaches

Criteria	Conventional	Discourse
Community experience	+ High level of experience and understanding.	- New approach; little experience.
Technology availability	+ Accessible through most academic libraries. Accessible through search engines.	+ Approach may be applied through reverse citation search. + Random selection for large theories is simple to implement. - Machine learning requires specialized software not widely available yet. Computationally more challenging.
Contextual factors	- Relies on consistent language in manuscripts. Not able to find research using different words for same concepts unless researcher is fully aware of these terms. + Community of reviewers who have had good experiences with the approach. - Unable to use statistical techniques to estimate coverage.	- Relies on discourse. It is particularly relevant for research discourse on a specific theory, such as TAM, that has clearly defined point(s) of origin. Not all research streams have such clearly defined points of origin, particularly those that are interdisciplinary, such as coordination theory. + Easy to explain + Enables estimation of percent of relevant manuscripts included in review due to random sampling (comprehensiveness). Confidence interval may be calculated.
Comprehensiveness and precision	- Boundary identification omitted 50-80% of potentially relevant manuscripts, depending on the specific keywords used. +/- 67-85% of the papers found were relevant for corpus construction, depending on the specific keywords used.	+ Boundary Identification omitted 15-20% of potentially relevant manuscripts + 80-85% of the papers found were relevant for corpus construction
Overall accuracy	- Mid-range: The harmonic mean of precision and comprehensiveness (F_1-score) in our controlled test ranged between 0.333 and 0.593. - Low: The self-reported precision and our estimated comprehensiveness yielded F_1-scores for reviews in real settings ranging between 0.006 and 0.184. This was often due to low precision, likely due to full-text manuscripts containing keywords in different contexts.	+ High: The harmonic mean of precision and comprehensiveness in our controlled test ranged between 0.790 and 0.811. + Likely to transition better into full-text evaluation as long as the algorithms know what constitutes title, abstract, keywords, and the body of the manuscript. + Will automatically assign a weight of zero to a keyword such as "usefulness" if its use in the body of a manuscript has a deleterious effect on the accuracy of the algorithm.
Theory size	+ Appropriate for small areas for which complete enumeration is feasible.	+ Appropriate for small areas for which complete enumeration is feasible as long as theory origination is clear. + Appropriate for larger areas for which complete enumeration is infeasible.

Table 4. Size of Information Systems Research Areas

Theory	Foundation article(s)	Number of manuscripts using term	Number of citations to foundation article(s)
"End-user computing"	Doll & Torkzadeh (1988)	30,000	2,886
"Information systems success"	DeLone & McLean (1992)	19,400	10,850
"Productivity paradox"	Brynjolfsson (1993)	16,500	3,113
"Adaptive structuration theory"	DeSanctis & Poole (1994)	5,080	3,921
"Task technology fit"	Goodhue & Thompson (1995)	11,800	4,272
"Computer self-efficacy"	Compeau & Higgins (1995)	20,800	5,671
"Knowledge management systems"	Davenport & Prusak (1998); Alavi & Leidner (2001)	72,100	5,148 10,559
"Virtual teams"	Jarvenpaa & Leidner (1998)	57,900	3,962
"E-commerce" *and* "trust"	McKnight, Choudhury, & Kacmar (2002); Gefen, Karahanna, & Straub (2003)	258,000	3,809 5,937
"Resource-based view" *and* "information systems"	Bharadwaj (2000)	27,500	4,332
"UTAUT"	Venkatesh et al. (2003)	21,100	18,632

This also opens the question of how comprehensive research review articles should be. Many past review articles using the conventional approach (including our own) have claimed to be a complete enumeration of the population, although our current analyses show that this is likely not true, due to inherent limitations of the conventional approach itself. Given that we can now estimate the size of the population, how large a sample should we take, if complete enumeration is not possible? If we are conducting a quantitative review, there are many good survey research techniques that we can use for determining a reasonable sample size for a review. In general, to assess both direct and mediated relationships of moderate size, these techniques suggest a sample of 75-150 (Fritz and MacKinnon 2007), but of course, this depends on the specific research area, and the desired effect size and power.

TAM is one of our field's most cited research areas, so one might argue that it is larger than other research areas. There are now over 70,000 research manuscripts that invoke the name of the theory. How does this compare to other information systems research areas? Table 4 shows that TAM is larger than some research areas but smaller than others. In other words, TAM is not an outlier. Most established research areas have thousands of potentially relevant manuscripts. In comparison to theories originating in other disciplines, TAM is still small in relation to diffusion of innovation theory (Rogers, 1962; Rogers, 1983; Rogers, 2003; Rogers, 2010) and about the same size as the theory of reasoned action (Fishbein & Ajzen, 1975) and the theory of planned behavior (Ajzen, 1991).

5   Limitations

One limitation of ADIT (though not of the discourse approach in general) is that it is primarily appropriate for large research domains. For small domains where complete enumeration of the identified manuscripts is possible, ADIT may add little value.

A second limitation is that ADIT requires a set of one or more foundation articles that subsequent research builds on. This is the case in most research areas but may not be the case for emerging areas that lack a well-accepted theoretical foundation (e.g., RFIDs, big data, IoT, smart cities). Emerging research areas are likely to be relatively small and thus unlikely to benefit from ADIT.

To assess the discourse approach, we had to implement a specific machine learning algorithm within the ADIT design science instantiation. Not all researchers may wish to invest the time and finances needed to create a comprehensive corpus of

manuscripts to be included in a research review article. For these authors, we advocate the use of a citation-based search technique followed by random selection to ensure their corpus is representative of the population of manuscripts. Random selection is simple and straightforward and requires little additional effort.

Finally, the goal of this article is not to evaluate the performance of different algorithms, but rather to demonstrate that the machine learning approach outperforms the conventional approach to boundary identification. This exclusion may represent an interesting avenue for future work.

6   Future Work on ADIT

The number of research publications continues to grow each year (vom Brocke et al., 2015; Webster & Watson, 2002), making the task of identifying relevant manuscripts harder. Proper boundary identification, especially with larger research areas, is a necessary component of a high-quality review (Templier & Paré, 2015; Webster & Watson, 2002). Thus, the problem of understanding the entire elephant will become more difficult over time as it becomes impossible to enumerate the entire population of prior research. We believe that the discourse approach using random sampling and machine learning techniques provides a more comprehensive, more precise, and less biased approach to identifying and selecting relevant manuscripts.

One of our goals is to make ADIT available to researchers as a web service so that anyone can use it. As ADIT is further evolved and new features are added, comprehensiveness, precision, and ultimately usefulness can be expected to improve. ADIT is currently composed of a web crawler, a database, domain experts, and machine learning analytics that constantly update the database as new manuscripts citing a specified theory are added to a literature database (e.g. MAS). Future work will include experimentation with other machine learning approaches, such as artificial neural networks (ANNs), or support vector machines (SVMs), which will make it possible to influence outputs by making choices in relation to domain knowledge (e.g., architecture, error measures, and outlier definition).

The goal of ADIT is to aid researchers in identifying ecosystems of domain knowledge (e.g., theories, concepts, and phenomena) and in determining which manuscripts in the ecosystem are relevant to a review of that domain. Domain knowledge may include specific network characteristics—for example, many relevant publications or a high degree of citation network overlap may indicate convergence into a single theory or divergence into competing theories. Future work on ADIT may include knowledge

management tools supporting implementation of forward and backward chaining in a meaningful manner. Future versions should also consider use of author names and their centrality in the theory ecosystem.

We also note that after our analysis of TAM reviews was concluded, Mortenson and Vidgen (2016) published a review based on a computational literature review (CLR) technique. CLR used a conventional approach (i.e., keyword search term "technology acceptance model") to construct a review corpus from SCOPUS. The corpus of 3,386 manuscripts was analyzed using latent Dirichlet allocation, a topic model, to illustrate the topic content, impact, and the social network of included manuscripts. It might be fruitful in the future to combine the CLR approach with ADIT to construct a more accurate set of TAM manuscripts available for a computational impact, content, and structure analysis. If our random sample of manuscripts from MAS generalizes to SCOPUS, the majority of manuscripts analyzed by Mortenson and Vidgen's (2016) were not TAM-contributing manuscripts, leaving their findings in question. Using ADIT prior to using CLR may address such concerns.

Ultimately, it is our hope that ADIT will track multiple theories of interest on a real-time basis and make these theory-specific corpora available through a web portal with integrated visualizations. Once the ecosystems, theory-citing manuscripts, and manuscripts relevant for analysis according to multiple theories are available, overlaps between theories may be empirically evaluated to further our understanding of theory creation, integration, and movement between disciplines.

7   Conclusion

As review papers both of theories and of research domains become increasingly important, the difficulty of constructing the corpus of prior research will increase as the volume of research continues to grow exponentially (Larsen, Voronovich, Cook, & Pedro, 2013; vom Brocke et al., 2015). There are many ways to conduct a review, depending on the goals of the project. In all cases, the construction of the corpus of literature to be analyzed is a critical step that often receives scant attention.

We investigated two approaches to corpus construction: the conventional approach and the discourse approach. We showed that in the case of TAM review articles, past uses of the conventional approach failed to identify *most* of the relevant research. Our applications of the two approaches found that use of the discourse approach produced a more comprehensive set of potentially relevant manuscripts that was as precise as the conventional approach. We further found that when applied in real

settings, the conventional approach did poorly on both precision and comprehensiveness.

Therefore, we recommend that researchers use the discourse approach, not the conventional approach. The discourse approach to corpus construction enables us to better understand the entire elephant that makes up past research, as well as how much of the elephant remains unexplored. The result will be better identification of knowledge, which will solidify the foundation for review articles or empirical manuscripts that draw on past theory and research.

References

Aamodt, M. (2015). *Industrial/organizational psychology: An applied approach*. Boston, MA: Cengage Learning.

Abbasi, A., Zhang, Z., Zimbra, D., Chen, H., & Nunamaker Jr, J. F. (2010). Detecting fake websites: The contribution of statistical learning theory. *MIS Quarterly*, *34*(3), 435-461.

Acton, Q. A. (2012). *Issues in sociology and social work: aging, medical, and missionary research and application*. Atlanta, GA: Scholarly Editions.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179-211.

Alavi, M., & Leidner, D. E. (2001). Research commentary: Technology-mediated learning—A call for greater depth and breadth of research. *Information Systems Research*, *12*(1), 1-10.

Babbie, E. R. (2013). *The basics of social research*. Boston, MA: Cengage Learning.

Bacchetti, P., Wolf, L. E., Segal, M. R.,& McCulloch, C. E. (2005). Ethics and sample size. *American Journal of Epidemiology*, *161*(2), 105-110.

Banks, G. C., Kepes, S., & McDaniel, M. A. (2015). Publication bias: Understanding the myths concerning threats to the advancement of science. In C. E. Lance & R. J. Vandenberg (Eds.), More statistical and methodological myths and urban legends (pp. 36-64). New York, NY: Routledge.

Baskerville, R., & Pries-Heje, J. (2010). Explanatory design theory. *Business & Information Systems Engineering*, *2*(5), 271-282.

Berger, K. S. (2003). *The developing person through childhood and adolescence*. London: Macmillan.

Berlin, J. A., & Ghersi, D. (2005). Preventing publication bias: Registries and prospective meta-analysis. In H. Rothstein, A. Sutton & M. Borenstein *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 35-48). Hoboken, NJ: Wiley.

Bharadwaj, A. S. (2000). A resource-based perspective on information technology capability and firm performance: an empirical investigation. *MIS Quarterly*, *24*(1), 169-196.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Boeker, M., Vach, W., & Motschall, E. (2013). Google Scholar as replacement for systematic literature searches: Good relative recall and precision are not enough. *BMC Medical Research Methodology, 13*(1), 1-12.

Boell, S. K., & Cecez-Kecmanovic, D. (2014). A hermeneutic approach for conducting literature reviews and literature searches. *Communications of the Association for Information Systems*, *34*, Paper 12.

Bostrom, R. P., Gupta, S., & Thomas, D. (2009). A meta-theory for understanding information systems within sociotechnical systems. *Journal of Management Information Systems*, *26*(1), 17-48.

Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2013). *Waikato Environment for Knowledge Analysis (WEKA) manual for version 3-7-8*. Retrieved from http://statweb.stanford.edu/~lpekelis/13_datafest_cart/WekaManual-3-7-8.pdf

Brynjolfsson, E. (1993). The productivity paradox of information technology. *Communications of the ACM*, *36*(12), 66-77.

Carson, E. R., & Cramp, D. G. (2013). *Computers and control in clinical medicine*. Berlin: Springer.

Charniak, E. (1991). Bayesian networks without tears. *AI magazine*, *12*(4), 50-50.

Chiasson, M. W. (2015). Avoiding methodological overdose: a declaration for independent ends. *Journal of Information Technology*, *30*(2), 174-176.

Chuttur, M. (2009). Overview of the technology acceptance model: origins, developments and future directions. Retrieved from https://aisel.aisnet.org/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1289&context=sprouts_all

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Compeau, D. R., & Higgins, C. A. (1995). Computer self-efficacy: Development of a measure and initial test. *MIS Quarterly*, *19*(2), 189-211.

Cortes, C., Jackel, L. D., Solla, S. A., Vapnik, V., & Denker, J. S. (1994). Learning curves: asymptotic values and rate of convergence. In M. Jordan, Y. LeCun, & S. Solla (Eds.), *Advances in Neural Information Processing Systems* (pp. 327-334). Boston, MA: MIT Press.

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: SAGE.

Davenport, T. H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know*. Cambridge, MA: Harvard Business Press.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319-340.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, *35*(8), 982-1003.

DeLone, W. H., & McLean, E. R. (1992). Information systems success: The quest for the dependent variable. *Information Systems Research*, *3*(1), 60-95.

Dennis, A. R., & Valacich, J. S. (2014). A Replication Manifesto. *Association for Information Systems Transactions on Replication Research*, 1, 1-5.

Dennis, A. R., Wixom, B. H., & Vandenberg, R. J. (2001). Understanding fit and appropriation effects in group support systems via meta-analysis. *MIS Quarterly, 25*(2), 167-197.

de Guinea, A. O., & Paré, G. (2017). What literature review type should I conduct? In R. Galliers & M.-K. Stein (Eds.), *The Routledge Companion to Management Information Systems* (pp. 73-82). New York, NY: Routledge.

DeSanctis, G., & Poole, M. S. (1994). Capturing the complexity in advanced technology use: Adaptive structuration theory. *Organization science*, *5*(2), 121-147.

Dohan, M. S., & Tan, J.(2013). Perceived usefulness and behavioral intention to use consumer-oriented web-based health tools: A meta-analysis," *Proceedings of the Nineteenth Americas Conference on Information Systems*.

Doll, W. J., & Torkzadeh, G. (1988). The measurement of end-user computing satisfaction. *MIS Quarterly*, *12*(2), 259-274.

Fan, W., Gordon, M. D., Pathak, P., & Pathak, P. (2005). Genetic programming-based discovery of ranking functions for effective web search. *Journal of Management Information Systems*, *21*(4), 37-56.

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC medical informatics and decision making 12*(1), 1-10.

Fishbein, M., & Ajzen, I. (1975). *Beliefs, attitude, intentions and behaviours*. Reading, MA: Addison-Wesley.

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, *18*(3), 233-239.

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: An integrated model. *MIS Quarterly*, *27*(1), 51-90.

Giddens, A. (2013). *The constitution of society: Outline of the theory of structuration*. Hoboken, NJ: Wiley.

Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS Quarterly, 19*(2), 213-236.

Garfield, E. 2006. The history and meaning of the journal impact factor, *Jama, 295*(1), 90-93.

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, *26*(2), 91-108.

Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly*, *30*(3), 611-642.

Nielsen, T. H., & Habermas, J. (1990). Jürgen Habermas: Morality, society and ethics: An interview with Torben Hviid Nielsen. *Acta Sociologica*, 33(2), 93-114.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10-18.

Han, L., & Jin, Y. (2009). A review of technology acceptance model in the e-commerce environment. *Proceedings of the ICMECG'09. International Conference on Management of e-Commerce and e-Government*.

Han, S. (2003). *Individual adoption of information systems in organizations: A literature review of technology acceptance model*. Turku, Finland: Turku Centre for Computer Science.

Holden, R. J. and B.-T. Karsh (2010). The technology acceptance model: Its past and its future in health care. *Journal of Biomedical Informatics, 43*(1), 159-172.

Harzing, A.W. (2007) *Publish or Perish*. Available from *https://harzing.com/resources/publish-or-perish*

Holden, R. J., & Karsh, B. T. (2010). The technology acceptance model: its past and its future in health care. *Journal of Biomedical Informatics*, *43*(1), 159-172.

Hsiao, C. H., & Yang, C. (2011). The intellectual development of the technology acceptance model: A co-citation analysis. *International Journal of Information Management*, *31*(2), 128-136.

Jarvenpaa, S. L., & Leidner, D. E. (1998). An information company in Mexico: Extending the resource-based view of the firm to a developing country context. *Information Systems Research*, *9*(4), 342-361.

Jones, M. R., & Karsten, H. (2008). Giddens's structuration theory and information systems research. *MIS Quarterly*, *32*(1), 127-157.

Martin, J. H., & Jurafsky, D. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Pearson/Prentice Hall.

Kalayeh, H., & Landgrebe, D. A. (1983). Predicting the required number of training samples, *IEEE Transactions on Pattern Analysis and Machine Learning, 5*(6), 664-667.

Kattan, M. W., Adams, D. A., & Parks, M. S. (1993). A comparison of machine learning with human judgment. *Journal of Management Information Systems*, *9*(4), 37-57.

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, *112*(24), 7426-7431.

Kepes, S., Banks, G. C., McDaniel, M., & Whetzel, D. L. (2012). Publication bias in the organizational sciences. *Organizational Research Methods*, *15*(4), 624-662.

Kepes, S., & McDaniel, M. A. (2015). The validity of conscientiousness is overestimated in the prediction of job performance. *PLoS One*, *10*(10), https:// journals.plos.org/plosone/ article?id=10.1371/journal.pone.0141468

King, W. R., & He, J. (2006). A meta-analysis of the technology acceptance model. *Information & Management*, *43*(6), 740-755.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174.

Larsen, K. R., & Bong, C. H. (2016). A tool for addressing construct identity in literature reviews and meta-analyses. *MIS Quarterly*, *40*(3), 529-551.

Larsen, K. R., Monarchi, D. E., Hovorka, D. S., & Bailey, C. N. (2008). Analyzing unstructured text data: Using latent categorization to identify intellectual communities in information systems. *Decision Support Systems*, *45*(4), 884-896.

Larsen, K. R., Voronovich, Z. A., Cook, P. F., & Pedro, L. W. (2013). Addicted to constructs: Science in reverse? *Addiction*, *108*(9), 1532-1533.

Lee, Y., Larsen, K.R, & Kozar, K. A. (2003). The technology acceptance model: Past, present, and future. *Communications of the Association for Information Systems, 12*(50), 752-780.

Lee, Y.-J., Ko, E., & Choo, H. J. (2015). Fashion consumer's acceptance of retail technology: a meta-analysis of tam in fashion retail context. *Global Fashion Management Conference at Florence Proceedings*.

Legris, P., Ingham, J., and Collerette, P. (2003) Why do people use information technology? A critical review of the technology acceptance model. *Information and Management, 40*, 191-204.

Li, Y., Qi, J., & Shu, H. (2007). A review on the relationship between new variables and classical tam structure. In L. Xu, A. Tjoa & S. Chaudhry (Eds.). *Ifip international federation for information processing*. Boston, MA: Springer.

Li, Y., Qi, J., & Shu, H. (2008). Review of relationships among variables in TAM, *Tsinghua Science and Technology, 13*(3), 273-278.

Ma, Q., and Liu, L. 2004. The technology acceptance model: A meta-analysis of empirical findings. *Journal of Organizational and End User Computing, 16* (1), 59-72.

Marangunić, N., & Granić, A. (2015). Technology acceptance model: A literature review from 1986 to 2013. *Universal Access in the Information Society, 14*(1), 81-95.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, *13*(3), 334-359.

Meadows, J. J. (1990). Theory in information science. *Journal of Information Science,* 16(1), 59-63.

Mortenson, M. J., & Vidgen, R. (2016). A computational literature review of the

technology acceptance model. *International Journal of Information Management*, *36*(6), 1248-1259.

Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T. R., & Mesirov, J. P. (2003). Estimating dataset size requirements for classifying DNA microarray data," *Journal of Computational Biology, 10*(2), 119-142.

Mullen, B. 2013. *Advanced Basic Meta-Analysis: Version 1.10*. New York, NY: Taylor & Francis.

Okoli, C. (2012). A critical realist guide to developing theory with systematic literature reviews. Available at https://ssrn.com/abstract =2115818

Page, L., Brin, S., Motwani, R., & Winograd, T. 1999. The pagerank citation ranking: bringing order to the web. Retrieved from http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Amsterdam: Elsevier.

Rivard, S. (2014). Editor's comments: The ions of theory construction. *MIS Quarterly*, *38*(2), 3-14.

Rogers, E. M. (1962). *Diffusion of innovations*. New York, NY: Free Press.

Rogers, E. M. (1983). *Diffusion of innovations* (3rd ed.). New York, NY: Free Press.

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York, NY: Free Press.

Rogers, E. M. (2010). *Diffusion of innovations*. NEW York, NY: Simon & Schuster.

Rosenthal, R. 1991. *Meta-Analytic Procedures for Social Research*. SAGE Publications.

Rowe, F. (2014). What literature review is not: Diversity, boundaries and recommendations. *European Journal of Information Systems, 23*(3), 241-255.

Schepers, J., & Wetzels, M. (2007). a meta-analysis of the technology acceptance model: investigating subjective norm and moderation effects, *Information & Management, 44*(1), 90-103.

Schryen, G., Benlian, A., Rowe, F., Gregor, S., Larsen, K., Petter, S., ... & Yasasin, E. (2017). Literature reviews in IS research: What can be learnt from the past and other fields? *Communications of the Association for Information Systems*, *41*, Paper 30.

Schultze, U. (2015). Skirting SLR's language trap: reframing the "systematic" vs. "traditional" literature review opposition as a continuum. *Journal of Information Technology*, *30*(2), 180-184.

Schwarz, A., Mehta, M., Johnson, N., & Chin, W. W. (2007). Understanding frameworks and reviews: A commentary to assist us in moving our field forward by analyzing our past. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, *38*(3), 29-50.

Sharma, R., and Yetton, P. 2001. An evaluation of a major validity threat to the technology acceptance model. *Proceedings of the 9th European Conference on Information Systems*.

Statistics Canada. (2013). *Non-Probability Sampling*. Retrieved from http://www.statcan.gc.ca/edu/power-pouvoir/ch13/nonprob/5214898-eng.htm#a3

Straub, D. W. (2009). Creating blue oceans of thought via highly citable articles. *MIS Quarterly*, *33*(4), 2-4.

Šumak, B., Heričko, M., & Pušnik. (2011). A meta-analysis of e-learning technology acceptance: The role of user types and e-learning technology types. *Computers in Human Behavior 27*(6), 2067-2077.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285-1293.

Tang, D., and Chen, L. 2011. A review of the evolution of research on information technology acceptance model. *International Conference on Business Management and Electronic Information Proceedings*.

Tate, M., Furtmueller, E., Evermann, J., & Bandara, W. (2015). Introduction to the special issue: The literature review in information systems. *Communications of the Association for Information Systems*, *37*, paper 5.

Templier, M., & Paré, G. (2015). A framework for guiding and evaluating literature reviews. *Communications of the Association for Information Systems*, *37*(1), 6.

Turner, M., Kitchenham, B., Brereton, P., Charters, S., & Budgen, D. (2010). Does the Technology Acceptance Model Predict Actual Use? A Systematic Literature Review. *Information and Software Technology*, *52*(5), 463-479.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425-478.

Vessey, I., Ramesh, V., & Glass, R. L. (2002). Research in information systems: An empirical study of diversity in the discipline and its journals. *Journal of Management Information Systems*, *19*(2), 129-174.

vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., & Cleven, A. (2009). Reconstructing the giant: On the importance of rigour in documenting the literature search process. *European Conference of Information Systems Proceedings*.

vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., & Cleven, A. (2015). Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. *Communications of the Association for Information Systems*, *37*, Paper 9.

Watson, R. T. (2015). Beyond being systematic in literature reviews in IS. *Journal of Information Technology*, *30*(2), 185-187.

Weber, R. (2012). Evaluating and developing theories in the information systems discipline. *Journal of the Association for Information Systems*, *13*(1), 1-8.

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, 26(2), 13-23.

Weick, K. E. (1995). What theory is not, theorizing is. *Administrative Science Quarterly*, *40*(3), 385-390.

West, J. D., Jensen, M. C., Dandrea, R. J., Gordon, G. J., & Bergstrom, C. T. (2013). Author-level Eigenfactor metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community. *Journal of the American Society for Information Science and Technology, 64*(4), 787-801.

West, J. D., Wesley-Smith, I., & Bergstrom, C. T. 2016. A recommendation system based on hierarchical clustering of an article-level citation network, *IEEE Transactions on Big Data, 2*(2), 113-123.

Willliam T. Grant Foundation (2009). *Request for research proposals: Understanding the acquisition, interpretation, and use of research evidence in policy and practice.* Retrieved from https://wtgrantfoundation.org/library/uploads/2016/01/2016-URE-Supplemental-Guidance. pdf

Williams, M. D., Rana, N., Dwivedi Y. K., & Lal, B. (2011). Is UTAUT really used or just cited for the sake of it? A systematic review of citations of UTAUT's originating article. *European Conference of Information Systems Proceedings.*

Wocjan, P., Janzing, D., & Beth, T. (2002). Required sample size for learning sparse Bayesian networks with many variables. arXiv preprint, cs/0204052.

Wu, J., & Lederer, A. (2009). A meta-analysis of the role of environment-based voluntariness in information technology acceptance. *MIS Quarterly, 33*(2), 419-432.

Wu, K., Zhao, Y., Zhu, Q., Tan, X., & Zheng, H. (2011). A meta-analysis of the impact of trust on technology acceptance model: Investigation of moderating influence of subject and context type. *International Journal of Information Management*, 31(6), 572-581.

Yong, E. (2012). Bad Copy. *Nature, 485*(7398), 298.

Yousafzai, S. Y., Foxall, G. R., & Pallister, J. G. (2007a). Technology acceptance: meta-analysis of the TAM—Part 1. *Journal of Modelling in Management, 2*(3), 251-280

Yousafzai, S. Y., Foxall, G. R., & Pallister, J. G. (2007b). Technology acceptance: A meta-analysis of the TAM—Part 2. *Journal of Modelling in Management, 2*(3), 281-304.

Zuk, O., Margel, S., & Domany, E. (2012). On the number of samples needed to learn the correct structure of a Bayesian Network. arXiv preprint, arXiv:1206.6862.

Appendix A. Existing TAM Reviews Using the Conventional Approach

A1. Existing TAM Reviews Using the Conventional Approach

We began by examining past review articles' use of the conventional approach (keyword search with backward chaining). We considered whether reviews using the conventional approach were comprehensive (i.e., identified the relevant manuscripts) and precise (i.e., how many of the identified manuscripts were not relevant) (Watson 2015).

We selected the technology acceptance model (TAM) because it is a well-established theory and a highly researched area of IS. According to Google Scholar, the foundational articles for TAM (Davis, 1989; Davis et al., 1989) had received 36,374 and 19,355 citations respectively by October 14, 2017. Of these, 4,020 and 2,550 citations, respectively, were received in 2016 alone, suggesting that TAM remains a highly cited theory. The extensive research on TAM has resulted in a substantial set of theory review articles to examine.

The past research review articles in this domain have predominantly been quantitative reviews (i.e., those using statistical results to draw conclusions; Schepers & Wetzels, 2007; Yousafzai et al, 2007b) rather than narrative reviews, (i.e., those that use descriptions to draw conclusions; Chuttur, 2009; Yousafzai , Foxall, & Pallister, 2007a). These articles thus focus on empirical manuscripts—manuscripts that provide quantitative data to support their conclusions. Theoretical commentaries that offer arguments not supported by data are important (especially for qualitative reviews), but are not generally included in the scope of these articles, because the purpose of a review is to drive what type of prior research is relevant. For these review articles, relevant manuscripts are those that provide an empirical contribution to the theory by testing, revising, or refuting it with quantitative empirical data.

To assess these past reviews, we had to adopt the same frame of reference: quantitative empirical manuscripts. While the specific operationalization of "empirical research" will differ based on the purpose of a research project, we developed a set of criteria to identify empirical manuscripts, based on our reading of these past review articles. To be considered an empirical contribution to a given theory, a manuscript must include at least one dependent variable and at least one independent variable from the focal theory; must report the statistical relationship between at least these two variables; and must *not* state that it is creating a new theory that is separate and distinct from the focal theory. These criteria (see Appendix B) are in line with theory reviews for meta-analytic research (Mullen 2013; Rosenthal 1991).

We began by examining published reviews of TAM. We employed the conventional approach (keyword search and backward chaining) to identify these articles because TAM and its components are at the core of the TAM review articles and thus will often appear in either the title or the abstract and also because the set was expected to be small, thereby allowing complete enumeration. Our sources were: ABI/Inform, ACM Digital Library, Business Source Premier, EBSCO, Google Scholar, IEEE Xplore, ISI Web of Knowledge (WOK), Microsoft Academic Search, and Science Direct. We employed the keywords "technology acceptance model," "TAM," "ease of use," "usefulness," "behavioral intention," and "intention to use" in combination with the keywords "review," "meta-analysis," and "meta-analysis." Once a review article was found, we read it to find references to other review articles. No additional review articles were found by reading citations. All but three review articles had "technology acceptance model" or "TAM" in the title, and the others contained one of these terms in the abstract. Our search was conducted on March 1, 2014. As an aside, we note that additional TAM review articles were published while this article was in preparation and under review (e.g. Lee, Ko, & Choo, 2015; Marangunić & Granić, 2015; Mortenson & Vidgen 2016). We address these articles in the Discussion section.

This search for TAM reviews identified 20 articles that contained a literature review or meta-analysis of TAM research. We selected only those articles whose goal it was to present a review of TAM in at least one setting. We excluded four review articles: Sharma & Yetton (2001), for not accounting for TAM manuscripts used; Yousafzai et al. (2007b), because it used a subset of another included study; Tang & Chen (2011), for including multiple theories without tracking membership; and Han & Jin (2009), for providing no comprehensive list of manuscripts used. For each article, we read the stated goal and coded the articles as having the goal of being comprehensive vs. some other narrower goal. This produced a set of 16 prior review articles for analysis, six of which were considered comprehensive and 10 designed to be narrow. Table A1 contains the articles that we categorized as comprehensive reviews. To be included, articles had to state goals such as performing an "exhaustive" (Lee, Kozar, & Larsen, 2003) or "comprehensive" (Yousafzai et al. 2007b) review that included "all available" manuscripts (Schepers & Wetzels 2007) or versions of such claims. To be included, articles also had to have search queries and inclusion criteria reflecting the goal of being comprehensive. Table A1 contains those studies that we classified as having the goal of being comprehensive.

Table A1. Comprehensive Past Literature Reviews

#	Article	Manuscripts included vs. found	Year range	Sources	Keywords	Inclusion criteria
[1]	Legris, Ingham, & Collerette. (2003)	22 of 80	1989-2001	Journals: MISQ, DS, MS, JMIS, ISR, and IM; "specialized databases and other sources on the web."	Not reported. Included backward chaining	TAM is used in an empirical study. The integrity of TAM is respected. The research methodology is well-described. The research results are available and complete.
[3]	Lee et al. (2003)	101 of ns	1986-2003	Social Science Citation Index, ABI/INFORM, and Business Source Premier. Also included were ICIS and HICSS conference proceedings and other papers published in interdisciplinary journals closely related to IS field.	Not Reported.	Not Reported.
[4]	Ma & Liu (2004)	26 of 91	1989-2003	"Top journals" (i.e., MISQ, ISR, IM, etc.) Search ACM and AIS libraries and major international conferences. ProQuest, EBSCO, and ResearchIndex at Google.	Not Reported.	Involved empirical testing of TAM directly or indirectly. Reported a sample size. Reported correlation coefficients between the constructs of TAM or other values that can be converted to correlations.
[7]	Schepers & Wetzels (2007)	51 of ns	1989-2006	ABI/INFORM, Scopus, ISI Web of Science, Google Scholar, and library catalogues.	Not reported.	TAM had to have been assessed in an empirical study. Integrity of the TAM concept had to have been respected: Relationships not justifiable by TAM reasoning were absent. The research methodology had to be well-described. Contained cross-sectional correlation matrix of the used TAM constructs.

Table A1. Comprehensive Past Literature Reviews

[8]	Yousafzai et al. (2007a)	95 of 36,463	1989-2004	ABI Inform, Academic Search Premier, Business Source Premier, Computer and Information Systems Abstracts, ERIC, Lexis-Nexis' Academic Universe, PsycINFO, Social Science Abstract, and SocioAbs. Plus a manual search of selected MIS, psychology, marketing and management journals. Plus backward chaining.	Keywords including but not limited to: "TAM," "technology acceptance," "perceived ease of use," "perceived usefulness," "usage behavior," "behavioral intentions," "Davis et al. (1989)"	"We adopt a comprehensive perspective and incorporate research pertaining to any of the methodological, technological, or process aspects of the TAM."
[10]	Wu & Lederer (2009)	71 of 1,550	1989-2006	"Studies from journals, books, dissertations, and conference proceedings…biblio-graphic databases and both electronic and hard copy bibliographies in journals, conference proceedings, and books"; ABI/ INFORM, Business Source Premier, ScienceDirect, ProQuest Dissertation and Thesis, WorldCat Dissertation and Thesis, and various conference proceedings such as the ICIS and AMCIS." "We did manual searches whenever back issues of the journals were unavailable in bibliographic databases. To find more studies, we also sent a general inquiry for working papers and conference proceedings to the IS community through the most popular mailing list in IS field, AISWorld."	Keywords such as "technology acceptance model," "TAM," "adoption," "acceptance," "behavioral intention," "use," "usage," "ease of use," and "usefulness." "The searches found more than 650 journal articles, 400 conference proceedings papers, and 400 unpublished dissertations. Those articles, proceedings papers, and dissertations were then examined to locate studies that could provide data to be included in the meta-analysis. Moreover, bibliographies of the articles identified were also scanned to locate additional studies. We thus identified over 100 studies and checked their potential for inclusion" (p. 424).	Operationalized PEoU, PU, and BI/usage. Reported reliabilities of measures. Described an information system-usage context in a way that gave enough information to code the measure of environment-based voluntariness. They reported sample sizes. They reported the correlations among PEoU, PU, and BI/usage, or they reported other values that could be converted to correlations.

Table A2. Noncomprehensive Past Literature Reviews[1]

#	Article	Manuscripts included vs. found	Year range	Sources	Keywords	Inclusion criteria
[2]	Han (2003)	42 of ns	1989-2003	Literature published in the "five top IS journals": ISR, MISQ, DS, MS, and JMIS.	Not Reported.	Articles that use PU as an internal belief to explore its role in end-user's behavior toward IS. Articles that used TAM as the theoretical basis to find the causal links between (1) external variables and PEoU to PU, (2) PU-A, (3) PU-BI, and (4) PU-usage. Relative advantage was treated as PU.
[5]	King & He (2006)	88 of 178	1989-2004	SSCI and Business Source Premier	"TAM" and "Technology acceptance model" as keywords; "article" as document type; excluded 55 articles that could not easily be retrieved.	Had to be empirical. Had to contain direct statistical test of TAM. Paper available online or through University of Pittsburgh Library.
[6]	King & He (2006)	30 of 108	1980s-2006	MISQ, DS, MS, JMIS, ISR, IM, JIT, IN, AMJ, CSI, GIQ, HCS, and DSS.	Not provided.	TAM is used in an empirical study. Some new variables were added in the research model. The research methodology is well-described, and the research results are available and complete.
[9]	Li, Qi, & Shu (2008)	34 of 198	1980-2005	Academic search engines like IEEE Xplore, Springer, Elsevier, EBSCO, and Blackwell.	"TAM" AND "technology acceptation model" as keywords.	TAM is used in at least one empirical study. Extended TAM models were built but contained the main classical TAM structure. The research methodology was well designed and the model results are credible and complete. The research covered a broad research domain.

Table A2. Noncomprehensive Past Literature Reviews[1]

[11]	Holden and Karsh (2010)	16	1999-2008	PubMed/ MEDLINE	Keywords: "technology acceptance model," "TAM," "TAM2," "UTAUT," and "universal theory of acceptance and use of technology"; also, the ABI/ INFORM Global database with the same keywords plus "health*," "physician*" and "nurs*." Based inclusion on reading of abstracts and articles— had to be available through university library.	Studies published on or before July 2008. Quantitatively tested relationships between variables specified by TAM. Studies of technologies that digitized information for the purpose of delivering (direct) patient care.
[12]	Turner, Kitchenham, Chartres, & Budgen (2010)	73 of 2,318	1989-2006	IEEE Xplore, ACM Portal, Google Scholar, CiteSeer library, Science Direct, and ISI Web of Science. Publications, technical reports, or "gray" literature that describe empirical studies, of any particular study design.	(Measurement OR measure OR empirical) AND ''technology acceptance model" AND usage AND (subjective OR ''self- reported" OR statistics OR questionnaire) OR objective OR validation) AND (year 1989 AND year 2006).	The TAM actual usage variable is measured, either objectively or subjectively. The version of the TAM being used must include measures of PEoU and/or PU, and the relationship (and the measure) to actual usage must be reported. Must include measure of BI and examine BI to actual usage linkage Each study was included only once.
[13]	Wu et al. (2011)	136 of 211	1992-2010	Academic Search Premier, ABI/Inform Global, Business Source Premier, Elsevier SDOS, LexisNexis Academic, JSTOR, Springer Link, Wiley InterScience, SAGE Journals Online, and Google Scholar.	Keywords including but not limited to "TAM," "technology acceptance," "perceived usefulness," "trust," and "actual use" are used to find potential relevant manuscripts. References of acquired manuscripts are further explored to identify additional manuscripts.	Had to be empirical. The research methodology has to be well-described, allowing evaluation of moderation effect. Included correlation matrix of constructs and reliability of variables.
[14]	Hsiao & Yang (2011)	72 of 518	1989-2006	ISI Web of Knowledge.	"Technology acceptance model" or "TAM."	1. Cited $\geq$ 20 times.
[15]	Šumak, Heričko, & Pušnik (2011)	38 (+4 non-TAM studies) of ns	unknown-2011	ScienceDirect, IEEE Xplore, ACM, etc., Google, Yahoo.	Combination of keywords, either related to acceptance theories (TAM, TTF, UTAUT, etc.) or keywords related to e-learning technologies (e.g., e-learning, eLearning, on-line learning, web learning, etc.).	Not Reported.

Table A2. Noncomprehensive Past Literature Reviews[1]

[16]	Dohan & Tan (2013)	16 out of ns	2002-2012	Google Scholar; PubMed; ISI Web of Knowledge; ACM Digital Library; Business Source Complete; CINAHL; MDConsult; AISeL; and the Cochrane Library. Further, several journals that are likely outputs for this type of research were included in this search. These journals included: JAMIA; IJMI; JMIR; TeH; IJHISI; HIJ; JMS; and MIM. Reference list of review articles were searched. Lastly, key researchers in the field were contacted for any feedback or assistance in this search.	Table with a number of query combinations included in article.	First, manuscripts must test the relationship between perceived usefulness (antecedent) and behavioral intention (determinant). Equivalent measures were included, specifically those of UTAUT (Venkatesh et al., 2003) and ISO (ISO, 1998), whose "performance expectancy" and "effectiveness" constructs are widely considered linearly equitable to perceived usefulness. Second, this article examines the performance of these variables in the context of only web-based tools. Third, this article restricts the focus to use of technology by patients, rather than any healthcare staff, such as doctors or nurses.

[1]Table A2 contains a list of the 10 studies we coded as not intended to be comprehensive. Han (2003) was not considered comprehensive because of a focus on five top journals. King and He (2006) focused only on journal articles, whereas Li, Qi, & Shu (2007) used only journals and only a small subset of journals. Li et al. (2008) used databases with primary focus on journals and only journal articles were retained. Holden and Karsh (2010) focused on a very specific type of context. Turner et al. (2010) was not considered comprehensive because it required actual use to be measured, which is not commonly done when testing TAM. Wu et al. (2011) was excluded for requiring information to test moderation effects. Hsiao and Yang (2011) was excluded for requiring an article to have been cited at least 20 times. Šumak et al. (2011) was excluded for focusing on the e-learning context. Dohan and Tan (2013) was excluded for focusing on medical patients and web-based tools.

Notes: AMJ: *Academy of Management Journal*; AMCIS: Americas Conference on Information Systems; CSI: *Computer Standards & Interfaces*; DS: *Decision Sciences*; DSS: *Decision Support Systems*; GIQ: *Government Information Quarterly*; HIJ: *Health Informatics Journal*; HCS: *Human-Computer Studies*; ICIS: International Conference on Information Systems; IM: *Information & Management*; IJHISI: *International Journal of Healthcare Information Systems and Informatics*; IJMI: *International Journal of Medical Informatics*; IN: *International Negotiation*; ISR: *Information Systems Research*; JIT: *Journal of Information Technology*; JMIS: *Journal of Management Information Systems*; JMIR: *Journal of Medical Internet Research*; JMS: *Journal of Medical Systems*; JAMIA: *Journal of the American Medical Informatics Association*; MIM: *Methods of Information in Medicine*; MISQ: *MIS Quarterly*; MS: *Management Science*; TeH: *Telemedicine and e-Health*.

All the articles in both Table A1 and A2 that described their boundary identification strategies used a conventional approach: keywords from the TAM theory to search manuscript databases. Several, but not all employed backward chaining, a practice recommended by Webster and Watson (2002). One (Lee et al., 2003) also followed Webster and Watson's (2002) recommendation to employ forward chaining.

The most commonly used keyword search term was the name of the theory ("technology acceptance model" or "TAM"), which was used by all eight articles reporting their search keywords. Three articles (33%) applied *usefulness* as a keyword; three articles (33%) used a variation of *use* (*use*, *usage*, or *actual use*); and one article (11%) added the keywords *behavioral intention, acceptance,* and *ease of use*.

Not all articles found by keyword search were relevant to the purposes of the review. As we looked at the prior review articles, we saw that not all articles specified how many manuscripts they found during boundary identification. Those that reported this information found between 80 and 36,463 manuscripts. All articles reported the number of manuscripts they selected for inclusion in their analyses (or were excluded from our evaluation); this ranged from 22 and 136 manuscripts. Thus, we can calculate some overall estimates for precision concerning the proportion of manuscripts identified during boundary identification that were determined to be appropriative for inclusion during corpus construction. On average, the set of review articles identified 4,171.5 prior manuscripts and selected 64.7 manuscripts for inclusion in their corpus for analyses, giving an average precision of 0.24 (24%). There is one outlier (Yousafzai et al., 2007b) that identified 36,463 manuscripts and kept 95 (precision of 0.0026, or 0.26%), suggesting that only one in every 384 manuscripts evaluated was relevant for their review.

Taken together, these prior review articles built their TAM reviews on a total of 448 unique manuscripts that the review article authors concluded provided quantitative empirical data that contributed to TAM. We excluded two of these 448 articles because we could not find them in the cited journal nor on the author's CV. We located full-text versions of 442 manuscripts (99.1%); we could not locate one accounting journal article and three doctoral dissertations. Excluding the two foundational TAM articles (Davis, 1989; Davis et al., 1989) resulted in a set of 440 research manuscripts. Of the 440 manuscripts, 418 cited either Davis (1989) or Davis et al. (1989), which left 22 manuscripts that cited neither. Two expert reviewers independently examined these 22 manuscripts and with 100% agreement judged that only two of the 22 manuscripts should be included in the corpus.

Thus, taken together, the 16 TAM review articles identified and constructed a population of 420 unique manuscripts that we categorized as warranting inclusion in the corpus for a review of TAM. However, an average of 60 manuscripts were identified and selected for inclusion by the TAM review, significantly fewer than the identified population of 420 ($t(14) = 35.12$, $p<.001$) for all 16 review articles.

The review articles were done at different points in time, so not all manuscripts in the population of 420 were published when each review conducted its analysis. So we selected the six comprehensive reviews that reported the information required to calculate both precision and comprehensiveness (Lee et al., 2003; Ma & Liu, 2004; Wu et al., 2011; Yousafzai et al., 2007a), and calculated what percent of manuscripts published within their review window they identified and selected.

We examined the six reviews that attempted to be comprehensive as a reality check of the conventional approach. Here, no review accomplishes a precision or comprehensiveness higher than 0.286. Some of the results reported here, in spite of our conservative (supportive) evaluation of their comprehensiveness are disheartening. The F_1-scores for the conventional approach in real reviews range between 0.006 and 0.184. We theorize that the difference between the performance of the conventional approach on our experiment and in past reviews is that our experiment was restricted to titles, abstracts, and author-supplied keywords, whereas past reviews, in many cases, were conducted on full-text databases. While precision was a strength of the conventional approach in the experiment, in that adding more relevant keywords drove up precision, in past reviews, as full-texts were introduced, one would expect that comprehensiveness would increase and precision decrease. However, we see no increase in comprehensiveness. Several of the reviews self-reported precision scores that are unsustainable for real use.

Therefore, these published, peer-reviewed, studies failed to identify and select *most* of the manuscripts identified and selected by other scholars using this same approach. This raises two important questions. First, if all six of these comprehensive review articles missed *most* of the relevant manuscripts, was this population of 420 the full set of theory-contributing manuscripts that should be included in a TAM review article by the time of the last review? Second, if every review article missed *most* of the relevant manuscripts, did these authors use the approach incorrectly, or is there something inherently flawed with the approach itself? We address each in turn.

Table A3. Traditional Review Evaluation

Method	Article	Comprehensiveness	Precision	F$_1$-score	AUC
Query not reported	Legris et al. (2003)	.275	.138	.184	n/a
Keywords including but not limited to "TAM," "technology acceptance," "perceived ease of use," "perceived usefulness," "usage behavior," "behavioral intentions," and "Davis et al. (1989)"	Yousafzai et al. (2007a)	.266	.002	.005	n/a
Keywords such as "technology acceptance model," "TAM," "adoption," "acceptance," "behavioral intention," "use," "usage," "ease of use," and "usefulness."	Wu & Lederer (2009)	.046	.120	.067	n/a
Query not reported	Ma & Liu (2004)	.093	.286	.140	n/a

A2. Estimating the Size of the Full Set of Theory-Contributing Manuscripts

To test whether 420 manuscripts is a good estimate of the full set of relevant manuscripts that should be included in a TAM review article, we created two nonoverlapping random samples (Random 1 and Random 2) of 300 publications each that were drawn from the 5,991 Microsoft Academic Search (MAS) manuscripts that cited either Davis (1989) or Davis et al. (1989) as of May 5, 2014.[6] We created a set of contribution criteria and coded each manuscript to identify whether the manuscript contributed to TAM, and would thus be one we would include in a review article if we were writing one (i.e., it provided empirical data to support or refute one of the theoretical relationships in TAM) (Holden & Karsh 2010; Ma & Liu 2004; Turner et al. 2010; Wu et al. 2011). See Appendix B for details on the inclusion criteria and coding. After cleaning the two samples, we combined them to provide a set of 516 manuscripts. Table A4 reports statistics on our four evaluation datasets.

Two domain experts independently coded these manuscripts to identify which manuscripts contributed to TAM (and thus were relevant to a review article) and which used it for other research. There were 137 TAM-contributing manuscripts among the 516 manuscripts (26.5%). Cohen's Kappas for interrater agreement were "substantial" to "almost perfect," and are available in Appendix B.

This analysis suggests that the number of manuscripts that should be identified and selected during corpus construction should be approximately 26.5% of the total of 5,991 manuscripts we found—in other words, approximately 1,590, with the 95% confidence interval between 1,378 and 1,797. This is significantly more than the 420 manuscript population identified by prior review articles using the conventional approach.

This 26.5% estimate has remained stable over time (we compared the three-year running average of contributing manuscripts per year to the total number of manuscripts for that year, resulting in a correlation of 0.982). The total of TAM-citing manuscripts has increased over time, but the total number of manuscripts included in TAM reviews has remained more constant, while the number of potentially relevant manuscripts has grown dramatically. By the end of our period of examination, 2012, that difference was an order of magnitude (i.e., ten times), but by that time no review attempted comprehensiveness.

---

[6] While Google Scholar may be the most inclusive academic database, it provides no application programming interfaces and forbids scraping. Searches are also capped at 1,000 results. At the time of this research, MAS represented the best alternative because of its API, and inclusion of gray literature.

Table A4. Evaluation Datasets

Name	Number of MS	Sample description	Examination process
16 review articles	420	All selected TAM manuscripts used in 16 identified TAM reviews and meta-analyses.	Trained research assistants were given the manuscripts and asked to create a table of the 16 articles as columns and the unique manuscripts as rows. Only manuscripts used as data or evaluated as part of the literature review were included.
6 comprehensive reviews (selected from the 16 review articles)	n/a	Six articles claiming to be comprehensive. Four of these reported numbers necessary to calculate precision and comprehensiveness.	Two authors examined all 16 review articles and removed all that did not claim to be comprehensive for both journals and other sources, and had equivalent inclusion criteria to those used in Random 1 and Random 2, above. This led to six comprehensive articles, four of which reported the information needed to evaluate their F_1-scores.
Random 1	264	300 manuscripts randomly selected from the set of 5,991 manuscripts that cited one or both foundational TAM articles.	295 manuscripts were retrieved (98.3%). One faculty member and one research assistant with five years of experience independently examined each manuscript. Of these, 9 were excluded for data quality problems in that they did not actually cite TAM, 13 were excluded because they were qualitative, and 9 were excluded because they were in a foreign language.
Random 2	252	300 manuscripts randomly selected from the set of 5,991 manuscripts that cited one or both foundational TAM articles. No overlap was allowed between Random 1 and Random 2 and both were part of the same random draw.	295 manuscripts were again retrieved (98.3%). Two faculty members independently examined each manuscript. Of these, 8 were excluded for data quality problems, 22 were excluded because they were qualitative, and 13 were excluded because they were in a foreign language.

Figure A1 shows the average number of expected TAM-contributing manuscripts per year (with 95% C.I. lines) relative to the number of manuscripts included in each of the 16 review articles. The articles are shown by their last year of article inclusion rather than their publication year.

Figure A1 shows that none of the TAM review articles that used the conventional approach reached the lower bound of the 95% confidence interval for the number of theory-contributing manuscripts available at the time of their publication. For 10 of the reviews (See Table A2), this is as expected given that they were not designed to be comprehensive. Surprisingly, for the six studies that claimed comprehensiveness (Table A1), after publication of the second of the six (Lee et al. 2003), while the literature base of TAM manuscripts grew exponentially, the reviews included fewer manuscripts. While the first six-year period (2001-2006) included five reviews that aimed to be comprehensive, the second six-year period (2007-2012) contained none.
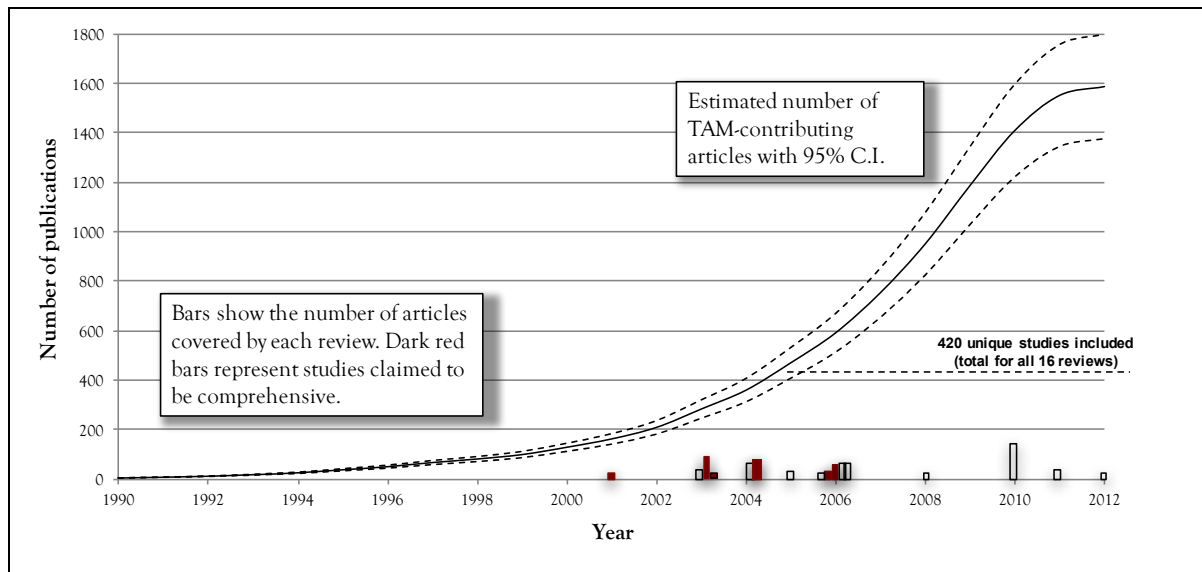
Figure A1. Coverage of Past TAM Reviews and Meta-Analyses

Our analyses are unable to assess the precision of the boundary identification approaches used by most prior review articles, because we have no access to the set of manuscripts that were initially identified by these authors. However, four of the six did report the total number of manuscripts found by their search strategy as well as the number of manuscripts included. We use this information to calculate precision, comprehensiveness, and the F_1-score of past reviews against our estimate of total number of relevant manuscripts.

Focusing then on only the four articles that claimed comprehensiveness *and* also reported both total manuscripts retrieved by search query and the total manuscripts included (Table A1), specifically Legris et al. (2003) Yousfzai, et al., (2007b), Wu & Lederer (2009), and Ma & Liu (2004), we are able to calculate their comprehensiveness relative to the estimated number of TAM manuscripts available at the end of their review window and found that, on average, they included 17.5% of the estimate of relevant manuscripts (range: 9.3% - 26.6%).[7]

There is some variability across the four articles, but this variability is noticeably lower than the difference from the benchmarks in Figure A1. A typical review article aiming to be comprehensive using the conventional approach failed to include 82.5% of prior research manuscripts that our analysis of MAS suggests are available (see Figure A1). The *most comprehensive* review article using the conventional approach included only 33.7% of prior manuscripts found in other review articles and 26.7% of prior research that our analysis suggests was available at the time. Given the undoubted efforts put into these articles by our colleagues, these results are worrisome and require further empirical examination.

One important question is whether the lack of comprehensiveness is inherent in the conventional approach itself, or whether prior authors have not used it appropriately. Part of this discrepancy may be explained by the way the conventional approach was implemented by specific author teams. For example, as shown in Table A2, different teams used different inclusion criteria, such as requiring a minimum number of citations (Hsiao & Yang, 2011), a measure of actual use (Han, 2003; Turner et al., 2010), a specific construct not traditionally associated with TAM such as Trust (Wu & Lederer, 2009), a type of system (Dohan & Tan, 2013), only manuscripts available through a specific university library (King & He, 2006), or included only journal articles (Li et al., 2007; Li et al., 2008). Conversely, some review articles went beyond TAM manuscripts and included related manuscripts such as those examining the *relative advantage* label for the *usefulness* construct (Han, 2003) or including results from UTAUT (Venkatesh et al., 2003) and manuscripts containing equivalent construct relationships (Dohan & Tan, 2013).

---

[7] It is worth noting that relative to the 5,991 MAS articles citing Davis (1989) or Davis et al. (1989) by May 4, 2014, a Google Scholar search on October 14, 2017 examining citations to only Davis (1989) in manuscripts published before the end of 2013, returned 21,400 articles, suggesting that any results we provide in this article are likely to be conservative by a factor of at least 3.5. In other words, our MAS-derived estimates of comprehensiveness may be at least 3.5 times higher than they should be, suggesting an actual average comprehensiveness around 5%.

A3. Bias in the Conventional Approach

Another important question is whether there is a bias in the manuscripts identified by the conventional approach. When research produces quantitative findings comparable across studies, as is true in quantitative meta-analysis, past research has consistently found that the journal publication process introduces bias (Banks et al. 2015; Berlin & Ghersi, 2005; Kepes et al., 2012; Kepes & McDaniel, 2015). We examined the distribution of journal articles relative to other sources (e.g., conference proceedings, book chapters, theses, and unpublished manuscripts) among the 420 manuscripts identified by the 16 review articles versus the 137 TAM contributing manuscripts in our benchmark samples. About 88% of the 420 manuscripts were journal articles compared to 66% in our 137 sample manuscripts, a significant difference (X^2 = 34.45; p < 0.01). This suggests that the conventional approach as currently practiced suffers from bias—a problem that takes on added significance for reviews that intentionally included only journals in their corpus (i.e., Han, 2003; Hsiao & Yang, 2011; King & He, 2006; Lee et al., 2003; Li et al., 2007) or only "top journals" (Han, 2003). While we here do not undertake the task of establishing that the conventional approach directly leads to bias, it is likely that any research approach that leads to low precision will force researchers to develop heuristics such as exclusive focus on journal articles or even articles in "top" journals.

Appendix B. Inclusion Criteria and Coding

B1. Inclusion Criteria

All of the research reviews in our TAM samples were quantitative reviews of empirical research, so we needed to adopt a similar frame of reference. To determine what constitutes an empirical contribution to a theory we turned to literature on theory, metatheory and theorization (Bostrom et al. 2009). A theory's domain is bounded by the name of the theory. We here make what may at first glance look like a controversial choice by stating that naming something TAM2 makes it different from TAM. We do this not because we believe TAM and TAM2 account for different domains or that the constructs are very different, nor because we believe that relationships tested in one could not represent contributions to the other. However, we argue that if their *authors* claim them to be different, we must begin with the assumption that they are different. Thus, authors of review articles are left with two options: (1) add the theory-originating manuscript for TAM2 to L_1 of the theory ecosystem, or (2) do a separate analysis and review for TAM2. We believe that theory reviews should be as "pure" as possible so that ontologies may then be used to integrate the theories and their findings later or even in the same review containing two separate studies. After all, if the authors decide to integrate two theories with different names, even if they are as similar as TAM and TAM2 they do in fact assess these theories to, on some dimension, be the same *before* collecting the empirical evidence arguably necessary to make such determinations.

A study making an empirical contribution to a specified theory must use a dependent construct consistent with the focal theory (for TAM, this would be *use* or its stand-in *behavioral intention to use technology*). Likewise, a contributing study must include at least one independent construct (e.g. for TAM, *ease of use*, *usefulness,* or *attitude toward using*), operationalized consistently with focal theory operationalization. Finally, it must provide empirical data testing the relationship(s) between the independent construct(s) and the dependent construct(s).

Based on the above arguments, we developed four inclusion criteria for identifying relevant manuscripts. We make no claim these are the ideal criteria that should be used by all review articles. Instead, we argue that authors of review articles need to make their own deliberate decisions about inclusion criteria and then be transparent by describing those criteria in their articles. We argue these four criteria are appropriate for the objectives of our article—identifying manuscripts that make an empirical contribution to TAM by providing empirical evidence to support or refute one or more of the relationships in TAM.

1. *Does not* claim to create a new theory that is separate from the current theory (e.g., by proposing a new theory with a new name).

2. *Must* be empirical. While non-empirical research can be very valuable, for the purposes of this test of ADIT, the focus is on empirical contributions to a theory. This criterion was also used by Ma & Liu (2004), Wu & Lederer (2009), Holden & Karsh (2010), Turner et al. (2010), and Wu et al. (2011).

3. *Must* use at least one dependent variable from the theory as a dependent variable, without materially changing its name or definition. This criterion is like that used by Turner et al. (2010) and Han (2003).

4. *Must* include empirical findings on the effects of at least one independent variable from the theory (without materially changing its name or definition) on the dependent variable. This criterion was also used by Han (2003) and Turner et al. (2010).

We recognize that Criterion 2 is covered by Criterion 4 (it is unlikely that any manuscript would meet Criterion 4 but not Criterion 2). Criterion 2 is kept for historical reasons given the many past studies that employed this criterion.

B2. Coding

The use of a random sample enabled us to make statistical conclusions about the larger population of manuscripts and is a key feature of our recommendations for future literature reviews of large theories. We created two random samples. Random 1 was coded independently by (1) one researcher with two decades of literature review experience, and (2) one research assistant with five years of review experience. After removal of excluded studies, the raters agreed in 240 cases and disagreed on 24 cases, leading to a Cohen's (1960) Kappa of 0.755, a level of agreement considered in the upper range of "substantial" by Landis and Koch (1977). Disagreements were resolved through discussion. Random 2 was coded independently by (1) one researcher with two decades of literature review experience, and (b=2) one researcher with 15 years of literature review experience. After removal of excluded studies, the coders agreed in 234 cases and disagreed on 18 cases, leading to a Cohen's Kappa of 0.832, a level of agreement considered in the low range of "almost perfect" by Landis and Koch (1977). Disagreements were again resolved through discussion. Coder one was the same for both exercises. In Random 1, 65 of 264 (24.6%) nonexcluded manuscripts were found to contribute to TAM, and in Random 2, 72 of 252 (28.6%) nonexcluded manuscripts were found to contribute to TAM.

Appendix C. Assessing the Performance of the Conventional Approach

C1. Full-Text Searches

The 16 review articles in Appendix A used a variety of academic research databases, including Web of Knowledge (used by six of the 16 review articles), Google Scholar (used by five), ABI/Inform (used by four), ACM Digital Library (used by four), Science Direct (used by four), EBSCO's Business Source Premier (used by two), and IEEE Xplore (used by two).

To evaluate the conventional approach on a standalone basis, we used the random sample estimated number of TAM contributing manuscripts to get a sense of how well common keyword search approaches work in various literature databases. We conducted our own keyword searches using some of the most common search terms from these 16 review articles on the full texts of manuscripts included in these seven databases. The number of manuscripts retrieved using four TAM concepts and are presented in Table C1. All searches were conducted on May 19, 2015.

Table C1. Manuscripts Retrieved by Keyword for Commonly Used Literature Databases

Search query	Google Scholar	MAS	Web of Knowledge	ABI/ Inform	ACM Digital Library	EBSCO	IEEE Xplore	Science Direct
Searching full text								
"Technology Acceptance Model"	44,100	1,280	2,726	3,058	607	4,172	2,119	2,918
Usefulness	2,740,000	121,635	128,146	318,774	21,097	247,316	95,585	454,328
"Ease of Use"	364,000	3,939	10,181	63,320	8,244	29,722	23,696	43,770
"Intention to Use"	54,000	1,483	1,534	2,424	487	5,546	1,804	5,178
Searching titles, abstracts, and manuscript keywords								
"Technology Acceptance Model"	N/A	N/A	N/A	770	N/A	1,591	512	536
Usefulness	N/A	N/A	N/A	9,272	N/A	61,775	14,182	41,527
"Ease of Use"	N/A	N/A	N/A	1,479	N/A	5,428	2,270	2,964
"Intention to Use"	N/A	N/A	N/A	471	N/A	1,260	244	575

Note: We used the EBSCO databases *Business Source Complete* and *Academic Search Premier*

Our benchmark analysis of MAS suggests that (1) there were roughly 1,591 research manuscripts that contributed to TAM (and thus relevant to a literature review) in 2012, and (2) there were 6,400 likely TAM-contributing manuscripts in Google Scholar on June 4, 2014. As seen from Table C1, common search strategies will generate between five and 407 times more manuscripts than are relevant. For MAS we see the opposite problem in that even the query "Technology Acceptance Model" produces fewer than the expected number of manuscripts. The search query "Technology Acceptance Model" found significantly fewer than even the MAS number of manuscripts for six of the seven databases (a comprehensiveness problem) and significantly more than this number for one database—Google Scholar with 33,400 manuscripts (a precision problem).

Using different search terms produces different results. Table C1 also shows the results of using selected constructs from TAM as keywords. Some of these searches lead to comprehensiveness problems, but most lead to precision problems; the search returns thousands more manuscripts than our benchmark suggests are relevant to a literature review of TAM. We conclude that the conventional approach is likely to suffer from precision problems especially when full-text searches are used on construct names alone. In conclusion, what Table C1 makes very clear is that a multi-database approach using any conventional search approach will return too many manuscripts for a human being to evaluate.

C2. Constrained Searches.

One search option is to constrain the search to require that keywords appear in some part of the manuscript, such as its title or abstract. Terms appearing in the title or abstract are intended to convey the central message of the manuscript (Larsen et al. 2008) and thus may be more likely to signal that the manuscript contributes to the theory. Table C1 contains the results of search for selected keywords in the title, abstract, or keywords. Only databases used by at least two of the 16 articles, as well as the new Microsoft Academic Search (MAS), were included. Not all databases allowed search in title and abstract only, which was denoted with "N/A" in C1.

The results in Table C1 *potentially* show increased precision as far fewer manuscripts are found compared to a full-text search. Because many manuscripts testing non-TAM theories would be likely to mention TAM in making various points, such manuscripts would be found in full-text searches, but be much less likely to surface in titles and abstract. However, now comprehensiveness becomes a problem because the number of manuscripts has dropped below the estimated number of manuscripts that contribute to the theory.

Appendix D: ADIT Technical Details

Figure D1 shows a partial view of the ADIT overview of theories specified in the system. Each theory is specified through another screen where the theory-initiating manuscripts are added. The system runs during the night every 24 hours unless the user forces an update on a new theory. The goal is to minimize impact on the MAS system that provides the data for ADIT.

Theory Networks Overview

Visualization	Network Name	Date Added	Last Run	Last Analysis	Changed Since Last Analysis?	Size of 2nd Level Network	Size of 3rd Level Network	Size of Theory Contributing Network	
⊕	Technology Acceptance Model	4/29/2013 8:46:34 PM	3/2/2014 2:13:58 AM	10/27/2013 11:58:53 PM	Unchanged	7570	458775	256516	Edit Settings
⊕	Theory of Planned Behavior/Theory of Reasoned Action Treating these two as the same theory for now. According to one paper: "...The TPB was developed by Icek Ajzen (1985, 1987) as an extension of the theory of reasoned action (Ajzen & Fishbein, 1980; Fishbein & Ajzen, 1975) to account for nonvolitional factors as determinants of a behavior"	5/4/2013 11:38:16 PM	3/2/2014 12:50:02 AM	11/12/2013 10:35:17 PM	Unchanged	17358	289324	0	Edit Settings
⊕	Unified Theory of Acceptance and Use of Technology (UTAUT)	5/6/2013 9:14:40 AM	3/1/2014 9:34:44 PM	6/17/2013 7:54:25 AM	Unchanged	1727	108042	0	Edit Settings
⊕	AIDS Risk Reduction Model (ARRM) The AIDS risk reduction model (ARRM) proposes that changing sexual behaviours related to HIV transmission occurs in 3 stages; (1) labeling, (2) making a commitment and (3) enacting the solutions. Movement through the stages is dependent on achievement of goals in previous stages.	6/29/2013 8:48:37 AM	3/2/2014 2:34:52 AM	7/29/2013 11:08:14 AM	Unchanged	225	5200	0	Edit Settings
⊕	Transtheoretical/Stages of Change v1 First version of theory (1983). See: http://en.wikipedia.org/wiki/Transtheoretical_model#cite_note-Prochaska2005-11	6/29/2013 10:31:50 AM	3/2/2014 2:32:08 AM	6/29/2013 12:40:00 PM	Unchanged	1804	2329	0	Edit Settings
⊕	Transtheoretical/Stages of Change v2 (1992)	6/29/2013 3:29:04 PM	3/2/2014 1:10:48 AM	6/29/2013 4:00:18 PM	Unchanged	2009	4734	0	Edit Settings
⊕	Transtheoretical/Stages of Change v3 (1997)	6/29/2013	3/2/2014 1:17:24	6/29/2013	Unchanged	627	1582	0	Edit

Figure D1. ADIT Theory Networks Overview Screen3F[8]

1. Theory Ecosystem Construction. The first step is to find the unique MAS identifiers for all foundational manuscripts. For TAM these were 1265954 (Davis 1989) and 1253523 (Davis et al. 1989). Once the foundational manuscripts for a given theory are selected, the MASCrawler is initiated. The MASCrawler class is an implementation of ICrawler specific to Microsoft Academic Search, which handles the retrieval of manuscripts, keywords, authors, citations, and references.

Once the relevant crawls have been enumerated, the crawler goes through the following process: if the crawl is a new crawl, the Crawler retrieves information regarding the canonical manuscripts, first checking if there is a current record of the canonical manuscripts (retrieved as part of a previous theory crawl) and retrieving them from the MAS record if they do not exist; if the crawl is a scheduled crawl that was interrupted during citation enumeration, the queued citations will be removed from the queue to avoid creating duplicate citations. Each canonical manuscript has its existing citations enumerated and compared to the latest citation data from the MAS record. If there are any additional citations, they are queued for processing.

Citations are dequeued, with the corresponding manuscript either being retrieved from the persistence model or, if the model does not yet exist, the MAS record. The retrieved manuscript is then set as citing the corresponding canonical manuscript. The references (manuscripts listed in the "References" section) of each manuscript in the previous step are compared to their existing references, with any new references being queued.

References are dequeued, with the corresponding manuscript either being retrieved from the persistence model or, if the model doesn't yet exist, the MAS record. The retrieved manuscript is then set as referencing the corresponding first-level manuscript. Once these manuscripts and their references have been stored in the ADIT database, each

---

[8] Because the system continues to download articles for the theory ecosystem, these numbers are different from those used in this article.

manuscript cited in these L_2 manuscripts is downloaded, leading to a full set of theory network manuscripts. At this point, all connections are enumerated in a network.

2. Coding of a Random Sample. Selecting the size of the random sample that will be used to train the machine learning algorithm requires judgment. It needs to be large enough to provide sufficient precision to discriminate between contributing and noncontributing manuscripts in the theory ecosystem. In general, larger samples produce better results, but with a declining marginal benefit (Bacchetti, Wolf, Segal & McCullough, 2005; Cortes, Jackal, Solla, & Vapnik, 1994; Wocjan, Janzing, & Beth, 2002) given that the larger the sample, the more work is required to code the manuscripts. The size of the sample also depends on the number of manuscript attributes that will be used to train the machine learning algorithm; the sample must be large enough so that there are no overspecification problems (Figueroa et al., 2012) (Mukherjee et al., 2003). In general, we recommend that the random sample is a minimum of 100-200 manuscripts (Figueroa et al., 2012; Kalayeh & Landgrebe, 1983; Zuk, Margel, & Domany, 2012). We used two samples of 300 manuscripts each, with a set of 23 attributes. We used a larger number than recommended because we did not expect to achieve a 98% pdf retrieval rate given that Williams et al. as mentioned achieved 52%. We also did not expect to find a 26.5% manuscripts that fit our inclusion criteria given that Williams et al. found that only 9.6% for UTAUT, a difference that matters for sample size requirements.

Appendix B provides additional details about the inclusion criteria we used to code each of the manuscripts in our sample as contributing or noncontributing. Two coders worked independently and then met to resolve differences so each manuscript in the random sample was coded for use by the machine learning algorithm. Selecting the manuscript attributes that will be used by the machine learning algorithm also requires judgment. Table D1 presents the attributes we used in our analysis of TAM which are either rhetorical attributes of the manuscripts themselves (e.g., use of the theory name in the manuscript title) or attributes of the manuscript within the theory ecosystem (e.g., Eigenfactor).

The rhetorical attributes are based on simple text pattern matching, and change for each theory examined, because they are the ways in which the manuscripts use elements of the theory in their rhetorical arguments. Determining these attributes requires an expert knowledge of the theory and the research discourse in the theory ecosystem. In general, it is better to err on having too many attributes rather than too few because feature reduction process in machine learning can detect those attributes that are most useful in categorizing manuscripts (Bishop 2006).

The second set of attributes comes from the theory ecosystem, which provides useful clues for delineating contributing and non-contributing manuscripts. Both the conventional approach and the ADIT method utilize the citation network. The difference is that ADIT identifies important manuscripts by traversing the full citation network rather than just forward/backward citations from the foundation theory manuscripts. It accounts for the citations to the foundation manuscripts but also the connections between L_2-L_2, L_3-L_3 manuscripts and L_2-L_3 manuscripts—taking into account the full citation structure.

The method is called the article-level Eigenfactor (ALEF). This approach accounts for the source of citation. In other words, citations from highly cited manuscripts are worth more than citations from less cited manuscripts. This may sound circular but the algorithm is well-defined (West, Jensen, Dandrea, & Gordon, 2013). A random walker under this model takes long paths from one point in the network to any other part of the network. The method for ranking nodes in networks is similar to the well-known PageRank algorithm for ranking webpages (Page, Motwani, & Winograd, 1999) where important websites receive links from other important websites. The difference is how the ALEF method deals with time-directed networks. The citation trails in these systems move inexorably backwards in time. The modifications of this algorithm—compared to standard PageRank—require shorter paths for the random walker. This corrects for disproportionately weighting older manuscripts with PageRank. We find that it improves the algorithm's ability to separate contributing manuscripts from noncontributing manuscripts, although we leave this analysis to a subsequent paper.

The mechanics of the algorithm are straightforward (West, Wesley-Smith, & Bergstrom, 2016). We construct an *n x n* adjacency matrix, **Z**, where the Z_{ij} entry is equal to 1 if there is a citation from manuscript *i* to manuscript *j*. Borrowing language from the original PageRank algorithm, you create a teleportation vector to each manuscript in the following way:

$$w_i = \sum_j^n (Z_{ij} + Z_{ij}^T)$$

The matrix **Z** is row normalized so that the row sums equal 1.

$$H_{ij} = \frac{Z_{ij}}{Zi}$$

The teleport vector is then multiplied by the *H_{ij} and then normalized by the number of papers in the corpus to give the following ALEF score.*

$$ALEF = n \, \frac{H_{ij}^T}{\sum_j [H_{ij}^T \cdot w_i]_{j_{\blacksquare \cdots}}}$$

More details for the calculation can be found at West et al. (2016).

3. Identification of Contributing Papers. This step begins with downloading the full texts of all L2 papers. We used Bayesnet, a versatile approach where nodes represent random variables, often with discrete sets of values. Bayesian Networks (BN) are generative, directed (acyclic) graph models where nodes represent random variables (r.v.) and the links represent probabilistic connections between the r.v. They are called BN because they use the famous Bayes' rule to infer link probabilities (not because Bayesian statistics are the only way to estimate the parameters). The strength and popularity of BN is the simple graphical representation of the random variables and the intuitive causal interpretation between the nodes. They have been used in medical diagnoses, business decision making and marketing, computer vision, speech recognition, and bioinformatics. For the approach described in the paper, we could have used other machine learning approaches (SVMs, ANNs, etc.), but we chose BN because of their widespread adoption and their intuitive interpretative appeal.

For the analysis, we used 10-fold cross-validation for examination of efficacy. In other words, each data set was split into ten folds (roughly equal-sized partitions). Each fold is treated as a validation sample in ten different runs of the algorithm where the other nine folds are used as training data. The results are based on average performance for the ten different runs. Table D1 specifies the attributes used as features in the machine learning. This is equivalent to the "keywords" used in conventional reviews (See Table A1), but are different in that in conventional searches the keywords have to be combined with OR or AND statements whereas the machine learning approach combines these attributes in hundreds or thousands of ways to detect and implement both inclusion criteria. Rerunning the analysis takes only a few seconds.

Table D1. Paper Attributes Used to Identify Theory Contributing Papers

Attribute	Description
Eigenfactor_Eco (EF)	The Eigenfactor_Eco score is reflective of the paper's importance in the theory ecosystem. Eigenfactor score was calculated using the network of all papers that cited TAM (Level 2) and then all papers that were cited in Level 2 (Level 3). A total of 65,000 papers were included, but then reduced down to the 5,991 papers in Level 2.
Theory-Attribution Ratio (TAR)	This feature examines each paper's references and sums up the Eigenfactor score for each L_2 paper the paper references (those that cite the foundational papers) divided by the number of references. This feature works under the assumption that papers that reference other L_2 papers may be more likely to indicate an intention to contribute to that theory.
Impact (I)	This feature calculates the impact of a paper, here defined as the count of the citations to a focal paper in a certain period (Garfield 2006).
Publication Year (PY)	Because theories tend to have a life cycle, knowing the year of publication should enable the system to more accurately evaluate whether a paper is intended to contribute to a theory.
Word count in Abstract (W_A)	Number of words in the abstract
Theory name in Title (Tt)	Does the theory name ("Technology Acceptance Model") exist in the title of the focal paper?
Theory name in Keywords (Tk)	Does the theory name ("Technology Acceptance Model") exist in the keywords of the focal paper?
Theory name in Abstract (Ta)	Does the theory name ("Technology Acceptance Model") exist in the abstract of the focal paper?
Theory acronym in Title (At)	Does the theory acronym ("TAM") exist in the title of the focal paper?
Theory acronym in Keywords (Ak)	Does the theory acronym ("TAM") exist in the keywords of the focal paper?
Theory acronym in Abstract (Aa)	Does the theory acronym ("TAM") exist in the abstract of the focal paper?
Usefulness construct in Title (U_T)	Does *usefulness* exist in the title of the focal paper?
Usefulness construct in Keywords (U_k)	Does *usefulness* exist in the keywords of the focal paper?
Usefulness construct in Abstract (U_T)	Does *usefulness* exist in the abstract of the focal paper?
Ease of Use construct in Title (EoU_T)	Does *ease of use* exist in the title of the focal paper?
Ease of Use construct in Keywords (EoU_T)	Does *ease of use* exist in the keywords of the focal paper?
Ease of Use construct in Abstract (EoU_T)	Does *ease of use* exist in the abstract of the focal paper?
Attitude construct in Title (At_T)	Does *attitude* exist in the title of the focal paper?
Attitude construct in Keywords (At_T)	Does *attitude* exist in the keywords of the focal paper?
Attitude construct in Abstract (At_T)	Does *attitude* exist in the abstract of the focal paper?
Behavioral Intention in Title (BI_K)	Does *behavioral intention* exist in the title of the focal paper?
Behavioral Intention construct in Abstract (BI_K)	Does *behavioral intention* exist in the abstract of the focal paper?

About the Authors

Kai R. Larsen is an associate professor of information systems in the division of Organizational Leadership and Information Analytics, Leeds School of Business, University of Colorado Boulder. He is a courtesy faculty member in the Department of Information Science of the College of Media, Communication and Information, a research advisor to Gallup, and a fellow of the Institute of Behavioral Science. Kai is most known for providing a practical solution to Edward Thorndike's (1904) Jingle Fallacy and for his contributions to the semantic theory of survey response (STSR), which holds that results of surveys using attitude scales primarily measure the linguistic relationships between survey questions.

Dirk S. Hovorka is an associate professor in Business Information Systems Discipline at the University of Sydney. His current research seeks to recenter the possible livable worlds that scientific practices bring forth through theory, design practices, and how we think about "the future" in terms of technology, society, and biophysical environments. His research interrogation of the philosophical foundations of information systems has been informed by his information systems PhD and interdisciplinary telecommunications and geology MS degrees (University of Colorado, USA). As the recipient of prestigious awards including the BGS (AACSB International Honor Society) 2018 Professor of the Year Award, and the University of Sydney Wayne Lonergan Award for Outstanding Teaching, Dirk is deeply committed to preparing students for future(s) challenges. Dirk co-authored the AIS 2011 Best Paper "Secondary Design: A Case of Behavioral Design Research." He is a JAIS Senior Editor, a CAIS Associate Editor, and HICSS co-chair of the "Knowing What We Know (Theories in IS)" minitrack.

Alan R. Dennis is professor of information systems and holds the John T. Chambers Chair of Internet Systems in the Kelley School of Business at Indiana University. He was named a fellow of the Association for Information Systems in 2012. Professor Dennis has written more than 150 research papers and has won numerous awards for his theoretical and applied research. His research focuses on three main themes: team collaboration; fake news on social media; and information security. He also has written four books (two on data communications and networking, and two on systems analysis and design). His research has been reported in the popular press over 500 times, including in the *Wall Street Journal*, *USA Today*, *The Atlantic*, CBS, PBS, Canada's CBC and CTV, UK's *Daily Mail* and the *Telegrap*h, Australia's ABC, France's *Le Figaro*, South Africa's *Sowetan Live*, Chile's *El Mercurio*, *China Daily*, India's *Hindustan Times*, and Indonesia's *Tribune News*. He is the co-editor in chief of *AIS Transactions on Replication Research* and the president of the Association for Information Systems.

Jevin West is an associate professor in the Information School at the University of Washington. He is the co-founder of the DataLab and director of the new Center for an Informed Public at UW. He holds an adjunct faculty position in the Paul G. Allen School of Computer Science & Engineering. He is also a data science fellow at the eScience Institute and affiliate faculty for the Center for Statistics & Social Sciences. His research and teaching focus on misinformation in and about science. He develops methods for mining the scientific literature in order to study the origins of disciplines, the social and economic biases that drive these disciplines, and the impact the current publication system has on the health of science. More information can be found at jevinwest.org.