Preprint:
Liu, M., Kitto, K. & Buckingham Shum, S. (In Press). Combining Factor Analysis with Writing Analytics for the Formative Assessment of Written Reflection. *Computers in Human Behavior* (Special Issue: Towards Strengthening Links between Learning Analytics and Assessment: Challenges and Potentials of a Promising New Bond).

Combining Factor Analysis with Writing Analytics for the Formative Assessment of Written Reflection

Ming Liu[1,2], Kirsty Kitto[2] & Simon Buckingham Shum[2]

[1] School of Educational Technology, Faculty of Education, Southwest University, CHINA. Email: mingliu@swu.edu.cn
[2] Connected Intelligence Centre, University of Technology Sydney, AUS. Email: Simon.BuckinghamShum@uts.edu.au; Kirsty.Kitto@uts.edu.au

Abstract

The formative assessment of written reflection provides opportunities for students to improve their practice in an iterative manner using reflective writing. However, manual formative assessment of written reflection is time consuming and subjective. While progress has been made in deploying writing analytics tools to provide automated, formative feedback, few approaches to automated assessment are grounded in a validated, theory-based, formative assessment model. To address this, we propose a five-factor model of the *Capability for Written Reflection (CWRef),* grounded in the scholarship of reflective writing pedagogy. This paper uses Confirmatory Factor Analysis to validate the CWRef model by examining the relative contributions of textual features, derived from writing analytics, to each factor in the model, and their contributions to CWRef. The model was evaluated with two reflective writing corpora, showing which textual features, derived using Academic Writing Analytics and Linguistic Inquiry & Word Count, were significant indicators of factors in both corpora. In addition, it was found that the reflective writing context was an important factor influencing the validity of the CWRef model. Finally, we consider how this new analytical assessment model could enable improved tracking of progression in reflective writing, providing the basis for improved formative feedback.

Keywords

Reflection Assessment, Writing Analytics, Factor Analysis

1. Introduction

Expectations are growing regarding the knowledge, skills and dispositions that university graduates should be able to demonstrate in readiness for a fast-changing job market. In response, universities seek increasingly to provide learners with more authentic assessments, that require them to display transferrable skill sets that are often referred to as "Graduate Attributes" (GAs), or "21st Century Competencies", in conjunction with the more discipline-specific abilities that we have come to expect from higher education. A range of approaches are being pursued in the sector, distinctive for the rich, embodied and complex challenges that they provide, in both the social and psychological realms. This raises the question of how to track the emerging competencies of our students, who may not even be directly observable (e.g. while on internship in a company, or working in teams across multiple locations and timezones). One approach is to ensure that assessment criteria incorporate GAs, so that they can be modelled, tracked and reported across diverse assignments (e.g. Thompson, 2016). Another approach is to design more authentic assessments, following principles such as the encouragement of reflexivity and the development of evaluative judgement identified by Herrington and Herrington (2005). A more technical approach is the use of activity-based Learning Analytics, combining mobile, multimodal sensors and analytics to track embodied activity and physiological data, in combination with more conventional data from platform-mediated interaction (e.g. Ochoa & Worsley, 2016). Each of these brings their own strengths and weaknesses to educators and learners, in terms of the insights they can offer, their technical complexity, and the literacies that students and educators require.

Applicable to all of these approaches is the well-known adage that summarizes Dewey's (1933) foundational work on teaching and learning, *"We do not learn from experience... we learn from reflecting on experience".* Critical self-reflection has been recognized increasingly as central to the development of agentic, self-regulated learners. When students engage meaningfully in reflection, they review the way they perceive events and issues, their beliefs, feelings and actions. Such reflective processes in learning have most impact when they are formative and future-oriented (Boud & Falchikov, 2006), which provides mechanisms to encourage meta-cognitive adaptation as students connect their thinking to the wider world (Gibson et al., 2016).

Reflection is a complex, internal process, which leads us to an important question: how can educators gather reliable evidence of student reflection? In this regard, *written reflection* (in private journals, shared blogs, or formal assignments) is by far the most common approach adopted in higher education (although we must acknowledge that students often express reflective thought in other modalities, including audio/video records, giving a presentation, or re-enacting a critical incident for discussion). Reflective writing can be a powerful process for the writer, as well as capturing evidence of significant, even transformative, learning for a different reader. Consider these examples from the literature:

> *"It was a great surprise to me to realize that coordination was such an important aspect of engineering"* (Reidsema, et al., 2010, p.9)

> *"Before I came to this class I had never really thought much about gender and what it means or that it is something that is fluid. Taking this course was completely eye opening and really made me think about things I have never had the chance to think about."* (Buckingham Shum, et al. 2017, p.76)

> *"I had never previously given thought to this idea, as I had thought that a patient's medications and medical conditions are fine to discuss with other family members."* (Lucas, et al. 2019, p.1267)

Despite its evident potential, a growing body of evidence shows that students find reflective writing hard to learn, and moreover, that educators (who often include casual tutors and teaching assistants) also find it hard to teach and assess (Ryan, 2013). Writing in the first person, acknowledging uncertainties and failures, disclosing emotions and feelings, and showing insight into how one is changing as a learner and professional, is an unfamiliar genre for many educators and students. Writing in this way challenges students to share their weaknesses, which goes against almost every other educational experience and form of assessment they have been schooled in. Furthermore, there are rarely clearly 'correct' answers as to how one should act in complex human dilemmas, or how one should make sense of an experience. On what basis, therefore, can written reflection be assessed, and how will students know what the difference is between good and poor reflection?

Research into written reflection for learning has devoted much attention to these questions. One strand of work has focused on the evaluation of individual written reflection on a single scale, such as Mezirow's (1991) three levels of reflection: *non-reflection, reflection* and *critical reflection* (and see also Plack et al., 2007; Wong et al., 1995). This evaluation is often based on the presence of multiple reflective elements, such as the *description, feelings* and *outcomes* elements in Boud et al.'s. (1985) reflection model, or in a modified Bloom's taxonomy (Plack et al., 2007). Other research adds a more formative assessment dimension, where a written reflection can be assessed based on the presence of several important reflective elements, and the assessment of the depth of each (Birney, 2012; Lucas et al., 2017; Poldner et al., 2014). These approaches seek indicators of both the *overall depth* of reflection, and *individual aspects* of reflection. These frameworks provide the language we need to talk more precisely about what good reflective writing looks like, as a proxy for the quality of the author's reflection. However, a significant limiting factor impedes both the empirical validation and the wider adoption of these frameworks in teaching practice: assessing reflective writing is extremely time-consuming.

Learning Analytics is defined in 2011 on the First LAK conference (https://tekri.athabascau.ca/analytics/) as "the measurement, collection, analysis and reporting to data about learners and their contexts, for purposes of understanding and optimizing learning the environments in which it occurs" . While it offers a new generation of tools for educational and learning science researchers to study learning processes, when deployed as an educational technology tool, it also enables new ways to augment learning and teaching as it unfolds, by closing the feedback loop to educators and students. Specifically, *Writing Analytics* (Buckingham Shum, Knight, et al., 2016) emphasises the analysis of written text for the purpose of generating automated feedback to support personal learning, and within that field, *Reflective Writing Analytics* (RWA) uses recent advances in text analytics (i) to automatically *identify reflective elements* at the level of sentence segment level (e.g. Kovanović et al., 2018) or sentences (e.g. Gibson et al., 2017; Ullmann, 2015, 2019), and (ii) to *evaluate reflection depth*, at either the sentence level (e.g. Ullmann, 2019) or document level (e.g. Liu, Buckingham Shum, Mantzourani, & Lucas, 2019). Compared to other established fields, such as automated essay evaluation, Writing Analytics focuses on not only the computational

evaluation of the students' written text, but also the learning design for better integration of the writing analytics tools into classrooms (Liu, Goldsmith, et al., 2019; Shibani et al., 2017).

Recently, Jung and Wise (2020) developed a multi-label classifier which extracted more than 100 textual features from a reflective statement, comparing them with the reflective elements that were identified and evaluated at the document level. These machine learning approaches corroborate earlier corpus-based studies reporting that some of these linguistic textual features were important indicators for the quality of written reflections (Birney, 2012). In particular, Cui et al. (2019) proposed a theoretical framework for reflective writing analytics which attempted to link textual features to conceptual elements of reflection. Despite this conceptual advance, that model fails to elaborate upon *how strongly* the identified textual features affect the quality of the reflective elements, or indeed, whether they impact upon the final quality of the overall reflection.

The aim of this paper is twofold. Firstly, we combine these two streams of work (written reflection assessment and writing analytics) by synthesizing a theoretical assessment model for what we term the *Capability for Written Reflection (CWRef)*. Secondly, this is evaluated using confirmatory factor analysis that links the textual features that can be extracted automatically from texts using writing analytics, to CWRef. We will argue that the analytic model proposed here is more explainable than reflective element classification (e.g. Ullmann, 2019) or depth detection (e.g. Jung & Wise, 2020) because the model measures not only the overall reflection depth of a document, but also the depth of the individual latent reflective factors underpinning this overall assessment — which parts of the writing are stronger and weaker. We will argue that this therefore provides new possibilities for the formative assessment of written reflection.

Two research questions drive the work reported here:

RQ1: How can we quantify and validate the relative contributions that textual features make to the different latent factors underpinning the quality of written reflection?

RQ2: To what degree does this model of reflective writing generalise to different reflective writing contexts?

We make two contributions in responding to these questions. Firstly, we contribute to writing analytics by extending Cui et al.'s (2019) work, which linked low level textual features to reflective elements. We add higher order *rhetorical move* textual features, and then *quantify and validate the relative contributions of these features* to the different reflective elements or factors through confirmatory factor analysis, which extends previous work on reflection detection (Jung & Wise, 2020; Kovanović et al., 2018; Liu, Buckingham Shum, et al., 2019; Ullmann, 2019). Secondly, we contribute to the assessment of written reflection by providing a method for automating this process. In comparison with Birney's (2012) work, which developed a reflective writing assessment instrument based on the judgement of human experts, we propose an automated writing analytics approach. We develop a model comprising five factors, whereby each factor is correlated with textual features that can be extracted using writing analytics. Both the model and the textual features it relies upon are then evaluated based on two reflective writing datasets.

In the remainder of the paper, Section 2 reviews in more depth the existing literature and frameworks related to reflection quality assessment and reflective writing analytics, and describes a more practical written reflection evaluation model, called *Capability of Written Reflection*, derived from this literature. Section 3 describes the methodology linking writing analytics to this new model. Sections 4-6 present the empirical validation of this model against two writing datasets from Pharmacy and Data Science postgraduate students. Section 7 discusses how this approach could in principle help to improve automated feedback, before Section 8 identifies directions for future work.

2. Synthesising the literature to derive a model of written reflection

This section reviews literature on reflection models for assessing the quality of written reflection, from which is synthesised a practical written reflection assessment model. This provides the conceptual foundation for making sense of the textual features that reflective writing analytics can identify.

2.1 Assessing the quality of Written Reflection

A range of models and rubrics tools for assessing the quality of reflection have been proposed, which can be classified into three kinds of formulations.

Firstly, the *single element* of depth model (Mezirow, 1991) defines the overall reflection quality assessed on a single scale ranging from non-reflective to highly reflective. Mezirow's model describes three levels of reflection: (1) Non-reflection (engagement in routine activity with little conscious thought); (2) Reflection (reflects on *what and how* s/he perceives, thinks, feels, or acts); (3) Critical Reflection (reflects on *why* s/he perceives, thinks, feels, or acts in particular ways). Mezirow's model was developed based on his transformative learning theory (Mezirow, 1991) which motivated students to utilise critical thinking and questioning skills to challenge their original assumptions and beliefs through an unpleasant learning experience. This model was useful for assessing short reflective texts, such as entries in personal journals (Plack et al., 2005) and blogs (Wright & Lundy, 2012), in a more summative assessment. However, this reflection assessment model has been critiqued on the grounds that the overall score of written reflection does not provide useful feedback on which aspects of reflections are missing and how to improve them (Koole et al., 2011).

Seeking a more nuanced approach than a single element of depth, *process* models consider reflection quality as resulting from the presence of multiple reflective elements (e.g. Mann, Gordon & MacLeod, 2009; Wong, Kember, Chung & Yan, 1995). The Boud et al. (1985) model, for instance, proposes three elements found in reflective journals or blogs, namely, (1) returning to experience, (2) attending to feelings, and (3) re-evaluation of experience. Boud et al. observe that these elements may not proceed in a linear progression, and the process may involve many cycles and repetitions. Boud's model was developed based on the experiential learning theories of predecessors (e.g. Dewey, 1933; Kolb, 1984) who regarded experience as the stimulus for learning. *Process-based* reflection models provide a formative link between the different elements of reflection, but struggle to assess reflective depth (Boud, 1985; Gibbs, 1988). For example, the Gibbs reflection model (1998) is simple to understand and is used widely in teaching, scaffolding learner reflection with a cycle of activities from *Description*

(what happened?), *Feelings* (what were you thinking?), *Evaluation* (what was good/bad about the situation?), *Analysis* (what sense can you make of the situation?), *Conclusion* (what else could you have done?), to *Action Plan* (if situation arose again what would you do?). However, this model suffers from the criticism that it fails to help students distinguish superficial description from more critical reflection.

Subsequently, several *hybrid models* combining single element and multiple element process models have been proposed, blending the best features of both approaches (e.g. Chirema, 2007; Gibson et al., 2017; Lucas et al., 2019; Moon, 2004; Tsingos, Bosnic-Anticevich, Lonie & Smith, 2015). These studies tend to combine Boud et al.'s reflection process model (Boud, 1985) with Mezirow's (1991) depth model to define a 2-stage process of reflection. The first process *identifies* the reflective elements including description, feelings, associations and outcomes, followed by a second process which *evaluates* the reflection quality as non-reflection, reflection or critical reflection. However, these hybrid models are often detailed and complex, which from a writing analytics perspective, makes it challenging to automate the extraction of useful textual indicators. For example, Lucas et al.'s model (2019) includes 9 reflective elements, while Birney (2012) included 12 elements. One of the criteria in Birney's model is *Links are made to broader social structures* which is not detected in current writing analytics since the detection of such links requires more contextual information. However, this criticism is not always valid, as this is not always required in reflective writing genres (e.g. a reflective project review).

For this reason, we distilled the most commonly used reflective elements/constructs in the hybrid approaches literature, to develop a simpler, but more generalizable five-factor model. Mezirow's reflection model (1991) for assessing reflective thinking capability based on the evidence of students' written work motivated the term *Capability of Written Reflection (CWRef).* As shown in Figure 1 (which will be gradually extended), we define CWRef as a function of four latent factors: *Context, Feelings, Challenges, Changes*.

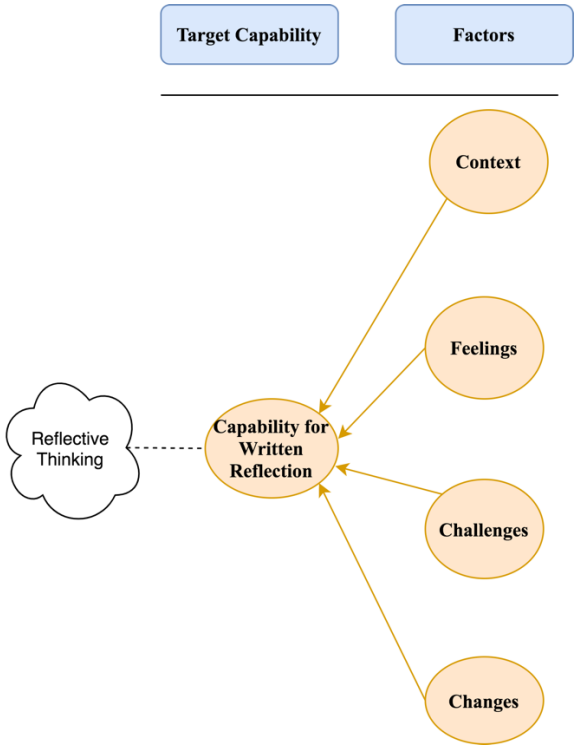

Figure 1: "Reflective Thinking" is operationalised as *Capability for Written Reflection (CWRef)* underpinned by four latent factors: *Context, Feelings, Challenges, Changes.*

Table 1 summarises how each factor is defined in theoretical and empirical reflection models in the literature (e.g. Boud, Keogh, & Walker, 1985; Gibbs, 1988; Gibson et al., 2017). For instance, the *Feelings* factor is acknowledged as an important aspect of reflective learning by researchers (Boud et al., 1985; Brookfield, 1995; Mezirow, 1990), which can be mapped to the *feelings* element defined by their reflection models. Similarly, the *Changes* factor can be mapped to the *change* or *outcome* element in Boud's reflection model (1985), or several elements defined in Birney's model such as *learning is evident*, *insightful understanding evident*, *changes in beliefs* and *revisions to future practices are discussed*.

Key factor (latent variable)	Basis in literature
1. *Capability for Written Reflection* serves as a proxy for reflective thinking	• Mezirow, 1991, The Level of Reflection, *Habitual actions, Reflective action, Premise Reflection*
2. *Context:* differences between learners in their initial thoughts about a learning event, linking their experience to their knowledge, beliefs or assumptions	• Gibson et al., 2017, Reflection Framework, *context* stage • Lucas et al., 2019, Reflective Rubric, *returning to experience* stage • Birney, 2012, Reflective Rubric, *Clear description of context* • Gibbs,1998, *description* stage
3. *Feelings:* the degree to which individuals feel positive or negative about their experience relating to future personal learning	• Gibson et al., 2017, reflection framework, *feelings* stage • Lucas et al., 2019, reflective rubric, *attending to feelings* stage • Birney, 2012, Reflective Rubric, *Self-awareness is evident* • Gibbs, 1998, *feelings* stage
4. *Challenges:* differences between learners in their critical analysis of difficulties experienced. Critical reflectors, for instance, tend to describe the impact of a problem on their goals and criticize themselves	• Gibson et al., 2017, Reflection Framework, *challenge* stage • Birney, 2012, Reflective Rubric, *Issues correctly identified* • Gibbs, 1998, *evaluation* and *analysis* stages
5. *Changes:* the extent to which individuals feel they learned from their experience, and how it may shape their future plans/behaviour	• Gibson et al., 2017, Reflection Framework, *challenge* stage • Lucas et al., 2019, Reflective Rubric, *outcomes of reflection* stage • Birney, 2012, Reflective Rubric, *Changes in beliefs or understanding are evident: Revisions to future practice are discussed* • Gibbs, 1998, *conclusion and action plan* stage

Table 1: Grounding the five key factors *Context, Feelings, Challenges* and *Changes* in prior scholarship in reflective writing

Next (Table 2) we show the learner's potential progression (left to right) in three levels of the *depth* of reflection, shown as the three columns *Non-Reflector, Reflector* and *Critical Reflector* (Mezirow, 1991). The cells are expressed in a form similar to an assessment rubric, drawing inspiration from the hybrid model of Lucas et al (2017; 2018; 2019) and Gibson et al (2017). Each row in Table 2 refers to a factor, while each cell articulates the level of reflection.

Constructs	Depth of Reflection		
***Capability for Written Reflection (CWRef)*: The ability to evidence, in writing, critical reflection on a challenging experience**	**Non-Reflector** **Habitual action:** Engages in routine activity with little conscious thought (context)	**Reflector** **Reflective action:** Reflects on what and how s/he perceives (context), thinks (challenges), feels (feelings), or acts (changes)	**Critical Reflector** **Premise reflection:** Reflects on *why* s/he perceives (context), thinks (challenges), feels (feelings), or acts (changes) in particular ways
…is a function of…			
***Context*: The observed learning experience**	Describes a learning event	Highlights a learning event, linking it to prior knowledge, beliefs or assumptions	Highlights a learning event, linking it to prior knowledge, beliefs or assumptions, and explains the reason for this association
***Feelings*: Feelings present during the initial experience**	Shows little or no evidence of personal feelings, thoughts, reactions	Evidences some feelings about an experience, but does not explain why I feel this way	Evidences personal feelings (positive and/or negative) about an experience, and explains the cause for such feelings, and connects them to challenges
***Challenges*: The difficulties/problems encountered during the experience**	Shares no evidence of any problems encountered	Evidences one or more problems and explains why and how they were challenging	Evidences the impact of one or more problems on goals, and shares ideas on how to address this
***Changes*: Lessons learned and future plans**	Shares no evidence of potential solutions or learning opportunities	Evidences potential solutions or learning opportunities	Evidences learning opportunities from own and other perspectives, and/or considers how change is likely to lead to future benefits

Table 2: *Capability for Written Reflection (CWRef)* is classified using three-levels of progression *(Non-Reflector, Reflector, Critical Reflector). CWRef* is underpinned by four latent factors *(Context, Feelings, Challenges, Changes),* with each of the three levels differentiated by what learners evidence in their writing.

Figure 2 now elaborates Figure 1 to specify a set of *behavioural indicators* corresponding to the cells in Table 2, and for each indicator, a set of *derived variables* as quantifiable textual features. The behavioural indicators were selected to distinguish deeper reflector from non-reflector behaviours (Table 2), as identified by Birney (2012) and Lucas et al. (2019). The rationale for the derived textual variables is as follows:

- **Context:** Deep reflectors tend to link more personal prior knowledge or belief to the highlighted learning events and explain the reason for this association (Lucas, Smith, et al., 2019). Thus, the derived variables include the number of first-person pronouns

(Ullmann, 2017), initial thoughts about highlighted events, beliefs and self-interpretation sentences (i.e. how the writer interpreted this learning event) (Gibson et al., 2017).

- **Challenges:** Describing problems (Gibson et al., 2017; Ullmann, 2017) is a good indicator that the writer is describing challenges. Further explanations including causal words (Birney, 2012) about why these problems are challenging, and self-critique are further responses to challenging experiences (Gibson et al., 2017; Lucas, Smith, et al., 2019).
- **Feelings:** Studies have shown that feelings in reflection are often evidenced by the presence of positive and negative emotional words, and first-person pronouns (Lin et al., 2016; Ullmann, 2017). Furthermore, providing explanations for one's feelings are indicators of greater depth (Gibson et al., 2017; Lucas, Smith, et al., 2019).
- **Change:** The use of first-person pronouns (Birney, 2012; Ullmann, 2017) and analytical thinking words (Kovanović et al., 2018) have been found to be good indicators of writing about personal change. Further clarification of potential solutions and learning opportunities (e.g. change sentences) are additional indicators of depth around how the author is changing (Gibson et al., 2017; Lucas, Smith, et al., 2019).

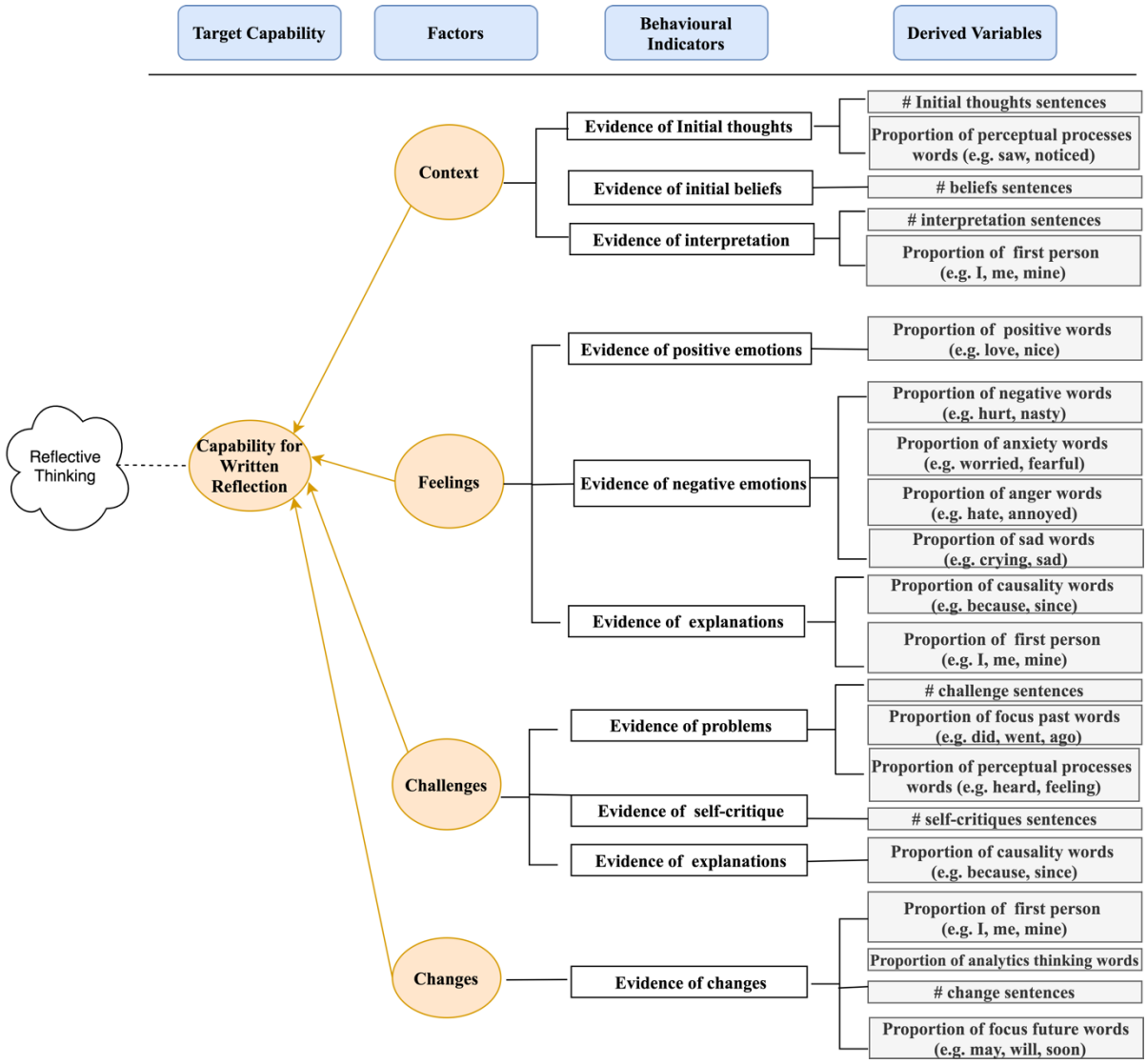

Figure 2: Extending Figure 1 with *behavioural indicators* and *derived variables* for each of the four latent factors: *Context, Feelings, Challenges, Changes.*

To summarise, the model defining *Capability for Written Reflection* is a synthesis of existing scholarship in reflective pedagogy, providing an explicit definition of the scope (breadth and depth) of the capability we want to assess, as a behavioural proxy for "reflective thinking". Critically, in this form it is still qualitative, a point to which we will return in Section 3 (Stage 3) using Confirmatory Factor Analysis (CFA).

Next, we introduce approaches to writing analytics that can identify automatically the textual features specified above.

2.2 Reflective Writing Analytics

Building on advances in text analytics, and the increasing availability of the technologies, Writing Analytics research has developed tools that can *classify the reflective elements* of text at a sentence level (Gibson et al., 2017; Kovanović et al., 2018; Ullmann, 2015, 2019), as the first step to *assessing the reflection quality*, at either the sentence level (Ullmann, 2019) or document level (Liu, Buckingham Shum, et al., 2019). Gibson et al. (2017) proposed a theoretical reflection model and developed concept mapping rules to identify three reflective rhetorical moves (*context, challenges* and *changes*), and three expression types (*emotions, beliefs* and *self-critique*) that frequently occur in reflective texts. These rules have been integrated into the Academic Writing Analytics (AWA) project, which has developed a text analytics platform powering a web-based reflective writing feedback tool (Buckingham Shum et al., 2017; Gibson et al., 2017; Knight et al., 2020). This type of rule-based approach requires linguistics experts and domain experts to work together to manually develop classification rules that are tested on a representative corpus.

An alternative approach that avoids such expensive effort, but requires a larger corpus, is to use machine learning to identify relevant reflective elements. Ullman (2019) developed a large annotated reflective writing dataset (around 5000 sentences) from the British Academic Writing English Corpus, identifying eight reflective elements using annotations at the sentence level: *experience, feelings, personal beliefs, recognizing difficulties, perspectives, lessons learned* and *future intentions.* He then used the most frequent words derived from the annotated dataset as features to train a statistical classifier, obtaining results with a moderate or higher reliability rating between machine and human (Cohen's kappa ranges between .53 and .85). One important limitation of this approach may be the model overfitting to the particular dataset used.

Moving beyond using keywords derived from the dataset to create features, another machine learning approach to reflective element detection is based on existing lexical dictionaries, such as *Linguistic Inquiry and Word Count (LIWC)* (Pennebaker, Booth, Boyd, & Francis, 2015), a text analytics tool that contains a list of words associated with psychological, cognitive and emotional processes. The default LIWC2015 Dictionary is composed of almost 6,400 words manually compiled by human experts over 15 years, collecting and validating different groups of psychological processing words from various lexicon resources, such as PANAS (Watson et al., 1988), Roget's Thesaurus, and standard English dictionaries. Similar to LIWC, Coh-Metrix is another text analytics tool, automatically extracting 109 features of text cohesion (i.e., referential, causal, co-reference, temporal, spatial, and structural cohesion), text complexity and readability from a document (Graesser et al., 2004). Kovanovic et al. (2018) combined LIWC features with Coh-Metrix cohesion and grammatical features to build a random-forest classifier capable of identifying 3 types of reflective element: *observation*, *motive* and *goal.* This produced a reflective writing dataset (containing 3324 annotated sentences), of arts students' reflections on their musical performance. They found that some LIWC features, such as LIWC.see (e.g. *view, saw*), LIWC.focuspast (e.g. *ago, did*), were important indicators of reflective elements. Subsequently, Liu et al. (2019) used a combination of LIWC features and

Academic Writing Analytics (AWA)[1] features to classify Pharmacy students' reflective statements at the document level, demonstrating that both LIWC and AWA features were important indicators of reflection quality (these are introduced in more detail below).

Several corpus-based approaches have investigated (using manual, not automated techniques) the relationship between the quality of written reflection and the linguistic features that they possess. Based on the analysis of six student reflective reports, Luk (2008) found that linguistic features such as linking devices (e.g. however, because, therefore) and hedges (e.g. might, could) are useful indicators to differentiate high grade reports from low grade reports. Similar to Luk's study, Reidsema and Mort (2009) analysed 20 reflective journals and found that high scoring reports used significantly more causal and appraisal resources, and slightly more temporal resources than the low scoring reports did. They concluded that texts scoring high on a reflective writing task were also linguistically richer. Birney (2012) analysed 27 reflective blogs and journals, and identified and weighted 12 constructs or indicators of reflection depth based on expert judgement. These 12 constructs include *context description, issues identification, analysis, implications of actions, multiple perspectives examination, learning* and *changes in beliefs.* She found moderate correlations between quality of the construct and some linguistic features defined by Ryan (2011). In sum, these corpus-based studies have demonstrated that linguistic features are important indicators for the quality of written reflections. Promisingly, some linguistic features (such as causal, appraisal, first person voice, future tense verbs, thinking and sensing verbs), can be automatically extracted in writing analytics (Liu, Buckingham Shum, Mantzourani, & Lucas, 2019).

Besides identifying the importance of textual features to the overall quality of written reflection mentioned in the corpus-based studies, researchers seek to align textual features to individual reflective elements, in order to provide insights for assessing these elements. Most recently, Cui et al. (2019) proposed a conceptual reflective writing analytics framework, which links the textual features derived from reflective writing analytics to reflective elements described in Gibbs' reflection model, using this alignment to analyse reflection variation across students, and over time. For example, the *Description* element is linked to LIWC perceptual processes (e.g. see, hear), and past-oriented features (e.g. ago, did), *Analysis* is linked to cognitive process features (e.g. causation, tentative), and *Feelings* is linked to affective process features (e.g. positive, anxiety). However, this alignment of LIWC features to corresponding reflection elements has not yet been validated, and the overall quality of students' written reflections was not quantified.

To summarise, in educational tools such as AcaWriter, automated, formative feedback is now possible based on the presence/absence, positioning, and sequencing of textual features. However, a key obstacle to providing better feedback is that the *depth* of reflection has, to date, remained opaque to machine processing, which in turn depends on more nuanced insight into which dimensions of writing need most attention. Progress in the field has established a sound rationale for mapping from particular textual features (as researched in writing analytics) to higher order constructs in models of reflective writing (as researched in the assessment of reflective writing). What has to date remained undefined is the *strength of mapping* between

[1] AWA features were extracted using the Text Analytics Pipeline (TAP) web services, which underpin the AcaWriter automated feedback tool, hosted at University of Technology Sydney (Knight et al., 2020; Liu et al., 2018). TAP and AcaWriter are available open source: https://cic.uts.edu.au/open-source-writing-analytics

(1) text features and model constructs (hence RQ1). Moreover, such a model needs to be tested in different reflective writing contexts (RQ2). To address these limitations, we now describe the methodology we developed to quantify the five-factor CWRef model, in the context of written reflection corpora from two different disciplines.

3. Methodology

This section first describes the concept of Confirmatory Factor Analysis (CFA), before detailing a methodology which shows how CFA was used to link recent advances in writing analytics to the formative assessment of reflective writing.

3.1 Confirmatory Factor Analysis

Confirmatory Factor Analysis is a theory-based sub-method of Structural Equation Modelling (SEM) that can be used to test how well theoretically grounded constructs are supported by observed data (Bollen, 1989; Mueller and Hancock, 2015). Thus, CFA evaluates the fit of observed data to a theoretically imposed model, often specifying assumed causal relations between latent factors and their observed indicator variables (Bollen, 1989; Satorra, 1990). The causal relations are expressed as a system of regression-like structural equations, which allow us to measure the latent factors from the observed variables.

In the CWRef model introduced above, the four reflection factors (*context*, *feelings*, *challenges* and *changes*) are measured from the observed textual features by using CFA, while the overall capability factor is derived from these four factors. The most common method used for parameter estimation in CFA models is maximum likelihood, which tries to minimize the differences between the model-implied covariances and the sample covariances of observed variables, based on the assumption that the variables follow a multivariate normal distribution (Bollen, 1989; Satorra, 1990); other methods can be considered if that assumption is violated (Li, 2016).

CFA has been applied to understanding latent constructs mainly based on questionnaire items, for example, Lethridge et al. (2013) used CFA to test the five-factor structure of the reflection questionnaire designed by Kember and Leung (2000). More recently, Fincham et al. (2019) used CFA to understand student engagement based on trace logs (e.g. the number of days students log in, or the number of unique videos watched), forum post sentiment and other textual features (e.g. post narrativity, syntactic simplicity and cohesion). The sample size recommended for CFA ranges from 100 to over 1,000 and the ratio of N to the number of variables should be greater than 10 (Myers et al., 2011). However, to our knowledge CFA has not yet been used to validate a model of reflection with respect to the output from automated writing analytics. While CFA is frequently used with questionnaire items in measurement theory, the method is general – it can be used to establish the factor relationships between any set of observed variables. As such, in this paper we will investigate the utility of extending this method to the analysis of semantic features of reflective writing. We consider this move justified because: (i) previous studies (e.g. Cui et al., 2019;Gibson et al.,2017) have pointed towards a markedly consistent mapping of reflective writing into a coherent number of latent variables. Each latent variable is mapped into a group of textual features indicators (illustrated in Figure 2); (ii) the textual feature scores have found to useful in measuring the level of

reflection (e.g. Liu, Buckingham Shum et al., 2019; Jung & Wise, 2020). In addition, our experimental results also showed the usefulness of textual indicators (see Table 4).

In the CWRef model introduced above, the link between the theoretically grounded four reflection factors (*context*, *feelings*, *challenges* and *changes*) and the observed textual features is measured using CFA. The overall capability factor is then derived from these four factors using a similar linkage.

3.2 Linking Writing Analytics to Formative Assessment

This section describes a four-stage process that we followed to: (i) extract text features from reflections using writing analytics and map them into the CWRef model; (ii) select features that contribute most to the quality of a written reflection; (iii) perform CFA to fit the CWRef model to a specified dataset, and (iv) finally validate the structure of the latent factors in CWRef model that were hypothesised in Table 2. As we focus on the evaluation of a theoretically justified and explainable model which is useful for automatic feedback generation, we did not make use of exploratory factor analysis to find out latent factors derived from a specific dataset. Instead, feature selection is utilised to link extracted features to a theoretically grounded model which is then evaluated using CFA. Thus, the many insights that have already been gained in this field are *imposed* upon the data, rather than a more ad hoc discovery process.

Stage 1. Textual feature extraction and mapping

The theoretically motivated model of reflective thinking (Table 2) was used to develop a "clicks to constructs" alignment (Buckingham Shum & Crick, 2016; Wise, Knight, & Buckingham Shum, In Press) that maps the textual features extracted from a reflective text to the latent factors proposed in the CWRef model. Buckingham Shum and Crick (2016) defined "from clicks to constructs" as the new version of making inferences from behaviour to constructs. For example, Milligan and Griffin (2016) constructed log file activity measures of MOOC-based learning capability's constituent sub-capabilities, such as *Critical Consumption, Production Orientation, and Risk Taking*, allowing a MOOC learner to be automatically evaluated on a scale from novice to expert.
Based on reflective writing analytics research (Cui et al., 2019; Kovanović et al., 2018; Ullmann, 2017), this mapping is shown in Figure 3, elaborating on Figure 2 by linking LIWC first person pronoun (LIWC.i – e.g. *I, me*), and perceptual features (LIWC.percept – e.g. *noticed, saw*) to the *Context* factor (Kovanović et al., 2018; Ullmann, 2017).

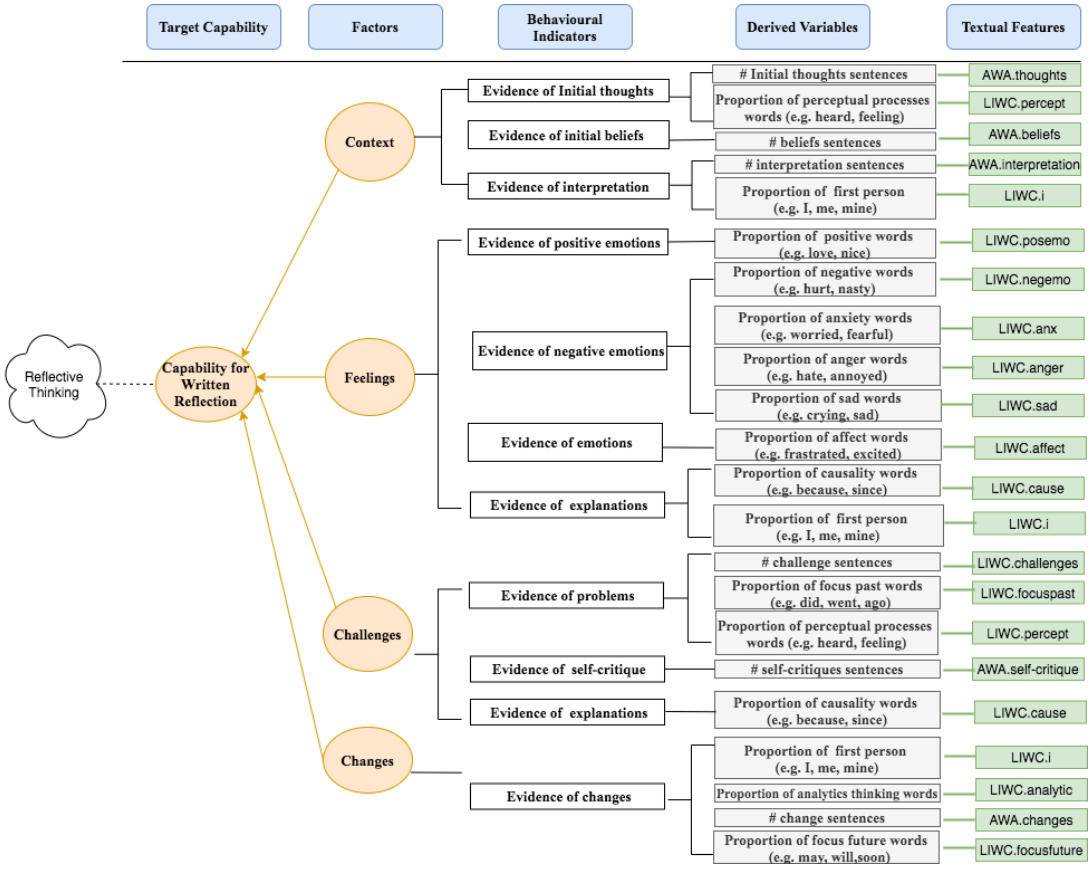

Figure 3: Factor map for *CWRef*, extending Figure 1 with AWA and LIWC textual features.

Similarly, LIWC emotional process features are linked to the *feelings* factor (Lin et al., 2016; Ullmann, 2017). These include the percentage of positive emotional words (LIWC.posemo) and the percentage of negative emotional words (LIWC.negemo, LIWC.sad, LIWC.anger and LIWC.anxiety) as well as LIWC first person pronoun. LIWC causality words (LIWC.Cause – e.g. *because, since*), perceptual features, as well as past-oriented features (LIWC.focuspast – e.g. *went, did*) are linked to the *Challenges* factor (Kovanović et al., 2018; Ullmann, 2017). Finally, the LIWC future-oriented feature (LIWC.focusfuture – e.g. *may, will, soon*), and analytics thinking words (LIWC.analytic – e.g. think, few) as well as first person pronouns, are linked to the *Changes* factor (Kovanović et al., 2018; Ullmann, 2017).

We extended this mapping by adding AWA features from Gibson et al.'s (2017) conceptual model of reflection. AWA *thoughts* and *interpretation* features (e.g. *That early role-play felt distant and impersonal, as I made a conscious effort not let my emotions interfere with the job I had been given.*) and *beliefs* features (e.g. *I believed that good teamwork is the key to success in design activities when time and resources are limited.*) were linked to the *context* factor, while AWA *challenges* (e.g. *I immediately froze as I dwelled upon the fact that I didn't take a patient-centred approach*.) and *self-critiques (e.g. Relating back to the situation I faced with the gentleman, there were certain improvements I could've made.)* features were mapped to the *challenges* factor. Those features are intended to capture the differences of reflection level under each factor of reflective thinking. With the addition of AWA features in addition to LIWC, this is a slight extension of Cui, et al.'s (2019) model.

Stage 2. Textual feature selection and normalization

At this stage the datasets (detailed in Section 4) were analysed and a feature selection process carried out using Principal Component Analysis (PCA). This step was necessary because reflective writing rubrics or prompts vary across learning contexts, which might lead to textual features specific to a particular assessment rubric, that are not strongly correlated to the CWRef factor. We selected features using exploratory factor analysis to ensure stability of the writing analytics features extracted from the data for CFA (Hurley et al., 1997; Schumm and Stevens, 1993). This method relies upon normally distributed data, so multivariate normality tests were performed to select normally distributed textual features. We applied the standard approximation of Hair et al. (2010) for this task, considering data to be normal if Skewness is between -2 to +2 and Kurtosis is between -5 to +5. A series of PCAs were then performed to select the most important features contributing to latent factors, choosing the higher weighted features (above 0.4) of the first PCA component, as suggested by Schumm and Stevens (1993). Lastly, because the scales of textual features are different, such as AWA.challenges (the number of challenges sentences) and LIWC.analytic (the proportion of number of analytics thinking words), z-score data normalization was performed.

Stage 3. Measurement modelling using CFA

To test the proposed five-factor CWRef capability model described in Figure 3, we performed CFA with maximum likelihood estimation (Kline, 2011). Based on the factor map shown in Figure 3, the selected textual features from Stage 2 were hypothesised to contribute to four latent reflection factors, context, feelings, challenges and changes. These four first-order latent factors were then conceptualised to contribute to a second-order latent factor, reflective thinking. Second-order CFAs were conducted to examine the contributions of the four reflective factors to an overall factor of reflective thinking. Data analysis was conducted using Lavaan (Rosseel, 2012), an R package for performing CFA, and the factor reliability test was performed using SemTools (2019).

Stage 4. Validity and Reliability investigations

Four measurements are generally used to test the validity and reliability of a CFA (Hair et al., 2010) and so adopted here. *Convergent validity* evaluates how strongly indicators converge on a single factor through an assessment of item factor loadings and their statistical significance, followed by an assessment of the factors' average variance extracted (AVE). Convergent validity is indicated by an item factor loading and the AVE of a factor greater than 0.5 with $p < .05$ (Hair et al., 2010). *Factor reliability* (CR) is a measure of internal consistency in scale items, which is calculated by the ratio of true score variance to total observed score variance (Fornell & Larcker, 1981). According to Hair et al. (2014), the minimum CR value should exceed 0.7. *Discriminant validity* refers to the extent to which factors are distinct and uncorrelated. Factors are considered discriminant when the square root of AVE values is greater than the correlations between any two factors (Fornell & Larcker, 1981). Finally, *criterion validity,* in this study, is evaluated in terms of the degree of correlation between the computed CWRef scores, and human teacher writing grades awarded to the reflective writing assignments.

A set of common *goodness of fit* indices were also used to evaluate the model's fit to the text corpora: Chi-squared; both Comparative Fit Index (CFI) and Tucker-Lewis index (TLI) above .90 indicate acceptable fit (Bentler, 1990). Root mean square error of approximation (RMSEA) below 0.08 indicates acceptable and good fit (MacCallum et al., 1996).

4. Reflection Contexts: Pharmacy and Data Science Masters

This section describes the empirical evaluation of the CWRef model that was performed using two independent datasets collected in authentic reflective learning environments. We followed the process described in the previous section to fit CWRef to these two datasets, each of which was generated from different learning designs and assessment regimes for reflective writing, described next.

4.1 Pharmacy Work Placement Reflection

The first reflection context comes from second-year Masters Pharmacy work placements in the United Kingdom. In total, 43 Pharmacy students participated in an experiential placement where they worked in a community Pharmacy or non-traditional setting such as an optician or a care home (Mantzourani et al., 2016; Mantzourani & Hughes, 2016). Students were asked to complete a reflective account, where prompts in a template were used to facilitate reflection. Examples of prompt questions include: *"Thinking about your professional development, what went well during placements? What was the highlight? What have you learned? How was this different to what you thought/expected? How did you feel at the time? Please tell us about something that happened in your placements that made you reflect on your role as a pharmacist in patient care and/or the role of other health and social care professionals?"* The template had been developed via multiple cycles of action research involving placement supervisor and student input (Deslandes et al., 2018). Each student wrote 7 reflective statements, producing a total number of 301 reflective statements, all of which were graded against a reflective rubric (Lucas et al., 2017; Tsingos et al., 2015): A score of 0 was assigned where the student had not demonstrated any reflective skills in the writing (classified as *Non-Reflective* – see Table 2, column 1), a score of 0.5 when an attempt was made to relate experiences or feelings to prior knowledge and identify learning (Table 2, column 2: *Reflective*), and a score of 1 when clear links were made between experiences, feelings, and learning, along with a demonstration of a change in behaviour (Table 2, column 3: *Critically Reflective*). Four human experts assessed the same set of reflective statements (Lucas et al., 2017; Tsingos et al., 2015). Human experts reached moderate to substantial agreement (intra-class correlation coefficient= 0.55-0.69, $p<0.001$) on rating these reflective elements.

4.2 Data Science Project Review Reflection

The second reflection context comes from a group project in a statistics course, delivered within an Australian Data Science Masters level degree. This course runs in both semesters of the academic calendar, and so this study was able to collect a total of 84 reflective texts, submitted for the same assessment task over three independent runs of the course between 2018 and 2019. For the task itself, students were asked to write a 700-1000 word performance review, where they reflected upon events that occurred during a group project that ran throughout the semester, and identified strategies for improvement in future Data Science team based projects. Students were instructed to reflect upon their contribution to the group project using the

following prompt questions: *"What went well? What did not work so well? What would you try next time to generate a better team dynamics? How did your team dynamics affect the statistical modelling process? Was your group dynamics "healthy"?"* Students were also instructed to consider their contribution to the broader community using the following additional prompts: *"How have you helped out people beyond your group? What responses have you made to people's questions in the forums and slack? Have you asked any questions that provoked an interesting discussion? How have you contributed to the fora?"* Finally, students were required to provide evidence supporting their claims in the reflection using an appendix. The submitted work evaluated according to the following two assessment criteria (each worth 50% of the mark for the associated task):

1. Depth of evidence demonstrating your contribution to your group and to the broader Statistical Thinking community.
2. Insightfulness and criticality in reviewing your contributions and identifying strategies for improvement in future collaborative work to achieve better outcomes.

The same one academic was responsible for marking all tasks, over the three runs of the course that are considered in this analysis. They took each of the above criteria as the "gold standard", with top marks being awarded for responses that best matched the requirement, and a sliding scale down for less well formulated submissions that failed to meet this top criterion. Note that neither the assessment task nor the rubric includes reference to feelings. The instructor who designed this assessment task framed it as a "performance review" to combat the negative reaction often displayed by Data Science students when confronted by a reflective task. Students were expected to use a more analytical and objective voice in this task. For the purposes of this analysis, we used the score given for the second assessment criterion to find its correlation to the CWRef factor.

4.3 Procedure

The Data Science context had only one reflective text recorded for every student considered. For the Pharmacy context, all the Pharmacy students' reflective statements were used together in the analysis even though each student produced multiple reflective texts. This was considered a reasonable simplification because the prompts for guiding each reflective text were different, and so they were assumed to be independent to facilitate analysis. This is a potential weakness of our analysis that remains to be validated in future work.

In Table 3 we see that 301 submissions were collected from the Pharmacy course, of which 243 were identified as either reflective or critically reflective. For Data Science, all 84 documents were awarded passing scores higher than 55 out of 100, a proxy which we took as indicative of reflective writing. These were further divided into high (top 42) and low graded groups (bottom 42) for our quality analysis (see Table 4). Table 3 also shows the average text length for both contexts, demonstrating that it is in Data Science is longer than in Pharmacy (an artefact that arose from differences in due to the assessment criteria).

Domain	Written task	Num. docs	Average words/doc
Pharmacy	Experiential Placement Review	301 including 243 reflective cases	229.75
Data Science	Project Review	84 reflective cases	1013.45

Table 3: Dataset Description

Table 4 further investigates the AWA and LIWC features mean scores obtained from different levels of written reflection. We see that these scores obtained from deeper written reflection (reflective in Pharmacy and high reflection in Data Science) are generally higher than the shallow reflective writing (non-reflective in Pharmacy and low scores in Data Science) across both datasets. But, in Data Science, some feature scores, such as LIWC.cause, and LIWC.focuspast and LIWC.analytic, are slightly higher (less than 0.11) in shallow reflective writing than those feature scores in deep reflective writing. This implies that these features may not be good indicators of the reflection level.

		AWA Features							LIWC Features								
		Tho	Bel	Int	Chl	Cri	Cha	Pos	Neg	Anx	Ang	Sad	Cau	Per	Pas	i	Ana
Pharm	Ref	2.85	2.60	2.37	2.21	2.95	.60	3.60	1.07	.52	.07	.12	2.50	1.85	5.40	4.80	74.4
	Non-Ref	1.47	1.38	1.33	1.17	1.22	.26	3.29	.53	.25	.06	.02	2.14	1.56	6.48	4.79	73.96
Data Science	high	12.07	7.62	8.95	9.83	11.19	2.36	2.99	.73	.19	.12	.13	2.31	1.51	5.53	4.47	85.03
	low	8.29	7.69	5.95	7.17	10.69	1.90	2.91	.71	.12	.09	.15	2.42	1.00	5.57	3.20	85.30

Note: Tho:thoughts, Bel:beliefs, Int:interpretation, Chl:challenges, Cri:Self-critiques, Cha:changes, Pos:posemo, Neg:negemo, Ang:anger, Cau:cause, Per:percept, Pas:focuspast and Ana:analytic

Table 4: Mean Score of Textual Features Organized by Different Levels of Reflection

Following the methodology described in section 3, both datasets were analysed independently. In stage 1 writing analytics was used to extract a total number of 18 AWA/LIWC features from each document collection. During stage 2, the normality test was performed. For the Pharmacy dataset, the skewness values were in the range of 0.443 and 1.981; and kurtosis values ranged from 0.137 to 4.891, while for the Data Science dataset, the skewness values were in the range of -1.375 and 1.957; and kurtosis values ranged from -0.392 to 4.520, which indicated normal distributions (Byrne, 2013; Hair et al., 2010). Then, PCA was used to filter out those features which contributed less than 0.4 to a factor for each dataset. As a result, 11 features were selected in the Pharmacy dataset, contrasting with 12 features from the Data Science dataset. For both datasets, LIWC.posemo, LIWC.sad, LIWC.cause, LIWC.focuspast were discarded. Among them, LIWC.sad and LIWC.posemo were removed since they did not follow normal distribution. In addition, it has been found they were less relevant to the Feeling factor after performing the PCA. Moreover, LIWC.percept and LIWC.analytic were removed in the Pharmacy dataset, while LIWC.focusfuture was removed in the DataScience dataset. In stage 3, based on the association between factors and textual variables shown in Figure 3, the unselected textual features derived from Stage 2 in each dataset were removed from the factor structure shown in Figure 3. Using this updated structure, CFA was then separately conducted on each dataset. These features were used to drive the generation of a CFA for each dataset.

We will now turn to a detailed discussion of the CFA results, specifically: goodness-of-fit, factor validity and reliability, and discriminant validity.

5. Results

The goodness-of-fit indices shown in Table 4 from the CFA demonstrate a strong fit of the data collected in the Pharmacy and Data Science contexts to the five-factor measurement model. The RMSEA values are .071 in Pharmacy and .062 in Data Science, which is considered an acceptable fit (Fabrigar et al., 1999). The CFI in Pharmacy and Data Science exceeds 0.9, and the TLI in Pharmacy, .865, is close to 0.9. Based on these indices, these two samples can be said to demonstrate acceptable fits to the five-factor model.

Model	X^2	df	p	RMSEA	CFI	TLI
Pharmacy	192.326	40	.000	.071	.902	.865
Data Science	78.145	50	.024	.062	.929	.906

Table 4: Fit indices for each 2nd-order CFA by reflection context: RMSEA = Root Mean Squared Error of Approximation; CFI = Comparative Fit Index; Tucker-Lewis index (TLI). See section 3 for details of these indices.

The CFA of the Pharmacy dataset shows AVE values ranging from 0.544 to 0.730 for the four latent factors: *context, feelings, challenge and reflective thinking*, which shows acceptable convergent validity. Similarly, the factor reliability values for these four factors range from 0.742 to 0.888 which shows acceptable internal reliability. Compared to these four factors, the changes factor shows a lower convergent validity (AVE:0.402). Figure 4 shows standardized factor loadings, which can be interpreted as correlation coefficients, for the five-factor model in this context. All 15 factor loadings are significant and 10 loadings are greater than 0.72 demonstrating good convergent validity. However, four factor loadings (Feelings, LIWC.affect, LIWC.focusfuture, LIWC.i) demonstrate low convergent validity scores between 0.28 and 0.43.

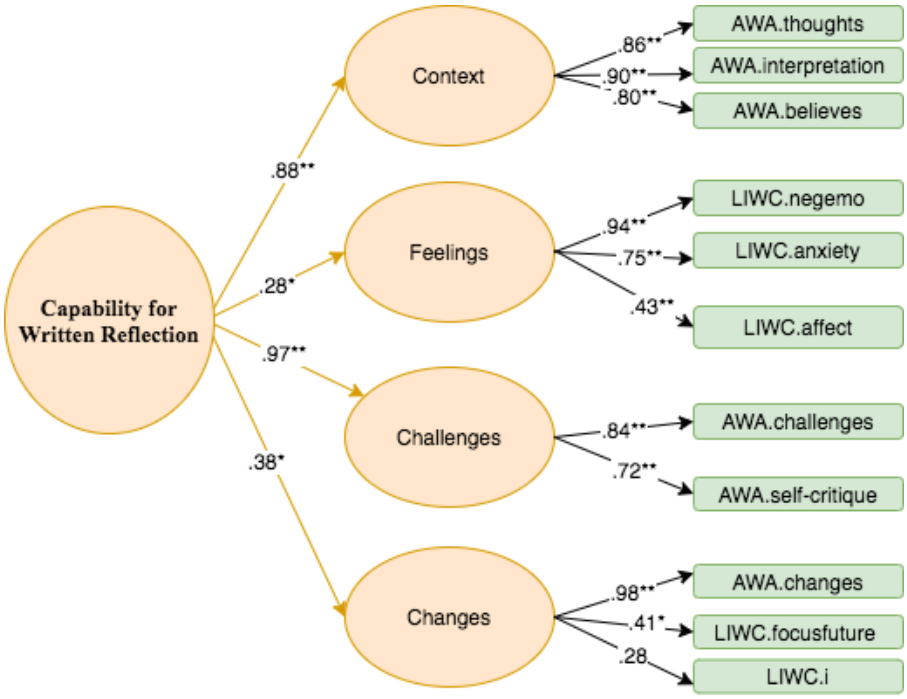

Figure 4: Reflection Skill 2nd-order CFA model: Pharmacy context. Standardised factor loadings between first order latent coefficients, and between first and second-order latent variables. Path with * and ** are statistically significant at the p <.05 and p <.001 level respectively.

Switching attention to the Data Science context, the CFA results give AVE values with acceptable convergent validity and internal consistency for two factors: context and reflective thinking, which are 0.661 and 0.899 respectively. Compared to these, the feelings (AVE:0.494) and changes (AVE:0.315) factors shows a lower convergent validity. The CR values for context and feelings factors are 0.844 and 0.741 respectively, which shows good internal consistency, while the challenges and changes are 0.618 and 0.476 indicating low internal consistency.

Figure 5 shows standardized factor loadings, for the five-factor model in this context. 15 of 16 factor loadings are significant. Among them, 11 factor loadings range from .58 to .98, while 4 factor loadings, including LIWC.analytic (-.27), LIWC.percept(.46), AWA.self-crique(.47), AWA.beliefs(.44), are below 0.50. The feeling factor loading value is not significant.

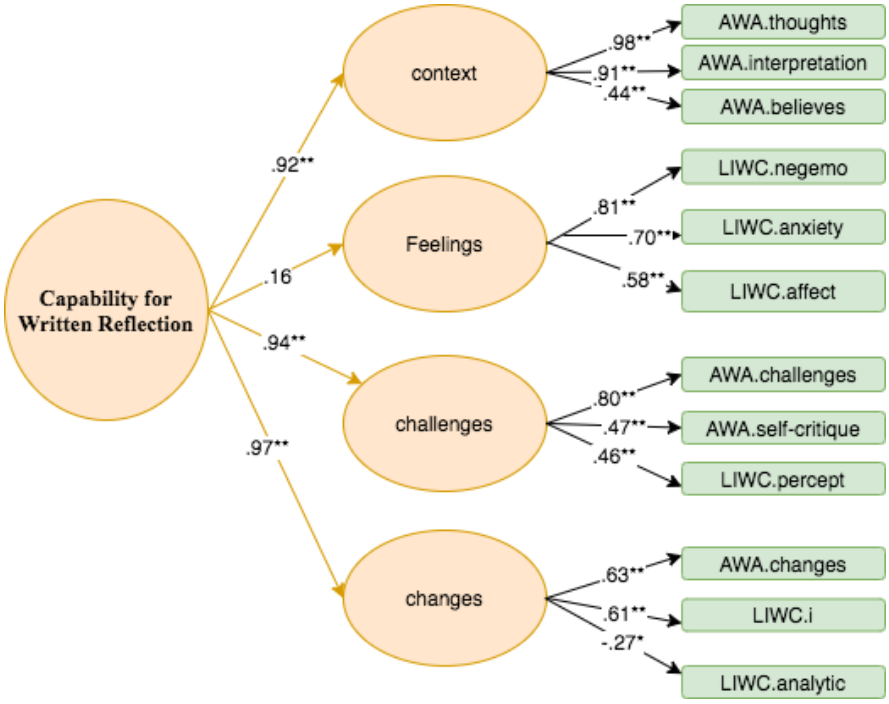

Figure 5: Reflection Skill 2nd-order CFA model: Data Science context. Standardised path coefficients between first order latent coefficients, and between first and second-order latent variables. Paths with * and ** are statistically significant at the p <.05 and p <.001 level respectively.

Table 5 shows that the square root of AVE values is generally greater than the correlation coefficients among the four latent variables in both contexts, which indicates an acceptable discriminant validity.

	Factor	Context	Feelings	Challenges	Changes
Pharmacy	**Context**	(.854)			
	Feelings	.101	(.737)		
	Challenges	.431	.113	(.782)	
	Changes	.338	.084	.375	(.634)
Data Science	**Context**	(.813)			
	Feelings	.262	(.703)		
	Challenges	.569	.252	(.598)	
	Changes	.548	.242	.489	(.561)

Table 5. Discriminant validity for the measurement model. Diagonal in parentheses: Square root of average of variance extracted from observed variables and off-diagonal: correlations between factors

Finally, the results of the Spearman's correlation indicated that there were significant positive weak associations between the writing score given by the lecturer and reflective thinking factor (CWRef) value, (r_s (301) =.378, p<.001) in Pharmacy and (r_s (83) =.420, p<.001) in Data Science. These results were consistent with Birney's study results (2012) which showed the

reflective score indicating the overall depth of reflection in a document was moderately correlated to the writing grade across different reflective writing genres, such as blogs and journals.

6. Discussion

We now return to our research questions to reflect upon what we have learned during our investigations of CWRef in two authentic learning contexts.

This paper has proposed a four-stage process that links writing behaviours extracted from student texts using writing analytics, to four latent written reflection factors. These have been demonstrated to contribute to a second-order reflective thinking factor, using confirmatory factor analysis to evaluate the validity and reliability of the five-factor CWRef capability model for two separate datasets.

Q1: How can we quantify and validate the relative contributions that textual features make to the different latent reflector factors underpinning quality of written reflection?

The CFA results for both datasets (see Table 7) indicated the usefulness of AWA reflective rhetorical features (e.g. AWA.thoughts, AWA.interpretation, AWA.beliefs, AWA.challenges, AWA.self-critique, AWA.changes), which significantly contributed to the Context, Challenges and Change factors described in Table 2. This validates the reflection model proposed by Gibson et al. (2017) along with the AWA features developed for detecting the reflective elements. In addition, the LIWC emotional features show substantial contributions to the Feeling factor shown in Table 2, a result consistent with Lin et al (2016). Regarding the contribution to Changes factor described in Table 2, the AWA.changes feature demonstrated usefulness for both datasets. Moreover, LIWC.i feature (indicating the frequency of using first person pronouns, such as I, me, and mine: Ullmann, 2019) was a useful indicator for Changes factor in the Data Science dataset. However, it proved less useful in the Pharmacy dataset, a difference that can be attributed to different writing style. Some Pharmacy students expressed future plans using 'we' instead of 'i'. For example*, "in the future, we would hope to think more from the patient's perspective to alter…"* Noticeably, the LIWC.analytic feature value (-.27*) was significantly negatively correlated to the Changes factor for the Data Science dataset. LIWC.analytic reveals the degree of analytical, logical and consistent thinking, which relates to the use of more articles and prepositions, and fewer personal pronouns, auxiliary verbs (Pennebaker, Chung, Frazee, Lavergne, & Beaver, 2014). The task design for the Data Science students led to a tendency towards describing the change factor with more first-person pronouns and auxiliary verbs, such as *I will, I could, I would*, which results in low LIWC.analytic scores.

Feature/Factor	Factor	Pharmacy	Data science
AWA.thoughts	Context	.86**	.98**
AWA.interpretation	Context	.90**	.91**
AWA.beliefs	Context	.80**	.44**
LIWC.negemo	Feelings	.94**	.81**
LIWC.anxiety	Feelings	.75**	.70**
LIWC.affect	Feelings	.43**	.58**
AWA.challenges	Challenges	.84**	.80**
AWA.self-critique	Challenges	.72**	.47**
LIWC.percept	Challenges	NA	.46**
AWA.changes	Changes	.98**	.63**
LIWC.i	Changes	.28	.61**
LIWC.focusfuture	Changes	.41*	NA
LIWC.analytic	Changes	NA	-.27*
Context	Reflective Thinking	.88**	.92**
Feelings	Reflective Thinking	.28*	.16
Challenges	Reflective Thinking	.97**	.94**
Changes	Reflective Thinking	.38*	.97**

Table 7: Summary of the factor loadings in both datasets.

Moreover, the experimental results indicated that some reflection level of these factors could be possibly assessed based on the AWA and LIWC feature sets regarding the fine-grained rubric defined in Table 2. Specifically, the combination of the AWA.thoughts, AWA.interpretation and AWA.beliefs features could assess the Premise Reflection level of the Context factor, while the LIWC emotional features (LIWC.negemo, LIWC.anxiety, LIWC.affect) could assess the Reflection level of the Feelings factor. The combination of AWA.challenges and AWA.self-critique could assess the Reflection level of the Challenge factor, while the AWA.changes and AWA.i could assess the Reflection level of the Change factor. To capture the Premise Reflection level of these factors; richer textual features could be investigated and developed, such as Explanation feature for Challenge factor and Prospective feature for Change factor (Ullmann, 2019); and the relationship between emotional features and Challenge/Change factor for Feelings factor.

Our results could be claimed to cast doubt on the importance of some text features previously described in the reflective writing literature (Kovanović et al., 2018; Ullmann, 2019), but it is important to discuss some caveats that require further investigation. Specifically, we note that in our current approach features such as LIWC.i, LIWC.analytic, LIWC.focusfuture and LIWC.percept, do not always display a significant correspondence to the theoretically motivated reflective factors that are listed in Table 2 and Figures 2-3. While this could be because of contextual factors associated with the learning design and assessment structure of classes encouraging different cohorts of students to write differently, this result might also be an artefact of the method applied (i.e. the CFA), or the relatively small nature of a dataset that covers only two courses. Future work will seek to expand upon our dataset, and to explore these results in more detail.

Q2: To what degree does this model of reflective writing generalise to different reflective writing contexts?

Regarding the contribution of four reflection factors to the second order reflective thinking factor, the CFA results showed that Context, Changes and Challenges factors significantly contributed to the CWRef factor for both datasets. However, the Feelings factor did not significantly contribute to the CWRef factor in the Data Science dataset. However, although the contribution of the Feelings factor is significant in Pharmacy dataset, it is still relatively weaker than other factors.

Together, this motivates a number of insights. Firstly, the use of only LIWC emotional features is insufficient to detect the depth of the Feeling factor. More work here remains to be completed, but an initial step could be taken by considering the position Lucas et al. (2019) have taken, emphasising that deeper critical reflection on feelings should go beyond merely articulating them, and seek to *connect* those feelings to changes in personal perspective.

Secondly, the *Learning Analytics/Learning Design coupling* is evident (Lockyer, Heathcote & Dawson, 2013). This principle reminds us that in order to make sense of any student activity data, one must understand the context that gave rise to it, which early work in the field failed to do. As learning analytics moves towards assessment, this principle converges with established practice in the learning sciences, such as Evidence-Centered Design (Behrens et al., 2019; Lockyer et al., 2013)) and measurement science, such as "metrolytics": (Milligan, 2020). These provide systematic methods to design and evaluate tasks that elicit data to evidence the capabilities being assessed. In this study, therefore, since the learning design (writing assignment, prompts and assessment criteria) inevitably shaped student reflections, we see that in the context of an assignment to write a "performance review", the lack of a writing prompt around *feelings* most likely caused the Data Science students to ignore this in their reflections.

Thirdly, it might be possible to drop specific factors in the CWRef model depending on the reflective writing requirements of the course (e.g. the feeling factor for a performance review). An alternative approach could be to co-design the reflective writing task with educators based on the five-factor model, ensuring that all elements are present. Indeed, the second approach is already being followed in the design of new curriculum offerings; we are currently co-designing reflective writing rubrics and prompts with lecturers for an internship reflection in Engineering and a critical reflection essay in the School of Business. We expect that both approaches will become the norm as the technology for delivering automated feedback on reflective writing, with the method chosen in accordance with the situation.

To summarise, the CFA results for the two datasets explored here indicate the promising nature of our approach, which integrates CFA with reflective writing analytics. This result confirmed the feasibility of using CFA in text analytics (Fincham et al., 2019), particularly reflective writing analytics. What has been learned from this new combination of two fields? Firstly, we see some support for the stability of the low-level textual features used in reflective writing analytics – across two datasets that were collected by two different academic teams, with no modification of their assessment structure. Some common features (e.g. AWA rhetorical features and LIWC emotional features) in Table 7 are important indicators for some reflection factors, *Context, Feelings* and *Challenges*. Given that the data analysis was performed upon each dataset separately, there was no reason to expect a similar set of variables to emerge –

there were 17 possible variables in the original feature set, so the agreement obtained between the two separate analyses is not something that could reasonably be expected by chance. This is a significant result, which starts to validate the approach that has emerged from the writing analytics community in recent years. It is important to note that this link to reflective writing work emerging from learning analytics has been provided "in the wild", by using a theoretically grounded model to explore two distinct datasets obtained from authentic teaching contexts with no modifications. The teaching has not been modified to fit the model, and yet the results are still highly promising.

It is particularly worth noting the *differences* between the two datasets, as they point to possible ways in which the underlying assessment and learning design of a course might be revealed through the use of writing analytics. For both datasets, the linking of writing analytics to a conceptually interpretable model enabled rich discussions with the academics involved in the design and delivery of the course – in both cases they confirmed that the findings made sense, linking them to the underlying assessment design of the reflective writing task. As we alluded to above, this result suggests that it might be possible to start categorising the underlying pedagogy of reflective writing tasks, perhaps using an approach that clusters feature sets that are theoretically validated according to accepted measurement models. This would potentially enable the *detection* of underlying teaching modes according to key indicators in writing analytics trace data – an exciting possibility that we reserve for future research.

Besides this intriguing possibility, many interesting avenues for future research still remain. Further investigation of the reflective writing rubrics and prompts defined by the educators are an essential first step, but requires (i) the construction of a larger reflective writing corpus consisting of student text samples from different reflective writing genres, (ii) the specification of the associated writing tasks, and (iii) any feedback that students obtained for their submissions (both formative and summative if available). The writing analytics community is engaging in the curation of such a resource that will likely prove essential to further substantive developments in the field. Beyond this first step, we also believe it is important to continue with a detailed investigation of the manner in which these underlying feature sets influence the mediating reflection factors, and hence the overarching CWRef factor. In particular, it seems likely that further writing analytics development is needed, particularly for the *feelings* factor, which does not appear to be well covered by the variables we have considered here – what writing analytic features might prove to be more reliable indicators of feelings if they are denoted as important to the quality of a reflection?

7. Implications for improving automated feedback

The principal goal of writing analytics is not just to automatically assess texts, but to deeply understand students' potential reflective thinking skills in a learning context, such that we are able to generate *actionable feedback* to improve their writing (Simon Buckingham Shum et al., 2016; Knight et al., 2018; Lucas, Gibson, et al., 2019). We now turn to a discussion of how the CWRef model might be used to improve feedback.

Hattie and Timperley (2007) identified three effective feedback levels: task, process and self-regulatory levels. *Task level feedback* helps to build surface learning knowledge about the task

being completed. An example of task level feedback generated by the AcaWriter[2] tool (Gibson et al., 2017) is: "*While it appears that you've reported on how you would change/prepare for the future, you don't seem to have reported first on what you found challenging. Perhaps you've reflected only on the positive aspects in your report?*". This type of feedback is generated based on the presence or absence of rhetorical moves, represented using textual features that are detected by the system (i.e. the absence of a *challenges* sentence in this case). However, the current implementation of AcaWriter makes very rudimentary judgments on the *quality* of the reflective writing in terms of the presence/absence of salient sentences, so it cannot track the progression (or regression) of highly nuanced writing. This limits the forms of feedback that can be offered. This paper used CFA to validate the model underpinning the CWRef construct, in terms of four important latent factors, which provides a more sophisticated measure of quality.

We have seen that the second-order regression model resulting from our CFA has acceptable validity across all four intermediate factors (context, feelings, challenges and changes) in the pharmacy context, which enables it to predict the target variable of CWRef with high levels of significance. Compared to the previous machine learning-based approach in measuring holistic reflective thinking scores (Liu, Buckingham Shum, et al., 2019) or rule-based approaches in classifying reflective elements at sentence level (Gibson et al., 2017), our new approach is both more explanatory (quantifying both the quality of overall written reflection and different latent factors based on textual features), and has a stronger theoretical alignment. Thus, this model enables the generation of more nuanced formative feedback regarding strengths and weaknesses. Once these multiple regression models are created through the four stages described in section 3, they can be applied to the formative assessment of a document, or indeed each draft, that is submitted to a feedback tool. This approach has potential for tracking progression and, when coupled with the revision history tracking capabilities of AcaWriter, it could be used to generate more powerful formative feedback. Thus, we believe that this approach can be used to generate *progress level* feedback, which has been linked to improving students' self-efficacy beliefs (Duijnhouwer et al., 2010). Figure 6 illustrates how one could visualize progress in terms of the underlying factors, by calculating the value of CWRef for each draft submitted to AcaWriter. This example tracks the revisions that a student in the Data Science dataset has made using AcaWriter, submitting 14 revisions to their draft along the way. The Y-axis shows normalized z-scores for each of the four latent factors, meaning that 0 represents the *average amount of written reflection expressed per document in the corpus*. It can be seen that from revision 1 to 13 this student is making general progress in improving the quality of their reflection, although sometimes the reflective thinking scores drops or increases because reflective textual features were added or removed.

The dotted vertical lines mark a number of behavioural transitions that are evident for this student. Specifically, during the first 5 revisions, although the writer made slight progress in improving the weighting of the challenge factor, this does not shift the overall CWRef score. However, in the second phase, the writer changed their approach, by adding more description and interpretation of the highlighted collaborative learning event and some problems that they faced. This is reflected in the increased *Context* and *Challenges* scores, which both reach positive values by draft 10, leading to a higher CWRef score (.41). In the next three revisions,

[2] *AcaWriter* (https://uts.edu.au/acawriter) is a web-based academic writing support tool using natural language processing technology to automatically generate personalized feedback on reflective writing(Gibson et al., 2017; Knight et al., 2020).

the writer added a future plan statement into the document, which brings the *Changes* factor to a positive value:

> *If this event was to occur in the future I would implement a better approach to team communication and change my behaviour to being direct and upfront.*

Interestingly, in the last phase of revisions, there is a dramatic decrease in the *Context* and *Challenges* scores. Here, the writer added new material, but in the process removed some description of the *Context* and *Challenges.*

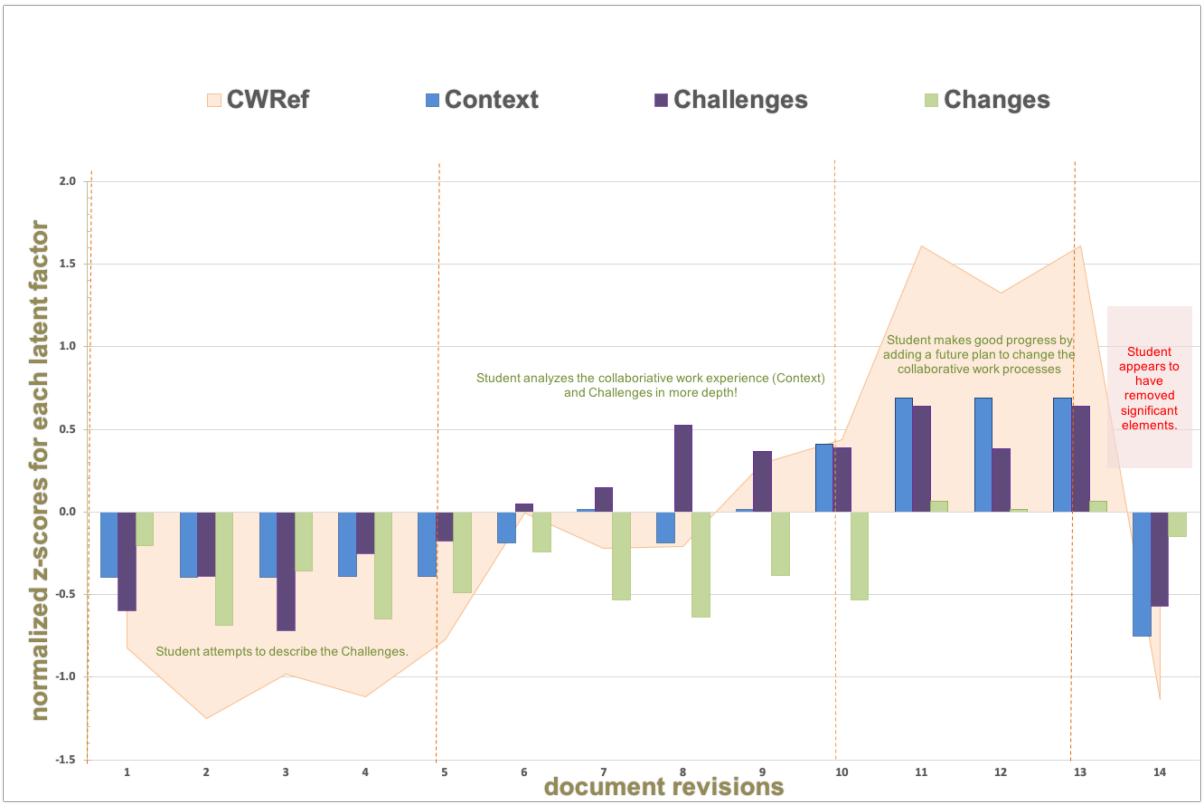

Figure 6: Visualizing CWRef-based progression across 14 drafts of a reflective report

This form of analysis opens up the possibility of providing effective progress level feedback defined by Hattie and Timperley (2007). The evidence collected from ongoing revisions could be used in automated messages that help students to recognize the relationships between the text/reflective elements added and the overall capability of written reflection, and reflect on what types of writing strategies might lead to further improvement. For instance, since the analytics showed that from revisions 5 to 10 the *Challenges* and *Context* factors increased dramatically, but the *Changes* factor still kept the same low scores, a feedback message shown could be generated to recognise the progress but also to suggest more attention to the future. An example message might take the following form:

> *<Revision 10> Well done, it appears that you have made some progress by analyzing your experience in more detail, and the challenges you encountered. However, one of the assessment criteria requires you to reflect on how you plan to change your processes in future collaborative work. It seems that you haven't commented on what you would do differently should the same event occur in the future. Perhaps think about changes in perspectives, strategies, tools, ideas, behaviour and/or approach.*

As noted, the third level of feedback identified by Hattie and Timperley (2007) is the *self-regulation* level. We cannot claim that Figure 6 is a suitable visualization to show directly to students, or even whether a simplified version would be understood. However, it serves to demonstrate a manner in which we might start to scaffold student reflection on the changing profile across their drafts once significant patterns are identified. If we imagine that AcaWriter's feedback on drafts was linked to each draft depicted in Figure 6, then the following example might help the writer to self-evaluate their writing strategy and how it has changed over a sequence of drafts:

> *<Revision 12> Well done! Looking back, you can see how your drafts have varied but overall, have really improved. Does it make sense to you why the scores for context, challenges and changes went up and down?*

We are cautious, however, about whether such a direct approach could work. On the one hand, we have evidence that students can learn, and moreover appreciate, the language of the constructs underpinning CWRef, since they have already been deployed in AcaWriter (Gibson, et al. 2017; Lucas et al, 2019) when rendered as visual annotations and feedback messages on their drafts. On the other hand, it may be that the insights from this new CWRef metric are best kept 'backstage', used to improve the quality of textual feedback on drafts rather than directly visualized, and/or used to provide educators with insights into individual student and cohort progression. We reserve our current uncertainty about how to resolve this tension for future work.

To summarise, students are already benefitting from fully automated formative feedback on their reflective writing, based on a much simpler model than the one presented in this paper. The results reported here move us closer to mapping a student's progression more precisely than is currently possible, in principle enabling more nuanced feedback, as discussed in this section. There is clearly still room to improve the performance of the model since the correlation between the writing score and CWRef score is still weak, ($r_s(301) = .378$, $p < .001$) in Pharmacy and ($r_s(83) = .420$, $p < .001$) in Data Science. Assessing the overall grade of a piece of writing remains a complex judgement for humans, in our view, since it depends on more factors than our writing analytics currently detect. The focus on formative feedback on drafts lowers the stakes for students and educators, compared to automated grading.

8. Conclusions, Limitations and Future Work

Reflective writing is a widespread practice to help learners reflect on challenging experiences. However, the evidence is that it is both challenging to teach, and to learn. Furthermore, the assessment of reflective writing is quite different to that of other genres. Central to improving reflective writing (as with any skill) is the provision of timely, actionable feedback (Lucas, Gibson, et al., 2019; Lucas, Smith, et al., 2019). Automating textual analysis approaches opens new possibilities to researchers studying reflective writing, the educators teaching it, and to the students learning it. *Reflective writing analytics* are being developed to automatically detect the *presence* of reflective elements (Gibson et al., 2017; Kovanović et al., 2018; Ullmann, 2019) and the *depth* of reflection (Liu, Buckingham Shum, et al., 2019; Ullmann, 2019) based on textual features in writing that NLP can extract. Recently, Cui et al (2019) aligned these textual features to factors in a model of reflection, and analysed reflection variation across students and over time, but did not quantify the relative weights of contribution between factors in the model, the relative contributions of different textual features to factors, or how these related to expert-assessed quality of written reflection.

We have presented a theoretically grounded five-factor model of student *Capability for Written Reflection (CWRef)* which acts as a proxy for the author's level of reflective thinking. Confirmatory Factor Analysis was used to quantify the relative contributions that these textual features make to these reflection factors, and these factors to the overall depth of written reflection. The validity of CWRef was evaluated by using CFA for two sets of reflective writing tasks in substantially different fields: Pharmacy and Data Science. Results indicated that a common set of textual features (AWA and LIWC) could be useful indicators for quantifying *context, feelings, challenges* and *changes* factors in both datasets, with good reliability (factor loadings of these features >= .44**). The five-factor model was more suitable for the Pharmacy dataset than the Data Science dataset in terms of the convergent validity and factor reliability. We have attributed this difference to the framing of the assessment task, which points to the importance of considering both learning design and task descriptions/criteria when attempting to automate feedback for reflective writing. Once again, we see that "one size does not fit all" in learning analytics (Gašević et al., 2016).

Although showing considerable promise for improving the state of the art for delivering automated feedback at scale to students about the quality of their reflective thinking, it is important to take into account the limitations of the current study. First, the sample size in the Data Science context may not be large enough for the approach used. According to Wolf et al. (2013), a typical sample size in studies where CFA is used is about 200 cases. Second, the possibilities concerning number of factors, factor structure and textual features have yet to be fully explored. Here, we based our CFA upon a theoretically justified model underlying the alignment shown in Table 3 between the textual features and latent factors. While justified by theory, more work must be completed to test other possibilities. For example, we currently assume that Context, Feelings, Challenges and Changes factors are independent. However, we acknowledge that these factors may be somehow correlated (Gibbs, 1988), or even dependent upon one another (Schumacker & Lomax, 2016), which may lead to different results for the feature importance analysis. We will use Structural Equation Modelling (Byrne, 2013) to examine the appropriateness of our current assumptions, and perhaps to extend them in the future.

The usefulness of textual features for evaluating the quality of individual submissions has not at present been fully investigated. In the current approach, this is only indicated by their factor loadings and their correlations to writing scores. However, if the features are stable enough then they could potentially be seen as indicators of quality. Human studies could start to evaluate this. In future work, we will seek to improve upon our results in this area by asking human experts to give a score on each factor for individuals, and evaluate the difference between human scores and the model prediction score for the factor based on the textual features.

Furthermore, in the previous data-driven approach (Liu, Buckingham Shum, et al., 2019), the LIWC and AWA textual features were selected purely based on the strong correlation between each feature and the level of reflection found in our dataset. It was found that some top ranked LIWC features (LIWC.Differ, LIWC.Quant, LIWC.Compare, LIWC.Adj), were negatively correlated with the reflective thinking level. For example, Liu et al. (2019) reported that non-reflectors tended to describe their learning context or changes in vague, general terms (e.g. "*Different* pharmacists had *different* attitudes and *different* processes in place to achieve this." or "The *whole* experience went well I really liked working there and definitely learnt a *lot* of new things."), rather than providing more specific details (e.g. about how exactly pharmacists differed from each other, and how these differences connected to their personal learning experiences). Thus, these textual features, while ignored in the CWRef model, could be useful indicators for describing the change factor in detail. Future work could combine *model-driven* textual features derived from theory (as with the CWRef model) and *data-driven* features derived from machine learning (as just illustrated). Together, these could provide richer sources of evidence for each factor in the model.

We will also consider adopting an Evidence-Centred Design approach (Mislevy et al., 2012) to improve the quality of observed writing behaviours or textual features contributing to each latent factor of the CWRef model, and the validity and reliability of the five-factor model. For example, we can extend our current practice of co-designing rubrics and prompts with educators who wish to use AcaWriter, to take into account the five-factor CWRef model. In addition, we will apply human-centred design methods in learning analytics to design the new kinds of formative feedback this CWRef analytics enables (Buckingham Shum, Ferguson, & Martinez-Maldonado, 2019).

In conclusion, one of the most significant contributions of this article is the demonstration of how writing analytics can support the evaluation of a theory-based model (CWRef), using automatically extracted textual features and CFA, which provides a great potential for formative reflection assessment and feedback generation.

9. References

Behrens, J. T., DiCerbo, K. E., & Foltz, P. W. (2019). Assessment of Complex Performances in Digital Environments. *The ANNALS of the American Academy of Political and Social Science*, *683*(1), 217–232. https://doi.org/10.1177/0002716219846850

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Birney, R. (2012). Reflective Writing: Quantitative Assessment and Identification of Linguistic Features [Waterford Institute of Technology]. In *PhD Thesis*. http://repository.wit.ie/2658/1/FinalThesis.pdf

Bollen, K. (1989). Structural equations with latent variables. In *Structural Equation Models*. John Wiley & Sons.

Boud, D., Keogh, R., & Walker, D. (1985). *Reflection: Turning experience into learning*. https://doi.org/10.1017/CBO9781107415324.004

Boud, D. (1985). *Reflection. Turning experience into learning*. London, Kogan.

Boud, David, & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment and Evaluation in Higher Education*, *31*(4), 399–413. https://doi.org/10.1080/02602930600679050

Boud, David, Keogh, R., & Walker, D. (1985). Promoting Reflection in Learning: a Model. In *Reflection: Turning Experience Into Learning* (pp. 18–40). Kogan Page.

Brookfield, S. (1995). Becoming a critically reflective teacher. In *Jossey-Bass*. https://doi.org/10.1007/s11606-010-1463-1

Buckingham Shum, S., Ferguson, R., & Martinez-Maldonado, R. (2019). Human-Centred Learning Analytics. *Journal of Learning Analytics*, *6*(2), 1–9. https://doi.org/https://doi.org/10.18608/jla.2019.62.1

Buckingham Shum, Simon, & Crick, R. D. (2016). Learning Analytics for 21st Century Competencies. *Journal of Learning Analytics*, *3*(2), 6–21.

Buckingham Shum, Simon, Knight, S., McNamara, D., Allen, L., Bektik, D., & Crossley, S. (2016). Critical perspectives on writing analytics. *LAK '16 Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 481–384. https://doi.org/10.1145/2883851.2883854

Buckingham Shum, Simon, Sándor, Á., Goldsmith, R., Bass, R., & McWilliams, M. (2017). Towards Reflective Writing Analytics: Rationale, Methodology and Preliminary Results. *Journal of Learning Analytics*, *4*(1), 58–84.

Byrne, B. M. (2013). Structural equation modeling with AMOS: Basic concepts, applications, and programming, second edition. *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming, Second Edition*. https://doi.org/10.4324/9780203805534

Chirema, K. D. (2007). The use of reflective journals in the promotion of reflection and learning in post-registration nursing students. *Nurse Education Today*, *27*(3), 192–202. https://doi.org/10.1016/j.nedt.2006.04.007

Contributors, S. (2019). *semTools: Useful tools for structural equation modeling* (p. Version 0.5-2). https://cran.r-project.org/web/packages/semTools/index.html

Cui, Y., Wise, A. F., & Allen, K. L. (2019). Developing reflection analytics for health professions education: A multi-dimensional framework to align critical concepts with data features. *Computers in Human Behavior*, *100*, 305–324. https://doi.org/10.1016/j.chb.2019.02.019

Deslandes, R., Lucas, C., Hughes, M., & Mantzourania, E. (2018). Development of a template to facilitate reflection among student pharmacists. *Research in Social and Administrative Pharmacy*, *14*(11), 1058–1063. https://doi.org/https://doi.org/10.1016/j.sapharm.2017.11.010

Dewey, J. (1933). How we think. In *Great books in philosophy*. Prometheus Books. https://doi.org/10.1037/10903-000

Duijnhouwer, H., Prins, F. J., & Stokking, K. M. (2010). Progress feedback effects on students' writing mastery goal, self-efficacy beliefs, and performance. *Educational Research and Evaluation*, *16*(1), 53–74. https://doi.org/10.1080/13803611003711393

Fabrigar, L. R., MacCallum, R. C., Wegener, D. T., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Fincham, E., Whitelock-Wainwright, A., Kovanović, V., Joksimović, S., van Staalduinen, J.-P., & Gašević, D. (2019). Counting Clicks is Not Enough: Validating a Theorized Model of Engagement in Learning Analytics. *The 9th International Learning Analytics & Knowledge Conference (LAK19)*, 501–510. https://doi.org/10.1145/3303772.3303775

Fornell, C., & Larcker, D. F. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research*, *18*, 39–50. https://doi.org/10.2307/3151312

Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *Internet and Higher Education*, *28*, 68–84. https://doi.org/10.1016/j.iheduc.2015.10.002

Gibbs, G. (1988). Learning by Doing: A guide to teaching and learning methods. In *Oxford Brookes University*. FEU. https://doi.org/978-1-873576-86-1

Gibson, A., Aitken, A., Sándor, Á., Buckingham Shum, S., Tsingos-Lucas, C., & Knight, S. (2017). Reflective writing analytics for actionable feedback. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK '17*, 153–162. https://doi.org/10.1145/3027385.3027436

Gibson, A., Kitto, K., & Bruza, P. (2016). Towards the Discovery of Learner Metacognition from Reflective Writing. *Journal of Learning Analytics*, *3*(2), 22–36.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers : A Journal of the Psychonomic Society, Inc*, *36*(2), 193–202.

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2014). Partial least squares structural equation modeling (PLS-SEM). In *Sage Publisher*. https://doi.org/10.1108/EBR-10-2013-0128

Hair, J., Black, W., Babin, B., & Anderson, R. (2010). Multivariate Data Analysis: A Global Perspective. In *Multivariate Data Analysis: A Global Perspective*. Prentice Hall.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Herrington, T., & Herrington, J. (2005). *Authentic learning environments in higher education*. Information Science Publishing. https://doi.org/10.1111/j.1467-8535.2008.00870_23.x

Hurley, A. E., Scandura, T. A., Schriesheim, C. A., Brannick, M. T., Seers, A., Vandenberg, R. J., & Williams, L. J. (1997). Exploratory and confirmatory factor analysis: Guidelines, issues, and alternatives. *Journal of Organizational Behavior*, *18*, 667–683. https://doi.org/10.1002/(SICI)1099-1379(199711)18:6<667::AID-JOB874>3.0.CO;2-T

Jung, Y., & Wise, A. F. (2020). How and How Well Do Students Reflect ?: Multi-Dimensional Automated Reflection Assessment in Health Professions Education. *Proceedings of the 10th International Conference on Learning Analytics & Knowledge (LAK'20)*, 595–604. https://doi.org/10.1145/3375462.3375528

Kember, D., Leung, D. Y. P., Jones, A., Loke, A. Y., McKay, J., Sinclair, K., Tse, H., Webb, C., Wong, F. K. Y., Wong, M., & Yeung, E. (2000). Development of a questionnaire to measure the level of reflective thinking. *Assessment and Evaluation in Higher Education*, *25*(4), 381–395. https://doi.org/10.1080/713611442

Kline, R. B. (2011). Principles and practice of structural equation modeling: Third Edition. In *Structural Equation Modeling*. Guilford Press. https://doi.org/10.1038/156278a0

Knight, S., Buckingham Shum, S., Ryan, P., Sándor, Á., & Wang, X. (2018). Designing Academic Writing Analytics for Civil Law Student Self-Assessment. *International Journal of Artificial Intelligence in Education*, *28*(1), 1–28. https://doi.org/10.1007/s40593-016-0121-0

Knight, S., Shibani, A., Abel, S., Gibson, A., Ryan, P., Sutton, N., Wight, R., Lucas, C., Sándor, Á., Kitto, K., Liu, M., Mogarkar, R. V., & Shum, S. B. (2020). AcaWriter: A learning analytics tool for formative feedback on academic writing. *Journal of Writing Research*, *12*(1), 141–186.

Kolb, D. A. (1984). Experiential Learning: Experience as The Source of Learning and Development. In *Prentice Hall*. https://doi.org/10.1016/B978-0-7506-7223-8.50017-4

Koole, S., Dornan, T., Aper, L., Scherpbier, A., Valcke, M., Cohen-Schotanus, J., & Derese, A. (2011). Factors confounding the assessment of reflection: A critical review. *BMC Medical Education*, *11*(104). https://doi.org/10.1186/1472-6920-11-104

Kovanović, V., Joksimović, S., Mirriahi, N., Blaine, E., Gašević, D., Siemens, G., & Dawson, S. (2018). Understand students' self-reflections through learning analytics. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge - LAK '18*, 389–398. https://doi.org/10.1145/3170358.3170374

Lethbridge, K., Andrusyszyn, M. A., Iwasiw, C., Laschinger, H. K. S., & Fernando, R. (2013). Assessing the psychometric properties of Kember and Leung's Reflection Questionnaire. *Assessment and Evaluation in Higher Education*, *38*(3), 303–325. https://doi.org/10.1080/02602938.2011.630977

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7

Lin, C. W., Lin, M. J., Wen, C. C., & Chu, S. Y. (2016). A word-count approach to analyze linguistic patterns in the reflective writings of medical students. *Medical Education Online*, *21*(1), 1–7. https://doi.org/10.3402/meo.v21.29522

Liu, M., Buckingham Shum, S., Mantzourani, E., & Lucas, C. (2019). Evaluating machine learning approaches to classify pharmacy students' reflective statements. *Proceedings of Artificial Intelligence in Education*, 220–230. https://doi.org/10.1007/978-3-030-23204-7_19

Liu, M., Goldsmith, R., Ahuja, S., & Huang, X. (2019). The Fifth Writing Analytics Workshop: Linking Reflective Writing Analytics to Learning Design. *Australian Learning Analytics Summer Institute*, 1–2.

Liu, M., Knight, S., Antonette, S., & Abel, S. (2018). *the Australian Learning Analytics Summer Institute ALASI'18: Writing Analytics Workshop*. http://wa.utscic.edu.au/events/alasi18-writing-analytics-workshop/

Lockyer, L., Heathcote, E., & Dawson, S. (2013). Informing Pedagogical Action: Aligning Learning Analytics with Learning Design. *American Behavioral Scientist*, *57*(10), 1439–1459. https://doi.org/10.1177/0002764213479367

Lucas, C., Bosnic-Anticevich, S., Schneider, C. R., Bartimote-Aufflick, K., McEntee, M., & Smith, L. (2017). Inter-rater reliability of a reflective rubric to assess pharmacy students' reflective thinking. *Currents in Pharmacy Teaching and Learning*, *9*(6), 989–995. https://doi.org/10.1016/j.cptl.2017.07.025

Lucas, C., Gibson, A., & Buckingham Shum, S. (2019). Pharmacy students' utilization ofc an online tool for immediate formative feedback on reflective writing tasks. *American Journal of Pharmaceutical Education*, *83*(6), 6800. https://doi.org/10.5688/ajpe6800

Lucas, C., Smith, L., Lonie, J. M., Hough, M., Rogers, K., & Mantzourani, E. (2019). Can a reflective rubric be applied consistently with raters globally? A study across three countries. *Currents in Pharmacy Teaching and Learning*, *11*(10), 987–994. https://doi.org/10.1016/j.cptl.2019.06.004

Luk, J. (2008). Assessing teaching practicum reflections: Distinguishing discourse features of the "high" and "low" grade reports. *System*, *36*(4), 624–641. https://doi.org/10.1016/j.system.2008.04.001

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149. https://doi.org/10.1037/1082-989X.1.2.130

Mann, K., Gordon, J., & MacLeod, A. (2009). Reflection and reflective practice in health professions education: A systematic review. *Advances in Health Sciences Education*, *14*(4), 595. https://doi.org/10.1007/s10459-007-9090-2

Mantzourani, E., Deslandes, R., Ellis, L., & Williams, G. (2016). Exposing Pharmacy Students to Challenges Surrounding Care of Young Children via a Novel Role-Emerging Placement. *Journal of Curriculum and Teaching*, *5*(124–134). https://doi.org/10.5430/jct.v5n1p124

Mantzourani, E., & Hughes, M. L. (2016). Role-emerging placements in pharmacy undergraduate education: Perceptions of students. *Pharmacy Education*, *16*(1), 88–91.

Mezirow, J. (1990). How Critical Reflection Triggers Transformative Learning. *Fostering Critical Reflection in Adulthood*. https://doi.org/10.1002/ace.7401

Mezirow, J. (1991). Transformative dimensions of adult learning. In *Jossey-Bass*. Jossey-Bass. https://doi.org/Retrieved from EBSCO HOST

Milligan, S. (2020). Standards for Developing Assessments of Learning Using Process Data. In M. Bearman, D. Boud, P. Dawson, J. Tai, & R. Ajjawi (Eds.), *Re-imagining University Assessment in a Digital World* (pp. 179–192). Springer. https://www.springer.com/gp/book/9783030419554

Milligan, S., & Griffin, P. (2016). Understanding Learning and Learning Design in MOOCs: A Measurement-Based Interpretation. *Journal of Learning Analytics*, *3*(2), 88–115. https://doi.org/10.18608/jla.2016.32.5

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and Educational Data Mining. *Journal of Educational Data Mining*, *4*(1), 11–48. http://www.educationaldatamining.org/JEDM/images/articles/vol4/issue1/MislevyEtAlVol4Issue1P11_48.pdf

Moon, J. (2004). *A Handbook of Reflective and Experiential Learning: Theory and Practice*. Routledge-Falmer.

Mueller, R. O., & Hancock, G. R. (2015). Factor Analysis and Latent Structure Analysis: Confirmatory Factor Analysis. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition* (pp. 5239–5244). https://doi.org/10.1016/B978-0-08-097086-8.25009-5

Myers, N. D., Ahn, S., & Jin, Y. (2011). Sample size and power estimates for a confirmatory factor analytic model in exercise and sport:a monte carlo approach. *Research Quarterly for Exercise and Sport*, *82*(3), 412–423. https://doi.org/10.1080/02701367.2011.10599773

Ochoa, X., & Worsley, M. (2016). Augmenting Learning Analytics with Multimodal Sensory Data. *Journal of Learning Analytics*, *3*(2), 213–219. https://doi.org/10.18608/jla.2016.32.10

Pennebaker, J W, Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic Inquiry and Word Count: LIWC2015*. Pennebaker Conglomerates. www.LIWC.net

Pennebaker, James W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PLoS ONE*, *9*(14), e115844. https://doi.org/10.1371/journal.pone.0115844

Plack, M. M., Driscoll, M., Blissett, S., McKenna, R., & Plack, T. P. (2005). A method for assessing reflective journal writing. *Journal of Allied Health*, *34*(4), 199–208. https://doi.org/10.1097/00001416-200607000-00014

Plack, M. M., Driscoll, M., Marquez, M., Cuppernull, L., Maring, J., & Greenberg, L. (2007). Assessing Reflective Writing on a Pediatric Clerkship by Using a Modified Bloom's Taxonomy. *Ambulatory Pediatrics*, *7*(4), 285–291. https://doi.org/10.1016/j.ambp.2007.04.006

Poldner, E., Schaaf, M. Van der, Simons, P. R.-J., Tartwijk, J. Van, & Wijngaards, G. (2014). Assessing student teachers' reflective writing through quantitative content analysis. *Europen Journal of Teacher Education*, *37*(3), 348–373.

Reidsema, C., Goldsmith, R., & Mort, P. (2010). Writing to learn: Reflective practice in engineering design. *Proceedings of the 9th Annual ASEE Global Colloquium (ASEE 2010)*.

Reidsema, C., & Mort, P. (2009). Assessing reflective writing: Analysis of reflective writing in an engineering design course. *Journal of Academic Language and Learning*, *3*(2), 117–129.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Ryan, M. (2011). Improving reflective writing in higher education: A social semiotic perspective. *Teaching in Higher Education*, *16*(1), 99–111. https://doi.org/10.1080/13562517.2010.507311

Ryan, M. (2013). The pedagogical balancing act: Teaching reflection in higher education. *Teaching in Higher Education*, *18*(2), 144–155. https://doi.org/10.1080/13562517.2012.694104

Satorra, A. (1990). Robustness issues in structural equation modeling: a review of recent developments. *Quality and Quantity*, *24*, 367–386. https://doi.org/10.1007/BF00152011

Schumacker, R. E., & Lomax, R. G. (2016). *A Beginner's Guide to Structural Equation Modeling*. 4th Edition, Routledge

Schumm, W. R., & Stevens, J. (1993). Applied Multivariate Statistics for the Social Sciences. *The American Statistician*, *47*(2), 155. https://doi.org/10.2307/2685203

Shibani, A., Knight, S., Buckingham Shum, S., & Ryan, P. (2017). Design and implementation of a pedagogic intervention using writing analytics. *Proceedings of the*

25th International Conference on Computers in Education, ICCE 2017 - Main Conference Proceedings, 306–315.
Shibani, A., Knight, S., & Shum, S. B. (2019). Contextualizable learning analytics design: A generic model and writing analytics evaluations. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK'19)*, 210–219. https://doi.org/10.1145/3303772.3303785
Tsingos, C., Bosnic-Anticevich, S., Lonie, J. M., & Smith, L. (2015). A model for assessing reflective practices in pharmacy education. *American Journal of Pharmaceutical Education*, *79*(8), 124. https://doi.org/10.5688/ajpe798124
Ullmann, T. D. (2015). Automated detection of reflection in texts - A machine learning based approach. In *The Open University* (Issue April). The Open University.
Ullmann, T. D. (2017). Reflective Writing Analytics - Empirically Determined Keywords of Written Reflection. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK '17*, 163–167. https://doi.org/10.1145/3027385.3027394
Ullmann, T. D. (2019). Automated Analysis of Reflection in Writing: Validating Machine Learning Approaches. *International Journal of Artificial Intelligence in Education*, *29*, 217–257. https://doi.org/10.1007/s40593-019-00174-2
Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect. *JOurnal of Personality and Social Psychology*, *54*, 1063--1070.
Wise, A., Knight, S., & Buckingham Shum, S. (n.d.). Collaborative Learning Analytics. In *International Handbook of Computer-Supported Collaborative Learning*. Springer.
Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety. *Educational and Psychological Measurement*, *73*(6), 913–934. https://doi.org/10.1177/0013164413495237
Wong, F. K., Kember, D., Chung, L. Y. F., & Yan, L. (1995). Assessing the level of student reflection from reflective journals. *Journal of Advanced Nursing*, *22*(1), 48–57. https://doi.org/10.1046/j.1365-2648.1995.22010048.x
Wright, L., & Lundy, M. (2012). Blogging as a tool to promote reflection among dietetic and physical therapy students during a multidisciplinary international service-learning experience. *Journal of Allied Health*, *41*(3), e73–e80.