An information system design theory for the comparative judgement of competences

An information system design theory for the comparative judgement of competences

Abstract

A Design Science Research project is presented, describing the creation of an Information System for the assessment of human competences while supporting learning. First, requirements that emanate from current mainstream competence evaluation practice are introduced. Then, design principles are presented to address the design requirements. Finally, design features are discussed that represent a concrete instantiation of the design principles in a working system prototype. The output of the design, development and evaluation of the artefact is presented as an Information Systems Design Theory. This theory provides principles that can be applied in different contexts where the evaluation of competences is needed.

Keywords

Design science research, comparative judgement, computer-based assessment

Introduction

The evaluation of competences is necessary and a recurrent activity in educational and human resources contexts. Traditionally, these evaluations are made through analytical reflections in which the representation (e.g. written text, video fragment,...) of a competence is scored on a set of predetermined categories or criteria. These are referred to as “rubric” evaluations. It is often assumed that such criteria assure that all judges assess the same, predefined aspects of the competence at hand (Jonsson & Svingby, 2007). However, this common practice raises issues regarding both the validity and the reliability of the assessment (Pollitt, 2004; Jonsson & Svingby, 2007; Sadler, 2009; Jones & Alcock, 2014).

The higher the stakes of an evaluation, the more damaging it can be to take action based on an unreliable or invalid evaluation. For example, it could be potentially damaging for an organisation to promote a specific employee to a more responsible function, based on an erroneous evaluation of his competence. Similarly, it’s undesirable from a human and societal point of view to make educational decisions to pass, fail or redo a study year based on an unreliable measurement of student competences. Thus, a system that makes the evaluation of human competences more credible would be of value both to the Human Resources and Educational sectors.

This is the **purpose and scope** (Gregor and Jones, 2007) of this paper and of the Information Systems Design Theory that it aims to build, which we will address by applying the principle of Comparative Judgement (CJ) to competence assessment. In such an approach, a person performs a competence according to a certain assignment, which results in a representation of this competence in a medium that can be stored (text, video, audio,…). For example, the competence of “argumentative writing” can be performed by a person, resulting in a representation of the competence in the form of a written text. CJ of this competence would then involve comparing representations of different people, iteratively indicating the representation that is of the highest quality.

Evaluation of competences is also a major way in which people learn. By learning from the results of their evaluation, they can improve their competences. As an IS that would improve the credibility of human competences would gather a great deal of data regarding the grounds of the evaluation, feeding this data back to support learning is a second aspect of the purpose and scope of this research.

We position this work as Design Science Research (DSR), which differs from other scientific approaches in the type of knowledge it produces. Natural and social sciences aim to describe, explain, and predict (Dresch et al, 2014). Taking a different approach, the goal of design science is to explain what works, by creating prescriptive knowledge (Hevner et al 2004) to find out how artefacts of a certain class should be constructed.

The contribution we aim to make in this paper is twofold, with a first contribution in the application domain of IS for CJ. A limited number of IS that leverage CJ concepts currently exist, like E-Scape, NoMoreMarking, Brightpath and ComPAIR. These are under active development and are yet to achieve mainstream adoption. Tarricone & Newhouse (2016) and Potter et al (2017) have discussed the use of IS to support CJ. They address the kernel theory of Comparative Judgement itself and discuss its advantages and disadvantages when applied to education, yet offer very little in terms of prescriptive knowledge, built on the practical development and evaluation of IS artefacts, to inform the creation of future IS for CJ support. Such an ISDT may also be of use to other areas, where group decision making needs tool support, like for example strategic decision making in organisations.

It is to fill this gap that we have studied an IS for CJ from a DSR perspective. To produce such practice-based theoretical insights, we have undertaken de creation of an Information System Design Theory (ISDT), as a main contribution of this paper. Gregor and Jones (2007) described ISDTs as the prescriptive knowledge type that is central to Design science. Such an ISDT combines knowledge of IT and human behaviour to prescribe guidelines for the creation of artefacts of the same type. According to Gregor and Jones (2007), an ISDT should be composed of the following parts: (1) the purpose and scope, (2) constructs, (3) principles of form and function, (4) artefact mutability, (5) testable propositions, (6) justificatory knowledge, (7) principles of implementation, and (8) an expository instantiation. Thus, this paper proposes an ISDT for building IS artefacts that permit the CJ of learning competences. The contribution of the paper is structured using the distinction between design requirements (DR), design principles (DP) and design features (DF) introduced in Meth et al (2015). Where the design requirements originate from the drawbacks inherent to current practice, the design principles address these requirements. The design features represent the instantiation of the design principles in a specific artefact.

A second contribution which the paper aims to make lies less in the application domain of IS for the support of CJ and more in DSR itself. The theoretical underpinnings of the field have now become well-established and many recent efforts have sought to differentiate DSR from design in practice. This has however resulted in a theoretical body of work that, if rigidly applied, may make the research paradigm less attractive, for example to new entrants in the field. Indeed, an essential process in the practice of DSR remains the design, development and evaluation of IS artefacts. As with most human activities that aim to create something new, these endeavours can be messy in their day-to-day practice. In dealing with the high number of concerns that determine the success of an IS development project (e.g. budget, scope definition, timing, politics, group dynamics) it can be hard to keep track of the elements that allow the practitioner

to contribute to DSR. In this paper, we aim to present an exemplar DSR study with an insight in a multi-disciplinary DSR project that was conducted over multiple years. While doing so, we will focus on the relationship between the operational aspects of IS artefact design and development and the creation of a DSR ISDT. We hope this study's structure will be useful as a template for others in the field.

The paper is structured as follows. First, we discuss the design requirements as they emerge from the current approach to competence assessment and insights from practitioners. Then, the kernel theory of CJ is presented as a framework for addressing the design requirements and research methodology is discussed. Next, design principles are formulated, based on the theory of CJ, which address the design requirements. Additionally, a set of design features that are embedded in an artefact is presented and evaluated. Finally, we discuss the findings in the light of DSR methodology and their relevance to the way competences are assessed. Throughout the text, the components of an ISDT (marked in bold) outlined above are mapped to the findings and an overview of the ISDT is presented in the discussion section.

Design requirements

In this section, we discuss the requirements for an artefact that would improve the credibility of competence assessment while supporting learning by the users of the artefact. These requirements mainly derive from the purpose and scope and have been refined through literature research, discussions with practitioners and evaluation of various versions of the artefact.

Design requirement 1: Valid assessments

In order to advance on current mainstream competence assessment practice, validity should be improved. To understand how this can be achieved, it is necessary to first discuss some of the weaknesses in the current mainstream evaluation practice, i.e. rubrics evaluation. When tasks, designed to gauge competences, are open-ended, they can be addressed in multiple ways. This is for example the case when creativity is a part of the competence being assessed. In such situations, a rubrics evaluation based on pre-set criteria becomes problematic, as the relevant dimensions of the competence increase or are unclear. Problems with validity arise, as it is almost impossible to discern all relevant criteria in advance (Jones & Alcock, 2014). Moreover, students sometimes receive the same overall final score while performing differently on individual criteria (Sadler, 2009). It is therefore questionable if the sum of scores on these criteria adequately represents a competence, as the weighing of the various criteria can be done in many different ways. In other words, criteria-based evaluations are too reductionist in nature (Pollitt, 2004) and questions exist regarding their validity (Jonsson & Svingby, 2007; Pollitt, 2012).

Design requirement 2: Time efficient assessments

Assessments should be time efficient. Assessing competence performance through rubrics is time-consuming. Indeed, the elements in the rubrics need to be carefully designed and assessors need to be trained in order to score the representations adequately (Jonsson & Svingby, 2007). Furthermore, rubrics-based scoring of

representations is time-consuming in itself. An alternative approach should improve on the time efficiency of the process.

Design requirement 3: Reduce cognitive load

Cognitive load should decrease. According to Bejar (2012), rubrics scoring can lead to a high cognitive load, as the assessor needs to take into account a relatively high number of dimensions. This in turn can lead to assessor fatigue, which can influence the quality of the rubrics scoring.

Design requirement 4: Increase reliability

Next to validity, reliability is at stake (Heldsinger & Humphry, 2010, Pollitt, 2012). Assessors differ in their internal standards, as some are stricter than others. Furthermore, assessors do not necessarily interpret the rubric criteria similarly. Consequently, the use of rubrics does not guarantee high inter-rater reliability (Jonsson & Svingby, 2007). Also, the moment in time at which an evaluation is performed can influence the scoring. This is for example the case when the first evaluated student cannot be compared to other students, yet subsequent students can.

Design requirement 5: Support competence development

In order to advance on current mainstream competence assessment practice, competence development should be supported. In education as in many organisations, assessing and monitoring competence development are closely intertwined goals. As assessment is often aimed at stimulating further competence development, feedback based on the assessment is of great importance for both the assessees and the organisations to which they belong. Therefore, effective feedback needs to be provided by the assessment tool.

Design requirement 6: Support accountability

Information should be available that makes it possible to trace the quality of the assessment and thereby support its accountability. Accountability in assessments has increasingly become important, especially when the assessment is "high stake", meaning the consequences of the assessment outcome are great. This is for example the case when evaluating candidates for important jobs. Therefore, providing information about the quality of the assessment is essential (Shaw et al, 2012).

The kernel theory of comparative judgement

The DR introduced above can be addressed through the kernel theory of comparative judgement. A kernel theory is "any descriptive theory that informs artefact construction" (Gregor & Hevner 2013, p340). It should explain why a design works. In an ISDT, this can be used as **justificatory knowledge,** i.e. "The underlying knowledge or theory from the natural or social or design sciences that gives a basis and explanation for the design." (Gregor & Jones 2007, p322).

Thurstone (1927) derived the Law of Comparative Judgement from the observation that an observer's response to a stimulus is not consistent from one occasion to the next. Thurstone (1927) and Laming (1990) both concluded that all human judgement is relative, i.e. humans need something to compare with in order to express the quality of a stimulus. Laming (2003) showed that when people are asked to make an absolute judgement, they still need a point of reference. When none is provided, they will choose their own point of reference, which is not necessarily the one chosen by others. We are therefore more reliable in comparing, than in assigning scores to single performances.

In the evaluation of human competences, this bias can be addressed by asking multiple assessors which of two representations of a competence (e.g. a text representing the competence of argumentative writing) is best. By applying such CJ, subjective differences in judgement are cancelled out. While judges are likely to debate whether representations pass the bar or how many points they deserve, they will more easily agree on which one is better. By applying CJ, the personal standard of assessors becomes less salient, improving the consistency of the judgments between assessors (Pollitt, 2012), which benefits the reliability of the assessment. Through repeated judgement of representations, a rank-order can be created, ranking the different competence representations.

When the CJ approach was first described, at the beginning of the 20th century, it was impossible to implement at a large scale. Indeed, competence representations need to be stored and managed, judgements between representations need to be coordinated, stored and aggregated. All this requires the use of an IS with an advanced statistical backend, able to process amounts of data that are virtually impossible to process manually. Therefore, it is only through the advent of wide-spread IS adoption that CJ has become feasible.

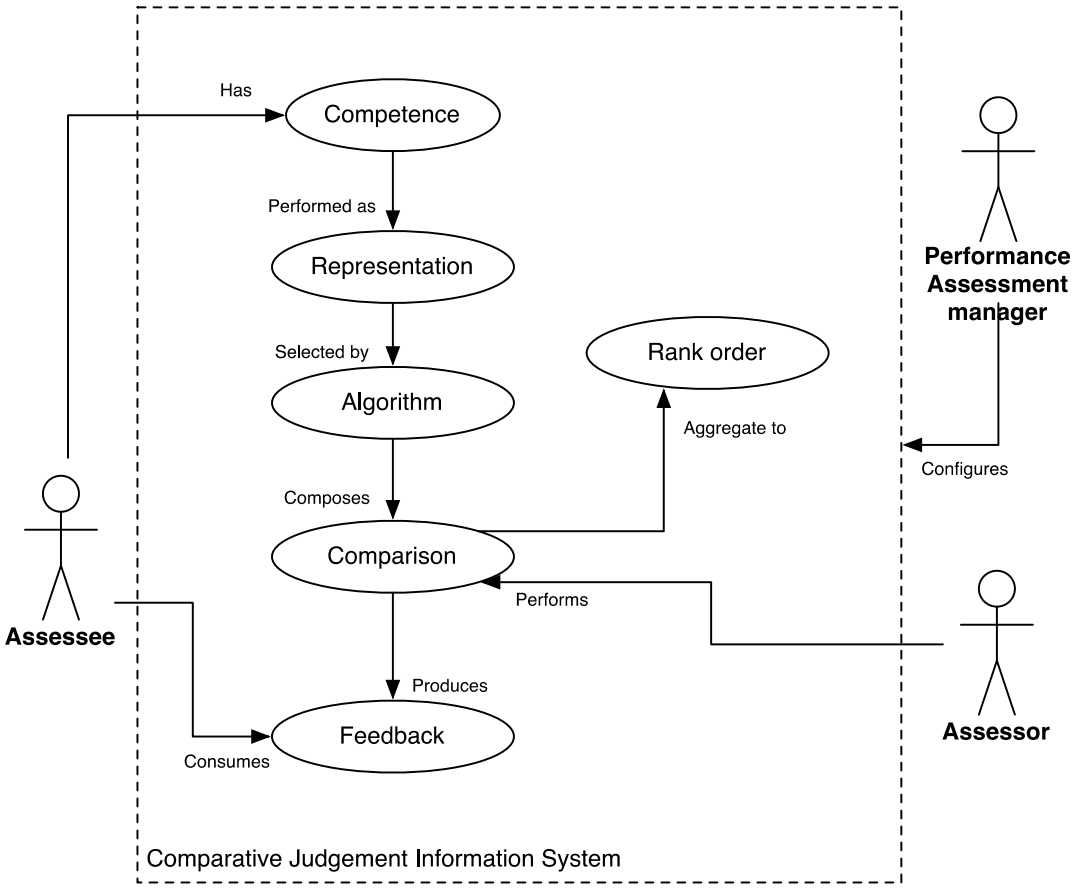

Figure 1: Overview of the constructs in the ISDT

In order to better understand the concepts that will be used throughout the paper, we provide the following description of the key **constructs** and their interdependence (graphically represented in **Figure 1**), based on CJ as a kernel theory (Thurstone, 1927; Laming, 1990; Pollitt, 2012). These constructs are an essential part of an ISDT. An assessee is the person (e.g. student, employee,...) of whom the performance of a competence is being assessed. This assessment is done by a number of assessors (e.g. judges, subject matter experts, a selection committee,...) who compare representations

(e.g. text, video, image or audio) of the competence at hand. A pair of representations of which the assessor must decide the winner is called a comparison. The assessment process (setting up an assessment, inviting assessors and assessees,…) is managed by a performance assessment manager (PAM), who can configure the assessment, which is the collection of comparisons made between the representations. A comparison is selected through a statistical algorithm that selects a pair of representations. This comparison selection can be done in different ways. One way is to select the pairs at random. Another way is through Adaptive Comparative Judgement (ACJ) (Pollit, 2012), in which more refined algorithms are used to make more efficient comparisons. For example, when it is reliably known from previous comparisons that one representation is of high quality and the other representation is of low quality, the decision of the assessor is highly predictable. Selecting such a comparison provides less information for the rank-order than sending out two representations for comparisons with a more equal quality. Thus, selecting what comparisons to send out can lead to a rank order that more quickly converges.

Methodology

Sein et al (2011) and Peffers et al (2007) propose DSR research methodologies that are concrete enough to offer practical guidance to the DSR practitioner. Both approaches have similarities, building on iterative cycles in which objectives are set, development is done and a new cycle begins based on what was learned from an evaluation of the artefact. In terms of differences, Peffers et al's (2007) Design Science Research Methodology (DSRM) does not necessarily advocate the building, intervention and evaluation in an organisational context, which is a central part of Sein et al's (2011) Action Design Research (ADR). ADR is different from stage-gate oriented approaches to DSR, where building, intervention and evaluation are seen as separate phases. Instead, these processes in ADR occur in parallel and are encapsulated by an organizational context.

Iivari (2015) identified two strategies for DSR. Strategy 1 is initiated by researchers to solve a class of problems, thereby making a DSR contribution that can be tested in an organisational context or not. Strategy 2 has a researcher solving a client problem by building a concrete IS artefact and from that experience builds a DSR contribution that addresses a problem class. In strategy 2, the impetus lies in the researcher aiming to solve the client's problem and not necessarily in contributing to DSR from the start. In the course of our research, we have shifted from working along the lines of DSRM in Strategy 1 to aligning more with ADR under Strategy 2. We will refer to DSRM performed as Strategy 1 DSR as mode 1 and ADR performed as strategy 2 DSR as mode 2.

In the DSR approach taken in the context of this paper, rigour (linking context to design) and design (designing and building the artefact) cycles (Hevner 2007) were conducted iteratively as follows. First, we defined the objectives of the Minimum Viable Product (MVP) that was to be created in the period to come. The requirements of each MVP were defined by the project team during a workshop, through the formulation of user stories on post-it notes. These SCRUM (Schwaber, 2004) user stories shared a common structure: as a <user role> I want to <action> in order to be able to <motivation>. After the user story generation phase, all user stories were individually discussed by the members of the project team and prioritized on a flip-chart following the priority levels of the MOSCOW method (Clegg & Barker, 1994). This method arranges user stories according to four different priorities: must have,

should have, could have and won't have. A main factor for determining the priority of the stories were the milestones that lay ahead as a result of the agreements, made between the project team and the various organisations in which field trials were to be organised. The MVP definition workshops lasted for about four hours each and were essential to steer the development efforts. In addition, as the team members came from various disciplines (education, psychology, organisational sciences, IT and DSR) these discussions yielded much insight and understanding in the issues that needed to be tackled in the MVP development cycle ahead.

After each milestone, evaluation of the MVP artefact was conducted using a mix of system log study, ex-post survey and interviews. In the first 3 MVP's, this was was done in non-organisational contexts in mode 1 DSR, while in MVP 4 and 5, this was done in an organisational context in mode 2 DSR.

In this paper, the MVPs represent the **expository instantiations** that are part of an ISDT and that were evaluated during the field trials. In MVP 1, the aim was to allow the comparative judgement of 3 written competence representations, produced by students from 10 different schools. This was done in computer classrooms on the university campus, i.e. in a controlled environment. MVP 2 focussed on addressing the issues that emerged from our evaluation of MVP 1, on developing an ACJ algorithm and on testing if this would benefit the reliability and efficiency of the system as a whole. In addition, we adapted the system for use in non-controlled environments, meaning the IS was accessible on a great variety of devices, wherever the user would choose to engage with it. MVP 3 focussed on providing assessees and assessors with feedback and thus on the learning aspect of the purpose and scope of the artefact. MVP 4 was focussed on addressing the feedback from the MVP 3 evaluation and allowed PAMs to more easily manage assessments. MVP 5 concentrated on allowing assessees to upload and manage their own representations and supporting CJ in peer-assessments.

Through the steps described above, we developed a better understanding of the system as a whole and how it could be constructed, based on the design requirements and the kernel theory of CJ. A deductive approach, based on the kernel theory of CJ, allowed us to identify design features to address the design requirements. In turn, the design principles were defined, starting from the design features, through inductive reasoning.

Design principles

An overview of the relationships between DRs and DPs can be found in Figure 2. The DPs constitute the **principles of form and function** of the ISDT.

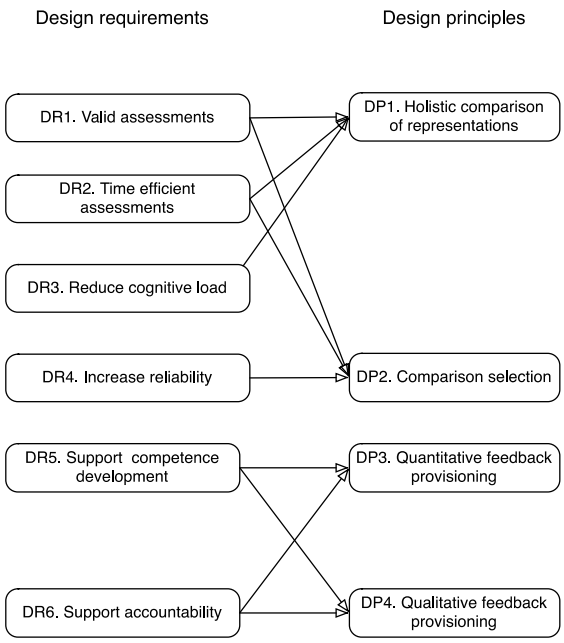

Figure 2: Relationships between Design Requirements and Design Principles.

Design principle 1: Holistic comparison of representations

DP1 addresses DR1 (Valid assessments), DR2 (Time efficient assessments) and DR3 (Reduce cognitive load). DR1 critiques the validity of judging representations using rubrics. In a holistic comparison, a representation is evaluated as a whole, instead of on a variety of subcriteria. By allowing holistic comparison of representations as part of a CJ assessment, more valid judgements are produced (Pollitt, 2012). Indeed, in contrast to rubrics evaluations, assessors can use their expertise to consider characteristics that may not have been thought up in advance, like a creative approach taken by the assessee.

DR2 states that the time efficiency of the assessment should be considered. The more time it costs to perform an assessment, the lower the odds that the IS will be adopted. Therefore, it is essential to allow the quick indication of which representations is best, which can be done more time efficiently though holistic evaluation than through rubric assessment.

According to DR3, reducing cognitive load on the assessors and thereby reducing the possibility of assessor fatigue, which would influence the results, is necessary. Bejar (2012) points to the fact that CJ can be less cognitively straining when done holistically, as no detailed analysis of the representation according to a set of rubric categories is necessary. Holistic evaluation represents a more intuitive, cognitively less demanding task (Greatorex, 2007; Jones et al, 2014).

Design principle 2: Comparison selection

Comparison selection occurs through the algorithms that select the comparisons to be made, determining how many judges are to evaluate each representation. DP2 addresses DR1 (Valid assessments), DR2 (Time efficient assessments) and DR4 (Increase reliability). DR1 is addressed by DP2 by relying on the expertise of multiple assessors, causing the final outcome to more adequately reflect the competence. As more judges shed their light on the representation, each with their own criteria, more aspects of the competence representation are evaluated, increasing validity. Also, one is no longer limited to assessing competences that are easy to evaluate by decomposing them into

subcategories. Therefore, a wider range of tasks can be assessed (Jones & Alcock, 2014) that are more authentic and open.

DP2 addresses DR2, as comparison selection can lead to better ways of selecting representations, reducing the number of comparisons needed. For example, Pollitt (2004) describes approaches in which this may be achieved, e.g. by re-using rank-orders from previous assessments in comparisons with new representations.

DR4 points to the fact that using rubrics scoring as a mode of assessment does not guarantee agreement among assessors, certainly when more open tasks are assessed (Jonsson & Svingby, 2007). Moreover, the moment on which one is judged is critical. For example the assessment of a particular competence representation might be influenced by the quality (e.g. strong or weak) of the previous assessed representation, implying that evaluations are not independent and may become biased. Using comparison selection, reliability is enhanced in several ways. Firstly, differences among assessors in their severity/leniency are filtered out because they only have to indicate which one is best. People are more reliable when doing CJ compared to assigning scores to criteria (Thurstone, 1927; Pollitt, 2012). Secondly, a final score is always based on the view of multiple assessors. Finally, through CJ, metrics on the reliability of the estimation for each representation can be provided. If results indicate that assessors differ in their views on a certain representation, other assessors can be asked to compare this representation to increase the overall reliability of the assessment.

Design principle 3: Quantitative feedback provisioning

DP3 addresses DR5 (Support competence development) and DR6 (Support accountability). The collection of data containing the decisions of all the assessors constitutes a dataset that can be analysed to produce a rank order, ranking the various representations according to quality. This allows comparison between various representations, implying that this rank-order can be used to generate and deliver feedback to assessees on how their competence representation compares to others. By doing this, the assessee becomes aware of how to improve on the competence. It can for example be insightful to compare a representation in the middle of the rank order with the ones at the top, to find out how a competence can be improved.

Quantitative feedback can also make the assessment more accountable (DR6). Indeed, having an insight in the rank order, reliability estimate and other quantitatively aggregated measures (e.g. number of times a representation of an assessee was compared, total time spent comparing a representation to others,…) can provide insight in why a certain outcome was produced.

Design principle 4: Qualitative feedback provisioning

DP4 addresses DR5 (Support competence development) and DR6 (Support accountability). This DP refers to the collection and presentation of the reasons why a certain decision was made. Such information, given by a multitude of assessors and presented in a way that provides the assessee with indications on how to improve, can be highly valuable to the learning process of the assessee (DR5). In addition, it produces an insight in the logic that was followed by each assessor per comparison, which supports the transparency and the accountability of the assessment (DR6).

Design features

The design features are the functionalities that were effectively implemented in a functioning artefact. They constitute the actual manifestation of the DPs and can be

evaluated as part of an expository instantiation of the IS artefact. The list of DFs that we discuss here are the ones that present the greatest differentiator of the artefact under study with respect to IS that reside outside of the purpose and scope covered by the proposed ISDT. Indeed, we could have discussed many other features with associated DRs and DPs that are part of the artefact, like e.g. user account management. Yet, these features are common to many IS, causing their possible discussion to only contribute to the ISDT at hand in a limited way.

MVP5 constitutes the most elaborated expository instantiation, bundling the DFs presented in this section. It has been made available as an open source project under the GPL3 license and can be downloaded or contributed to on GitHub (hyperlink to be added after anonymous review phase). Figure 3 provides an overview of the DRs, DPs and DFs that constitutes the conceptual model forming the core meta-artefact of the ISDT.

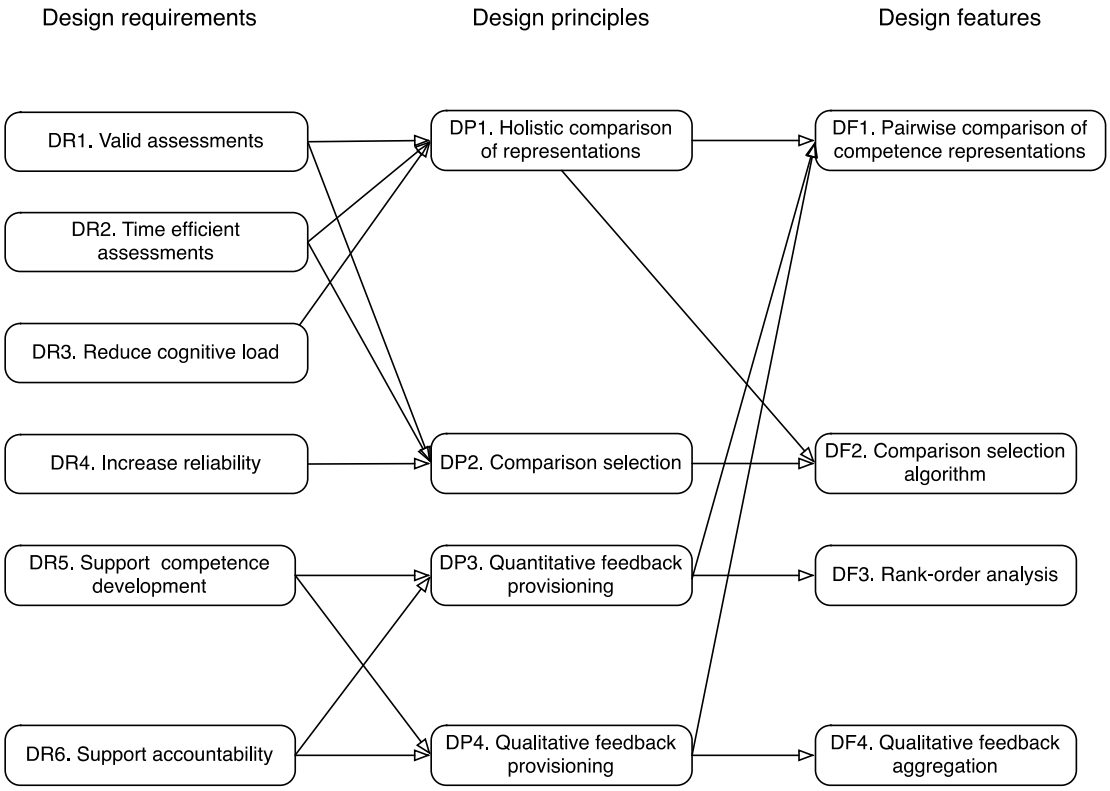

Figure 3: Conceptual model, containing Design requirements, Design principles and Design features.

Design feature 1: Pairwise comparison of text and image representations

We implemented pairwise comparison of text and image representations, allowing their holistic comparison (DP1) and learned that assessors like to see representations next to each other, permitting a better comparison of e.g. text structure. The resulting UI for the MVP3 prototype can be seen in Figure 4: In building this feature, we separated the decision on what constitutes the best competence representation from other data entry steps related to the comparison. In this way, decisions can be made intuitively and holistically, as the assessor only has to decide which one is best without having to make the cognitive effort to elaborate more complicated elements like why one representation is better than the other. This DF also provides the data for the provisioning of quantitative (DP3) and qualitative (DP4) feedback, as the assessors are presented with

subsequent data entry screens, separate from the decision on which representation is best, in which they can motivate their decision.

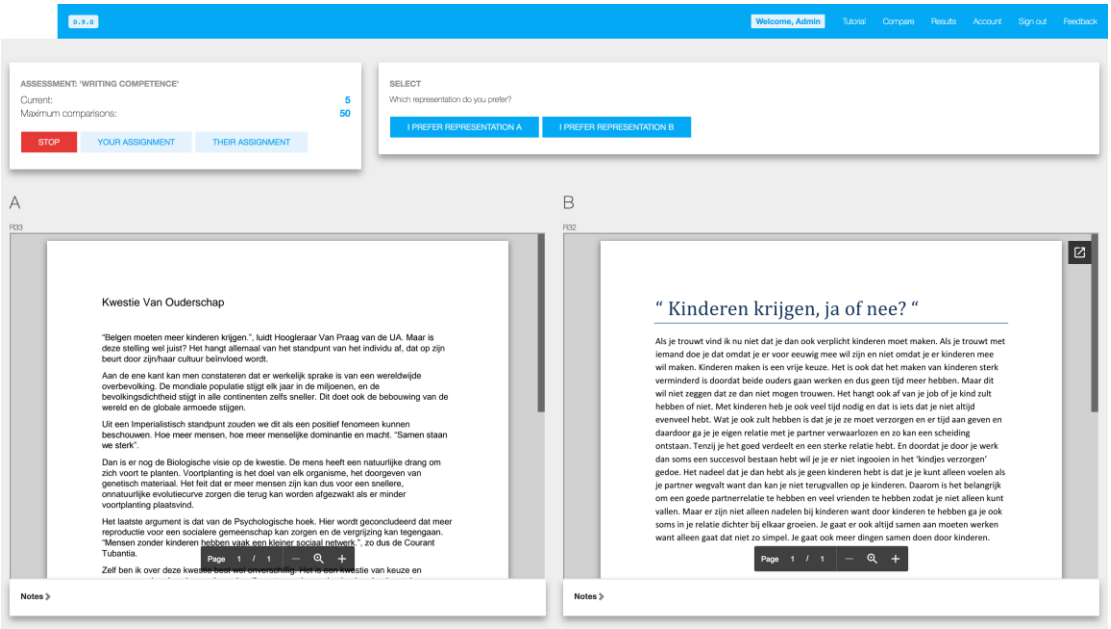

Figure 4: MVP5 UI for the pairwise comparison of text representations.

Design feature 2: Comparison selection algorithm

Comparison selection (DP2) can be done either randomly or adaptively. In the former case, pairs of representations are selected at random and presented to assessors for comparison. In the latter case, ability estimates calculated from decisions in previous comparisons are used to inform the selection of pairs. We have implemented both. An ACJ algorithm was described by Pollitt (2004), but Bramley (2015) concluded that this algorithm is likely to artificially boost reliability, because of a large uncertainty in the reliability metrics, early in the CJ process. Therefore, we developed an alternative adaptive algorithm that uses a previously created rank order as a benchmark and of which the goal was to efficiently place new representations in predetermined categories.

Design feature 3: Rank-order analysis

When enough comparisons are made, a rank-order can be created by using statistical models, attuned to the binary (wins and losses) characteristics of the data (Pollitt, 2012). This rank-order can be used to provide quantitative feedback to the assessee (DP3) through an interval scale, ordering the representations by quality. Figure 5 shows an example rank-order as implemented in MVP3, depicting data collected in evaluation 3. The position of a representation in the rank-order depends on how often it has won from the representations it was compared with.

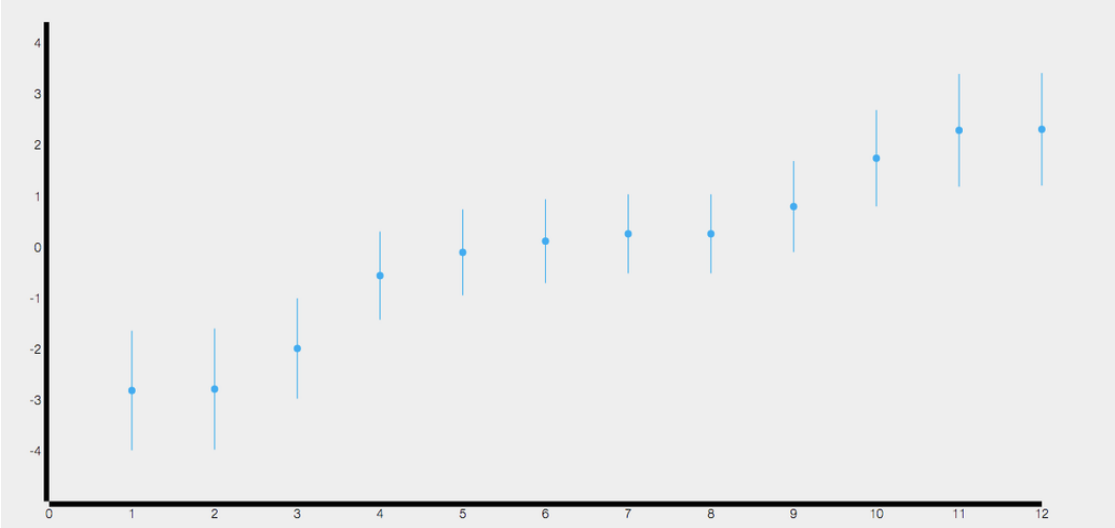

Figure 5: Rank order with 95% confidence intervals produced in evaluation 3. Y-axis represents ability as a measure of quality. X-axis shows individual competence representations.

Design feature 4: Qualitative feedback aggregation

Besides acting as an assessment tool, the IS can also be used as a learning instrument for the assessee, by including the opportunity to add feedback to the comparison and thus providing the assessee with qualitative feedback (DP 4). Qualitative feedback collection was implemented by asking the assessors to formulate positive and negative comments for each representation, as can be seen in Figure 6. By presenting such feedback in an aggregated way per representation, the assessee obtains a nuanced set of feedback items formulated by multiple assessors, indicating how the representation can be improved as well as what was good about it.

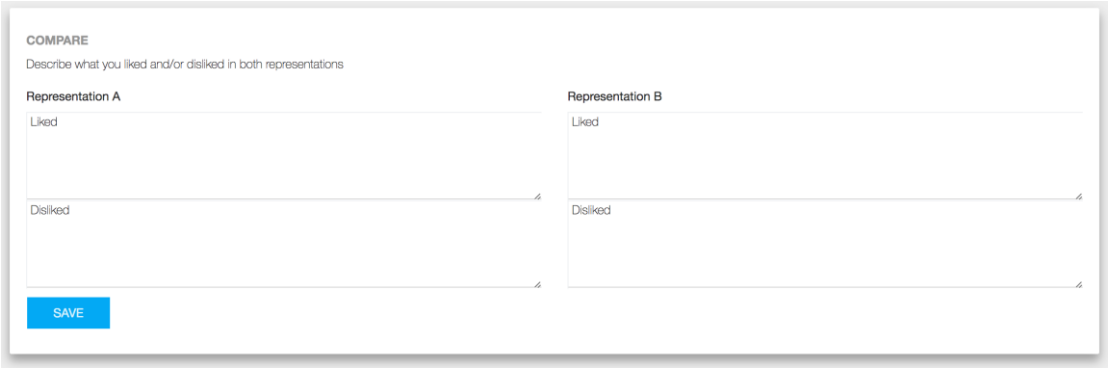

Figure 6: UI element for capturing positive and negative specific feedback.

Evaluation

We assessed the DRs, DPs and DFs in 11 evaluations of the 5 MVPs, of which an overview can be found in Table I and on which we report next. The 15 assessors in evaluation 2 are a subset of the 68 assessors that participated in evaluation 1. The participants in the other evaluations are exclusive to those evaluations, i.e. there was no reuse of participants between evaluations. The SSR is discussed later, on p14.

Eval.	Competence	MVP	Domain	Assessees	Assessors	Algorithm	TAM eval.	DSR mode	Outcome
1	Argumentative writing	1	Education	135 High-school students	68 Teachers and teachers in training	Random	No	Mode 1	SSR Task 1= 0.69 SSR Task 2= 0.61 SSR Task 3= 0.82
2	Argumentative writing	2	Education	30 High-school students	15 Teachers and teachers in training	Adaptive	Yes	Mode 1	80% classified correctly
3	Writing formal letters	2	Education	12 High-school students	11 Teachers in training	Random	No	Mode 1	SSR = 0.68
4	Mathematical problem solving	3	Education	58 High-school students	10 mathematics teachers + 4 mathematics teachers in training	Random	No	Mode 1	SSR Task 1= 0.81 SSR Task 2= 0.80
5	Capability of visual representation in the arts domain	3	Education	11 High-school students (147 representations)	13 teachers	Random	No	Mode 1	SSR= 0.77
6	Interpreting statistical output using peer-evaluation	3	Education	44 Master students	33 Master students	Random	No	Mode 1	SSR= 0.80
7	Entity-relationship modelling	4	Education	30 Bachelor students	30 Bachelor students	Random	Yes	Mode 2	SSR= 0.79
8	Evidence-based practice evaluation	4	Education	93 Bachelor students	93 Bachelor students	Random	Yes	Mode 2	SSR= 0.80
9	Comparison of project proposals for internal university funding	4	Education	20 Applicants for internal university funding	5 Project evaluators	Random	No	Mode 2	SSR= 0.71
10	Evaluation of student papers	5	Education	84 Post-graduate students	4 Lecturers	Random	No	Mode 2	SSR= 0.71
11	CV-screening	5	HR	42 CVs from job applicants	7 recruiters	Random	No	Mode 2	SSR= 0.88

Table I: Evaluations performed throughout the development process

Usability and technology acceptance

Usability is an important precondition to technology acceptance (Davis, 1989) and was especially of concern in the early MVPs. Indeed, we had no clear indication during the early stages of the project that the user would be at ease when using the system. In later MVP evaluations, the usability measurement became less important, as we became more confident of the system's usability. In early MVPs, we assessed the system's usability through the 10-point System Usability Scale (SUS) (Brooke, 1996). The SUS score for MVP 1 was 73.08, placing it in the 67% percentile rank, meaning that 33% of SUS studies analysed for by Sauro & Lewis (2012) had a better score than MVP1. In evaluation 3, which was performed on MVP2, the SUS was 77.5. The SUS survey was combined with observations made during the evaluation of MVP1, which took place in a controlled environment. In addition, ex-post interviews were conducted to identify the main usability issues, which were addressed in later MVPs.

In 3 interventions, we evaluated technology acceptance through the TAM scale introduced by Davis (1989). The TAM questionnaire contains items that relate to the constructs of Ease of Use and Usefulness. We averaged the scores that are associated with each of the constructs in Table II. The evaluation clearly conveys a picture of a system that is found to be easy to use and is also perceived as useful, yet still has some progress to make to further prove its usefulness.

That two of the TAM evaluations took place in MVP 4, which was conducted using mode 2 DSR is notable, as these were performed in realistic organization settings. The lower usefulness of the IS expository instantiation in the mode 2 study when compared to mode 1 told us that the system still needed to better align with the organizational context in which it was deployed. This is something we worked on in the subsequent mode 2 development cycle of MVP 5 by for example adding features that allowed assessees to upload and manage their own representations.

Eval.	Evaluation	MVP	Average Ease of use	Average Usefulness
2	Argumentative writing	2	5.52 (N=12, Stdev=1.67)	4.35 (N=15, Stdev=1,56)
7	Entity-relationship modelling	4	4.9 (N=11, Stdev = 1.39)	4.25 (N=11, Stdev=1.46)
8	Evidence-based practice evaluation	4	4.64 (N=5, Stdev=0.91)	4.17 (N=5, Stdev=1.02)

Table II: Average Ease of use and Usefulness measured by the TAM questionnaire on a 7 point Likert scale ranging from 1=very strongly disagree to 7=very strongly agree

Reliability, validity and efficiency

As shown in Table I, the reliability of the rank order was verified in each evaluation except evaluation 2, through the Scale Separation Reliability (SSR) (Bramley, 2015). The SSR indicates to what degree the spread in the results is not due to measurement error. As can be seen in Table I, these reliabilities were mostly around 0.70 or higher, implying that the relative position of the items on the scale is quite fixed. In other words, if the assessment were to be repeated, we are relatively sure that this would result in a similar rank-order. The validity of the rank-order was analysed for evaluation 1, using collected feedback. The results showed that assessors based their decisions on content-relevant features and that all dimensions that are related to the writing competence were addressed (reference to be added after anonymous review phase). Both are preconditions for content validity (Messick, 1989), meaning that the final rank-orders represent a valid scale in argumentative writing. During evaluation 1, we also compared the CJ rank-orders with the rank-orders generated through rubric evaluations. This resulted in correlations of 0.77 (task 1, $p<0.005$), 0.79 (task 2, $p<0.005$) and 0.85 (task

3, $p<0.005$). The strong correlations demonstrate concurrent validity. Indeed, similar constructs are measured with the different methods, as students who performed well when evaluated through the rubrics, also did well when evaluated through CJ (reference to be added after anonymous review phase).

In evaluation 2, we tested the ACJ algorithm and found that it substantially reduces the number of comparisons that need to be made. With a reduction of 50% in number of comparisons, this algorithm was able to reach a proportion of 0.80 correctly classified representations (reference to be added after anonymous review phase). As the random CJ algorithm outperformed rubrics evaluation and the ACJ algorithm improved on the efficiency of random CJ, ACJ also outperforms rubrics. ACJ is a highly relevant feature that can improve the efficiency of the CJ approach, yet requires an existing rank order, providing benchmarked representations to which new representations can be compared. Therefore, it is only applicable in recurring competence assessments where the same type of representations are evaluated. This reduces the applicability of the algorithm and explains why we did not further test it in subsequent evaluations: such pre-existing scales were not yet available.

In terms of efficiency and as reported in (reference conference paper to be added after anonymous review phase), we found that for the short essays in evaluation 1, a time investment of 19 minutes per representation using random CJ provided a reliability of above 0.70. With double the invested time per presentations (38 minutes) and using rubrics, the reliability level was 0.54. These results suggest that for open-ended tasks and when higher reliability levels are desired, CJ is more efficient than rubrics and that, even given a much higher time investment, rubrics evaluation is not able to reach the reliability of CJ.

Mode 1 and mode 2 DSR

As shown by Table I, earlier evaluations were taking place in mode 1 conditions, while the latter ones took place in mode 2 conditions. Indeed, the latter interventions where situated in existing processes that were already part of the organisation and caused design changes to the system, due to organisational reasons. The main change was the need to hide the ranking in feedback reports, voiced by practitioners in lower education. They saw it as a bad practice for younger students to be compared to each other in terms of their abilities. As this is a point of view that varied between the contexts in which the assessments took place, we decided to let the PAM configure the system to decide if the assessees should be able to compare the rank of their representation to others.

Another concern that surfaced in mode 2 settings was for the privacy of the assessees and specifically for the way in which the representations are stored and who can access them. Certain representations, for example video footage capturing the ability to speak French, proved to be very sensitive. In one video, the assessee started crying halfway through the footage. This underlines the organisational need for security and privacy of assessee data. Part of the way of addressing this is to make the representations anonymous. Still, as the video example shows, this is not straightforward for all media types, as automatically obfuscating the face and voice of assessees in a video requires advanced manipulation techniques.

Discussion

Methodology

In the development and evaluation of MVP1 through MVP3, the nature of the user stories that drove the development was mainly determined by the research objectives that had been identified from literature. However, as the system's development progressed, organisations increasingly became interested in using the IS by seeing it in operation and by learning of its results. As a consequence, the requirements for new MVPs more and more came from organisations themselves and the objectives of the next MVP originated from the evaluation of interventions in their midst. As discussed in the section on mode 1 and mode 2 evaluation, during mode 2 research, organizational requirements became more salient and the artefact was therefore increasingly shaped by the organisational context in which it operated.

Linking this to the MVPs already discussed, our DSR methodology followed mode 1 in MVP 1 to 3 and gradually evolved to mode 2 in MVP4 to MVP5. The transition took place over a period of 30 months. Still, the ADR in mode 2 was not performed in one long stretch at a single organisation as described in Sein et al (2011), but in multiple, shorter building, implement and evaluation phases in various organisations.

The conceptual model (Figure 3) that we have presented in this paper and which constitutes the meta-artifact contributing to DSR discussed by Iivari (2015) was mainly the result of the work in the first 3 MVPs and therefore can be seen as the result of mode 1 research. The model was created by the project team in MVP cycles that were primarily focused on applying the kernel theory of CJ to the evaluation of competences. The subsequent MVPs 4 and 5 yielded more insight in the way organisations wanted to apply the system. Thus, the major advances to the DSR state of the art were produced by mode 1 type research, while the mode 2 type research produced contributions that were "fairly light" when compared to the conceptual model, as Iivari (2015) calls the typical output of Strategy 2 research. However, mode 2 research also resulted in the addition of a DR: DR6 (Support accountability) was a DR that appeared through conversations with practitioners who had seen the artifact in action and reflected on its use in their own context.

ISDT

Applying the structure proposed by Meth et al (2015), has allowed us to propose a conceptual model that serves as a meta-artefact in the ISDT. Many of the components of the ISDT have already been discussed above, and are summarized in Table III. Yet, some of its aspects still remain to be explored, in the remainder of this section.

ISDT component	**Contribution**
Purpose and scope	Improve the credibility of the evaluation of human competences and support learning as a result of the evaluation
Justificatory knowledge	Comparative judgement as a kernel theory applied to the evaluation of competences
Constructs	Derived form the kernel theory of CJ applied to evaluation of competences,

	like assessee, assessor or comparison, as represented in Figure 1
Principles of form and function	Design Principles in the conceptual model (Figure 3)
Testable proposition	Relationships in the conceptual model (Figure 3)
Artefact mutability	Aspects derived from SCRUM backlogs as to be solved in the future
Principles of implementation	Pedagogical aim and feedback type
Expository instantiation	URL to GitHub repo to be added after anonymous review phase

Table III: An overview of the ISDT components

Principles of implementation

The principles of implementation describe the processes for creating an artefact. Gregor and Jones (2007) present examples of such principles, for instance referring to guidelines on the process of normalizing databases. These principles constitute the steps needed to implement an abstract artefact into practice. In our research, such principles became more apparent in the mode 2 phase of our research, where we actually implemented and evaluated the system in real-life organisational settings. We have identified two main principles of implementation: pedagogical aim and feedback type. When implementing the system in an organisation, these principles need to be taken into account to decide on the actual functional instantiation of the artefact.

Pedagogical aim is related to the relationship between DR5 (Support competence development) and DP4 (Qualitative feedback provisioning), and results from the notion that an assessment tool is not necessarily used as a learning instrument. Therefore, when no pedagogical aims are set, the collection of qualitative feedback may be dropped, leading to a considerable efficiency gain, as providing such feedback is the most time-demanding task during comparison.

In terms of feedback type, as stated in the section on naturalistic evaluation, we learned that in some organisations, it is out of the question to show assessees how they have performed in comparison to their peers. This is why we have included the possibility to configure the way in which feedback is shown to whom, allowing the PAM not to show the rank-order to assessees. This principle of implementation is related to the relationship between DR5 (Support competence development) and DF3 (Rank-order analysis).

Testable propositions

Testable propositions can be derived from the various relationships in the conceptual model (Figure 3), as a combination of DRs, DPs and DFs. For example, one could test the proposition that holistic comparison of representations (DP1) through the pairwise comparison of competence representations (DF1) leads to assessments with a higher validity (DR1). Due to the high number of such propositions that can be derived and space limitations, we can not go into detail on them. However, future research will elaborate on these testable propositions, based on the research data we have gathered in the various evaluations of our MVPs.

Artefact mutability

Artefact mutability is about the type of artefact evolution that is anticipated by the ISDT (Gregor & Jones, 2007). Pöppelbuß & Goeken (2015) have shown that artifact mutability is a complex concept that can be interpreted in various ways along 19 different dimensions. One way to approach it, is as the way in which future incarnations of the artifacts in the ISDT will evolve. As we see the conceptual model (Figure 3) and the expository instantiation to be the main artifacts in this work, this is where we position the anticipated change. The main reflections on artefact mutability occurred by looking at the user stories that guided the various SCRUM sprints, and that were grouped by MVP development cycles. As was pointed out before, the objectives of each new MVP were defined by the research team along the lines of mode 1 DSR, mainly driven by the kernel theory of CJ. In the later MVPs, following mode 2 DSR, these user stories were highly influenced by discussions with organisations that would be using the expository artifact within the time-frame of the coming MVP cycle.

As the user stories were prioritized, there were stories that made the backlog of the MVP cycle and others that did not. The ones that did not or the user stories that were not addressed in the previous MVP cycle, formed an important data-source from which to distill the possible mutations that the artifact could be subjected to in the future. We see such application of SCRUM stories as particularly useful to the DSR researcher that operates in mode 2, where one starts with addressing a concrete client problem and the DSR contribution only becomes an important concern later in the process (Iivari, 2015). In such a case, SCRUM sprints and their contents represent a valuable resource to inquire on artifact mutability over time and in doing so contribute to DSR.

As a concrete source of mutability in our IS artefact, we identified a relaxation of the current latent assumption that the assessor and the assessee roles cannot be combined. Indeed, the IS under a different mutation would be useful in a situation where this is not the case, such as in peer-assessment, where assessees can also be assessors. More and more potential peer-evaluation interventions appeared as we progressed towards mode 2 research and we expect these to remain a major driver of our artefact's mutability in the future, impacting DR2 (Time efficient assessments) through DP2 (Comparison selection). Indeed, as the workload for performing comparisons is distributed among assessees, the time that needs to be spent by assessors is reduced. This is especially useful in cases where there are large numbers of assessees and few assessors or teachers, like in Massive Open Online Courses (MOOCs).

Another source of mutability, driven by DR2 (Time efficiency), is the possibility to perform comparisons between more than two representations. In such a case, instead of making a binary judgement on which one is best, representations have to be ranked according to quality. This has the potential advantage of being more efficient, as the amount of information produced is higher for the amount of representation assimilation (e.g. reading of a text) per assessor. However, a trade-off may exist with the cognitive load of the assessors, represented by DR3 (Reduce cognitive load). Such a mutation would manifest itself as a new DF in the conceptual model, complementing DF1 (pairwise comparison).

Conclusion

In this paper, we have aimed to make both a contribution to the application domain of IS to support competence evaluation and to the discipline of DSR. Addressing the former, we have explained that, when assessments are high-impact and the assessed competences are open-ended (creativity, leadership, social skills,…), using rubric

evaluations that assign scores to various sub-dimensions of a competence poses validity and reliability problems. In addition, it imposes a high cognitive load and can be time consuming. Often, assessment is not only summative, but also formative, aiming to support learning. Also, due to an increasing need for transparency, evidence should be provided on the quality of the assessment process for the sake of accountability. To meet these design requirements and aiming to improve the current assessment practice, we have discussed how the kernel theory of Comparative Judgement can help through a number of design principles: performing holistic comparison of competence representations, comparison selection, quantitative and qualitative feedback provisioning.

These principles can be instantiated through the design features that we have presented as part of the expositionary IS artefact under discussion. The structure proposed by Meth et al (2015) was leveraged to show how a meta-artefact can be presented as part of an ISDT while being based on a kernel theory from the behavioural sciences. The expository IS artefact is available as an open-source system for download and extension. The presented work is multidisciplinary, combining the fields of educational sciences, psychology, IT and DSR. We hope that the ISDT will inform builders of similar classes of IS, when applying CJ to the evaluation of competences. A limitation of this work is that we did not investigate the testable propositions that were presented as the different relations that can be deduced from the conceptual model in Figure 3. Future work could address this limitation and focus on testing the various propositions. For example, it would be interesting to study the testable proposition that the pairwise comparison of competence representations through holistic comparison reduces cognitive load.

Another limitation of this work is that the ISDT is targeted on the evaluation of competences. However, we have experienced throughout this project that other domains, requiring complex group decision making, could also benefit from the use of CJ tool support. We therefore see an opportunity for future research to apply the current ISDT for the design and development of CJ IS artefacts in other areas and to expand the ISDT to a broader application domain.

A second aim of the paper was to contribute to the DSR body of knowledge. To guide the DSR practitioner who needs to combine the daily practice of building IS artefacts with theorizing in the DSR paradigm, we have aimed to provide both an insight in how we addressed both tasks. Methodologically, the project shifted from the application of Peffers et al's (2007) Design Science Research Methodology under Iivari's (2015) Strategy 1 (which we called mode 1 DSR) to Sein et al's (2011) Action Design Research under Strategy 2 (that we termed mode 2 DSR), as the development effort became increasingly steered by organisational requirements. The main lines of the ISDT were determined by mode 1 research and complemented by smaller additions through mode 2.

Although we engaged in ADR under mode 2 DSR in the latter MVP cycles, a limitation of this research is that the IS artefact is still under development and has not become a fully embedded part of daily operations in any organisation. As a result, we have not yet been able to explore a mature ensemble artefact that was shaped by the daily demands of professional life. As the artefact matures and its appeal towards organisational inclusion grows, we will be able to study such an artefact in more detail and investigate the organisational ramifications on the CJ IS artefact in future work.

One area that we found promising for the practitioner of mode 2 DSR is the use of artefacts that are created to guide development in the SCRUM agile project management process. User stories and the way in which they progress over the various SCRUM sprints that make out an IS development project are very useful to inform a DSR practitioner on how to formulate certain aspects of an ISDT. Certainly, they can be the basis for the inductive formulation of the principles of form and function. Yet, we discussed that they can also be used to reflect on artefact mutability. In future research, more attention could be given to the way in which the artefacts that support the software development process can inform the formulation of the various components of an ISDT when the DSR practitioner works from practice to theory, as is the case in Iivari's (2015) Strategy 2 DSR.

References

BEJAR, I (2012) Rater cognition: Implications for validity. Educational Measurement: Issues and Practice 31(3), 2–9.

BRAMLEY, T (2015) Investigating the reliability of Adaptive Comparative Judgment. Cambridge Assessment Research Report. Cambridge Assessment , Cambridge.

BROOKE J (1996) SUS-A quick and dirty usability scale. Usability Evaluation in Industry 189(194), 4–7.

CLEGG, D, and BARKER, R (1994) Case Method Fast-Track: A Rad Approach. Addison-Wesley, Boston.

DAVIS, FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology, *Management Information Systems Quarterly* 13(3), 319-340.

DRESCH, A, LACERDA, DP, and ANTUNES J (2014) Design Science Research: A Method for Science and Technology Advancement. Springer, New York.

GREATOREX, J (2007) Contemporary GCSE and A-level awarding: A psychological perspective on the decision-making process used to judge the quality of candidates' work. In Proceedings of the 2007 BERA conference.

GREGOR, S and JONES, D (2007) The anatomy of a design theory. Journal of the Association for Information Systems 8(5), 312–335.

GREGOR, S and HEVNER, AR (2013) Positioning and presenting design science research for maximum impact. Management Information Systems Quarterly 37(2), 337-355.

HELDSINGER, SA and HUMPHRY, SM (2010) Using the Method of Pairwise Comparison to Obtain Reliable Teacher Assessments. The Australian Educational Researcher 37(2), 1-19.

HEVNER, AR, MARCH, ST, PARK, J, and RAM, S (2004) Design science in information systems research. Management Information Systems Quarterly 28(1), 75–105.

HEVNER, AR (2007) A Three Cycle View of Design Science Research. Scandinavian Journal of Information Systems, 19(2), 87-92.

IIVARI, J (2015) Distinguishing and contrasting two strategies for design science research. European Journal of Information Systems 24, 107–115.

JONES, I and ALCOCK, L (2014) Peer assessment without assessment criteria. Studies in Higher Education 3*9*, 1774-1787.

JONES, I, SWAN, M, and POLLITT, A (2014) Assessing mathematical problem solving using comparative judgement. International Journal of Science and Mathematics Education 13(1), 151-177.

JONSSON, A and SVINGBY, G (2007) The use of scoring rubrics: Reliability, validity and educational consequences. Educational research review *2*(2), 130-144.

LAMING, D (1990) The reliability of a certain university examination compared with the precision of absolute judgements. The Quarterly Journal of Experimental Psychology *42*(2), 239-254.

LAMING, D (2003) Human judgment: the eye of the beholder. Thomson Learning, London.

MESSICK, S (1989) Meaning and values in test validation: The science and ethics of assessment. Educational Researcher 18(2), 5-11.

METH, H, MUELLER, B, & MAEDCHE, A (2015) Designing a Requirement Mining System. Journal of the Association for Information Systems 16(9), 799–837.

PEFFERS, K, TUUNANEN, T, ROTHENBERGER, M A, and CHATTERJEE, S (2007) A design science research methodology for information systems research. Journal of Management Information Systems 24(3), 45–77.

POLLITT, A (2004) Let's stop marking exams. Proceedings of the 2004 IAEA Conference.

POLLITT, A (2012) Comparative judgement for assessment. International Journal of Technology and Design Education 22, 157–170.

PÖPPELBUß, J, and GOEKEN, M (2015) Understanding the Elusive Black Box of Artifact Mutability. 12th International Conference on Wirtschaftsinformatik, 1557–1571.

POTTER, T, ENGLUND, L, CHARBONNEAU, J, MACLEAN, MT, NEWELL, J, & ROLL, I (2017) ComPAIR: A New Online Tool Using Adaptive Comparative Judgement to Support Learning with Peer Feedback. Teaching & Learning Inquiry 5(2), 89-113.

SADLER, DR (2009) Indeterminacy in the use of preset criteria for assessment and grading. Assessment & Evaluation in Higher Education 34(2), 159-179.

SAURO, J, and LEWIS, JR (2012) Quantifying the user experience: Practical statistics for user research. Elsevier, Amsterdam.

SCHWABER, K, Agile project management with Scrum. Microsoft Press, Redmont.

SEIN, MK, HENFRIDSSON, O, PURAO, S, ROSSI, M, and LINDGREN, R (2011) Action design research. Management Information Systems Quarterly 35(1), 37–56.

SHAW, S, CRISP, V and JOHNSON, N (2012) A framework for evidencing assessment validity in large-scale, high-stakes international examinations. Assessment in Education: Principles, Policy & Practice 19(2), 159-176.

THURSTONE, LL (1927) A law of comparative judgment. Psychological review 34(4), 273-286.

TARRICONE, P, NEWHOUSE, CP (2016) Using comparative judgement and online technologies in the assessment and measurement of creative performance and capability. International Journal of Educational Technology in Higher Education *13*(1), 16.